

PITCH ESTIMATION FOR NOISY SPEECH

AZAR KHURSHID

DOCTOR OF PHILOSOPHY

2002

Copyright © Azar Khurshid, October, 2002.

This copy of the thesis has been supplied on the condition that anyone who consults it is understood to recognise that its copyright rests with its author and that no quotation from the thesis and that no information derived from it may be published without the authors prior consent.

Supervision Committee

1st Supervisor:
2nd Supervisor
3rd Supervisor

Dr. Susan Denham
Prof. Michael Denham
Dr. Guido Bugmann

PITCH ESTIMATION FOR NOISY SPEECH

by

AZAR KHURSHID

A thesis submitted in partial fulfilment of the
requirements for the degree of

DOCTOR OF PHILOSOPHY

Plymouth Institute of Neuroscience

School of Computing, University of Plymouth

October, 2002

LIBRARY STORE

UNIVERSITY OF PLYMOUTH	
Item No.	9005473670
Date	16 JUL 2003 T
Class No.	THESIS 534 KHU
Cont. No.	X704605286
PLYMOUTH LIBRARY	

REFERENCE ONLY

To My Parents and Teachers

PITCH ESTIMATION FOR NOISY SPEECH

Azar Khurshid

Abstract

In this dissertation a biologically plausible system of pitch estimation is proposed. The system is designed from the bottom up to be robust to challenging noise conditions. This robustness to the presence of noise in the signal is achieved by developing a new representation of the speech signal, based on the operation of damped harmonic oscillators, and temporal mode analysis of their output. This resulting representation is shown to possess qualities which are not degraded in presence of noise. A harmonic grouping based system is used to estimate the pitch frequency. A detailed statistical analysis is performed on the system, and performance compared with some of the most established and recent pitch estimation and tracking systems. The detailed analysis includes results of experiments with a variety of noises with a large range of signal to noise ratios, under different signal conditions. Situations where the interfering “noise” is speech from another speaker are also considered. The proposed system is able to estimate the pitch of both the main speaker, and the interfering speaker, thus emulating the phenomena of auditory streaming and “cocktail party effect” in terms of pitch perception. The results of the extensive statistical analysis show that the proposed system exhibits some very interesting properties in its ability of handling noise. The results also show that the proposed system’s overall performance is much better than any of the other systems tested, especially in presence of very large amounts of noise. The system is also shown to successfully simulate some very interesting psychoacoustical pitch perception phenomena. Through a detailed and comparative computational requirements analysis, it is also demonstrated that the proposed system is comparatively inexpensive in terms of processing and memory requirements.

CONTENTS

LIST OF FIGURES.....	IV
LIST OF TABLES.....	VI
1. INTRODUCTION.....	1
1.1. CONCEPTS RELATED TO SPEECH SIGNALS AND AUDITORY PROCESSING.....	2
1.2. ADVANCES AND PROBLEMS IN MODELING PITCH PERCEPTION.....	4
1.3. USES AND MOTIVATION FOR PITCH ANALYSIS.....	6
1.4. DEALING WITH ADVERSE NOISE AND MULTIPLE SPEAKER ENVIRONMENTS.....	8
1.5. ORIGINAL CONTRIBUTIONS AND RESULTS.....	9
1.6. OUTLINE OF THE THESIS BASED ON CHAPTERS.....	10
2. CASE STUDY AND BACKGROUND.....	12
2.1. A HISTORICAL ACCOUNT.....	12
2.2. TEMPORAL MODE OF PITCH DETERMINATION.....	14
2.3. PLACE OR FREQUENCY MODEL OF PITCH DETERMINATION.....	17
2.4. MIXED MODE PITCH DETERMINATION.....	18
2.5. MULTIPLE PITCH TRACKS FROM SIMULTANEOUS SPEECH.....	19
2.6. DETAILED DESCRIPTION OF PITCH ESTIMATION SYSTEMS USED IN THIS STUDY.....	20
3. DAMPED HARMONIC OSCILLATORS BASED SIGNAL PROCESSING.....	31
3.1. PRE-PROCESSING METHODS FOR SPEECH ANALYSIS.....	32
3.2. THE DAMPED HARMONIC OSCILLATOR BASED ANALYSIS.....	37
3.3. DESIGN OF THE BANK OF DHO UNITS AS A SIGNAL PROCESSING FRONT-END.....	51
3.4. ANALYSIS OF DHO BANK USING TEST SIGNALS.....	54
3.5. CONSTRUCTION OF TIME-FREQUENCY ENERGY MAPS.....	63
3.6. COMPARISON WITH OTHER TEMPORAL ANALYSIS SCHEMES.....	66
3.7. NOISE ROBUSTNESS AND BIOLOGICAL PLAUSIBILITY OF THE PROPOSED SYSTEM.....	69
4. THE PROPOSED PITCH ESTIMATION AND TRACKING SYSTEM.....	72
4.1. PERCEPTUAL ARGUMENTS FOR HARMONIC GROUPING FOR PITCH PERCEPTION.....	74
4.2. COMPUTATION OF SINGLE PITCH TRACKS THROUGH HARMONIC GROUPING.....	78
4.3. EXTENSION TO THE ESTIMATION OF MULTIPLE PITCH TRACKS FOR SIMULTANEOUS SPEECH.....	89
4.4. MODELING OF SOME PERCEPTUAL PITCH PHENOMENA.....	92
4.5. KEY ASPECTS OF THE PROPOSED SYSTEM.....	96
5. EXPERIMENTAL SETUP AND RESULTS.....	98
5.1. THE APPARATUS USED.....	99
5.2. THE EVALUATION PROCEDURE AND ERROR METRICS.....	105
5.3. THE SUMMARY OF RESULTS FOR SINGLE PITCH TRACKS EXPERIMENTS.....	112
5.4. SUMMARY OF RESULTS FOR TWO PITCH TRACKS EVALUATION.....	116
6. DISCUSSION.....	119
6.1. THE SINGLE PITCH ESTIMATION PERFORMANCE FOR HIGH RESOLUTION SIGNALS.....	119

6.2. THE SINGLE PITCH ESTIMATION PERFORMANCE FOR LOW RESOLUTION SIGNALS	124
6.3. THE MULTIPLE PITCH ESTIMATION PERFORMANCE FOR HIGH RESOLUTION SIGNALS	130
6.4. THE MULTIPLE PITCH ESTIMATION PERFORMANCE FOR LOW RESOLUTION SIGNALS	134
6.5. COMPARISON OF COMPUTATIONAL REQUIREMENTS.....	136
6.6. COMPARISON AND ANALYSIS OF THE OVERALL PITCH ESTIMATION PERFORMANCE .	137
7. CONCLUSIONS.....	142
7.1. SUMMARY OF RESULTS.....	142
7.2. CONTRIBUTIONS	144
7.3. SCOPE FOR FUTURE RESEARCH WORK AND EXTENSIONS.....	147
APPENDIX 1. IMPLEMENTATION DETAILS – COMPUTATIONAL COMPLEXITY AND MEMORY REQUIREMENT ANALYSIS.....	148
A.1. THE DHO AND HARMONIC GROUPING BASED SYSTEM	148
A.2. THE AUTOCORRELATION BASED SYSTEMS.....	150
A.3. THE PROBABILISTIC MULTIPLE PITCH ESTIMATION AND TRACKING SYSTEM (PMPT)	152
A.4. THE COMPARATIVE ANALYSIS	153
APPENDIX 2. THE COMPLETE RESULTS FOR SINGLE PITCH TRACK EXPERIMENTS.....	155
APPENDIX 3. THE COMPLETE RESULTS FOR MULTIPLE PITCH TRACK EXPERIMENTS.....	200
REFERENCES.....	237

LIST OF FIGURES

Figure 2.1. A simple Autocorrelation based PDA.....	16
Figure 3.1. The filter characteristics as reported by Slaney and Lyon.....	36
Figure 3.2 Speech Synthesis model of LPC analysis.....	37
Figure 3.3. The normalised magnitude response of a single DHO unit.....	43
Figure 3.4. The normalised magnitude response of a second order filter.....	43
Figure 3.5. The normalised magnitude response of a gammatone filter.....	44
Figure 3.6. The phase response of a DHO unit.....	45
Figure 3.7. The phase response of a second order filter.....	45
Figure 3.8. The impulse response of a DHO unit.....	46
Figure 3.9. The unit step response of the DHO unit.....	47
Figure 3.10. The Impulse response of the second order filter.....	47
Figure 3.11. The step response of the second order filter.....	48
Figure 3.12. The experimental noise response of the DHO units and second order filters.....	50
Figure 3.13. The Magnitude response of the bank of DHO units.....	53
Figure 3.14. The phase response of the designed bank of DHO units.....	54
Figure 3.15. The output response of a DHO unit to a chirp signal.....	55
Figure 3.16. The output response of a DHO unit to a chirp signal with added noise.....	56
Figure 3.17. The response of an example bank of three DHO units to sinusoidal input.....	57
Figure 3.18. The output of the bank of DHO units for speech like input.....	58
Figure 3.19. The output of the bank of DHO units, for noisy speech like input.....	59
Figure 3.20. The output of the bank of DHO units for speech signal.....	61
Figure 3.21. The power spectrum of speech signal calculated using FFT.....	61
Figure 3.22. The bank of DHO output for speech signal with noise.....	62
Figure 3.13. Spectrogram calculated using FFT for speech signal with noise.....	62
Figure 3.24. Diagram to describe the calculation of periodicity of DHO outputs.....	64
Figure 3.25. The Time-Frequency energy plot with chirp signal used as input.....	65
Figure 3.26. The time-frequency energy plot noisy speech signal.....	66
Figure 4.1. A neural autocorrelator by Licklider.....	76
Figure 4.2. The system design and overview for single pitch track estimation.....	79
Figure 4.4. An illustration of estimation for speech like stimulus.....	87
Figure 4.5. An illustration of pitch estimation for noisy speech like stimulus.....	88
Figure 4.6. An example pitch track produced using KEELE data.....	88
Figure 4.7. An example pitch track produced using KEELE data with added noise.....	89
Figure 4.8. An example of multiple pitch tracking.....	91
Figure 4.9. System output for the virtual pitch experiment : example 1.....	93
Figure 4.10. System output for the virtual pitch experiment : example 2.....	93
Figure 4.11. The Output of the system in response to IRN input.....	95
Figure 4.12. The standard deviation of the pitch tracks for increasing iterations IRN input.....	96
Figure 5.1. The spectrogram view of the three different noise environments.....	103
Figure 5.2. Example of preprocessed reference pitch track.....	110
Figure 6.1. Average Gross Error Rates for clean speech for high resolution signals.....	120

Figure 6.2. The GEE20 Error contributions in clean speech conditions for various systems...	121
Figure 6.3. Fine error rates for clean speech signals.....	122
Figure 6.4. Gross error rates (GEE20), for different levels of white noise for high resolution signals.....	123
Figure 6.5. Gross error rates (GEE20), for different levels of Environmental noise for high resolution signals.....	123
Figure 6.6. Gross error rates (GEE20), for different levels of Music noise for high resolution signals.....	123
Figure 6.7. Average Gross Error Rates for Clean speech for low resolution signals.....	125
Figure 6.8. An illustration of the catastrophic failure of YIN system for low resolution clean speech signals.....	126
Figure 6.9. The TLE and THE error metrics for clean low resolution signals.....	127
Figure 6.10. Gross error rates (GEE20), for different levels of White noise for low resolution signals.....	128
Figure 6.11. Gross error rates (GEE20), for different levels of Environmental noise for low resolution signals.....	128
Figure 6.12. Gross error rates (GEE20), for different levels of Music noise for low resolution signals.....	129
Figure 6.13. The GEE20 performance for the foreground pitch track estimation for different pitch estimation systems.....	132
Figure 6.14. The foreground and background pitch track estimation errors (GEE20) for the DHO system (BG = background, FG = foreground).....	133
Figure 6.15. The foreground and background pitch track estimation errors (GEE20) for the PMPT system.....	133
Figure 6.16. The foreground track average gross estimation error (GEE20) for the various systems at different SNRs.....	134
Figure 6.17. The foreground and background pitch track average error estimates, as computed using the GEE20 error metric.....	135
Figure 6.18. The number of mathematical operations requirement for the various systems analysed.....	136
Figure 6.19. The memory requirements for processing one second of speech sampled at 8000 Hz, in kilo-words.....	137

LIST OF TABLES

Table 3.1. The parameters used the bank of DHO design.....	52
Table 4.1. The parameters of the harmonic grouping algorithm with typical values.....	84
Table 4.2. The state variables of the pitch tracking algorithm.....	85
Table 4.3. The parameters of the pitch tracking algorithm and their description.....	85
Table 5.1. Summary of results for “clean” high resolution speech signals.....	113
Table 5.2. Summary of results for “low noise” conditions for high resolution signals.....	114
Table 5.3. Summary of results for “medium to high” noise conditions for high resolution signals.....	114
Table 5.4. Summary of results for “very high” noise conditions for high resolution signals.....	115
Table 5.5. Summary of results for “clean” low resolution speech signals.....	115
Table 5.6. Summary of results for “low” noise conditions for low resolution speech signals.....	115
Table 5.7. Summary of results for “medium to high” noise conditions for low resolution speech signals.....	116
Table 5.8. Summary of results for “very high” noise conditions for low resolution speech signals.....	116
Table 5.9. Summary of results for 5 dB SNR for high resolution speech signals.....	117
Table 5.10. Summary of results for 0 dB SNR for high resolution speech signals.....	117
Table 5.11. Summary of results for -5 dB SNR for high resolution speech signals.....	117
Table 5.11. Summary of results for 5 dB SNR for low resolution speech signals.....	118
Table 5.12. Summary of results for 0 dB SNR for low resolution speech signals.....	118
Table 5.13. Summary of results for -5 dB SNR for low resolution speech signals.....	118
Table A.2.1. The clean speech high resolution signal results for evaluated systems.....	156
Table A.2.2. Results for 25 dB White noise for high resolution speech signals.....	157
Table A.2.3. Results for 20 dB White noise for high resolution speech signals.....	158
Table A.2.4. Results for 15 dB White noise for high resolution speech signals.....	159
Table A.2.5. Results for 10 dB White noise for high resolution speech signals.....	160
Table A.2.6. Results for 5 dB White noise for high resolution speech signals.....	161
Table A.2.7. Results for 0 dB White noise for high resolution speech signals.....	162
Table A.2.8. Results for -5 dB White noise for high resolution speech signals.....	163
Table A.2.9. Results for 25 dB Environment noise for high resolution speech signals.....	164
Table A.2.10. Results for 20 dB Environment noise for high resolution speech signals.....	165
Table A.2.11. Results for 15 dB Environment noise for high resolution speech signals.....	166
Table A.2.12. Results for 10 dB Environment noise for high resolution speech signals.....	167
Table A.2.13. Results for 5 dB Environment noise for high resolution speech signals.....	168
Table A.2.14. Results for 0 dB Environment noise for high resolution speech signals.....	169
Table A.2.15. Results for -5 dB Environment noise for high resolution speech signals.....	170
Table A.2.16. Results for 25 dB Music noise for high resolution speech signals.....	171
Table A.2.17. Results for 20 dB Music noise for high resolution speech signals.....	172
Table A.2.18. Results for 15 dB Music noise for high resolution speech signals.....	173
Table A.2.19. Results for 10 dB Music noise for high resolution speech signals.....	174
Table A.2.20. Results for 5 dB Music noise for high resolution speech signals.....	175
Table A.2.21. Results for 0 dB Music noise for high resolution speech signals.....	176
Table A.2.22. Results for 0 dB Music noise for high resolution speech signals.....	177

Table A.2.23. Results for clean speech pitch estimation for low resolution signals.....	178
Table A.2.24. Results for 25 dB White noise for low resolution speech signals.....	179
Table A.2.25. Results for 20 dB White noise for low resolution speech signals.....	180
Table A.2.26. Results for 15 dB White noise for low resolution speech signals.....	181
Table A.2.27. Results for 10 dB White noise for low resolution speech signals.....	182
Table A.2.28. Results for 5 dB White noise for low resolution speech signals.....	183
Table A.2.29. Results for 0 dB White noise for low resolution speech signals.....	184
Table A.2.30. Results for -5 dB White noise for low resolution speech signals.....	185
Table A.2.31. Results for 25 dB Environment noise for low resolution speech signals.....	186
Table A.2.32. Results for 20 dB Environment noise for low resolution speech signals.....	187
Table A.2.33. Results for 15 dB Environment noise for low resolution speech signals.....	188
Table A.2.34. Results for 10 dB Environment noise for low resolution speech signals.....	189
Table A.2.35. Results for 5 dB Environment noise for low resolution speech signals.....	190
Table A.2.36. Results for 0 dB Environment noise for low resolution speech signals.....	191
Table A.2.37. Results for -5 dB Environment noise for low resolution speech signals.....	192
Table A.2.38. Results for 25 dB Music noise for low resolution speech signals.....	193
Table A.2.39. Results for 20 dB Music noise for low resolution speech signals.....	194
Table A.2.40. Results for 15 dB Music noise for low resolution speech signals.....	195
Table A.2.41. Results for 10 dB Music noise for low resolution speech signals.....	196
Table A.2.42. Results for 5 dB Music noise for low resolution speech signals.....	197
Table A.2.43. Results for 0 dB Music noise for low resolution speech signals.....	198
Table A.2.44. Results for -5 dB Music noise for low resolution speech signals.....	198
Table A.3.1. Performance with background utterance n7 at 5 dB SNR (high resolution signal).	202
Table A.3.2. Performance with background utterance n8 at 5 dB SNR (high resolution signal).	204
Table A.3.3. Performance with background utterance n9 at 5 dB SNR (high resolution signal).	206
Table A.3.4. Performance with background utterance n7 at 0 dB SNR (high resolution signal).	208
Table A.3.5. Performance with background utterance n8 at 0 dB SNR (high resolution signal).	210
Table A.3.6. Performance with background utterance n9 at 0 dB SNR (high resolution signal).	212
Table A.3.7. Performance with background utterance n7 at -5 dB SNR (high resolution signal).	214
Table A.3.8. Performance with background utterance n8 at -5 dB SNR (high resolution signal).	216
Table A.3.9. Performance with background utterance n9 at -5 dB SNR (high resolution signal).	218
Table A.3.10. Performance with background utterance n7 at 5 dB SNR (low resolution signal).	220
Table A.3.11. Performance with background utterance n8 at 5 dB SNR (low resolution signal).	222
Table A.3.12. Performance with background utterance n9 at 5 dB SNR (low resolution signal).	224
Table A.3.13. Performance with background utterance n7 at 0 dB SNR (low resolution signal).	226

Table A.3.14. Performance with background utterance n8 at 0 dB SNR (low resolution signal).	228
Table A.3.15. Performance with background utterance n9 at 0 dB SNR (low resolution signal).	230
Table A.3.16. Performance with background utterance n7 at -5 dB SNR (low resolution signal).	232
Table A.3.17. Performance with background utterance n8 at -5 dB SNR (low resolution signal).	234
Table A.3.18. Performance with background utterance n9 at -5 dB SNR (low resolution signal).	236

Acknowledgements

Large amounts of thank are due to a very large number of people, without whose support this work could not have been undertaken or completed. Inevitably, in this brief endeavour of thanking them all, many are likely to be missed.

I would like to thank Dr. Sue Denham for all the support that she continues to provide, and for her patience. I would like to thank Prof. Mike Denham for his valued ideas and insights throughout the length of my academic stay at the University of Plymouth. I would also like to thank Miss Carol Watson for her kindness and concern not only to me, but to all the research students. Many thanks are also due to the whole research community of the Plymouth Institute of Neuroscience, and the Centre for Neural and Adaptive studies at the University of Plymouth.

I would also like to thank all the staff at NeuVoice Limited, for making it possible to work on my dissertation and research interests, and indeed encouraging me to work and providing all the support when needed.

I would like to also thank my friends, my parents, and my family for all the support and encouragement.

Author's Declaration

At no time during the registration for the degree of Doctor of Philosophy has the author been registered for any other University award.

This study was financed by the Centre for Neural and Adaptive Studies, University of Plymouth, United Kingdom.

A program of advanced study was undertaken, and relevant scientific seminars and conferences were regularly attended, at which work was presented.

Patents

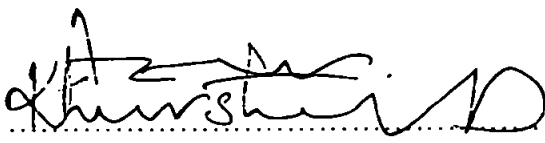
"Noise reduction for speech recognisers," Patent applied for. Application number 0206349.2, filed 18th March, 2002.

"Fundamental frequency/pitch estimation system for noisy speech signals," UK patent applied for, October, 2002.

Relevant Scientific Seminars and Conferences Attended

Eurospeech, 2001, Alborg, Denmark, September, 2001.

CNAS Research seminar series, for years 1999, 2000, and 2001.

Signed 

Dated 05 May '03

Chapter 1

INTRODUCTION

The work that is described in this dissertation relates to the perception of pitch when listening to speech. The topics that are explored include the perception of pitch in speech by humans and emulation of this perception through measurement, by computers. The main thrust of this research is to build a computational model of pitch measurement in speech this is both biologically plausible and more importantly, has a robust performance, especially in challenging high noise environments.

Sensory systems like the auditory system have evolved because they provide an advantage in the evolutionary process of selection. The advantage of any sensory system is that it provides information about what is going on in the environment and as such enables appropriate action and reaction. One can be sure about that because any sensory system that has once existed but was a failure in this respect was inevitably “filtered out” by evolutionary selection [Terhardt, 1991]. In view of these notions it is not surprising that higher animals, such as man, possess an auditory system that is robust to the distortions that may diminish its ability to perform. Therefore, by learning from the way information is processed in the auditory system, one can develop models which are similarly robust in their ability to handle challenging noise conditions. These “biologically inspired” techniques could then be used to make the computer based systems of speech processing similarly robust. On the other hand, such computer models of the auditory system help us to understand the biological auditory processes much better.

The perception of pitch is related to the periodicity of the sound signal. A very concise definition of pitch is elusive, but most agree that “Pitch is the perceptual correlate of the frequency of a simple tone” [Shroeder, 1999]. It is the feature of sound by which listeners can arrange sounds on a scale from “lowest” to “highest” on a frequency scale. For complex sounds like speech, the sensation of pitch is more difficult to define, but is closely related to the lowest frequency of the sound, or the fundamental frequency. Complex periodic sounds in a normal speech or music

context have a pitch that covaries with the fundamental frequency^{*} independently of the presence or absence of energy in the sound spectrum at this frequency [Shouten, 1940]. In the signal processing sense, the pitch or fundamental frequency of a periodic sound is defined as “the greatest common divisor of its harmonic components”. Before continuing with the presentation of more advanced concepts, a few basic definitions are presented in the next section.

1.1. Concepts related to speech signals and auditory processing

This section briefly introduces concepts regarding speech signals that are used in this dissertation. Speech is produced by forcing air from the lungs through the oral and nasal cavities. Different sounds are created by changing the shape of the vocal tract (oral + nasal cavities) and/or changing the characteristics of the airflow from the lungs. The latter can roughly be either periodic, or noise-like, which leads to periodic (voiced) or noise-like (unvoiced) speech. An example of a voiced sound is /a/ in *father* and an unvoiced sound is /s/ in *see*. An important parameter for voiced speech is the repetition frequency which is called the *pitch frequency*. Male speakers have in general a lower pitch frequency (creating a deeper voice) than female and child speakers. Usually, the pitch frequency component is accompanied by several components at the frequencies which are integer multiples of the pitch frequency. These components are called the *harmonics*. The vocal tract can be viewed as an acoustic tube of varying diameter. Depending on the shape of the acoustic tube (mainly influenced by the tongue position), a sound wave traveling through it will be reflected in a certain way so that interferences will generate stronger *resonances* at certain frequencies than others. These stronger resonances are called *formants*. Their frequency determines the speech sound that is heard, and generally is not influenced by the pitch of the sound. The prosody of speech is the qualitative variation of stress, duration and pitch of the utterance, and is usually a carrier of supplementary information in speech, like emotion and context.

Audio data is generally present in the form of electric oscillations. These can either come from a microphone recording acoustic sound waves, from a tape or other storage device, or an electronic instrument or device. To convert these oscillations to a form that can be treated by a

^{*} In this dissertation, the terms “fundamental frequency” and “pitch” are sometimes used interchangeably. For most signals, and conditions that we deal with, this is true. However, this is not true for all signals under all conditions, and we shall point out in the text if such is the case.

computer they have to be *sampled*, which is a process of converting the analogue electrical signal into a digital signal. The sampling of sounds is performed at fixed rates called the *sampling frequency* or *sampling rate*. A higher sampling rate corresponds to a better quality signal, because it can contain more details about the original signal. The *bandwidth* of a sampled signal is the largest frequency information it may contain, and is half the sampling frequency, sometimes also called the *Nyquist frequency*. The sample resolution is equal to the number of computer bits used to store a single signal sample. Usually, this is a multiple of 8 (1 byte).

The biological auditory systems employ a variety of strategies and stages to analyse sounds. When the sound enters the ear canal, it hits the eardrum, which vibrates with a motion corresponding to the ripple of the sound wave. The motion of the eardrum is transported to the *cochlea*, where there is a membranous structure called the *basilar membrane* [Moore, 1997]. This is attached to special sensory cells called the *hair cells*. The cochlea provides a very fast response sound analyser, and can distinguish a very large range of frequencies, ranging from 20 to 20000 Hz. The hair cells are connected to the auditory nerve, which passes through several neural pathways to the *auditory cortex*. Different frequencies of sound are represented in particular regions of the cortex. This arrangement is called *tonotopic representation*, and is found throughout the auditory pathway, starting from the basilar membrane in the inner ear. The most medial portion of the auditory cortex contains the representation of the basal end of the basilar membrane of the inner ear, while the apical end of the basilar membrane is represented in the lateral portion of the auditory cortex. This arrangement is called *place coding*. The *tonotopic* organization of each auditory cortical sub-region remains unclear, but in general it appears that in each subfield that has been mapped, low tones are represented posteriorly, while the high tones more anteriorly.

Although the human auditory system is capable of analyzing a large range of frequencies, most of the useful information in the speech signal has been shown to lie within the first three formants of the signal, which corresponds to a frequency range of 150 Hz to 5000 Hz [Klatt, 1980]. In signal processing terms, the speech signal is defined as quasi-periodic and non-stationary. A *quasi-periodic* signal is periodic for short amounts of time, but the periodicity changes over longer periods of time. Another property resulting from the processing of the speech signals is that of *amplitude modulation*. If the bandwidth of a single processing unit covers more than one frequency

component or harmonic, then these components interact to produce an output which is a result of an interaction of these components, which itself contains components equal to the difference between these frequencies. This phenomenon of amplitude modulation at higher frequency components by in analysis of voiced speech sounds is used in some pitch estimation systems. The reader is referred to [Strope et al., 2001] for review of these ideas.

1.2. Advances and Problems in Modeling Pitch Perception

In the process of building and testing pitch perception models, one can draw on the wealth of existing knowledge about the types of signals that generate a pitch sensation, and the large amount of psychoacoustics data that has been collected by researchers. Psychoacoustical study as a recognized branch of science has been in existence for more than a century. Some of the earliest studies recognizable as psychological science in the 19th century were concerned with the perception of the loudness and pitch of sounds. However, even before scientific methods developed, philosophers engaged in speculation about the nature of sound. Psychoacoustic thinking dates all the way back to Pythagoras, who is credited with recognizing that strings whose lengths are related as the ratio of small integers sound good when plucked at the same time [Singh, 1997]. Modern psychoacoustics, since the work of Wegel, Fletcher and others in the early 20th century [Allen, 1996], has evolved a sophisticated understanding of the early stages of hearing. Robust and well-tested models have been developed, especially of single perceptual features (such as pitch and loudness) of simple stimuli, and the way in which one simple sound “masks” (hides) another depending on the time-frequency relationship between the two sounds [Plomp, 1976].

Although the field of psychoacoustics has given us a lot of insight into the perception of pitch and other related phenomenon, the computational models that are derived from the understanding of the data cannot always reproduce the results in complex sounds and environments. The early days of pitch research dealt primarily with understanding the exact capabilities and psychophysical discrimination accuracy for pitch; more recently, research has focused on the construction of computational (or at least functional) models that mimic the human ability to determine pitch from acoustic signals [Slaney, Lyon, 1990], [Meddis, O’Mard, 1997], [de Cheveigné, 1998]. In these latest attempts, there is also an attempt to reproduce the various psychoacoustic phenomena related to pitch. These include the *missing-fundamental*

phenomenon, the percept of *dichotic pitch*, *musical intelligibility* and the *multiple pitch stimuli effect*. The *missing fundamental* [Shouten, et al, 1962] effect is experienced in its simplest form by most of us during a telephone conversation. The telephone signal is usually band-limited to frequencies above the voiced sounds pitch range, but we still perceive the pitch effortlessly [Hess, 1983]. The *dichotic pitch* phenomenon is demonstrated by presenting different sound signals to each ear. On their own, neither of the signals produce a perception of pitch, but presented simultaneously to both the ears, listeners experience a definite pitch [van der Brink, 1974]. *Music intelligibility* relates to the phenomena in which the pitch of complex tones made up of a random number of harmonics can be heard equally well whether the subject is presented with them monotically (all in one ear) or dichotically (different harmonics sent to each ear) [Moore, 1997]. There have been psychoacoustical tests in which the subjects are presented with two or more vowels at one time, and depending on the differences in pitch and the identity of the vowels, different results are obtained. These experiments explored the role of differences between the fundamental frequencies of concurrent voices on the perceptual separation of those voices; all the frequency components of one vowel share the same fundamental frequency, allowing them to be perceptually grouped, but differ from those of the competing vowel, which can therefore be perceptually segregated. A difference in fundamental frequency between two concurrent vowels (a "double-vowel" stimulus) is known to facilitate identification of the vowels [Scheffers, 1983], [Meddis, Hewitt, 1992]. The improvement in identification was assumed to arise from a perceptual segregation process, which allowed the characteristic spectral features of the two vowels to be analysed separately, rather than as a confusing mixture. The role of the fundamental frequency in segregating natural speech is probably large, since the voiced portions of speech are important for intelligibility and since competing voices will differ in fundamental frequency most of the time [Culling et al, 1994]. Several computational models for the perceptual segregation of simultaneous voiced speech have been proposed, which measure the two fundamental frequencies using models of pitch perception and then separate the two vowels by selecting the frequency components which are harmonics of those two fundamentals. These simultaneous speech stimuli are known as *multiple pitch stimuli effects*. These phenomena are described in detail in [Terhardt, 1980].

1.3. Uses and Motivation for Pitch Analysis

The percept of pitch is fundamental to the understanding of the hearing processes in the ear and brain. Equally, a computational system for pitch detection and measurement is fundamental to the processing of the signal in an efficient way. Although pitch is one of the most understood parts of speech processing in humans, with many different theories and computational systems, the systems cannot always provide an accurate measure of the pitch of the presented speech signal when it is adulterated with significant amount of noise. With the presence of other, interfering speaker(s) in the speech signal, the problem is compounded, and usually the systems designed for single pitch estimates have much reduced performance [Hermes, 1993]. Resolution of these problems is an active area of research. Due to the problems faced by even the best pitch detection systems, the use of pitch as an important analysis parameter for real-world applications of speech processing, apart from use in voicing detection, has remained dormant.

Use of Pitch information in Robust Automatic Speech Recognition Systems: Most automatic speech recognition (ASR) systems use a hierarchy of non-stationary stochastic models operating at progressively longer intervals of speech analysis and statistical modeling at different representational levels (phonetic, word, phrase etc), in order to decipher the “what” content of the speech signal being analysed. However, ASR systems rarely use pitch or voicing information in this process. Linear prediction is used, for example, with a predictor polynomial that is significantly shorter than the pitch period. In Mel Frequency cepstral coefficients computation, the initial spectral estimate is averaged over multiple pitch periods, and then integrated across frequency, providing the approximation of auditory frequency selectivity [Rabiner, Juang, 1993]. In all these signal processing systems for ASR, there is a deconvolution of the vocal tract transfer function and the driving function (periodic glottal pulses). Although deconvolution in this way is justified as segregation of “what” information from the “how” information, doing so in the first stages of analysis does not give the benefit of this lost information to the later stages of the recognition process, where this information would be useful in the presence of noise. Speech communication has evolved to be robust to noise, and although the pitch information may be “redundant” in the tasks of ASR, it plays a large part in defining the saliency of speech and is robust to high levels of noise. Therefore, elimination of this information in the first stages may not be optimal [Strope et al, 2001]. ASR systems in noisy speech have much lower performance when compared to natural speech recognition. Apart from analysis and recognition, pitch

information, properly incorporated should also benefit the training of ASR systems. Currently, most ASR systems require a large amount of speech from lots of different speakers for statistical modeling of speech units. If, however, these systems made use of the noise robust representations by better utilization of redundancy in the speech signal, it may be possible to train these systems with much less data. Pitch processing also helps in the speech/ non-speech decision, which can be very hard in challenging noise conditions.

Use of Pitch information in Speech Synthesis Systems: In speech production systems, the prosody of the speech to be produced has a huge impact on the intelligibility of the result. A vast body of research has been devoted to the human speaking process, including high-speed motion pictures of the vocal tract, x-rays of moving articulators, myographic recording from the muscles that control the articulators, amongst many others. In addition neural networks have been trained to speak in order to gain further understanding of the processes involved in the production of speech [Sejnowski, Rosenberg, 1986], [Guenther, 1995]. One of the most popular (and successful) ideas for machine based speech production is the concatenative speech synthesis paradigm. Although whole-word concatenation is least complicated in terms of co-articulatory effects, it suffers from the problems of prosody in the larger sentence and context structure, and size-of-dictionary constraints. Typically, sub-word units are used in the process of generating concatenated speech. Units which minimize the co-articulation effect have been designed for this purpose, including demisyllables [Fujimura, 1979] (where the boundary for each unit is a steady state vowel), and diphone (vowel to post-vocalic consonant transitions). However, these units have been found to be language dependent, and usually have to be re-engineered for different languages. In current speech synthesis systems, like the festival system [Dusterhoff, Black, 1997], the quality of the speech is quite high, but the mechanical nature of the speech sounds produced remains an anathema to most systems. These problems have most often been associated with prosody (or lack of it). Prosody in turn depends on the pitch for a large part, and researchers have been making use of this information to produce natural sounding speech [Silipo, Greenberg, 1999]. The use of prosody typically involves rules of pitch manipulation and constraints as an utterance evolves over time, based on the overall context. However, the use of pitch for these models of synthesis is not very well understood, especially the contribution of pitch to stress, and to take effective advantage of the various variables involved (i.e. the role of

amplitude, duration and pitch, and their interplay in determination of prosody), [Bergem, 1993] in order to improve quality and intelligibility of synthesised speech remains a difficult task.

Use of Pitch in Computational Auditory Scene Analysis: Since the 1970s, the work of Bregman [Bregman, 1990], his collaborators, and others has resulted in a new body of psychoacoustic knowledge collectively known as *auditory scene analysis* (ASA). The goal of this field is to understand the way the auditory system and brain process complex sound scenes, where multiple sources that change independently over time are present. Two sub-fields are dominant: auditory *grouping* theory, which attempts to explain how multiple simultaneous sounds are partitioned to form multiple “auditory images”; and auditory *streaming* theory, which attempts to explain how multiple sequential sounds are associated over time into individual cohering entities, called streams of sound. Both these groups of study in ASA make use of pitch information for the respective tasks.

Pitch as a speaker characteristic: The pitch of voiced speech varies with the speaker who produces it. Pitch for female speakers is significantly higher on average, compared to male speakers. Other speaker characteristics like vocal tract length (VTL) are used in conjunction with pitch in some speaker identification tasks. The problem of speaker identification based on the speech signal is hard to tackle with pitch alone because of the variability of pitch based on the prosodic requirements of speech production. However, it can be seen as one of the dimensions in the multi-dimensional speaker space [Furui, 1996].

Pitch and Music: Although not the domain of our research, pitch has been traditionally a central concept in the studies of western music. This includes the way multiple notes group horizontally into melodies, vertically into chords, and in both directions into larger-scale structures such as “harmonies” and “keys.” The preponderance of formal music theory deals with the subsumption of notes into melodies and harmonic structures, and harmonic structures into areas of “key” or “tonality.” It is believed that an understanding of pitch in speech within challenging noise environments will have fruition in the study and processing of music as well [Scheirer, 2000].

1.4. Dealing with adverse noise and multiple speaker environments

The primary problem that faces the real word computer based speech analysis technology is to deal with noisy signals. Noise can be attributed to transmission and digitization noise, and to

interfering sounds or environmental noise. The transmission and digitization processes are well controlled and compensation techniques have been developed to deal with these problems to some extent. Environmental noise without any a-priori knowledge of the source of such noise has proved to be a very hard problem, even with moderate levels of noise. In challenging noise conditions, where the signal to noise ratio (SNR) falls below 5 decibels, most analysis systems simply break down. Most ASR systems can only tackle the noise problem by including noisy data in training, which requires a-priori assumptions about the noise environment in general.

Speech interface based applications in noisy environments provide a primary challenge for auditory research because of the move towards mobile devices with small form-factor, with very limited input/output interfaces. Although speech forms a very natural mode of input/output for these small devices like mobile phones or “pocket computers”, these devices have the added requirement of being used in everyday noisy environments like in railway stations or in cars. There is a great need to make the speech based interfaces more robust to noise.

Although humans can handle multiple speaker environments quite effectively, pitch detection systems that can achieve this in real world noisy signals have not yet been developed to a satisfactory accuracy level or can handle only certain types of noise. Pitch determination for signals with multiple speakers has traditionally been used in the computational auditory scene analysis research for segregation or *streaming* (more on these systems in the next chapter). However, it also has application in the field of regular speech processing, by treating one of the speakers as the foreground speaker and the rest of the signal components, including other speakers, as noise.

1.5. Original Contributions and Results

In this dissertation, a new system of pitch estimation and tracking is proposed. The system is very simple in design and computationally efficient.

The proposed pitch estimation system uses damped harmonic oscillators to model the tonotopic ordering of sounds by the basilar membrane. A temporal representation, based on the treatment of the output of the damped harmonic oscillator units on a “temporal pattern coding” basis rather than the more commonly used “channel based coding” is developed. The temporal pattern coding uses the fine temporal structure of the output signals, rather than the channel

frequency, as is the case in Fourier transform and filter-bank based systems. This processing produces a representation that demonstrates properties similar to the “noise masking” properties observed in the auditory system (see chapter 3 for detailed discussion). A harmonic grouping based system for pitch frequency estimation is proposed that uses this representation’s high frequency resolution. The system is based on the Barlowian approach to perception for minimization of representation [Barlow, 1959], and is not dependent on a-priori knowledge of the pitch frequency. In the Barlowian approach to the problem, the pitch would be an emergent property of the auditory system, in order to achieve some sort of minimal representation of the information contained in the sound stimulus. Similarly, in the proposed system, the pitch frequency arises as a product of the need to group lower harmonics for coding and representation efficiency. The system is easily extended to multiple pitch frequency estimation for simultaneous speech from two speakers with different fundamental frequency.

A detailed statistical analysis, and performance comparisons with selected pitch estimation systems in a large variety of noise conditions and different signal conditions is performed. This large scale analysis involving different error metrics for comparison of different pitch estimation systems on a single database, especially for noisy speech, has not been reported before, especially with different kinds of noises at many different SNR values. Some recent studies have attempted to address the dearth of comparative performance of pitch estimation systems in [de Cheveigné, Kawahara, 2002] and [Hermes, 1993], among others, but they have usually stopped short of publishing detailed results on all the error metrics so that a detailed comparative analysis on the merits of these systems may be carried out.

This research has dealt with the modeling of pitch perception in very tough environmental conditions, with high noise levels and low signal resolution. It may be pointed out that no attempts are made to put forward a new theory of auditory perception. The focus of this work is to present a system that is developed from the ground up to be robust to challenging noise conditions, has reasonable computational requirements, and is suitable for practical applications.

1.6. Outline of the thesis based on chapters

In the current chapter, the problem of pitch determination by computers was introduced. The motivation behind pitch analysis in speech was also presented. A brief introduction of pitch determination and related concepts was presented. The aims of the research carried out and

original contributions to this dynamic area of research were outlined, with the details addressed in the following sections of the dissertation.

In chapter two, *Case Studies and Background*, a detailed survey of different pitch analysis models and theories behind them is presented. The systems chosen for analysis and evaluation of performance are described in detail.

Chapter three presents a detailed analysis of the front end of the proposed system, along with the derivation of the equations that govern the dynamic behaviour of this part of the system. We also present an analysis of controlled synthetic signals and speech signals to demonstrate the properties of the system, and explain the major reasons for the noise robustness of the system. An empirical analysis of the properties of the systems is also presented.

Chapter four gives a detailed account of the proposed pitch measurement and detection algorithms, as well as motivations for their design. The multiple pitch tracking system extension to the model is also presented. Some typical psychoacoustical phenomena are discussed, and a general discussion of the system's capabilities is presented.

Chapter five comprises of a detailed benchmark study of the performance of the proposed system, and some of the chosen pitch estimation systems. Their performance in noise is compared for different types and levels of noise, and a detailed error analysis of the results is presented. Multiple pitch track experimental results are also presented with detailed error analysis.

Chapter six, *Discussion*, takes account of the results that are presented in the previous chapter, and provides arguments that explain the results.

Chapter seven concludes this dissertation and highlights the main conclusions that can be drawn from this work and the results presented.

The appendix 1 to this dissertation contains a computational complexity and requirements analysis of the various systems evaluated in the text and compares the results with the proposed system. Appendixes 2 and 3 contain detailed tabulated results for all the experiments discussed in the main text.

Chapter 2

CASE STUDY AND BACKGROUND

Most people are familiar with the general principles of hearing and speaking. By means of the speech organs, vibrations are produced, and depending on what needs to be conveyed, the vibrations have different characteristics. These signals are then transmitted through the air to the ears of the listener. There, the speech signal transfers its vibrational energy to the ear drum, and through an intricate mechanical coupling of the bones of the middle ear, the vibrations reach the inner ear. In the inner ear, the cochlea transforms the signal into neural impulses, which are sent to the brain for analysis and recognition.

To trace the historical acquisition of this knowledge, and the current state of research in the area of auditory processing and pitch analysis, we present a brief historical account of the various stages of these developments to the current day.

2.1. A Historical Account

Pythagoras, who is credited with recognizing that strings whose lengths are related as the ratio of small integers sound good when plucked at the same time [Singh, 1997], is a fine example of our earliest fascination with the subject of auditory perception. However, a comprehensive theory of auditory analysis did not exist before Helmholtz (a translation of his works is available '*Sensations of tones*') [Helmholtz, 1870]). His theory dominated the field for some six decades. He realised that we have no difficulty in following the individual instruments in a concert, thus it follows that different streams of sounds are propagated without mutual disturbance, and that the ear can analyse a complex mix of these streams into its original constituents [Boer, 1977].

Helmholtz's explanation of how the ear performs this complex analysis task was based on two earlier theories. He used Ohm's law of hearing to suggest that the ear performs a type of Fourier analysis to separate a complex tone into its corresponding sinusoidal components. He extended this concept to account for phenomena like 'difference tones' by suggesting a non-linearity in the analysis, that introduces new sinusoids corresponding to the difference tone, not present in the original signal. The other part was based on Müller's doctrine of specific nerve energies. This

doctrine states that a particular type of sensation is related to the type of nerve fibres that are stimulated by the sound stimulus. Helmholtz explained this by assuming that every pitch that is discernable, corresponds to a different nerve, or a small group of nerves. Further assumption of a connection between these nerves and the cochlear segments resonating to specific tones enabled him to formulate a theory of pitch perception [Moore, 1997].

Helmholtz's simple explanations on auditory perception soon started to show inconsistencies. The greatest difficulty was the selectivity of resonators. The fine frequency discrimination by the human ear would imply that the resonators in the inner ear were highly selective. However, selectivity and damping of resonators are inversely proportional and high selectivity implies low damping. Lower damping would mean that tones presented even for short periods would have long persistence. This is certainly not the case, and doubts about the theory presented by Helmholtz started to grow, until 1900, when Gray suggested that nerve fibres maximally excited by the segments on the basilar membrane give rise to a sensation of pitch, while the rest are suppressed.

Von Békésy made the first known measurements of the vibrations of the basilar membrane in 1960 [von Békésy, 1960]. He found that there is mechanical analysis in the cochlea, so that sinusoids are distributed along its length according to their frequency, i.e., particular sections of the basilar membrane respond to frequencies associated with them. He also found that based on the amplitude of the tone, each stimulating sinusoid displaces a large part of the membrane. In light of these findings, Helmholtz's theory became untenable, and new ideas about auditory analysis were sought, and found [von Békésy, 1963].

Explanations of the perception of pitch can be found at the centre of the auditory analysis theories, and fall into two groups. These are the *temporal* models and the *frequency* models or the *place* models.

The temporal model assumes that the frequencies in the lower and middle regions of the spectrum are determined by the timing of the neural impulses rather than the place of vibration on the basilar membrane. The main evidence for these theories comes from experiments which show that periodicity of the waveform may give rise to a pitch sensation, even though there is no corresponding frequency component present in the original signal [Licklider, 1956], [Boer, 1977].

The place models postulate that the preliminary mechanical analysis at the basilar membrane is supplemented by a neural sharpening process that limits pitch perception to a small group of nerve fibres. Arguments for this model were invoked to explain the well-defined pitch of very short tones. However, it was later argued that the analysis time required by the place theory for inhibition processes would lead to difficulties in explaining the pitch perception of a simultaneous tone or a tone of different frequency just after the first small duration tone. This is because the establishment of inhibition requires a duration of analysis larger than the time delay between two tones that have a different pitch [Whitefield, 1970].

Temporal analysis for pitch estimation and full auditory analysis has been encouraged by physiological studies demonstrating phase locking of the auditory nerve fibre activity to stimulus tone period [Kiang et al, 1965]. The sensitivity of the place methods to the formant structure of high amplitude speech sounds has also encouraged the detailed development of models of temporal representation of auditory nerve fibre activity. The major aim of these representations and models is to simulate a wide range of physiological phenomena linked to the perception of pitch, such as virtual pitch or the pitch of the missing fundamental [Shouten, 1940], the pitch of inharmonic complexes [Plomp, 1976], [Moore et al, 1985], and repetition pitch [Bilsen, 1966]. These various psychophysical effects of pitch perception are discussed in chapter 4.

In the present study, the emphasis is on speech signals, and their perceived pitch. The non-stationary speech signal is much more interesting than stationary tonal signals and measurement of pitch has much practical use in computer based speech analysis systems. Although measurement of pitch in speech signals is much more difficult than in pure tonal complexes, there are some features algorithms usually take advantage of, including a well known existence range of pitch for speech (usually 60 Hz to 350 Hz), and local continuity constraints inherent in speech production systems. The different systems that employ these models for measurement of pitch in speech signals shall be discussed in the following sections.

2.2. Temporal Mode of Pitch Determination

The methods and algorithms that are described in this section are time-domain pitch detectors that operate directly on the speech waveform to estimate the pitch period. For these pitch detectors the measurements most often made are peak and valley measurements [Dubnowski, et al, 1976], zero-crossing measurements [Sondhi, 1968], and autocorrelation measurements [Hess,

1983]. The basic assumption in these systems is that the signal has been previously suitably pre-processed to remove any effects of formant structure so that the time domain structure provides good estimates of the period (the formants affect the peak in the ACF calculation by sometimes producing peaks larger than those due to the fundamental frequency of the signal). These techniques include methods like centre-clipping [Noll, 1967].

2.2.1. Time Domain Autocorrelation Based Methods

A typical autocorrelation based pitch determination algorithm (PDA) consists of three stages. The first stage is the pre-processing stage that operates on the original signal. The aim is to make the resulting signal spectrally flat. The second stage is windowing the signal and the calculation of the autocorrelation function over different lags for each windowed section. The third stage is the calculation of the maximum autocorrelation peak in the pitch range, and based on the strength of the peak, determining the period and the salience of the period. If the salience is high, the windowed section of the speech is termed voiced, otherwise it is termed un-voiced or silent. These three stages vary slightly for different methods. We shall give a brief example of one of the systems described by Rabiner [Rabiner et al, 1976].

The speech waveform is low pass filtered to a low cut-off frequency near 1000 Hz frequency. The low-pass speech is then processed in sections of 30 ms with an overlap between segments of no larger than 15 ms. The next stage of the process is centre-clipping. Centre-clipping is a simple technique that makes the values of the signal zero, when its absolute value lies below a predetermined level. A clipping level C_L is determined from the current segment of speech. The C_L is usually set near to 60 % of the maximum peak of this portion of the signal. Following the determination of the clipping level, the signal is centre clipped so that the resulting signal has three possible values of +1, 0, or -1. If the signal sample is greater than the C_L then its assigned a value of 1, if its below $-C_L$, it is assigned a value of -1, and zero otherwise.

Following centre-clipping, the signal autocorrelation function is evaluated over a range of lags, usually ranging for speech signals from 2 ms to 20 ms. Additionally, the autocorrelation is also computed at 0 lag for normalisation purposes. The autocorrelation values at various lags are then searched for a maximum normalised value. If this maximum value exceeds a certain threshold, the current segment of speech is classified as voiced, and its fundamental frequency computed,

which is proportional to the lag at which the maximum occurs. The block diagram of the system is presented in figure 2.1.

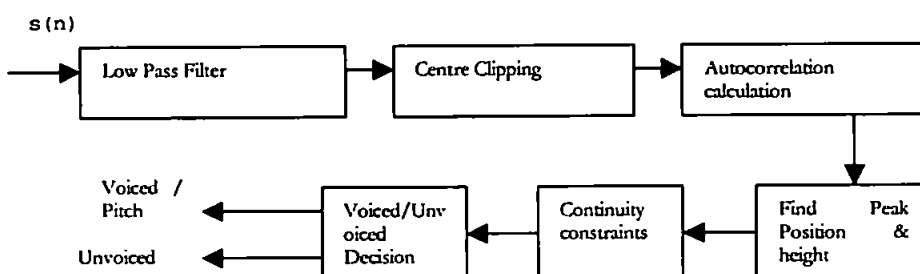


Figure 2.1. A simple Autocorrelation based PDA..

A variant of this system uses inverse filtering instead of centre clipping in order to achieve a spectrally flat signal for the autocorrelation analysis. In this case the signal is first low pass filtered to about 1000 Hz, and then decimated. The resulting 2 kHz signal is then inverse-filtered to give a spectrally flattened signal which is then auto-correlated.

2.2.2. Zero Crossing and Peak and Valley Measurements

The pitch detection algorithms discussed here, place pitch markers (mark the position of pitch related peaks in the time waveform) on the low pass filtered speech and are therefore also called phase synchronous pitch detection algorithms [Sondhi, 1968].

In the pre-processing stage, speech signal is low pass filtered to about 1000 Hz. To obtain the appropriate pitch markers, the excursion cycles in the signal are detected by measurement of the intervals between major zero-crossings. Then a heuristic approach is used to identify those excursion cycles that correspond to true pitch periods. This includes a series of steps involving the measurement of short time energy, and anticipates syllabic rate pitch changes in the signal. An error measure is used to provide continuity constraints in the pitch markers. Regions of unvoiced pitch are identifiable in this method by the lack of pitch markers in the processed portion of speech.

Another phase synchronous method of pitch detection uses autocorrelation based pitch estimates, and then uses the zero-crossings in the original signal for the measurement of the

phase synchronous pitch periods [Goodwin, 1992]. Firstly, the output of the autocorrelation system is assumed to be a running estimate of the pitch periods of voiced speech with proper time indices of the evaluated window. The algorithm then searches for the first major zero crossing after a non-zero pitch period is detected (zero pitch period value is assumed for unvoiced). Then, the pitch period is interpolated between two pitch periods by the time instance of the zero-crossing, at the zero crossing moment. This leads to an estimate of the pitch at the zero-crossing moment, thus phase locking the pitch estimate with the signal itself. Mathematically, if P_a is the first pitch period, and P_b is the next pitch period, and the zero crossing measured at t_0 , then the period calculation for P_0 is given by the equation below.

$$P(t_0) = \frac{P_a(t_b - t_0) + P_b(t_0 - t_a)}{t_b - t_a} \quad \dots 2.1.$$

2.3. Place or Frequency Model of Pitch Determination

The systems described here are a class of frequency-domain pitch detectors which use the property that if the signal is periodic in the time domain, its frequency representation will have a series of bands at the fundamental frequency and its harmonics. Thus simple measurements can be made on the frequency-domain representation of the signal (or a non-linearly transformed version of it as in the cepstral pitch detector) to estimate the pitch of the signal.

The first deliberate attempt to develop a pitch estimation system based on place theory was made by Duifhuis in [Duifhuis et al, 1982]. He implemented his system based on the Goldstein optimum processor [Goldstein, 1973]. Exact implementation of the system as envisaged by Goldstein was very computationally demanding as it needed maximum likelihood estimation of the high resolution spectrum of the signal. The Gaussian widening required by Goldstein's model to explain pitch perception in tones with nonharmonic partials, was replaced by a rectangular widening, in what is now termed a *Harmonic Sieve*. All spectral peaks contributed with an all-or-none principle to the pitch estimate, based on certain audibility and masking constraints. Making the all-or-none principle more relaxed with a gradual decrease in contribution of the estimate based on the position of the peaks, and their amplitudes, results in systems similar to a spectral comb [Goldstein, 1973], or the Sub-Harmonic Summation (SHS) [Scheffers, 1983].

However, the most popular pitch detection technique belonging to this class of models is cepstral pitch estimation, first proposed by Noll in [Noll, 1967]. The Cepstrum is defined as the inverse Fourier transform of the logarithm of the Fourier transform of the windowed signal. In the case of cepstral pitch estimation, the signal is first windowed with a Hamming window of at least 512 points. The cepstrum of this signal is computed, and the peak cepstral value and its position determined. If this peak exceeds a pre-determined threshold, the portion of the signal is declared voiced, and the position is used to compute the pitch period. If the peak does not exceed the threshold, a zero-crossings count is made, and if this exceeds a certain threshold, the section is termed unvoiced. Otherwise, it is called voiced and the period of the cepstrum is determined by the maximum value position of the cepstrum. A heuristic decision loop is suggested and often used for prevention of *pitch doubling* and *pitch halving* errors.

Another system which is used in practical speech recognition systems (with some variations) is based on the Linear Predictive Coding (LPC) theory by Atal [Atal et al, 1971]. The first step in this pitch detection system is the voiced/unvoiced decision, which is actually performed by a trained pattern recognition system. The sampled speech is low pass filtered and decimated to about 2 kHz sampling rate. A 41-pole LPC analysis is done on this signal, and the resulting coefficients are spectrally flattened using a Newton transformation. The position of the peaks in this representation gives the pitch of the signal at the 2 kHz rate, which is then interpolated to give a higher resolution.

2.4. Mixed Mode Pitch Determination

In 1951 Licklider proposed a *duplex* theory of pitch perception [Licklider, 1951], to account for many properties of pitch perception, including the perception of the missing fundamental as well as the pitch of modulated noise. Licklider imagined a neural system that measured the running temporal autocorrelation in each auditory frequency channel. A frequency channel is simply a section of the basilar membrane that can be assigned its own characteristic frequency to which it maximally responds. The sensation of pitch is then determined by the common periodicities observed across channels.

In 1983, Lyon simulated an implementation of the duplex theory [Lyon, 1983], and labelled the simulation output the *correlogram*. In general, simulations using these models provide a graphical output that correlates well with pitch. The time lag of the peak in the summary correlogram

(summation of the autocorrelation function analysis across channels) is usually found to be the reciprocal of the frequency of the perceived pitch and the height of the peak is often correlated with pitch salience.

Since the attempts by Lyon, this area of research in modelling the pitch perception has seen great activity. Meddis [Meddis, Hewitt, 1991], [Meddis, O'Mard, 1997], formalised the simulations and added the final summary autocorrelation stage. Cariani [Cariani, Delgutte, 1996] also showed that similar processing of measured auditory nerve impulses is sufficient to predict many pitch perception phenomena. Other researchers have replaced the autocorrelation function with different mechanisms that measure the temporal intervals in each channel [de Cheveigné, 1998].

2.5. Multiple Pitch tracks from simultaneous speech

A reliable algorithm for multiple pitch tracking is critical for many auditory processing tasks such as computational auditory scene analysis (CASA), prosody analysis, speech enhancement and recognition. This is because pitch is an important cue in the segregation of vowel sounds in speech [Meddis, Hewitt, 1992]. However, due to the difficulty of dealing with the interference from noise and mutual interference among multiple harmonic structures, the design of such an algorithm is very challenging and most existing pitch determination algorithms are limited to clean speech or a single pitch track in modest noise. Certain speech related applications, like speech and speaker recognition, would greatly benefit from a system which is able to detect speech from a target speaker, in the presence of other speaker(s). This target speaker's speech may not always be dominant (discontinuous background speech may be detected as foreground speech in certain instances). Therefore, a pitch estimation system that is able to detect and track pitch from one speaker may be used in these applications to ensure continuity of performance in the presence of background speech.

An ideal pitch estimation system for most applications should perform robustly in a variety of acoustic environments. However, the restriction to a single pitch track limits the types of background noise in which these algorithms can perform usefully. For example, if the noise background contains harmonic structures such as background music or voiced speech, a multiple pitch tracking algorithm is required for providing useful results. The background concurrent speech may be treated as noise in single pitch tracking systems, however, a streaming architecture is needed to make sure that the system can track the pitch of a target speaker, and assign the

right pitch to each stream. This problem lies outside the scope of single pitch tracking systems, as there is no inherent concept of target, foreground or background streams.

There have been several proposals made for multiple pitch tracking systems, for example see [Assmann, Summerfield, 1990], [Tolonen et al, 2000], [de Cheveigne et al,] and [Weintraub, 1986]. However, these systems are designed and tested for either vowels or synthetically generated signals, and have not been extensively evaluated in terms of statistical performance on a longer speech signals. Many of these systems are perceptual models of the segregation of sound streams, and not necessarily multiple pitch period estimation systems.

Most of the multiple pitch estimation systems, including the one proposed in this dissertation, are designed to estimate two simultaneous pitch tracks and have several processing stages in common. Here is a brief account of the various stages involved in the estimation of two pitch tracks from simultaneous speech from two speakers with different fundamental frequencies.

The first stage is to simulate the middle and outer ear low and high frequency attenuation effects. This is followed by simulation of mechanical frequency-selectivity of the basilar membrane. The third stage is simulation of mechanical to neural transduction at the inner hair cells. This lead to an output comprising multiple “channels” of activity, almost always tonotopically distributed with frequency overlap between channels. The fourth stage is the calculation of a running autocorrelation function in each individual channel. The fifth and the sixth stage differ in different models. In some models, like [Meddis, Hewitt, 1992], these stages involve first computation of a summary autocorrelation function by summing activity in all channels, and then picking up peaks as candidates of pitch (first and second peaks which are not harmonically related in case of Meddis and Hewitt). In other models, like that of Wu et al [Wu et al, 2002], this process is more involved. It consists of picking up peaks and channels selectively, based on a certain statistical hypothesis, and then evaluation of that hypothesis, operating under some global constraint (explained later in this chapter in detail).

2.6. Detailed Description Of Pitch Estimation Systems Used In This Study

There are many different pitch estimation algorithms, and new ones continue to appear. However, most of them are based on similar principles and therefore share the same strengths and weaknesses. Nevertheless, the evaluation of most of these systems in different noise

conditions and signal resolutions appears to be lacking. Indeed, to the best of our knowledge, the most prominent of these systems have yet to be tested under the above mentioned conditions.

The main reason for the testing and analysis, such as that carried out in this thesis is to evaluate the practical usefulness of these systems as noise robust pitch estimators. The auditory system in mammals has evolved over millions of years. The ability to perceive threats, prey and predators, and to communicate with others, are essential to survival. The barn owl for example, perceives the slightest rustle in the field, in the presence of other noises, in order to hunt, with a very advanced and specially adapted auditory system [Konishi, 1993]. The robustness of auditory perception under such conditions is not only an advantage, it's a requirement.

It is therefore quite surprising that most of the models of auditory pitch perception have not been tested for their robustness under challenging noise conditions. One of the reasons for this situation is the fact that it has proved quite difficult to develop a model which can perform well (in comparison to the human subjects) even in situations where no noise is present. Also, too often too much effort has been spent on explaining all the different psychoacoustical quirks of human pitch perception, and robustness in noise has been paid too little attention. Most recent auditory pitch perception models have not been extensively tested in noise. Practical pitch perception systems, like the ones we shall describe next, have also not been extensively tested for robustness in challenging noise environments.

Multiple pitch tracking has proved to be a very difficult problem in its own right. Although there are systems that have been used in the context of computational auditory scene analysis to handle these problems, the approach has been to carry out streaming experiments, and the systems have not been extensively tested for their ability to produce accurate multiple pitch tracks, until recently [Wu et al, 2002]. This problem is also an interesting one for music applications. Therefore one of the aims of this work is to evaluate the reference systems that are describe below for both robustness in noise, and their ability to track pitch in the presence of interfering speech.

The systems that are described next are among the fore-front of research in the area of pitch estimation systems. Special attention to their description is paid here as their performance is evaluated and compared with the performance of the proposed system.

2.6.1. PRAAT – Window Normalised Autocorrelation Based Pitch Estimation

The PRAAT pitch estimation algorithm proposed by Paul Boersma in [Boersma, 1993] is implemented in the publicly available package on the internet [Boersma, 2002]. It is a simple periodicity detection system that operates in the autocorrelation (lag) domain, and produces accurate pitch estimates and harmonic to noise ratios (HNR). It is better than other systems, in terms of finding the right peak in the lag domain, based on the autocorrelation principle because it performs normalisation of the autocorrelation domain representation of the signal with the window function, before computation of the pitch estimates. It claims robustness in noise in principle, but in the original paper, only synthetic signals like pulse trains and sine waves were used to evaluate performance in noise, and not speech. It has been evaluated for speech signals elsewhere [de Cheveigné, Kawahara, 2002], however in those studies, no noise evaluation was carried out.

Computation of Normalised Autocorrelation:

The autocorrelation of a periodic and stationary signal $x(t)$ is given by the equation 2.2.

$$r_x(\tau) \equiv \int x(t)x(t + \tau)dt \quad \dots 2.2$$

The function $r_x(\tau)$ evaluated at lag τ , corresponds to a frequency given by $1/\tau$. The function $r_x(\tau)$ at lag 0 is the power of the signal, and the normalised autocorrelation, $r'_x(t)$, is given by equation 2.3.

$$r'_x(t) = \frac{r_x(\tau)}{r_x(0)} \quad \dots 2.3$$

However, most of the speech signals are not stationary, and in order to use autocorrelation in this context, the signal is windowed. The window function used is normally maximum in the middle portion of the signal and tapered to zero towards the edges. In this system, in addition to using the normalised autocorrelation of the windowed signal, the system also normalises the autocorrelation function $r'_x(\tau)$ with the normalised autocorrelation function of the window ($r_w(\tau)$) which was used to make the portion of the signal under consideration stationary.

$$r_x(\tau) \approx \frac{r_x(\tau)}{r_w(\tau)} \quad \dots 2.4$$

The reason for this further normalisation is to undo the tapering effect the windowing process has on the higher lags. In situations where the second harmonic of the fundamental is high in energy, the lack of this normalisation leads to octave jumps upwards on the frequency scale. This is a serious problem especially for telephone quality speech, where the fundamental frequency component is usually missing. The actual computation of autocorrelation is done through a Fast Fourier Transform (FFT).

Calculation of Harmonics to Noise Ratio and Parabolic Interpolation:

The autocorrelation of a signal at zero lag equals the power of the signal. For normalised autocorrelation, the maximum at τ_{max} represents the relative power of the periodic or the harmonic part of the signal, and its complement represents the relative power of the noise component. Thus the harmonic to noise ratio is given by equation 2.5.

$$HNR = 10 \text{Log}_{10} \left(\frac{r_x(\tau_{max})}{1 - r_x(\tau_{max})} \right) \quad \dots 2.5$$

For perfectly periodic sounds, the HNR is infinite.

The HNR is used in this system to determine, based on the value of the ratio and a fixed threshold, whether a portion (frame over which HNR is calculated) is voiced (has a higher HNR with respect to the fixed threshold), or unvoiced (lower HNR with respect to the fixed threshold).

This pitch detector aims to detect the fundamental frequency very accurately. However, the sampling rate of the digital signal places inherent restrictions on the accuracy with which the frequency of any of its components can be measured, proportional to the sampling rate. These problems are overcome in the lag domain by up-sampling the signal in the frequency domain (which is the intermediate domain for calculation here). This interpolation is done both for the frequency domain representation of the window function, and the frequency domain representation of the signal, according to the parabolic interpolation of $\sin(x) / x$ equation, where

x is the frequency domain variable. This interpolation makes the system more robust to aliasing problems in higher frequency regions (close to the Nyquist limit).

Precision and Figures of Performance:

For pure periodic functions, the accuracy of the algorithm is claimed to be 10^{-8} in the lag domain. Undoubtedly, this makes the system the most accurate system for sampled data. However, for a sine wave with a frequency of 206 Hz and a window of 40 ms, with noise (white) added at 20 dB SNR level, there are 40% local octave errors. If the proposed global path finder is used (the global path finder is a system that weighs the potential pitch candidates with cost functions for voiced unvoiced transitions and octave-jumps), the octave errors are reported to reduce to 0%. However it is also acknowledged that for a dynamically changing signal, this may not be possible.

2.6.2. YIN: A recent fundamental frequency estimator for speech and music

The YIN algorithm is a recent periodicity estimator for speech that we have come across in the literature [de Cheveigné, Kawahara, 2002]. It is also well evaluated (by the authors) on different databases, claiming impressive gross error rates for clean speech in several speech databases, hence our motivation for including this system in this evaluation and comparison. The system is comprised of techniques used by many other systems in a unique way. Like the system described above, it is based on the autocorrelation calculations, and relies on parabolic interpolation for accuracy. It uses a cumulative mean normalised difference function (explained later in the text) for a local search for the best peak in the lag domain representation of the signal to look for better (more accurate) estimates. The system is described with an outline of the processing steps and claimed performance below.

The Method:

The first step in the method used for computing pitch in the YIN system is that of autocorrelation, which forms the front end of the system. There are no references to the normalization of the autocorrelation function with the autocorrelation of the window used in the paper [de Cheveigné, Kawahara, 2002], and this does not appear to be the case on code inspection.

The next stage is the calculation of the difference function. Mathematically, the difference function is defined in terms of the autocorrelation function $r_i(\tau)$ in equation 2.6. below.

$$d_i(\tau) = r_i(0) + r_{i+\tau}(0) - 2r_i(\tau) \quad \dots 2.6$$

However, according to de Cheveigne, the difference function is quite sensitive to amplitude changes, and therefore, in the next stage of processing, a “cumulative mean normalized difference function” is computed, as below.

$$d'_i(t) = \left\{ \begin{array}{l} 1, \quad \text{if } \tau = 0 \\ d_i(\tau) / \left[\frac{1}{\tau} \sum_{j=1}^{\tau} d_i(j) \right] \text{ otherwise} \end{array} \right\} \quad \dots 2.7$$

The next stage is to set an absolute threshold and choose the smallest value of τ that gives minimum of d' (less than the absolute threshold). This threshold was set to a value of 0.1.

Parabolic interpolation is implemented by choosing the minimum value of $d'(\tau)$, and interpolating this value by fitting a parabola to the neighbouring values and choosing the minimum point on the parabola.

The last stage of the algorithm finds the best local estimate of $d'(\tau)$. For each time index t , search is performed for the minimum of the function $d'_0(T_0)$, for parameter θ in the vicinity of t , i.e., in the range $T_0 = [t - T_{\max}/2, t + T_{\max}/2]$, where T_0 is the estimate at time θ and T_{\max} is the largest expected period. The typical value of T_{\max} was reported to be 25 ms.

The Performance Claims:

The YIN system is quite well tested, and performance figures were reported for several databases, with comparisons to many other similar systems. However, the evaluation was carried out without any additive noise, or variations in the signal resolution. The system was reported to have the best performance compared to the other evaluated methods. The averaged figures indicate a better performance by about a factor of 3. Over 99% of the estimates were accurate within a tolerance of 20% of the reference pitch data, 94% within 5% tolerance, and 60% within a tolerance of 1%.

2.6.3. The Auditory Toolbox – Slaney’s Correlogram based Pitch Estimation

This is the classic correlogram pitch estimation system proposed and implemented by Slaney and Lyon [Slaney, Lyon, 1990]. Similar, but more advanced system is presented in [Meddis, Hewitt, 1991] by Meddis and Hewitt, based on the auto-correlogram computation. However, only the auditory toolbox based system was evaluated because the system described by Meddis and Hewitt is much more computationally expensive. The system included in the auditory toolbox by Slaney is based on Licklider's [Licklider, 1951] “Duplex Theory” of pitch perception, and tested on a variety of stimuli from human perceptual tests. It is aimed to accurately model the way humans perceive pitch. They claim that it correctly identifies the pitch of complex harmonic and inharmonic stimuli, and that it is robust in the face of noise and phase changes. This perceptual pitch detector combines a cochlear model with a bank of autocorrelator units. By performing an independent autocorrelation for each channel, the pitch detector is relatively insensitive to phase changes across channels. The information in the correlogram is filtered, nonlinearly enhanced, and summed across channels. Peaks are identified and a pitch is then proposed that is consistent with the peaks.

The representation used by the pitch detector, which corresponds to the output of Licklider's duplex theory, is the correlogram. This representation shows the spectral content and time structure of a sound on independent axes of an animated display. A pitch detection algorithm analyses the information in the correlogram and chooses a single best pitch. The system does not address the decision of whether there is a valid pitch (the voiced/unvoiced decision), nor does it attempt to enforce or utilize frame - by - frame continuity of pitch.

The Model:

According to Slaney and Lyon [Slaney, Lyon, 1990], the human auditory system can be simplified to three processing stages, and the model they implemented, is broadly based on these stages. These are described below.

The Cochlear Model: A cascade of second order filters is used to model the propagation of sound along the Basilar Membrane (BM.) At each point along the cochlea the BM responds best to a broad range of frequencies and it is this movement that is sensed by the Inner Hair Cells. The “best” frequency of the cochlea varies smoothly from high frequencies at the base to low frequencies at the apex. Inner Hair Cells only respond to movement of the BM in one direction.

This is simulated in the cochlear model with an array of Half Wave Rectifiers (HWRs) that detect the output of each second order filter. The HWR non-linearity serves to convert the motion of the BM at each point along the cochlea into a signal that represents both the envelope and fine time structure. Finally, four stages of Automatic Gain Control (AGC) allow the cochlear model to compress the dynamic range of the input to a level that can be carried on the auditory nerve. The multiple channel coupled AGC used simulates the ear's adaptation to spectral tilt as well as to loudness.

The Correlogram: If a sound is periodic, the autocorrelation functions for all cochlear channels can be assumed to show a peak at the horizontal position that corresponds to a correlation delay equal to the period of repetition. This is generally equal to the perceived pitch period. Since the peaks in all channels, or rows of the image, occur at the same delay, or horizontal position, they form a vertical line in the image. This is based on the “duplex” theory, which says that sounds with a perceived pitch, even if they are not periodic, will produce a vertical structure in the correlogram at the delay related to the perceptual pitch. On the other hand, formants, or narrow resonances in the frequency domain, are displayed as horizontal bands in the correlogram. The correlogram is computed by finding the (short-time, windowed) autocorrelation of the output of each cochlear frequency channel.

The Pitch Estimation System: The pitch estimator consists of four steps. A preprocessing step modifies the correlogram to enhance the peaks. In the second step, the values at each time lag in the enhanced correlogram are then summed across all frequencies. Peak locations at this stage give estimates of all the possible periodicities in the correlogram. The third step is to combine evidence at the subharmonics of each pitch to make the pitch estimate more robust. Finally, the largest peak is picked, being careful to avoid octave errors, and a numerical value of the pitch is determined based on the location of the peak. The system uses a technique described by Nishihara [Nishihara, Crossley, 1988] to judge the location of the pitch peaks. In general the peaks in the pitch function are symmetric and an accurate estimate of their centre is made by fitting a polynomial to the points near the peak. Using multiple points to determine the location of the peaks allows the pitch period to be determined with a resolution finer than the sampling interval (in low noise situations), and a more robust estimate to be made when noise is present.

The Performance Claims:

The original model was not evaluated in a systematic way for noise or varying signal resolutions. It was however demonstrated that the system can emulate various perceptual effects such as the phenomenon of *virtual pitch* when the fundamental frequency is missing or of very low energy.

2.6.4. The Probabilistic Multiple Pitch Tracking System

The probabilistic multiple pitch tracking system that is described here was published most recently [Wu et al, 2002], and is one of the few multiple pitch tracking systems whose performance has been tested on speech signals for two simultaneous speakers. The software to simulate the model was obtained by a request to the authors. However, in its current configuration, the system supplied works only on signals sampled at 16 kHz sampling rate, and was not used for all of the experiments that were performed, but only for the two simultaneous speakers test, sampled at the required 16 kHz sampling rate.

The system is based on the processing of a summary autocorrelation function of a large number of channels. This is followed by a probabilistic processing step on the output of the autocorrelation function computation, using Hidden Markov Models (HMM) to estimate the two (or less) pitch tracks. The model is described in detail below.

The Model:

The algorithm consists of four stages.

In the first stage, the input signal is sampled at 16 kHz and then passed through a bank of 128 fourth order gamma-tone filters [Patterson et al., 1987]. The frequency channels are classified based on their centre frequencies as either belonging to the low-frequency group, or the high frequency group, with the channels having their centre frequency below 800 Hz belonging to former (channels 1–55), and the rest to the latter group (56–128). The high frequency channels have their output passed to an envelop estimation system. A normalized autocorrelation function is then computed on the envelopes of the higher frequency channels and the output of the low frequency channels (computation is performed in each channel separately at the rate of 10 ms, using a window size of 16 ms).

Channel and peak selection comprise the second stage. For low frequency channels, the strength of the autocorrelation function output is considered. If the autocorrelation function in a channel has a peak greater than a fixed threshold at the lag corresponding to the centre frequency of the channel, it is concluded that the corresponding frequency component is present in the signal, and the channel is selected for further processing, otherwise, it is rejected. For high frequency channels, another autocorrelation function is evaluated, this time with a larger window. If the difference between the two outputs (standard autocorrelation and larger window autocorrelation) is not large, the channel is selected, otherwise it is rejected. A local search method is used for the selection of peaks in the autocorrelation function output of high frequency channels. If the peak is above a certain threshold, and there is another peak at a lag corresponding to double the period, the original peak is kept, otherwise it is removed. Another method is used in conjunction to this method. If a strong peak in the high frequency channels is found, all the multiple peaks corresponding to other periods are removed. This process is aimed at reducing the errors due to multiple and sub-multiple pitch peaks in the autocorrelation functions.

The third stage comprises the probabilistic calculation of pitch periods and time lags of selected peaks, achieved by studying the statistical fit of the selected peaks to a particular pitch period hypothesis, based on the normalised autocorrelation function. First, the contribution of each frequency channel to a pitch hypothesis is calculated. Then, the contributions from all channels are combined into a single score. By studying the statistical relationship between the ideal pitch periods and the time lags of selected peaks obtained from the last stage, attempts are made to formulate the probability of a channel supporting a pitch hypothesis, using a statistical integration method for producing the conditional probability of observing the signal in a time frame given a hypothesized pitch period.

The final stage acts as a probabilistic pitch tracking system, given the different pitch hypotheses from previous steps. The system uses an HMM for approximating the generation process of harmonic structure in natural environments. The hidden nodes represent the possible pitch states (one pitch, two pitches, no periodicity) given the observation nodes. The observation nodes are represented by the set of selected peaks and lags from each time frame. In the final step the state-spaces for one, two or zero pitch states are discretised and the Viterbi algorithm is employed for finding the optimal sequence of states.

Performance Claims:

The system is claimed to reliably track pitch in various situations, one speaker, speech mixed with other acoustic sources, and multiple (two) speakers' speech. The system was compared to another multiple pitch tracking system by Tolonen et al [Tolonen, et al, 2000]. The results show huge advantage in terms of performance improvement over this system, with gross error rates improved by nearly four times. The absolute gross error rates for the dominant pitch for the multiple pitch case (two simultaneous talkers) was 0.93%, while the gross error rates for the non-dominant pitch are not given.

In this chapter, a number of pitch estimation systems and techniques were presented. Their performance and claims were discussed, and put in a historical context in terms of the development of advanced pitch estimation algorithms. It was realized that none of the systems have been evaluated in a statistical sense for varying noise conditions, or for different signal resolutions. Although attempts have been made to carry out comparative evaluations of different pitch estimation algorithms, these factors have generally been omitted from the analysis.

Chapter 3

DAMPED HARMONIC OSCILLATORS BASED SIGNAL PROCESSING

The Front-End Analysis

Speech signal processing plays a fundamental role in all speech related research, whether it is modelling psychoacoustical data, or coding for compression and transmission or storage of the signal. During the course of long years of speech processing, a variety of signal processing techniques have been developed and used. These include the Fourier transform and the Fast Fourier Transform and related techniques, digital filters and banks of digital filters, the linear predictive coding technique, and other stochastic signal analysis techniques. All of these techniques are suitably adapted for speech processing and collectively called front-end processing techniques. The bank of damped harmonic oscillators is a novel front-end signal processing technique, inspired by Helmholtz's model [Helmholtz, 1870] of basilar membrane processing, but grounded in the theory of damped oscillators and signal processing.

One of the most useful ways of characterizing speech is in terms of an acoustic waveform, called the speech signal. For the processing of this signal by means of computers, the acoustic waveform is converted into a current waveform, using a microphone or other such device. This continuous time current waveform representing the acoustic analogue waveform is achieved through a transducer like a microphone. The current waveform is then *digitised* by sampling the signal at fixed (and very short) intervals of time, to obtain the digital speech signal, using an electronic A/D (analogue to digital) converter device. The quality of the digital signal thus obtained is a function of the sampling interval, and the number of digital bits used to represent a single sampled value of the signal, the smaller the sampling interval, and larger the number of bits used for representation, the higher the quality of the digital signal. Most of the speech processing research is done on this representation of the acoustic signal. Using an inverse mechanism, involving the application of D/A (digital to analogue) converter, and a transducer like a sound speaker, the processed speech waveform can be converted back into an acoustic waveform.

Once a digital signal representation of the acoustic waveform is obtained, the representation is a one-dimensional time-amplitude signal, which is very difficult to analyse in this 'raw' form. The difficulty arises because the actual speech sounds are pseudo-periodic signals composed of different frequencies, each of these frequency components being present in varying degrees of strength, which may change over time. It is pseudo-periodic because the frequency components described above change with time (in both the frequency of oscillation and the amplitude with which they oscillate). Mathematically, we can state this by equation 3.1.

$$s(t) = a_1(t)\Psi(f_1) + a_2(t)\Psi(f_2) + a_3(t)\Psi(f_3) \dots + a_n(t)\Psi(f_n) \quad \dots 3.1$$

It is the task of all the front-end techniques to estimate the contribution of components of frequencies f_i , given a_i at a time instant denoted by variable t . We shall take a brief look at the most commonly used techniques to do this analysis, and establish the necessary signal processing background before presenting the proposed technique.

3.1. Pre-processing Methods for Speech Analysis

Research into efficient and robust front ends is a very active area of research on its own. In speech processing, there are three different techniques on which most of the speech research is based. All these techniques, at least as an intermediate stage, transform the signal to a representation in the frequency domain, in order to calculate the coefficients a_i in equation 3.1.

3.1.1. Fourier Analysis and Related Techniques

For Fourier analysis of any signal, the frequency based function Ψ in equation 3.1 is modelled as a complex exponential. Thus the Fourier model of the signal is represented by equation 3.2.

$$s(t) = \int A_f e^{2\pi f t} dt \quad \dots 3.2$$

The complex exponentials are called the basis functions in Fourier terminology [Kammler, 2000]. The term A_f is the amount of discrete exponential $e^{2\pi f t}$ that must be used in the recipe for the signal $s(t)$. The summation operation in equation 3.1 is replaced by an integral operation in equation 3.2. For sampled signals, the Fourier synthesis equation 3.2 is given by equation 3.3, and is called the Discrete Fourier Transform (DFT) pair.

$$s[n] = \sum_{k=0}^{N-1} A[k] e^{2\pi i k n / N}$$

$$A[k] = \frac{1}{N} \sum_{n=0}^{N-1} s[n] e^{-2\pi i k n / N} \quad \dots 3.3$$

The equation 3.3 is called the DFT analysis/synthesis pair. The top equation is the signal model, and the bottom equation defines the coefficients $A[k]$. The variables t (time) and f (frequency) have been replaced in equation 3.3 by $[n]$ and $[k]$ to emphasise the discrete nature of these calculations. The DFT analysis is computationally expensive in the form described by equation 3.3. The FFT (Fast Fourier Transform) algorithm uses the redundancies in this calculation to make the computation of the DFT much more efficient.

The Fourier Transform based analysis can also be used for the computation of the cepstrum of the signal. Cepstrum is defined as the Fourier transform of the log of the power spectrum of the signal [Oppenheim, Schaffer, 1975]. It was observed that for periodic signals the spectrum is itself periodic, and that the Fourier transform of the power spectrum provides this period. The cepstrum serves as a log compressed representation, thus reducing the difference in energy of the various frequency bands. FFT computation can be used for computing the Mel Scaled Cepstral Coefficients (MFCC) computation. MFCC representation is similar to cepstral representation, but the cepstral output is scaled according to a warped frequency scale, which is based on the place model of the basilar membrane frequency selectivity [Rabiner, Juang, 1993]. FFT is also used for efficient computation of the autocorrelation function of the signal. In general, place models of pitch estimation also work on the Fourier representation of the signal. Computation of the instantaneous frequency based pitch estimation also depends on the FFT analysis presented above.

The main advantage of using the Fast Fourier transform is its computational efficiency. However, the signal model assumed in the analysis is that of a stationary, periodic signal. For non-stationary pseudo-periodic signals like speech, the Fourier analysis of the raw speech signal would give erroneous results. In order to perform Fourier analysis of the speech signal, it is first split into small sections or frames, and multiplied by a window to make it stationary and remove the effects of splitting the signal into smaller parts (the window functions used generally have tapering edges to remove these effects). This technique of Fourier analysis is called the Short

Time Fourier Transform (STFT). However, this process limits the frequency and temporal resolution of the signal. The frame size determines the frequency resolution (the number of frequencies for which the analysis can be performed). The temporal resolution is reduced depending on the frame size and the overlap between frames, by limiting the number of instants at which the spectral estimate is available. However, making the frame size smaller reduces the frequency points at which the estimate of the signal energy is available. The STFT also introduces inherent errors in the frequency analysis, namely the short term and long term spectral leakage. These are effects of sampling, and the fact that the estimate is only available for discrete frequencies [Kammler, 2000]. Therefore there is a trade-off between accuracy and computational efficiency in the STFT of the signal.

3.1.2. Filter Bank based Analysis

A filter bank is a collection of band-pass filters, with each filter output giving a measure of the energy in the frequency band it is designed for. Filter design theory and their realisation is a subject of great depth, and the reader is referred to [Rorabaugh, 1997] for a complete treatment.

The sampled signal $s(t)$ is passed through a bank of P band-pass filters, giving the band-pass filtered outputs $s_i(t)$, given by the equation 3.4,

$$s_i(t) = s(t) * h_i(t), \quad 1 \leq i \leq P$$

$$s_i(t) = \sum_{m=0}^{M_i-1} h_i(m)s(t-m) \quad \dots 3.4$$

where we have assumed that the impulse response of the i^{th} band-pass filter is $h_i(m)$ with a duration of M_i samples, and the * symbol represents the convolution operation.

Since the purpose of the filter bank is to give a measure of energy of the speech signal in a given frequency band, each of the band-pass signals $s_i(t)$, is first half-wave or full-wave rectified. This non-linear treatment concentrates the energy in the lower frequency region of the output, as well as creating high frequency images. Following this step, the output is low-pass filtered to remove the high frequency images and maintain the DC component of the output. The resulting signals give an estimate of the energy of the frequency components of the original signal in each of the frequency bands of the bank of filters. The mathematical basis of these operations is dealt in full in [Oppenheim, Schaffer, 1975] and [Rabiner, Juang, 1993].

The bank of filters can be realised as an FIR (Finite Impulse Response), IIR (Infinite Impulse Response), or through FFT. A treatment of all these different techniques is beyond the current scope, and not relevant to the current discussion.

The more important part of filter-bank design for speech processing, which is also a subject of much research is the spacing of the bands on the frequency spectrum, the number of such bands (= number of filters), and the bandwidth of each of these filters. The most popular filter banks are designed on the various perceptually relevant scales, like the Mel Scale [Pickles, 1988], the bark scale [Zwicker, Fastl, 1990], or the ERB scale [Glasberg, Moore, 1990]. With some minor differences, the bark scale and the Mel scales arrange the centre frequencies and bandwidths of the filters in the same manner as the experimentally determined frequency profile along the basilar membrane. The result is close to linear spacing for frequencies below 1000 Hz, and approximately logarithmic for frequencies above 1000 Hz. The ERB scale has the filter bandwidths nearly multiplicative as a function of centre frequency of the band-pass filters for frequencies above 1000 Hz. In its most general form, each band-pass filter is implemented via a direct convolution, i.e., no efficient FFT structure can be used. This makes the implementation of these non-uniform filter banks at least 6 to 7 times slower than a uniform filter bank, which may be realised using an FFT computation. Although some “quick and dirty” methods for realising the non-uniform filter banks are available, there are generally not used [Rabiner, Juang, 1993]. Figure 3.1 shows the typical frequency characteristics of filters used in Lyon’s cochlear model [Slaney, 1988], [Slaney, Lyon, 1993].

3.1.3. Linear Predictive Coding Analysis

Linear Predictive Coding (LPC) for speech analysis/synthesis which originated about three decades ago [Atal et al, 1971], is based on an all pole model of the speech signals. LPC is used in most of today’s commercial speech analysis/synthesis systems. With LPC, formant like analysis of the speech signal was made possible, without the need for explicit format tracking, which has proven to be quite problematic. This is because the peaks in the LPC spectrum are usually linked with the active format region of speech. LPC computation was made very efficient with the method of partial correlations [Itakura, 1975].

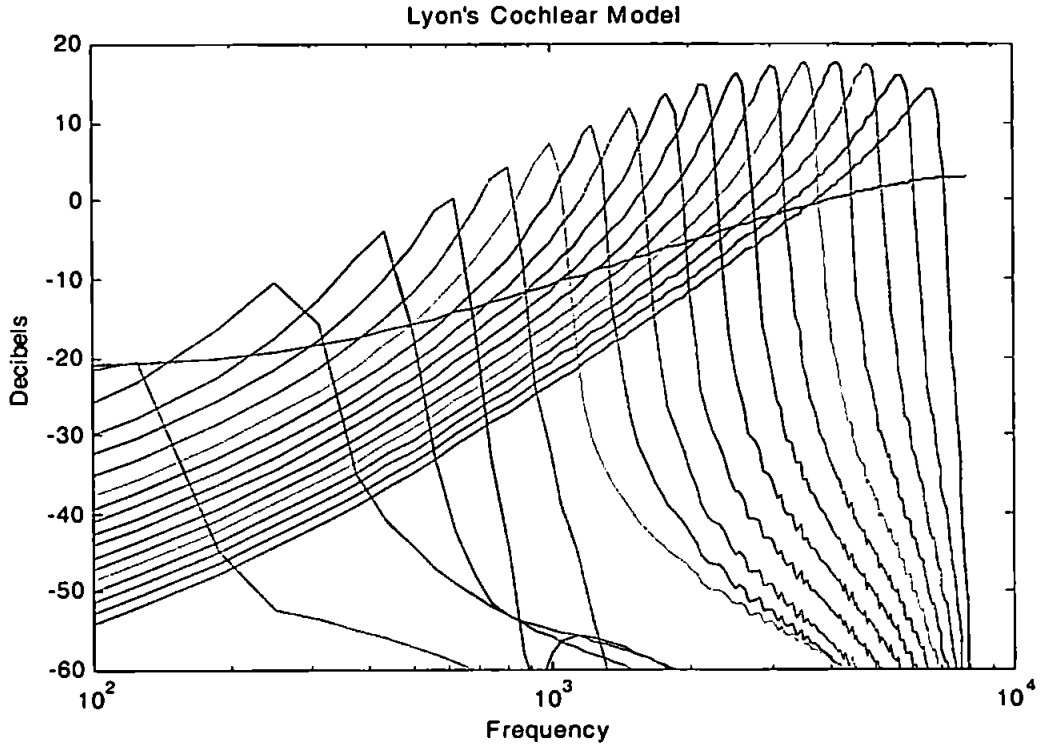


Figure 3.1. The filter characteristics as reported by Slaney and Lyon in [Slaney, Lyon, 1993]. The x-axis represents the frequency, and the y-axis is the gain of each filter in the filter-bank, as a function of the frequency.

The basic idea behind the LPC analysis method is that of the auto-regressive model. That is, given the past p values of the signal, the current sample $s(n)$ can be modelled as the weighted sum of these past p values. The basic assumption is that this model remains constant for the length of the frame (i.e. assuming stationarity for that frame). The mathematical equation for this analysis is given by equation 3.5,

$$s(n) = a_1 s(n-p-1) + a_2 s(n-p-2) + a_3 s(n-p-3) \dots + a_p s(n-1) + \varepsilon \quad \dots 3.5$$

where the ε is the error in the model, and is also modelled as a constant term for each frame.

If the error term ε is modelled as $G.u(n)$, where $u(n)$ is the input to the model, and G is gain, equation 3.5 can be represented as equation 3.6.

$$G.u(n) = s(n) - \sum_{k=1}^p a_k s(n-k) \quad \dots 3.6$$

The transfer function of the model in the z domain is given by equation 3.7b.

$$G.u(z) = s(z) \left(1 - \sum_{k=1}^p a_k z^{-k} \right) \quad \dots 3.7a$$

$$H(z) = \frac{S(z)}{G \cdot U(z)} = \frac{1}{1 - \sum_{k=1}^p a_k z^{-k}} = \frac{1}{A(z)} \quad \dots 3.7b$$

This is the most significant step in the LPC analysis, because not only does it tell us that LPC model can be viewed as an all-pole model, but also as a model of speech production. In this view of the model, the $G.u(n)$ is the excitation signal or the glottal pulse, and the LPC coefficients model the shape of the vocal tract. This also acts as a justification for the stationarity assumption of the model inside the frame, as it is well known that the shape of the vocal tract changes slowly with time. Figure 3.2 shows this speech synthesis model of the LPC analysis.

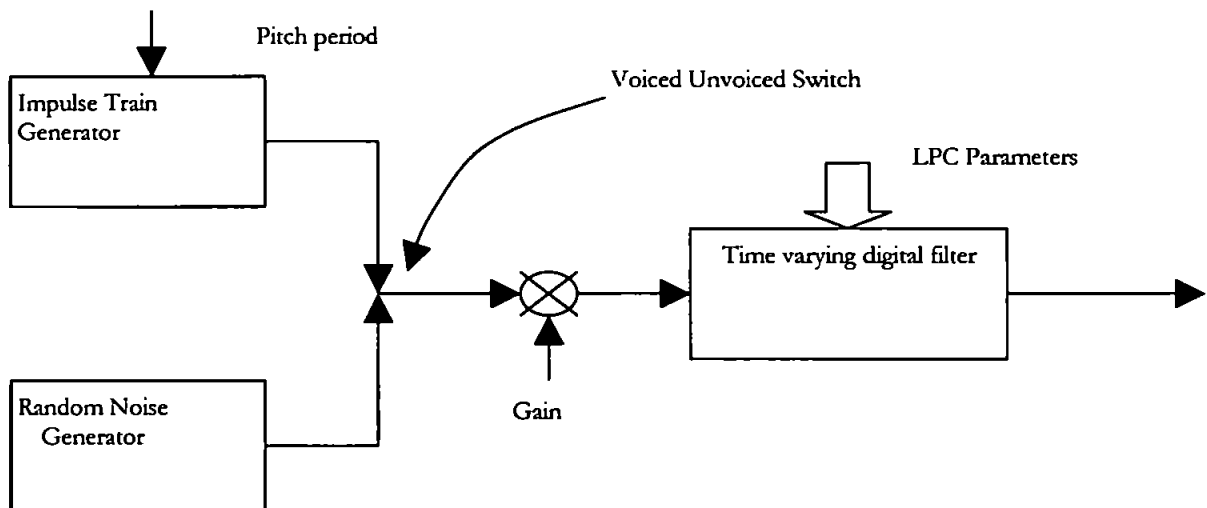


Figure 3.2 Speech Synthesis model of LPC analysis. The pitch period and voiced/unvoiced switch are important parts of this model.

3.2. The Damped Harmonic Oscillator Based Analysis

In this dissertation, a new signal processing front end for the purpose of periodicity analysis is proposed. The proposed system is similar to the bank-of-filters based approach, but is much

more computationally efficient (an analysis of the computational costs and efficiency is included in the appendix A.1. of this thesis). The results of this processing are used to develop a signal representation based on the temporal analysis of the output. The resulting system affords a much higher frequency resolution. The most important claim however, is that this model of speech signal processing and the associated representation that is developed, is robust to added noise, and can perform an effective separation of the periodic part and the a-periodic noise. These claims shall be substantiated by means of examples in this chapter, and by means of an extensive empirical study of performance of the complete system in chapters 5 and 6. It may be noted however, that the property of noise robustness is not derived from the operation of the damped harmonic oscillators alone, but, is a property of the whole system that is proposed in this and the following chapter.

The operation of a damped harmonic oscillator is a conceptually advantageous construct that yields itself to further analysis in terms of temporal properties of the oscillatory behaviour of a mechanical system in order to derive the components of a complex tonal signal. The principles of the operation of the damped harmonic oscillator are described next, followed by an analysis of their frequency and temporal characteristics.

3.2.1. Principle of Operation

The fundamental units of the proposed system, the damped harmonic oscillators, are not an explicit model of the basilar membrane vibration as a result of the acoustic stimulus, but are inspired by its mechanical analogue. From the days of Helmholtz [Helmholtz, 1870], it was widely agreed that the cochlea in the mammalian inner ear acts as a frequency analyser. From the experiments carried out by von Békésy [von Békésy, 1960] it was further inferred that the cochlea acts a mechanical tuning device, with various sections of the basilar membrane selective for particular frequencies of the stimulus. It was also discovered that at least at moderate sound pressure levels, the frequency tuning of the basilar membrane is quite broad, and that there is also significant damping. At the time, these results came as quite a surprise because psychoacoustical tests [Licklider, 1956] had shown that the frequency selectivity in human subjects was much higher than the experimentally measured selectivity at the cochlea. However, currently held views [Yates et al., 1985] based on further experiments on the tuning of the basilar

membrane have shown the selectivity of the basilar membrane to be at par with the selectivity observed in the auditory nerve fibre tuning curves.

From these experimental results, it is clear that the basilar membrane acts as a frequency selection system with the oscillatory movement of the basilar membrane sharply tuned to the frequency of the stimulus [Russel, 1987]. Therefore, the fine frequency distinctions measured in the psychoacoustical tests, must be inferred as encoded in the dynamic nature of the oscillations of the basilar membrane, which are then transmitted to the auditory nerve through the inner hair cells in the inner ear. This is the basic principle of operation of our system as well.

A damped harmonic oscillator [Pain, 1976], [Park, 1964], is a very simple device, that has two design parameters, the characteristic frequency of oscillation, and the damping constant. However, the characteristic frequency only controls the amplitude of oscillation in response to the stimulus, and the oscillator actually oscillates with the frequency of the applied stimulus, with the amplitude of these oscillations inversely proportional to the difference between the characteristic frequency and the frequency of the stimulus, and directly proportional to the amplitude of the original stimulus. This simple design principle means that the system composed of such units is data driven, as the system's dynamic behaviour is driven largely by the stimulus, and not so much by the properties (or parameters) of the system itself. The input stimulus "drives" the damped harmonic oscillator at a frequency that is determined by its own frequency, and the design parameters influence the amplitude of the activity. This situation is analogous to that of the basilar membrane with broad with the fine dynamic behaviour (fine in terms of temporal mechanical oscillations) encoding the periodicity of the actual signal. The damping constant of the damped harmonic oscillator determines the effective bandwidth of the oscillator, by enforcing a time constant related to the oscillatory behaviour, and is analogous to the stiffness of the basilar membrane.

3.2.2. Dynamic Operation and Derivation of the System Equations

The Damped Harmonic Oscillator (DHO) is a unit that oscillates preferentially to a signal with frequency close to its characteristic frequency. In the absence of a close frequency component, the oscillations are very small in amplitude and decay asymptotically towards zero amplitude. The state variable of a DHO can be described concisely as a complex number z as in equation 3.8.

$$z = x + iy \quad \dots 3.8$$

The dynamic operation of the DHO unit is controlled by equation 3.9.

$$\frac{dz}{dt} = (b + iw)z \quad \dots 3.9$$

For clarity let,

$$k = b + iw \quad \dots 3.10$$

Then, the solution to equation 3.9 can be written as the standard solution in equation 3.11.

$$z = e^{kt} \quad \dots 3.11$$

Therefore, using equations 3.11 and 3.10,

$$z = e^{bt} e^{i\omega t} \quad \dots 3.12$$

Then, if

$$\begin{aligned} x &= \operatorname{Re}(z) \\ y &= \operatorname{Im}(z) \end{aligned} \quad \dots 3.13$$

we get,

$$\begin{aligned} x &= e^{bt} \cos(\omega t) \\ y &= e^{bt} \sin(\omega t) \end{aligned} \quad \dots 3.14$$

using 3.9 and 3.13,

$$\begin{aligned} \frac{dx}{dt} &= \operatorname{Re}(z) = \operatorname{Re}((b + iw)(x + iy)) \\ \frac{dy}{dt} &= \operatorname{Im}(z) = \operatorname{Im}((b + iw)(x + iy)) \end{aligned} \quad \dots 3.15$$

Therefore the final dynamical system takes the form of equation 3.16.

$$\begin{aligned}\frac{dx}{dt} &= bx - wy \\ \frac{dy}{dt} &= by + wx\end{aligned}\quad \dots 3.16$$

In the above solution, the time subscripts have been ignored for clarity. The parameter b is the damping constant of the oscillator, and determines the rate of decay of the variables x and y with time. The parameter w is the characteristic frequency of the damped oscillator. From equation 3.14 it is clear that the variable b should be negative for damped oscillations.

The equation 3.12 defines the behaviour of the impulse response. In the presence of a continuous time real valued signal s , the system can be treated as a forced damped oscillator. In that case, the dynamic equation takes the form of equation 3.17.

$$\begin{aligned}\frac{dx}{dt} &= bx - wy + s \\ \frac{dy}{dt} &= by - wx\end{aligned}\quad \dots 3.17$$

In the case of digital signals, the signal is only specified at a certain rate, known as the sampling rate. To implement the system described by equation 3.17, we need to have a discrete time version of the equation. These equations depend on the sampling frequency of the input signal, f_r . Under these assumptions, the equation 3.17 can be modified as equation 3.18.

$$\begin{aligned}\frac{x(t+1) - x(t)}{\Delta} &= bx(t) - wy(t) \\ \frac{y(t+1) - y(t)}{\Delta} &= by(t) + wx(t)\end{aligned}\quad \dots 3.18$$

Leading to the final system of difference equations 3.19.

$$\begin{aligned}x(t+1) &= (1 + \Delta b)x(t) - w\Delta y(t) + s(t) \\ y(t+1) &= (1 + \Delta b)y(t) + w\Delta x(t)\end{aligned}\quad \dots 3.19$$

The term Δ is the time interval between two samples, and is determined as the reciprocal of the sampling rate and ω is the angular frequency. The output of the DHO units is the corresponding y variable in equation 3.19.

3.2.3. Magnitude and Phase Response of the DHO Units

The DHO unit as described in equations 3.17 has its Laplace transform as described by equation 3.20, when the initial values are taken as zero.

$$\begin{aligned} sX &= bX - \omega Y + I \\ sY &= bY + \omega X \end{aligned} \quad \dots 3.20$$

Where I is the input variable and Y is the output variable. The complex frequency variable s for the variables X and Y is implicit. Rearranging the variables, and substituting yields the transfer function of the DHO unit, as described by equation 3.21.

$$\frac{Y}{I} = \frac{\omega}{(s-b)^2 + \omega^2} = H(s) \quad \dots 3.21$$

The system defined by the transfer function of the form in equation 3.21 has a steady-state response $H(j\omega)$ which can be obtained by evaluating the transfer function $H(s)$ at $s = j\omega$. The magnitude response is simply the magnitude of $H(j\omega)$, i.e.,

$$|H(j\omega)| = [\text{Re}(j\omega)^2 + \text{Im}(j\omega)^2]^{1/2} \quad \dots 3.22$$

Figure 3.3 shows the magnitude response of a typical DHO unit. The magnitude response of a comparative second order filters with Q factors of 20 and 50 are shown in figure 3.4 for comparative purposes. The second order filters used for comparison were designed using the Auditory Toolbox provided by Slaney [Slaney, 1998]. The transfer function of the filters used is given by equation 3.23.

$$H(s) = \frac{1}{s^2 + a_0s + \omega_0^2} \quad \dots 3.23$$

* The Q factor or the Quality factor of the second order filter determines its selectivity, a higher Q being more selective than a lower Q . Q factor can be seen as the ratio of the energy of the system to the energy dissipated in one cycle. For more on second order filters theory, please see [Rorabaugh, 1997]

where a_0 is given by w_0/Q , where w_0 is the centre frequency and Q is the quality factor.

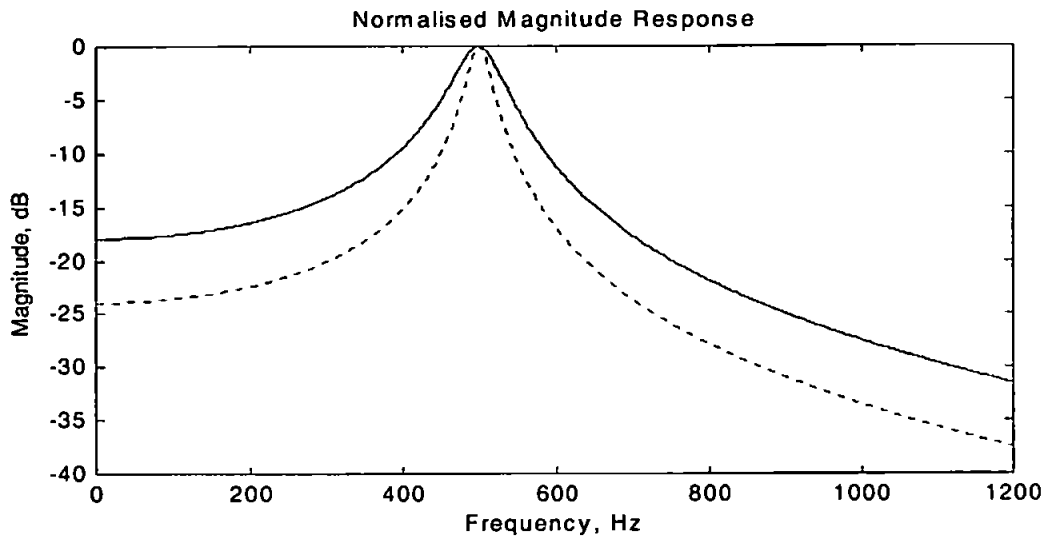


Figure 3.3. The normalised magnitude response of a single DHO unit with design frequency of 500 Hz. The continuous line is for $b = -60$, and the dashed line is for $b = -30$.

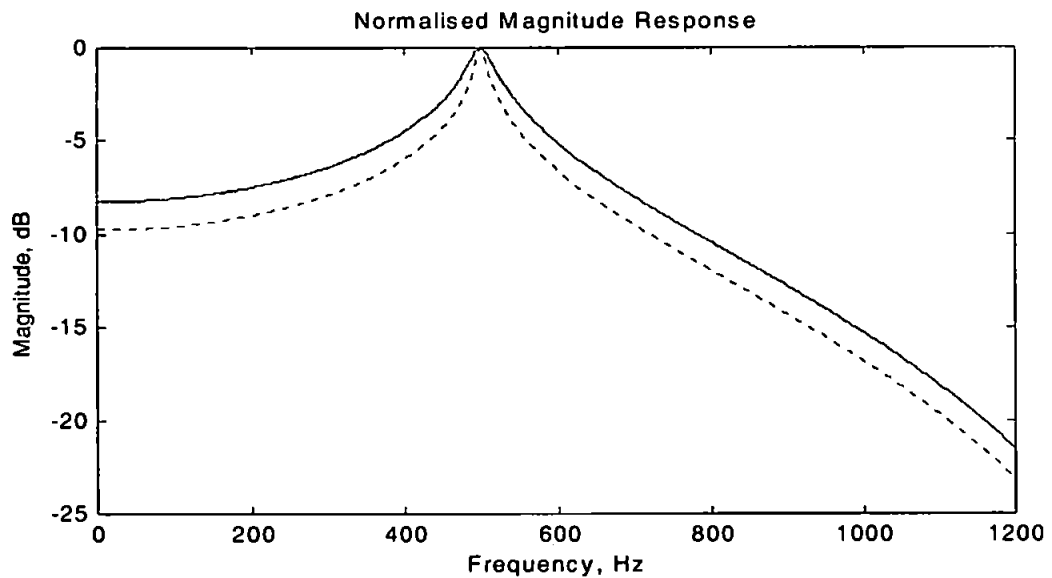


Figure 3.4. The normalised magnitude response of a second order filter, with design frequency of 500 Hz. The continuous line is for a quality factor(Q) of 20, while the dashed line is for $Q = 50$.

Several important conclusions may be drawn about the nature of the magnitude response of a DHO unit. The response is shallower at frequencies below the design frequency of the unit than

for higher frequencies. The magnitude response is also continuous and has no discontinuities, or local peaks. Compared with the second order filters, the response is different in two ways. Firstly, it may be observed that the region around the design frequencies for the DHO response is broader, and secondly, the shape of the curve is different, with the second order filter response falling sharply, but not to the same extent as that of the DHO magnitude response.

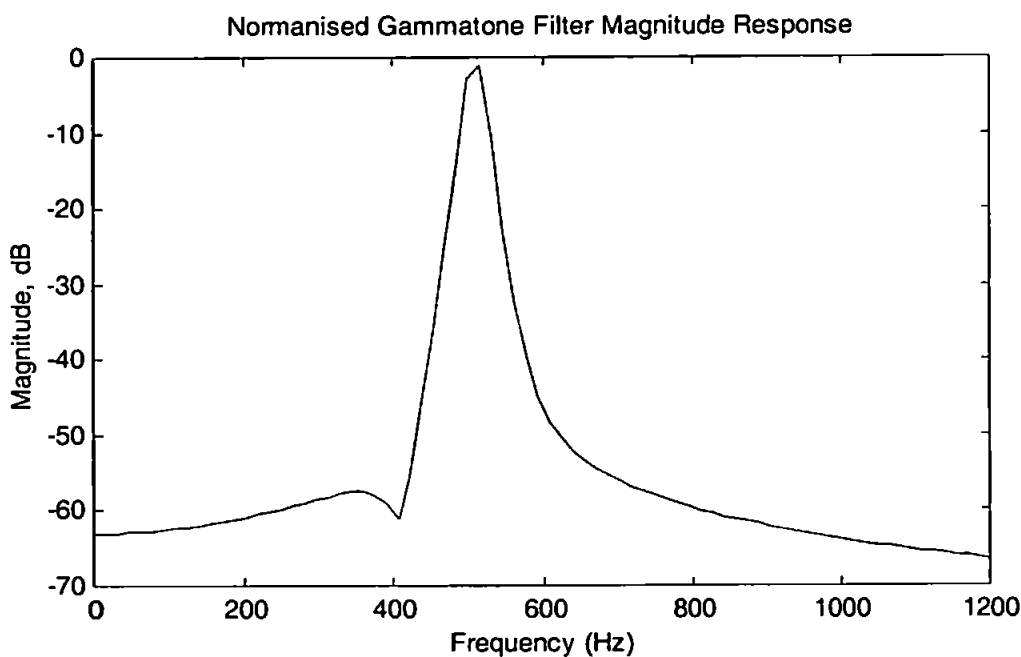


Figure 3.5. The normalised magnitude response of a gammatone filter, with design frequency of 500 Hz. The response is derived from a filter which was selected from a bank designed with 80 filters to cover the design range of 60 to 4000 Hz, using the Auditory toolbox Matlab function [Slaney, 1998] *MakeErbFilters*.

Figure 3.5 shows the normalised magnitude response of a gammatone filter with its characteristic frequency at 500 Hz. It is clear from the figure that the filter is highly selective, compared to the second order filter and the DHO unit. The gammatone filter is a fourth order filter. As can be seen from the figure, the magnitude response of a gammatone filter is much sharper, it is not guaranteed to be a smooth function of the frequency. High selectivity means that the response will be substantially attenuated at frequencies different from the characteristic frequency.

The phase response of a system with the transfer function as in equation 3.21 is given by equation 3.24.

$$\theta(\omega) = \tan^{-1} \left(\frac{\text{Im}(j\omega)}{\text{Re}(j\omega)} \right)$$

... 3.24

The phase response of the bank of DHO units is presented in figure 3.6. The phase response of a comparative bank of second order filters is presented in figure 3.7.

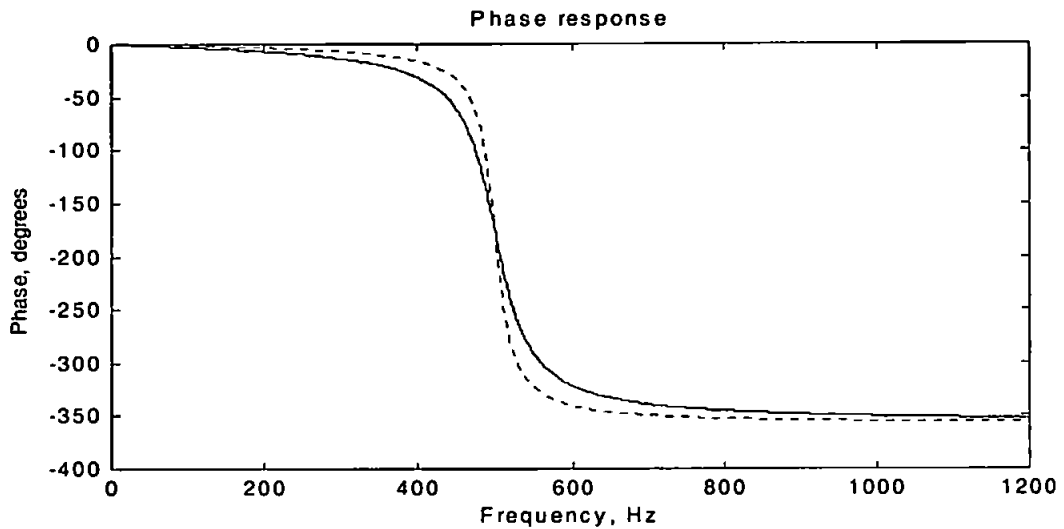


Figure 3.6. The phase response of a DHO unit with design frequency of 500 Hz. The continuous line is for $b = -60$, and the dashed line is for $b = -30$.

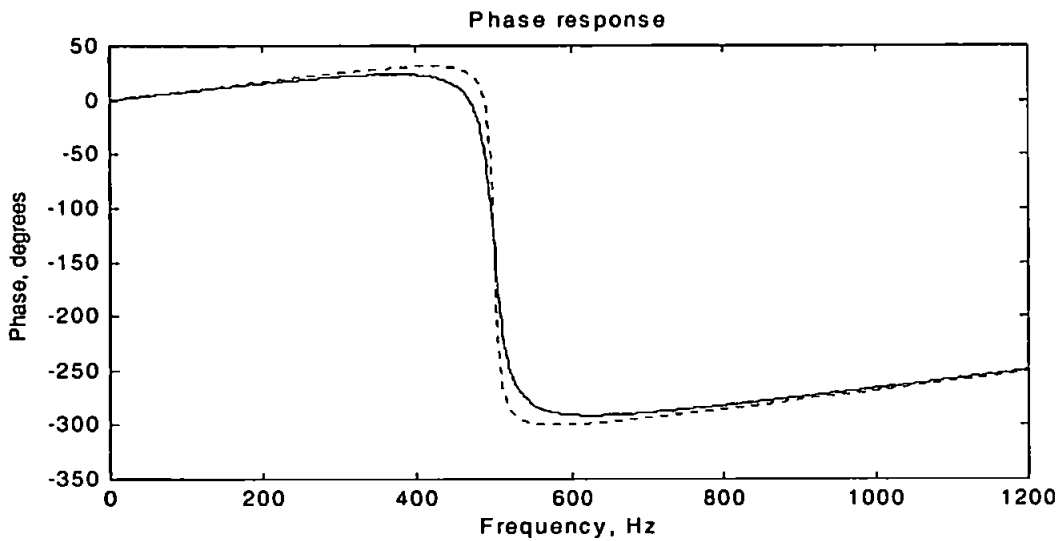


Figure 3.7. The phase response of a second order filter with design frequency of 500 Hz. The continuous line is for a $Q = 20$, while the dashed line is for $Q = 50$.

From the phase plots of the DHO unit presented above, it may be observed that the response lags by half a cycle at the design frequency for both the systems. However, for frequencies substantially below the design frequency, the delay is close to zero, and for frequencies substantially higher than design frequency, the delay is nearly one cycle. From the second order filter phase response, it may be observed that the response also lags by half a cycle at the design frequency, however, unlike the DHO unit, the delay does not approach a constant value for frequencies substantially lower or higher than the design frequency. For both the systems, the values of b and Q respectively have little effect on the phase response apart from frequencies around the design frequency. If the phase response is plotted from -180° to 180° ($-\pi$ to π), then the response is in phase with the input at characteristic frequency, leads the signal by 180° for frequencies lower than the design frequency, and lag by 180° for frequencies higher than the design frequency for both the systems.

3.2.4. Temporal Response Analysis and Transient Behaviour

In the previous section, the steady-state response of the proposed system was presented. In this section, the aim is to present the temporal response in order to analyse the transient response of a system of DHO units. The transient analysis is performed by analysing the response of a DHO unit to a delayed impulse, and to a delayed step function. Comparisons are made with the transient response of a typical second order filter.

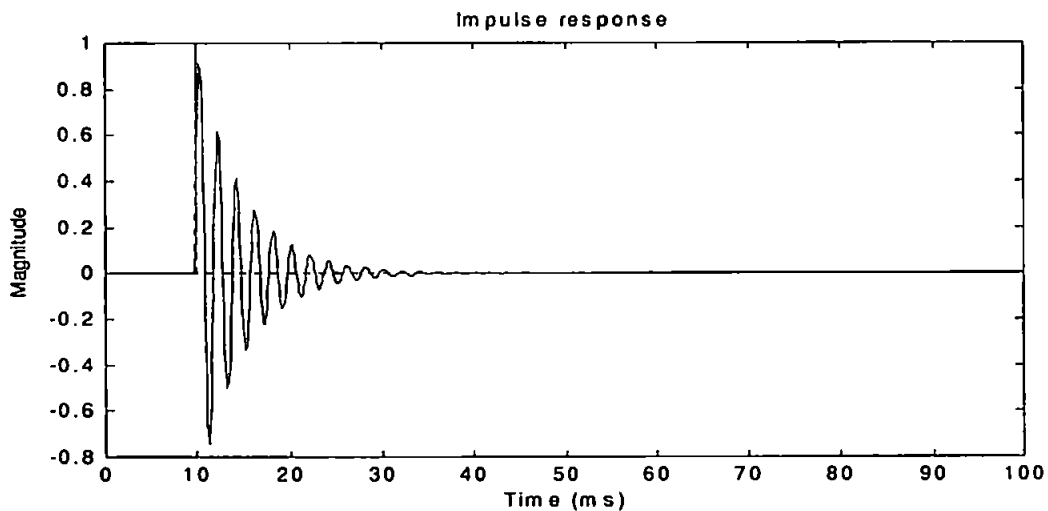


Figure 3.8. The impulse response of a DHO unit, design frequency = 500 Hz, $b = -30$. The impulse at 10 ms time instant is represented by the dashed line. The delay in response is measured to be 0.37 ms.

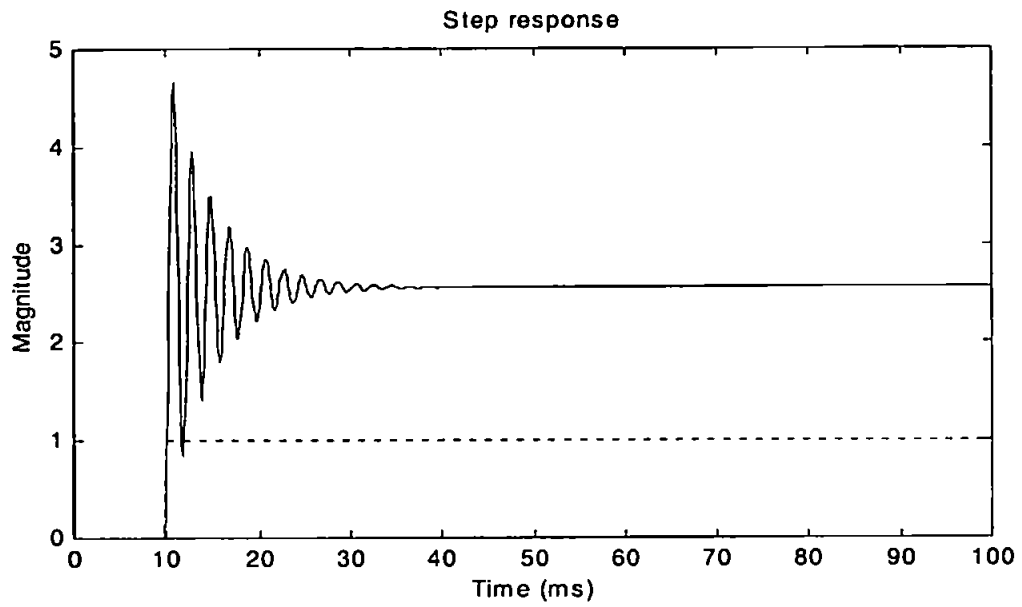


Figure 3.9. The unit step response of a DHO unit with design frequency = 500 Hz, $b = -30$. The response settles down into steady state after about 25 ms of the initial stimulus.

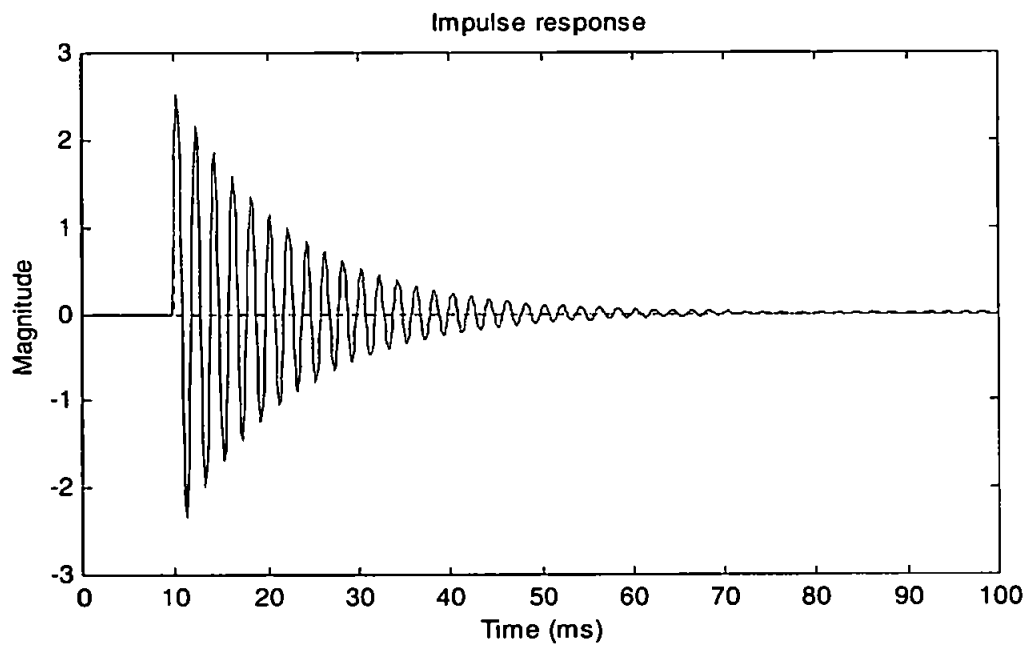


Figure 3.10. The Impulse response of the second order filter with design frequency = 500 Hz, $Q = 20$. The impulse at 10 ms time instant is represented by the dashed line. The delay in response is measured to be 0.27 ms.

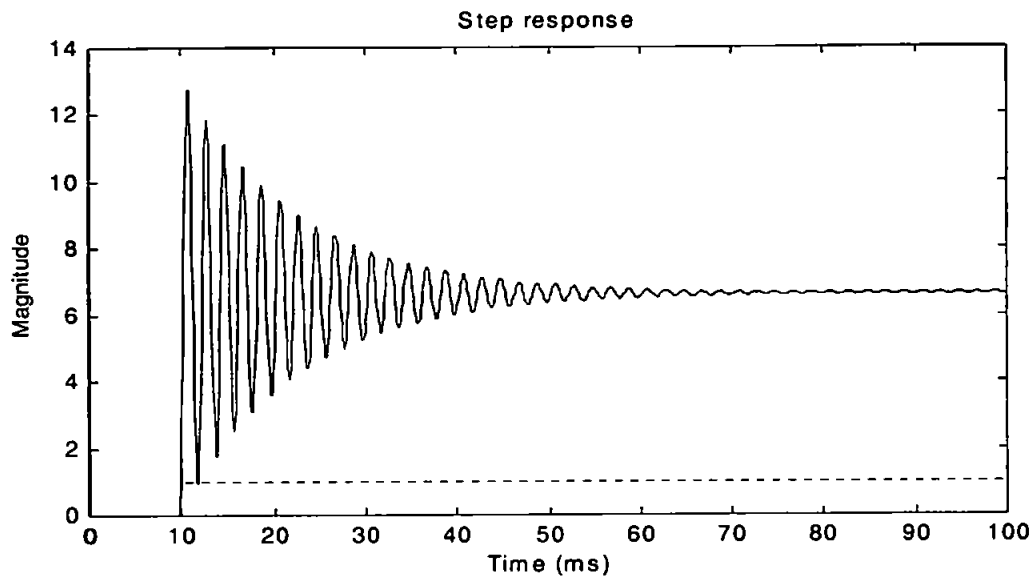


Figure 3.11. The step response of the second order filter with design frequency = 500 Hz, $Q = 20$. The response settles down into steady state after about 60 ms after the initial stimulus.

From comparison of figures 3.8 and 3.10, the time delay involved in the impulse response is of the system gives a measure of the delays involved in the system response. When compared with a second order filter, the delays involved (0.37 ms) are larger (compared to the second order filter delay of 0.27 ms, figure 3.9). However, the transient response of the system, as measured by the time taken by the response to reach steady state when the input is a unit step response is much shorter (25 ms) compared to that of the second order filter (60 ms, figure 3.11). This indicates that the comparable DHO system has a more damped response than the second order filter, with proportionally lesser delay in its impulse response.

3.2.4. Analysis of DHO Response to Noise

As described by equation 3.21, the DHO unit is a linear filter. The shape of the magnitude response of a typical DHO unit, as shown in figure 3.3 shows that the response is similar in shape in some senses to a corresponding second order filter. However, as brought out in the previous sections, there are also differences that can be observed in the response, when compared to second order filters. In this section, an analysis is presented that is aimed to provide an insight into the effect of these differences on the response to noise being present in the input signal.

The response of an analogue or a digital filter when a zero mean white noise process with a power spectral density (two-sided) $N_0/2$ is applied, is a finite average power N , given by equation 3.25, where $H(f)$ is the magnitude response of the system.

$$N = \frac{N_0}{2} \int_{-\infty}^{\infty} |H(f)|^2 df \quad \dots 3.25$$

Given the equation above, in presence of similar noise processes, the response is completely characterized by the shape of the magnitude response, as defined in $H(f)$. For a zero mean white noise process based signal, the power spectral density (PSD) is more or less uniform (i.e. equal power in each frequency “bin”). The magnitude response that decides the total power N of the response, are not only different in case of the comparison of DHO units and second order filters in terms of the general profile, but they are also different in terms of levels of attenuation away from the design frequency. For a typical DHO unit, this level of attenuation is much more compared to a similar second order filter. Therefore, one can expect to see much less power in the output signal (N in equation 3.25) derived from DHO response, as compared to a similar second order filter.

In response to a pure tonal signal, the integration in equation 3.25 is zero for all points but for the sinusoidal components (tones) present in the signal. In this case, the output power of the signal is governed by the response curve values at these points. For the case when the signal contains a pure tone with frequency close to the design frequency of the unit/filter, the response will be the same as input signal, multiplied by a certain gain value and slightly delayed (barring the initial transient behaviour). The power N of the response will be $H(f_c)$ where f_c is the signal frequency. For the case when the input signal is a mixture of a tone and zero mean white noise, the actual response signal will be dependent on the actual short term nature of the noise process, but generally, the response power would be equal to that of the input without the noise. However, the actual short term structure of the response will vary depending on the short term structure of noise. The effect would be to slightly increase or decrease the signal power associated with the tonal frequency component. By comparing the shape of the magnitude response curves of typical DHO units and second order filters, it may be observed that the DHO unit response is broader around the design frequency as compared to the second order filter. Therefore, these slight perturbations to the signal power at the tonal component frequency

due to the presence of white noise would have less effect on the overall response of a DHO unit compared to a similar second order filter. The figure 3.12 below compares the actual responses of the DHO units and second order filters to white noise, and to a tonal signal with added white noise. The responses are scaled so that the maximum response in steady state to the pure tonal signal is taken as one. This is done because the gains of the two systems being compared might be in general different. Only steady state portion of the responses is shown to facilitate analysis and comparison.

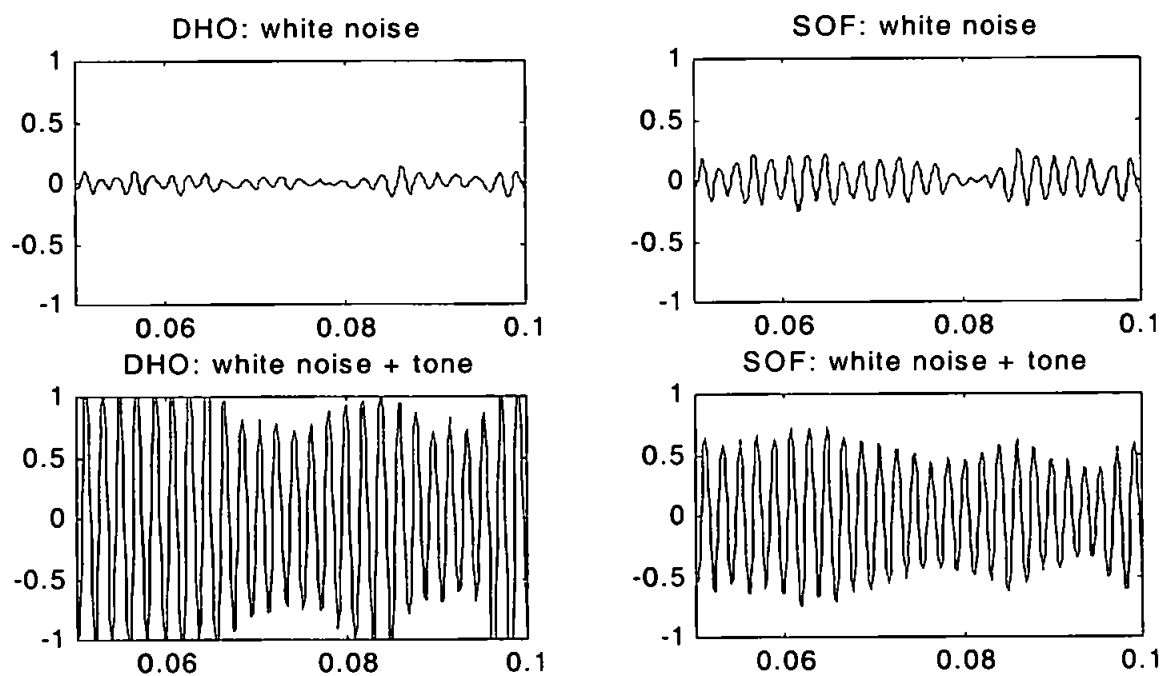


Figure 3.12. The experimental noise response of the DHO units and second order filters is compared here. The DHO unit is designed at a design frequency ω of 500 Hz and a $b = -30$. The second order filter (SOF) is designed with centre frequency of 500 Hz, with a quality factor of 20. All the responses are scaled with respect to the steady state response at design frequency. The top two plots compare the response to zero mean white noise. Please note the difference in scales of the two plots with the DHO response (top left) being four times smaller than the SOF response. The bottom two plots compare the response to a mixture of tone and noise. The tonal component was present at 510 Hz. Again, please note that the DHO response is twice as large when compared to the SOF response in bottom right hand plot.

3.3. Design of the Bank of DHO Units as a Signal Processing Front-End

A single DHO unit has its behaviour described by equation 3.19 in the digital domain, given a sampled signal. In this equation $(1 + \Delta.b)$ is the normalized damping (or more appropriately in this equation, the leakage factor), and $\Delta.w$ is the normalized radial frequency. To design a bank of such units that covers the frequency range of interest, we can use the same principles as those used in the design of a bank of digital filters for speech analysis, as described in section 3.1.2.

The cochlear processing essentially acts like a bank of overlapping band-pass mechanical oscillating units. The bandwidths of these units are arranged in a specific (tonotopic) order. Below 500 Hz, the critical bandwidth is a constant around 100 Hz. For higher frequencies, the bandwidths are roughly one fifth of the centre frequencies [Zwicker, Fastl, 1990]. Thus the relationship between the Bark, which is the critical bandwidth unit, and the linear frequencies is linear for low frequencies and exponential for large frequencies. Since in our pitch estimation experiments characteristic frequencies ranging up to 1000 Hz are used, in those experiments only the linear scale for determination of characteristic frequencies should suffice. For full scale illustrations, the ERB scale of frequencies is used, which is given by equation 3.26 [Glasberg, Moore, 1990]. The first part of the equation gives the centre frequency on the ERB scale given the centre frequency F_c and the second part gives the associated bandwidth.

$$ERB_{cr}(F_c) = 21.4 \times \text{Log}_{10} \left(\frac{4.37 \times F_c}{1000} + 1 \right) \quad \dots 3.26$$

$$ERB_{bw}(F_c) = 24.7 \times \left(\frac{4.37 \times F_c}{1000} + 1 \right)$$

Due to the digital approximations of equation 3.17, as w approaches the Nyquist frequency, the system tends to get unstable, i.e. instead of damped oscillations, un-damped oscillations increasing in amplitude are observed. This is due to the relatively large step size used in the simulations of the equation 3.17, with a step size equal to Δ . In implementing equation 3.19, the characteristic frequency, specified by w is varied between 0 and f_s / π where f_s is the sampling frequency. The relationship between the actual characteristic frequency f'_n and the design parameter f_n is given by the equation 3.27 below.

$$f'_n = f_n + e^{f_n / 365} \quad \dots 3.27$$

The above equation was obtained by considering the response of full spread (nearly up to the Nyquist limit), and then fitting a curve to the response frequencies, compared with the characteristic frequencies. This equation is not used in the pitch estimation system, as the highest frequency (1000 Hz.) is much lower than the signal bandwidth (4000 Hz.). Another way to remove the need for this approximation is to decrease the step-size to $\Delta/2$ or $\Delta/3$.

To design the bank of DHO units of order N (i.e. the number of DHO units in the bank), the characteristic frequencies are varied from f_L (the lowest frequency of interest), to f_H , the highest frequency of interest, on a linear scale. The damping constant b is kept constant to have equal bandwidth for all the N units. Thus these four parameters specify the complete design of the bank of DHO units. The values used in the system for rest of the thesis are given in table 3.1 below.

Parameter Name	Description	Value Used
f_H	Highest centre frequency	1000 Hz
f_L	Lowest centre frequency	60 Hz
N	Number of DHO units	40
B	Controls bandwidth	-40

Table 3.1. The parameters used in bank of DHO design for the proposed pitch estimation system.

The lowest and the highest frequencies were chosen to reflect the range of resolvable harmonics of the speech signal in the auditory system [Plomp, 1965] (the issue of resolvability is discussed further in chapter 4) given the normal speech pitch range, since the proposed system considers only resolved harmonics for computation of the pitch estimate. The number of resolved/resolvable harmonics is a function of the fundamental frequency, but given the pitch range of human speech in normal conditions, this range of frequencies appears to be adequate. However a slightly larger or smaller range of frequencies could also be used. The number of units and the bandwidth is chosen so as to design units with overlapping frequency response with complete resolution and frequency range coverage.

The magnitude and phase responses of the designed bank of DHO units is provided in figures 3.13 and figure 3.14. The magnitude response is normalised with respect to the single largest gain (i.e. the maximum gain has a value of 1). The phase responses are provided over the entire cycle.

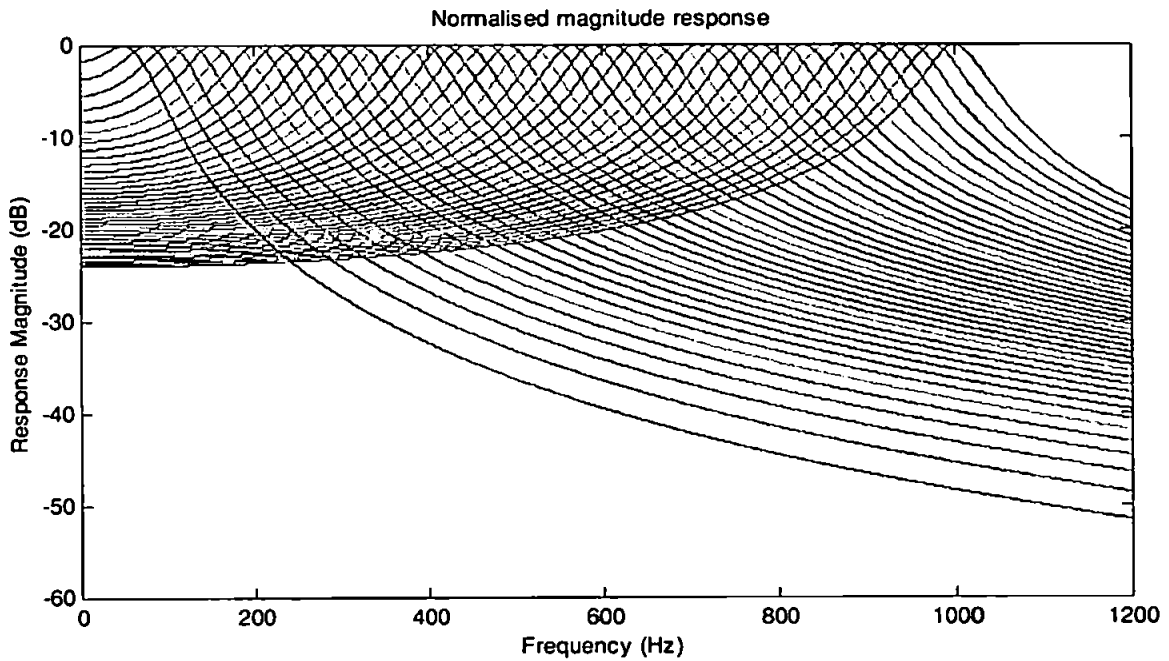


Figure 3.13. The Magnitude response of the bank of DHO units. The magnitude is normalised with respect to the single largest gain value.

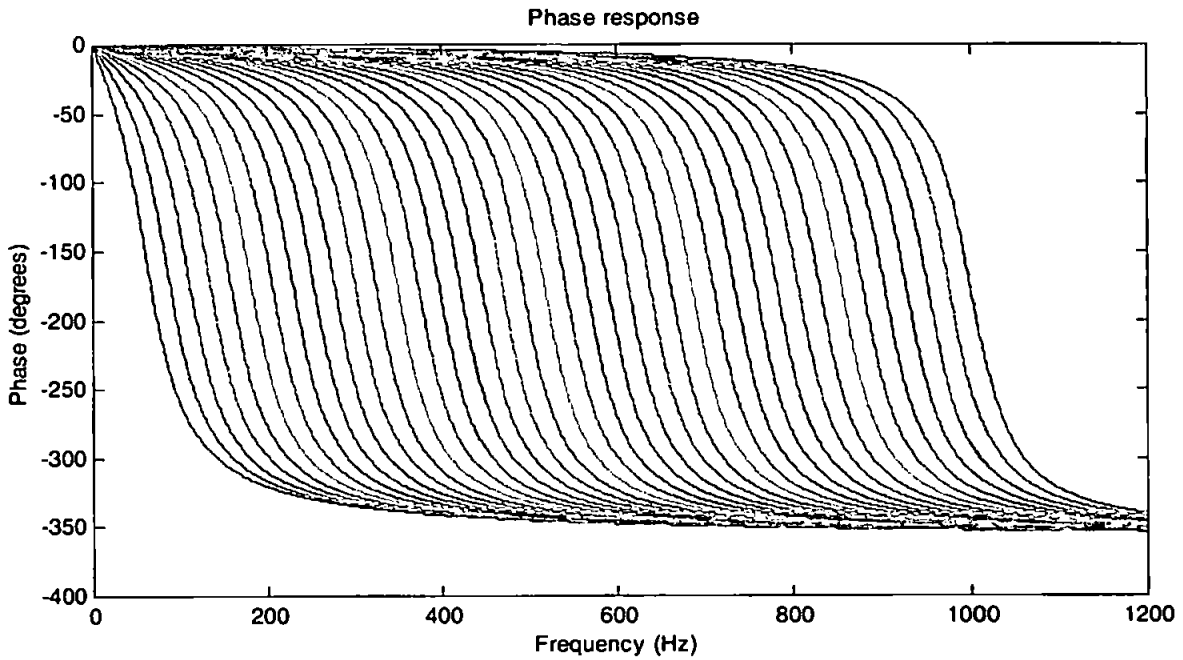


Figure 3.14. The phase response of the bank of DHO units. The phase is measured over one cycle.

3.4. Analysis of DHO Bank Using Test Signals

As mentioned earlier, each unit in the bank of DHO units has its own design frequency to which it maximally responds. Figure 3.8 provides an example of the impulse response of a typical unit. In this section the aim is to provide a deeper insight into the working of the bank of DHO units. It is hoped that this can be achieved by further empirical analysis of the output response and the dynamic behaviour and presentation of the results of this analysis in graphical form. To make things clear, sinusoidal signals and speech are used as the input stimuli, with white noise at zero decibels signal to noise ratio as additive noise.

Single Unit Chirp Response:

Let us start with a DHO unit of characteristic frequency of 500 Hz. As an input, a chirp signal is applied, with instantaneous frequency increasing linearly with time, starting with 100 Hz, and finishing with 1000 Hz, over a period of one second, at the sampling rate of 8 kHz. It is quite clear from the figure 3.15 that the single unit picks up the relevant part of the signal (which is in the middle) and amplifies it preferentially. It is also clear that this selectivity is quite broad. Figure

3.16 shows the response to the same signal, but with white noise added at zero decibels signal to noise ratio (equal amplitude noise is added to the signal). The output signal amplitude is slightly lower for the signal portion with frequency in the neighbourhood of 500 Hz; however, the response profile in general remains the same, with some low amplitude noise. The 500 Hz and the neighbouring portions of the output signal are periodic with the same frequency as the input signal, even in the presence of noise. This is an important property, which makes the system output SNR (signal to noise ratio) much higher than the input signal SNR. Another property is that output signal in the entire frequency of the DHO output is same as input signal frequency, but with a varying amplitude, the maximum of which occurs at the characteristic frequency of the DHO unit. In this sense the DHO unit behaves like a linear filter.

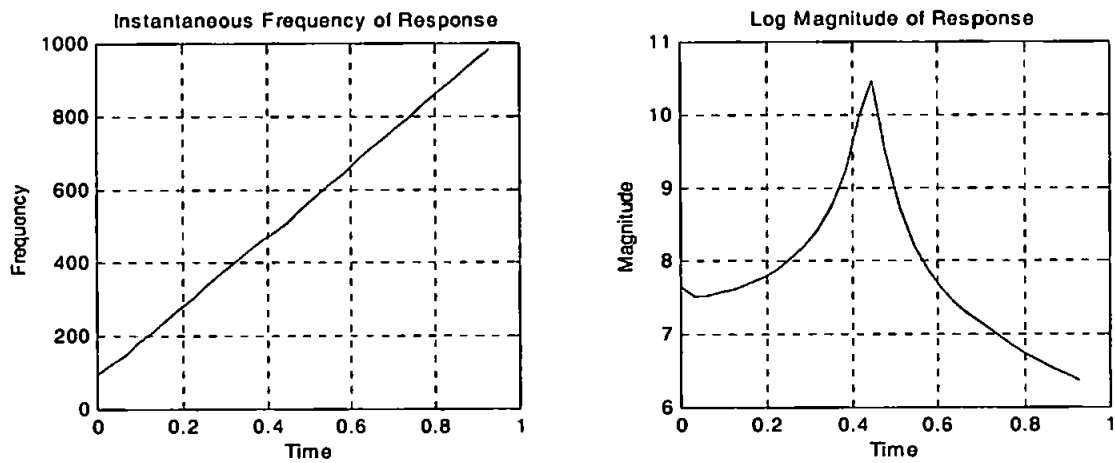


Figure 3.15. The output response of a DHO unit with $CF=500$ Hz, and $b = -30$. The input is a chirp signal, with frequency increasing linearly from 60 Hz to 1020 Hz over the period of one second. The graph on the left is the instantaneous frequency of the response, calculated using the Hilbert transform. The graph on the right is the log magnitude of the response. A slight non-linearity in the plot on the left reflects the transient behaviour of the unit response near the design frequency (500 Hz).

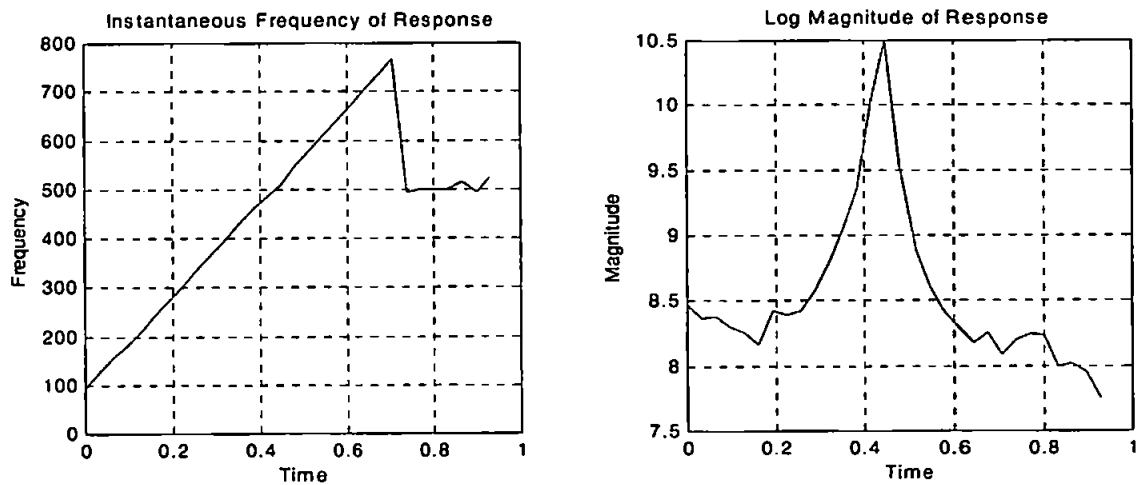


Figure 3.16. DHO unit response to a chirp signal with added noise at 0 dB SNR. The DHO unit is same as used in figure 3.15. The instantaneous frequency of graph on the left shows that the frequency of oscillations of the unit is its characteristic frequency for noisy signals at frequencies much higher than the design frequency. The log magnitude graph on the right as compared to that in figure 3.15 is more noisy, but with the location and size of the peak magnitude unchanged.

Response of a Bank of Three Units:

The input stimulus consists of a mixture of 300 Hz and 900 Hz tones. The analysis is based on the output of DHOs with characteristic frequencies of 300 Hz, 600 Hz, and 900 Hz. Figure 3.17 illustrates the output of these DHO units for the given stimulus. The DHO units with characteristic frequencies of 300 Hz and 900 Hz respond with much higher amplitude than the DHO with the characteristic frequency of 600 Hz, the component which is not present in the input signal. Also, the units oscillate at the same frequency as those present in the input stimulus. A combination of the correct frequency of oscillations and their higher amplitude indicates that the system is able to “pull out” or separate the components of the input signal.

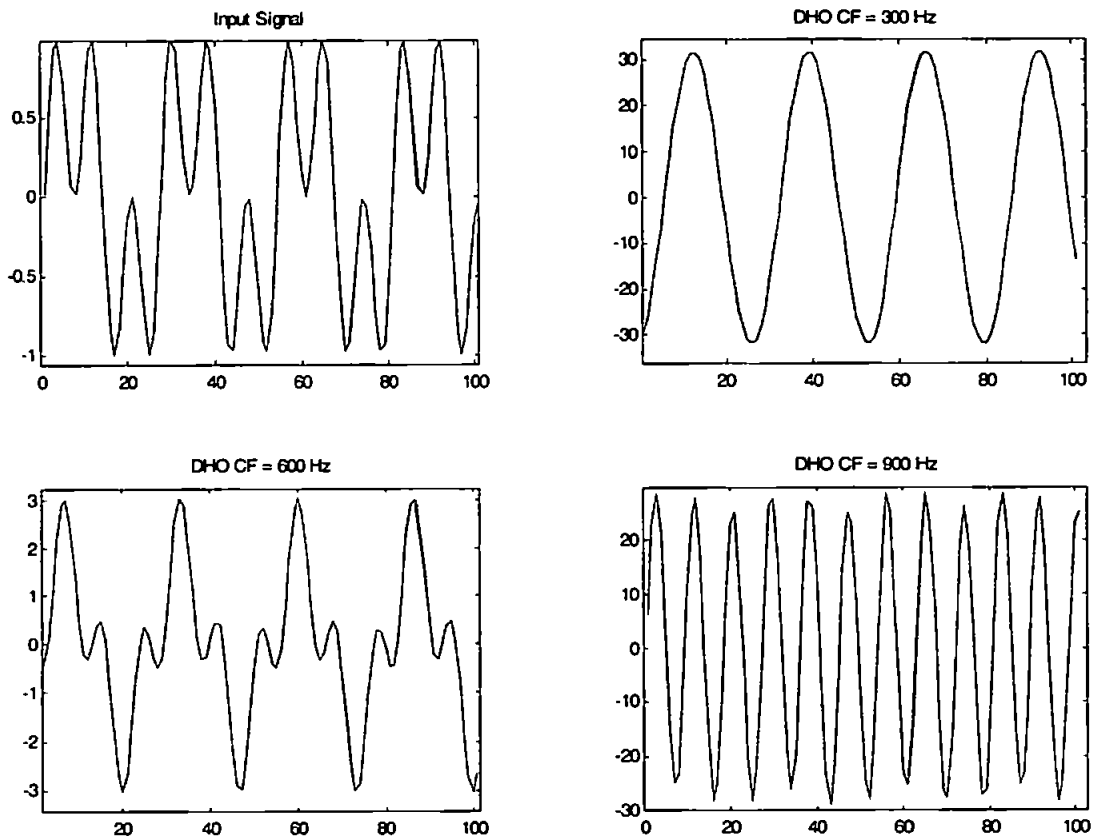


Figure 3.17. The top left plot is the input signal comprising of 300 Hz and 900 Hz components, sampled at 8000 Hz, the top right is the output of the DHO with characteristic frequency of 300 Hz, the bottom left is the output of the DHO with characteristic frequency of 600 Hz, and the bottom right is the output of the DHO with characteristic frequency of 900 Hz. The x-axis is time, with the number of samples; the y-axis is the amplitude of oscillations.

Response to Speech Like Synthetic Signal:

Let us consider the speech like signal given by equation 3.28.

$$x(t) = (1 + D \cdot \sin(2\pi \cdot f_0 \cdot t)) \cdot \sin(2\pi \cdot 2 \cdot f_0 \cdot t) \quad \dots 3.28$$

The signal $x(t)$ is speech like in the sense that the strong $2 \cdot f_0$ Hz component can be considered the first formant component of voiced speech, and this component is modulated with a weaker component, analogous to the fundamental frequency, f_0 . The parameter D specifies the “depth” of modulation. A value close to zero means very weak modulation, and a value close to one

means very strong modulation. A value of 0.3, for example, indicates quite weak modulation, as may be expected in speech signal where the fundamental frequency component is lower in power. The output of the bank of DHO units is presented in figure 3.18. From the figure, it is clear that the component with the fundamental frequency (120 Hz in this example) is quite active, even though the depth of modulation for the formant frequency is low. Figure 3.19 gives the output of the same signal, but with zero decibels of added white noise. Again, it may be observed that a even very high noise level does not destroy the structure of the response, although the amplitude of the response is slightly reduced.

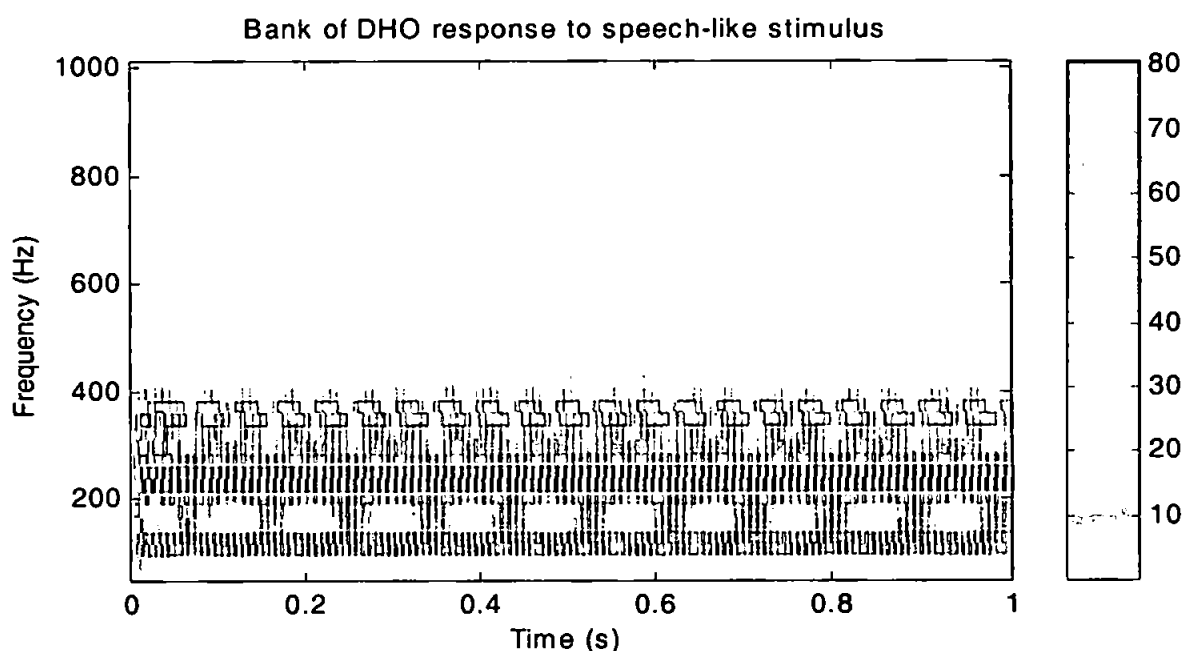


Figure 3.18. The output of the bank of DHO unit, to the input given by the equation 3.28, with $D = 0.3$, and $f_0 = 120$ Hz. The x-axis is time (in seconds), and the y-axis is frequency in Hz.

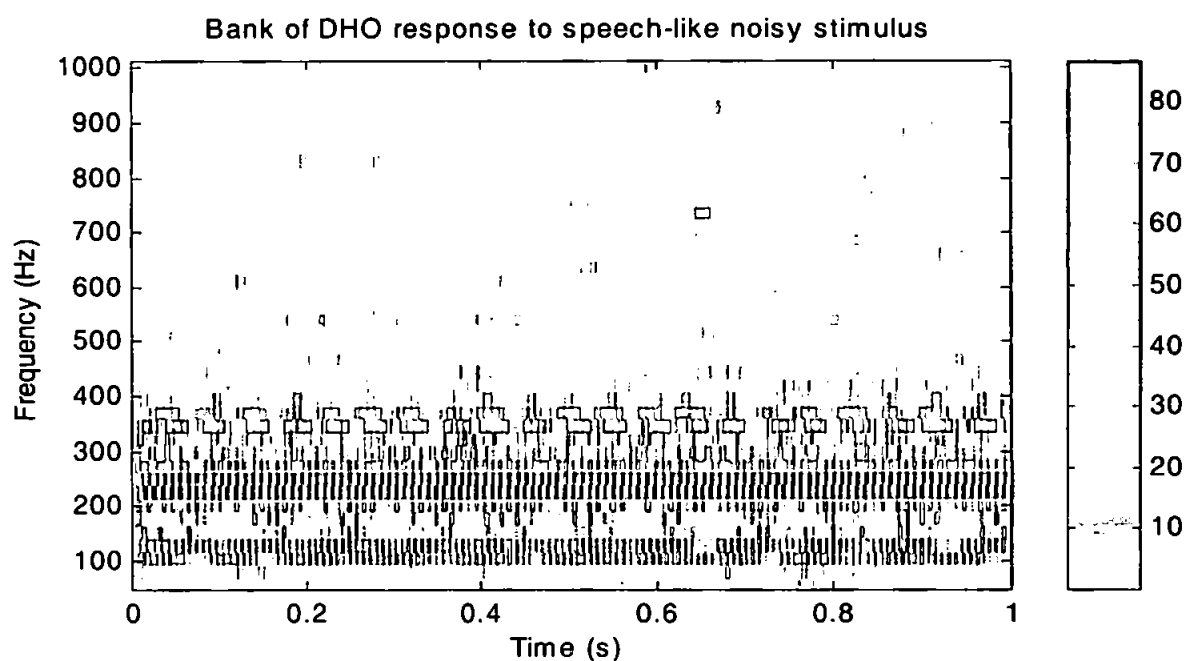


Figure 3.19. The output of the bank of DHO units, to the input given by the equation 3.28, but with white noise added at 0 dB. The x-axis is time (in seconds), and the y-axis is frequency in Hz.

Speech Signal and Spectrogram Representations:

The representation of the figures 3.18 and 3.19 is the spectral representation, with the frequency on the y-axis and the time on the x-axis. As in the case of filter-bank output, in this representation, all the waveforms (DHO outputs) are stacked on top of each other (in order of increasing centre frequency).

The comparison of DHO output with the FFT power spectrum is a valuable one, as it enables us to analyse the temporal and frequency resolution and discrimination, in the presence of noise. Figures 3.20 and 3.21 are respectively the output response of the DHO units and the FFT power spectrum given a clean speech signal. Figures 3.22 and 3.23 are respectively the output response of the DHO units and the FFT power spectrum given a speech signal with added noise as the input. The aim of this analysis is to demonstrate the robustness to noise of the DHO system compared with the FFT power spectrum representation. An overall noise estimate is constructed by comparing the output of the clean signal representation with the output obtained by adding white noise to the signal at zero decibels SNR. The estimated noise is the difference in the average power of the signals. The SNR is computed as the ratio of the average power level of the

signal and the average estimated noise level. If the average power of the clean signal representation is P_{clean} and the average power of the estimated noise is P_{noise} , then we have the estimated SNR by equation 3.29.

$$SNR = 10 \cdot \log_{10} \left(\frac{P_{\text{clean}}}{P_{\text{noise}}} \right) \quad \dots 3.29$$

The estimated SNR for the DHO output, calculated using the clean signal based output, and the speech signal with added white noise at 0 dB SNR is **-1.55 dB**. For the FFT based power spectrum, the estimate, using the same signals is found to be **-8.05 dB**. The FFT used for the computation was computed using 512 point calculation (256 frequency estimates are available for the complete frame and full frequency range, 65 points for up to 1000 Hz range, with no interpolation or smoothing), a Hamming window was used to compute the signal frames, and the overlap between frames was 64 signal samples. Only the first 65 values, corresponding to the analysis range of 1000 Hz upper frequency were considered. The SNR calculations are negative because for large number of points in the clean signal with low power in the output, the corresponding regions under noise have comparatively higher energy. However, comparing the performance of the two systems, the DHO system handles the noise better than the FFT computation. The operation of the DHO units, as illustrated in figure 3.15 is such that even in the absence of a frequency component close to its characteristic frequency in the input signal, the unit produces activity at a reduced amplitude (power) with the frequency of oscillations corresponding to the component in the input signal closest to its characteristic frequency. This property of the DHO units influences the SNR calculations as when noise is present in the input signal, even though the unit may not respond with the same frequency, the overall power of the output signal remains similar.

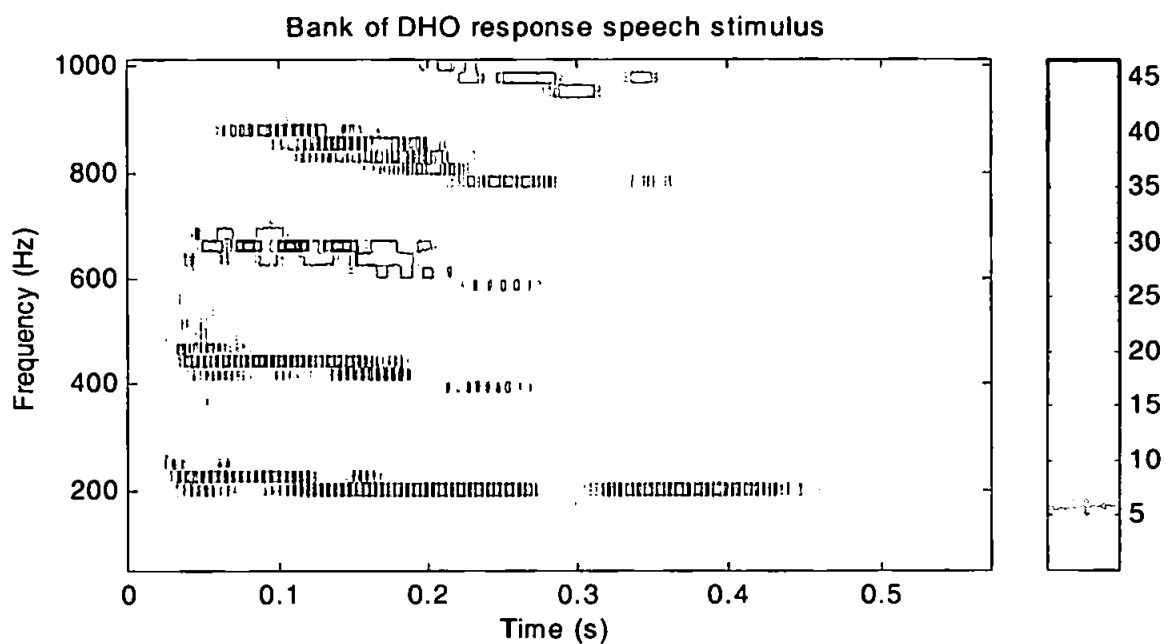


Figure 3.20. The output of the bank of DHO units, to the input is speech from a female (telephone quality), saying the word "brown". The x-axis is time (in seconds), and the y-axis is frequency in Hz.

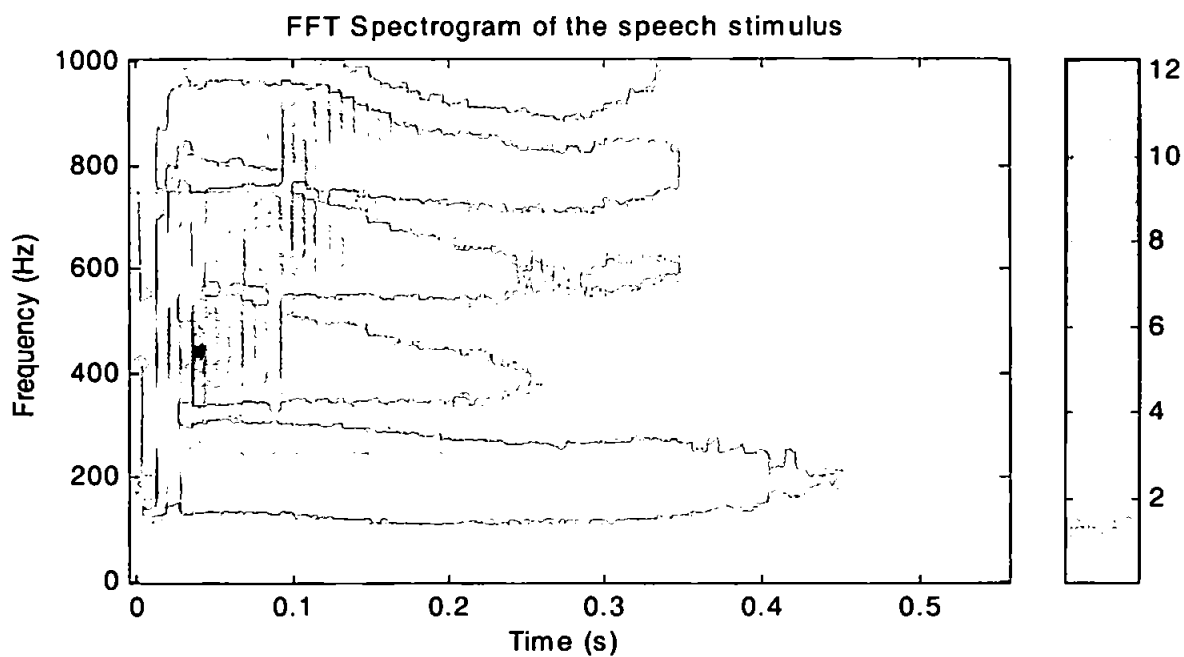


Figure 3.21. The same signal input as in figure 3.20, showing the power spectrum produced using a 512 point FFT and hamming window.

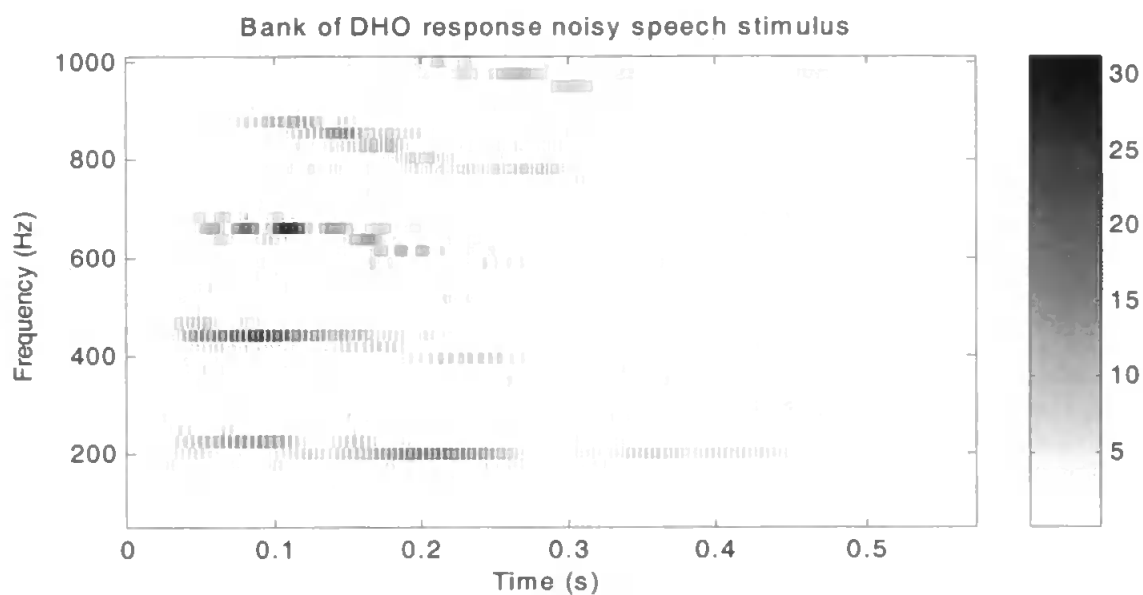


Figure 3.22. The bank of DHO output for speech signal with noise. The input is same signal as in figure 3.20, but with added white noise at 0 dB SNR. Estimated SNR w.r.t the clean representation = - 1.55 dB.

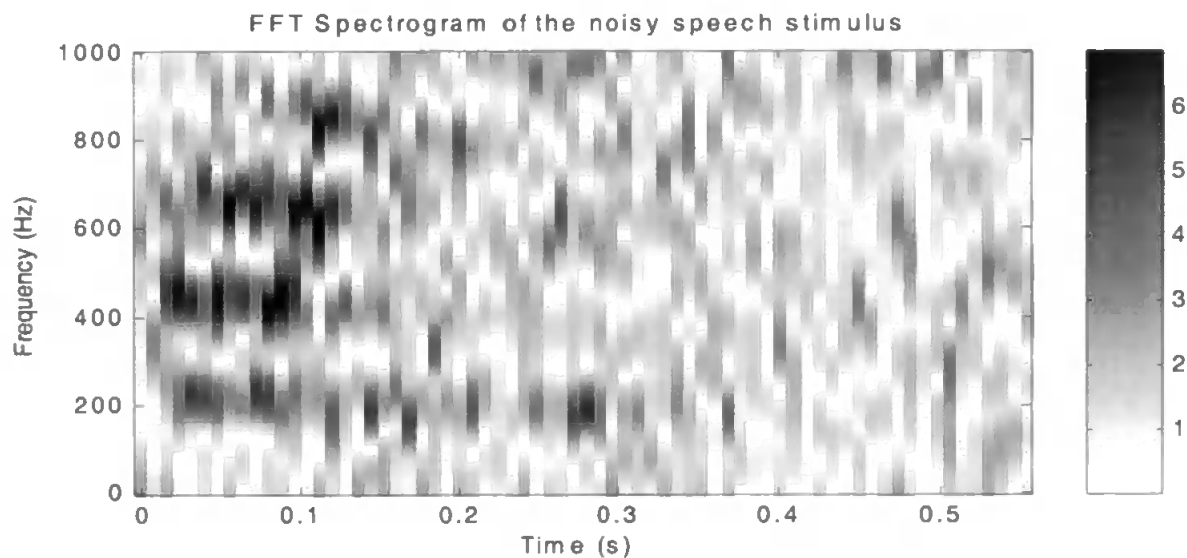


Figure 3.13. Spectrogram representation using FFT for speech signal with noise. Input is same as in figure 3.21, but the input signal with added white noise at 0 dB SNR. Estimated SNR w.r.t the clean representation = - 8.05 dB.

3.5. Construction of Time-Frequency Energy Maps

The discussion and analysis above involved using the “raw” output of the DHO units. However, this representation does not take full advantage of the fine temporal structure of the output. The frequency of oscillations is a very useful factor in the construction of the time-frequency energy maps, and it is the frequency of oscillations that is used in the construction of the time frequency energy maps, rather than the characteristic frequencies of the units. By measurement of the true oscillation frequency of the outputs of the DHO units, the representation achieves a much higher frequency resolution. In contrast to the usage of the traditional filter-bank outputs, where the energy of the output is assigned to fixed channels based on the centre frequencies of the digital filters, the representation that is proposed is constructed by measuring the true frequency and energy of the output of each unit and combining the output of units with the same oscillating frequency.

3.5.1 Measurement of Frequency for Each Output

The measurement of frequency is performed for each unit of the bank of DHO units. A peak is defined as a point when the amplitude is locally highest and greater than some positive constant value. This value depends on the maximum range of the input signal (a pre-processing step was used to make the signal range the same).

The maximum amplitude for the first of the two peaks is measured as A_a , and the time of the first peak stored as t_a . After the detection of t_a , the time of positive slope zero-crossing just after the point t_a is measured as t_0 . The next positive peak time t_b is measured. The period of the signal is determined at point t_0 by interpolating the period at t_0 between t_a and t_b . The period P_0 is determined by equation 3.30.

$$P_0 = \frac{A_a(t_b - t_0) + A_b(t_0 - t_a)}{A_a + A_b} \quad \dots 3.30$$

The process is repeated by assigning A_b to A_a and t_b to t_a .

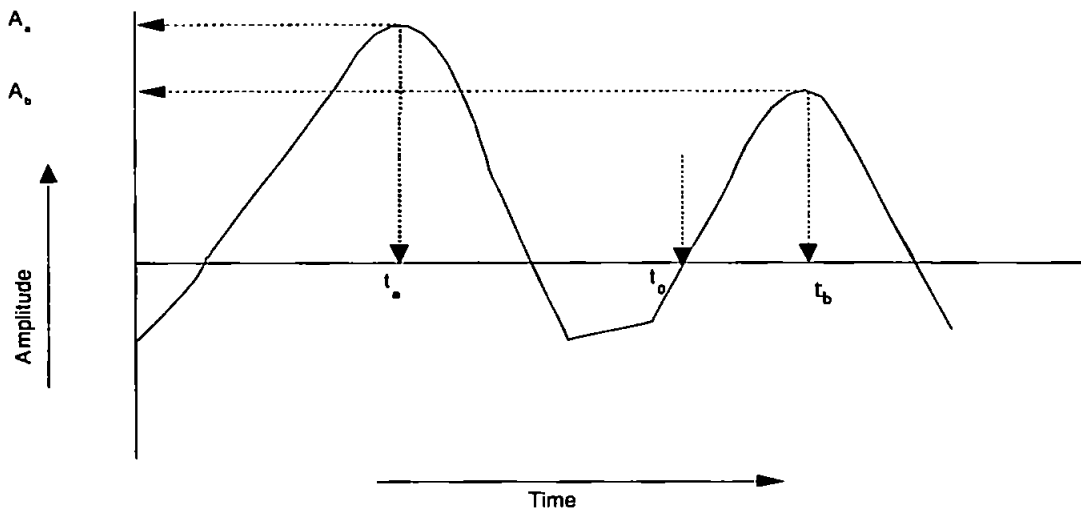


Figure 3.24. Diagram to describe the calculation of A_a , A_b , t_a , t_b , and t_0 , and their relationships.

For times when $A_a > 0$ and $A_b = 0$, P_0 is assigned a value of zero to prevent incongruous pitch estimates. Given the sampling frequency is given by F_s , the frequency of the output is determined as F_s/P_0 for non-zero values of P_0 and zero when the period is zero or unassigned. The calculation and relationships between the variables is described in figure 3.24.

3.5.2 Measurement of Energy for Each Output

The energy in each output is found at the same time as the periodicity. Due to the thresholded nature of the periodicity calculation, the energy calculation algorithm assigns a value of zero energy to any period of any channel whose current activity (maximum value between zero crossings) is below a certain threshold. The measurement of the energy is calculated using the following algorithm.

- Initialise a vector before processing begins of size N , where N is the number of DHO units in the bank. Let this be named `stored_max(n)`.
- For each channel, and each sample, consider the output of the DHO is $y(n,t)$. Then, if the value of $y(n,t) > \text{stored_max}(n)$, replace the value of `stored_max(n)` with the value of $y(n,t)$.
- If for unit n , the current position is positive slope zero-crossing, then assign the current energy of the unit the value `stored_max(n)`. Reinitialise `stored_max(n)` to zero.

- Wait for the next input sample to be processed by the bank of DHOs, and then repeat all the steps once again.

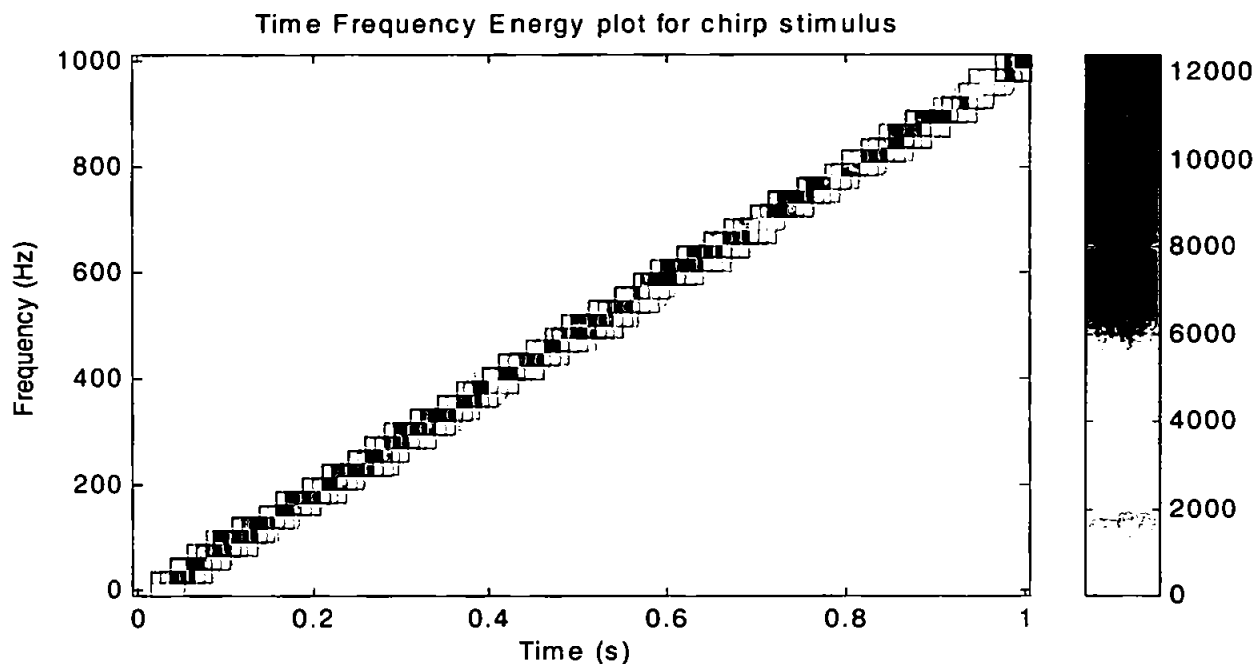


Figure 3.25. The Time-Frequency energy plot. This output was produced in response to an input chirp signal with low frequency of 1 Hz, and the highest frequency of 1000 Hz. The x-axis is time in seconds, and the y-axis is frequency.

A sample output, with input a chirp signal is presented in figure 3.25. The figure illustrates the output for the algorithms described above. The representation in this form is similar to the spectral representation, although with a very fine frequency resolution.

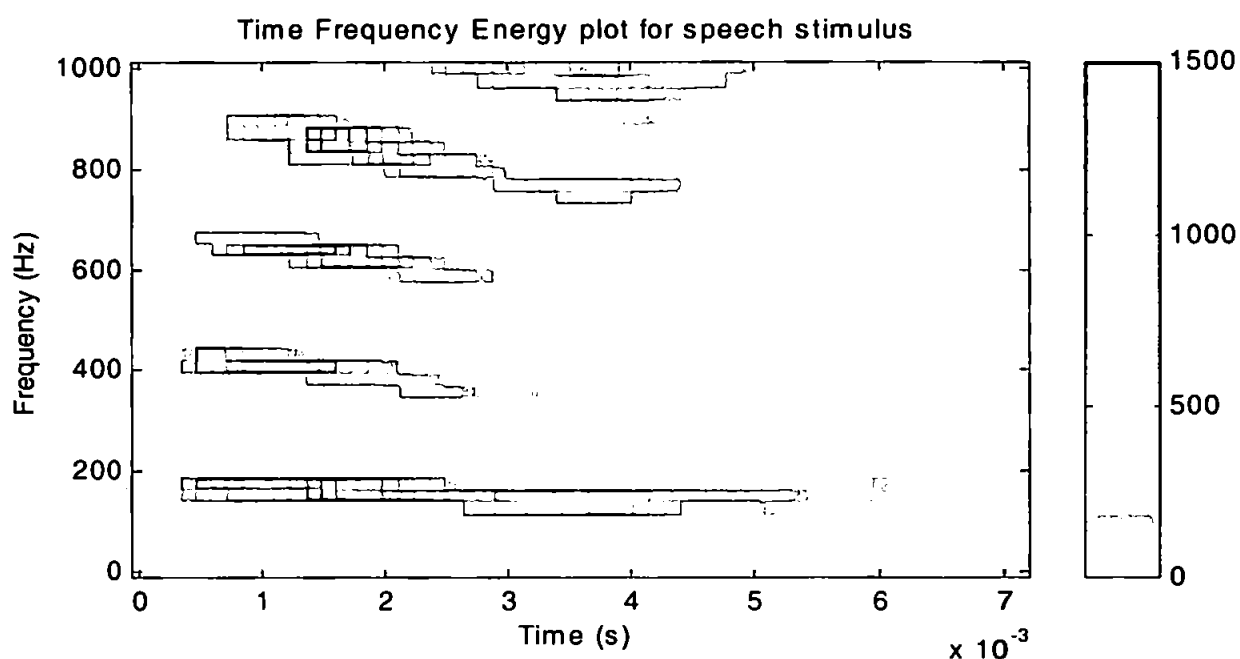


Figure 3.26. The time-frequency energy plot produced by processing the DHO output, with the input of speech mixed with white noise at 0 dB SNR (same as figure 3.22). The average SNR calculated for this output is 1.01 dB (when compared with the output from a clean signal).

The resulting representation improves the SNR of the output when compared to the raw DHO output representations. For example, in figure 3.26, the SNR is improved in response to same signal and signal conditions, from -1.55 dB (in figure 3.22) to 1.01 dB in figure 3.26. The figure also illustrates the point about the systems “noise masking” abilities. The portions of the response which are closer to the strong continuous inputs (between time points 0.1 to 0.3 seconds in figure 3.26), have much less noisy components at lower amplitude, compared with regions before and after.

3.6. Comparison with other temporal analysis schemes

Over the course of many years of auditory perception research, many different temporal processing schemes have been proposed. This section briefly discusses the different approaches, and highlights the differences and similarities with the proposed system.

The optimum processor theory of the central formation of pitch by Goldstein [Goldstein, 1973] proposes the idea of a central processor acting on the stochastic estimates of harmonic frequencies, without any phase or amplitude information, in order to estimate the pitch that best

explains the stimulus. The cochlear output leads to a frequency analysis of the stimulus in this model of processing. Following this, independent noisy channel based processing are assumed to produce a maximum likelihood estimate of the harmonic numbers associated with the harmonic frequencies. This is followed by a maximum likelihood estimation of the pitch frequency, given the estimates of the harmonic numbers present in the stimulus. The theory itself is independent of whether the mode of processing is channel based or completely temporal. Thus this model provides a mechanism of estimation of the pitch frequency, given the information about the harmonic frequencies. In the proposed system, the pitch frequency is also estimated from the harmonic frequencies, using a temporal mode of operation. However, in the proposed system, the amplitude information is also made use of, in estimation of the pitch period. Moreover, the estimation of the harmonic numbers is deterministic rather than stochastic, although presence of noise would add a stochastic component in any calculation.

The Ensemble Interval Histogram (EIH) model of auditory processing was proposed by Ghitza in [Ghitza, 1991], [Ghitza, 1992]. In this model, the temporal processing occurs after the cochlear filtering stage (with channels equally spaced on a log-frequency scale), through an array of level detectors corresponding to each channel. These level crossings are roughly arranged on a logarithmic scale. The level crossing detectors measure the time interval between each of the level crossings of the output of their corresponding channel, and contribute a count of the level crossings to an EIH corresponding to the time interval between level crossings. Since the level crossings are equally distributed on a logarithmic scale, the magnitude of any EIH bin is related in some fashion to decibel units. The major differences between the EIH model and the proposed model is that the number of level detectors in the proposed case is 1, and that unlike the several discrete level crossing events that each channel may contribute to in case of EIH, in the proposed system each channel can contribute to at most one frequency bin. Also, the periodicity calculation in the proposed system is based on the zero crossings rather than level crossings, making the system less susceptible to level changes in the signal due to the presence of noise. The effect of these differences is that the proposed system may be more robust to small perturbations in the amplitude of the response due to noise, as there is only one, relatively low amplitude level detector. On the other hand, slight changes in the amplitude in the short term may lead to increased or decreased counts in the frequency bins of the EIH. These perturbations would then be multiplied because of the multiplicity of the level detectors. This multiplicity of

level detectors would not always help the system, as the different level detectors would generally be contributing to different frequency bins. The main advantage of a multiplicity of level detectors would be in higher frequency regions, where the signals are heavily modulated by the fundamental period, resulting in beating in the speech signal spectra at the rate of the fundamental. However, for our analysis of limited frequency range, this would not be of any particular advantage. Moreover, the costs involved in terms of computing these multiple level crossings would be quite large.

The “Generalised Synchrony Detector” (GSD) proposed by Seneff [Seneff, 1988], also makes use of the temporal characteristics of the cochlea filter output stage. In this scheme, the output of the cochlear filtering stage first passes through several intermediate stages including low-pass filtering, half-wave rectification and adaptation. This adapted output is then processed through a GSD tuned to the centre frequency of the corresponding auditory filter. Thus, if there is a prominent peak in the signal corresponding to a frequency f , it is claimed that the result of GSD processing is such that the channel whose centre frequency is closest to f would specifically detect the “correct” periodicity, and the output of the adjacent channels would be comparatively small. However, the output of such a system is not particularly suited to pitch estimation. The author describes the results of the GSD processing as:

“Harmonic structure is completely obliterated in the synchrony spectrogram for male voices, but typically retained in the first-formant region of female voices. Harmonics between F1 and F2 are typically suppressed, because prominent energy in the first-formant frequency in the channel output destroys synchrony to the intermediate harmonics. Pitch striations over time are usually absent, due to the amplitude normalisation process.”

These characteristics may make the estimation of pitch from the GSD output difficult. The major difference between GSD process itself and the proposed system is that the contribution to any frequency component is a summation of contributions over all channels, rather than just from the design frequency channel. Since there are no intermediate stages, the fine temporal structure is analysed in its raw form in the proposed system, which results in the preservation of the maximum amount of information for the temporal processing.

The DOMIN system of Carlson and Granström in [Carlson, Granström, 1982] is also a temporal mode processing model that explicitly measures the frequency of the output of the cochlear filtering stage. However, unlike the proposed system, the DOMIN system only measures the number of channels that are oscillating at the same frequency, rather than the actual strength of the frequency component, based on multiple outputs at the same frequency. Thus the output of the DOMIN system is far more discrete, and may not be able to make fine distinctions between the relative strength of different harmonics, which is vital for the estimation of the pitch frequency in a harmonic grouping based system.

Cooke [Cooke, 1991] uses instantaneous frequency computation on the output of the auditory filters in order to estimate slow variations in the frequency components of a speech signal (or any other generic sound) over time. The estimates of these slowly varying estimates are used to form “place-groups” corresponding to the centre-frequencies of the filters which have similar instantaneous frequency. The final frequency estimate is made by a weighting function on these groups. The amplitude of each place group is measured by summing up the amplitudes of all filters making up the place group. This method is similar to the proposed system, however, the details of how the frequency and the amplitude calculated are quite different.

3.7. Noise Robustness and Biological Plausibility of the Proposed System

This chapter presents a novel way in which the processing of the input sound stimulus results in a representation that was demonstrated to be robust to noise. The robustness to noise arises due to two factors. The first is the broad selectivity of the DHO units. It was shown that this broad selectivity helps the system to handle noise. When presented with a signal with no noise, resulting activity increases the dynamic range of the system by providing an estimate of the frequency of the signal component which lies close to its characteristic frequency. When presented with an uncorrelated noise source and a tonal signal, the system still responds to the strongest frequency in the vicinity of its characteristic frequency, thus providing an effect similar to noise-masking, where the stronger frequency component masks the noise around the frequency region where the energy in the input is higher. The second reason for noise robustness comes from the treatment of the output of the DHO units on a “temporal pattern coding” basis rather than the more commonly used “channel based coding”. In the proposed temporal pattern coding scheme that was developed, the frequency information of the output is calculated from

the fine temporal structure, rather than the characteristic frequency of the signal. This leads to a superior time-frequency resolution, with the DHO system only providing a rough ordering of the frequency components, but a detailed temporal description of the signal, driven by the actual signal timing and form.

It is also claimed that the proposed scheme is biologically plausible. Timings of discharges in auditory nerve fibres reflect the time structure of acoustic waveforms, such that the inter-spike intervals that are produced precisely convey information concerning stimulus periodicities. Similar stimulus-driven temporal discharge patterns are observed in major neuronal populations of the cochlear nucleus [Richmond, Gawne, 1998]. Channel based models of frequency estimation depend upon the activation of specific neural channels or of configurations of channels. So constructed, channel-based schemes depend critically upon the extent to which particular neurons are activated. In such a model, if the connectivities of neurons are suddenly rearranged in the system, or the neurons are damaged, the coherence of neural representations will be disrupted, at least until the system can be adaptively rearranged to reflect the new channel-identities. Also, under the channel-based schemes, spatial patterns of channel-activation have to represent arbitrary combinations of stimulus properties in order to create signal feature maps, which may then be used for pattern recognition or learning purposes.

However, in the auditory system, as in many other sensory systems, receptor cells depolarise when stereocilia are deflected in a particular direction, such that the timings of spikes predominantly occur during one phase of the stimulus waveform as it presents itself to the individual receptor (for example, after having been mechanically filtered by the cochlea). This form of stimulus-locking is known as *phase-locking*, and is observed in the auditory system. Moreover, given that phase-locking exists, then the time intervals between the spikes that are produced (inter-spike intervals) reflect stimulus periodicities, such that time intervals themselves can serve as neural representations of stimulus form [Delgutte, 1995]. This is the argument which is used for the choice of temporal-code based representation of the signal information in the proposed system. Inter-spike interval information is extremely precise, permitting the fundamental period to be reliably estimated with a high degree of accuracy. This case is also analogous to the proposed system, where the DHO output information is used to estimate the periods of the outputs to a very high accuracy, increasing the frequency resolution of the system.

Although there is a large body of evidence to support these arguments [Cariani, 1997], [Rose et al, 1967], [Ryugo, 1992], whether such a temporal analysis is in fact implemented in the central auditory system, what form it might take, and where it might occur are issues that are presently under investigation [Cariani, 1999].

This chapter considered a new auditory information processing model, based on the operation of damped harmonically oscillating units. A system design was presented that was shown to produce a noise robust and high resolution time-frequency representation. This gain in performance in noise and resolution was shown to arise from the way the output of the DHO units is processed. In the next chapter, a pitch estimation system is presented. This system uses the representation produced by processing of the DHO outputs, to find reliable pitch estimates.

Chapter 4

THE PROPOSED PITCH ESTIMATION AND TRACKING SYSTEM

Pitch Tracking through Harmonic Grouping

In proposing a pitch estimation system, the aim of this dissertation is not to put forward a new auditory theory of pitch perception. The aim is to propose a very robust pitch estimation system, that is biologically plausible, and in accordance with principles of modern auditory perception theory and experimental data. To this end, this chapter describes the proposed system of pitch estimation and tracking based on harmonic grouping.

The idea of contributions to the pitch percept by some sort of auditory grouping is not new and is related to the binding problem in auditory scene analysis literature [Bregman, 1990] [Brown, Cooke, 1994]. The binding problem concerns the auditory scene analysis task of assigning a common identity or source to different auditory components. In the case of speech, this could be achieved by detecting common amplitude modulation or common fundamental frequency of the different spectral components. Indeed the models of 'harmonic sieve' and 'summary autocorrelation' implicitly group harmonics of a common fundamental. The 'harmonic sieve' perceptual experiments by Moore et al [Moore et al, 1985] for example demonstrated the "mistuned harmonic" principle which shows that the frequency components when shifted by more than 3% with respect to the fundamental period, gave rise to the perception of separate sounds. However, in most models of auditory processing for speech signals, the "binding problem" is solved by assuming a-priori availability of pitch information [Brown, Cooke, 1995], [Brown, Wang, 1997]. Usually, autocorrelation or summary autocorrelation based methods are used to provide an estimate of pitch in these models.

However, explicit grouping of spectral frequency components has remained problematic for practical fundamental frequency estimation algorithms because of the very high spectral resolution requirements. Such high spectral resolution demands great computational cost for any practical system.

In the proposed method, explicit grouping of harmonic components in the time-frequency plane (the frequency is measured in the outputs of the DHO units) is carried out, which leads to greater tolerance to noise and also provides a chance to carry out multiple pitch estimation and tracking. Other harmonic grouping systems, like the one described by Brown and Cooke [Brown, Cooke, 1995], rely on an estimate of the fundamental frequency to be computed first by autocorrelation, in order to determine if a particular harmonic belongs to a particular group.

Explicit grouping of harmonics has been proposed before, in systems related to the idea of the harmonic sieve [Duifhuis et al., 1982], [Scheffers, 1983]. These models, based on the findings of the perceptual experiments with mistuned harmonics [Moore et al., 1985], construct a template for all possible harmonics given a particular pitch hypothesis. The width of the sieve is derived from the findings of the perceptual experiments on mistuned harmonics [Moore et al., 1985]. These experiments establish that mistuning any harmonic by less than 3% has little effect on its contribution to the pitch percept, and that the contribution starts decreasing up to a mistuning of 8%, after which the mistuned component has no effect at all. Cooke in [Cooke, 1991], also uses a harmonic grouping scheme based on a common fundamental hypothesis, which produces separate harmonic groupings when there is more than one fundamental frequency present in the tonal complex. This system [Cooke, 1991] is quite similar to the proposed system, both, in terms of the initial hypothesis of pitch candidates, and in the assignment of saliency to each of these pitch candidates.

Some researchers have come to the conclusion that pitch perception results two mechanisms, one based on the contribution of the lower order (resolved) harmonics, and the other depending on the higher (unresolved) harmonics. Systems designed to take advantage of one of the aspects cannot, in general, simulate the contribution of the other aspect. It is thought that the temporal mechanism is used only for the processing of *amplitude envelopes*, and coexists with a completely different central processor of spectral cues [Terhardt, 1974], [Ohgushi, 1978], [Carlyon, Shackleton, 1994]. The theoretical differences arise from the relative importance ascribed to each signal aspect (i.e., temporal versus spectral cues, first order versus all-order inter-spike differences). In the next section, we shall present a brief look at the experimental and physiological data that motivates us to use explicit harmonic grouping to estimate pitch.

4.1. Perceptual Arguments for Harmonic Grouping for Pitch Perception

The original ideas of Helmholtz [Helmholtz, 1870] gave rise to immediate objections, based on insight gained by perceptual experiments with complex stimuli. When complex stimuli contain sounds that are harmonically related, they give rise to the perception of a tone that is largely indistinguishable from the pitches of an isolated tone of the same frequency. However, if the specific nerve energies of Helmholtz's models were to exist, all the frequencies would be expected to be heard separately.

In order to rectify the model of perception as put forward by Helmholtz, two scenarios were envisaged. The first one was that originally put forward by Gray [Gray, 1990], in which it was proposed that the perception of pitch arises from the principle of maximum stimulation. The second account for the perception of pitch is that a difference tone is generated [Plomp, 1965]. The principle of maximum stimulation was negated by the experiments that showed that the pitch of a harmonic complex could be perceived even in the absence of the fundamental pitch component [Fletcher, 1934]. Evidence against the difference tone theory was conclusively put forward by Shouten [Shouten, 1940]. He reasoned that if a sinusoidal signal is present in a stimulus, then it can be cancelled by adding a second sinusoidal component of exactly the same amplitude, with a phase difference of 180 degrees. He used this argument to cancel the difference pitch component in the stimuli he presented to his subjects, and found that even after the cancellation of the difference component, the corresponding pitch is clearly perceived, thus the difference tone could not be responsible for perception of pitch. Experiments by Licklider [Licklider, 1956], further demonstrated the absence of evidence for the difference pitch.

Another technique used for pitch estimation is the detection of periodicity in the modulation spectrogram of the signal [Dau et al, 1997], [Strope et al, 2001]. In these models, the higher frequency channels are used to predict the pitch period, usually through autocorrelation, by discovering the rate of amplitude modulation. Previous to 1956, most data indicated that the perceived pitch in complex stimuli corresponds to the frequency of amplitude modulation [Small, 1970]. According to Small, De Boer [Boer, 1956] showed that this correspondence is true only approximately. In perceptual experiments that they carried out, they presented their subjects with stimuli where the modulating frequency was kept constant and the carrier frequency was varied. Under the hypothesis that the pitch perceived corresponds to the rate of amplitude

modulation, the perceived pitch would be constant as the carrier frequency was varied. However, this was not found to be true in all conditions. In fact, what De Boer found was that as the carrier frequency increases, the pitch first increases, then abruptly jumps to a lower value, rises again, and repeats the same process. This variation in pitch is oscillatory about the modulation frequency. If the carrier frequency is kept constant and the modulation frequency is increased, a similar effect was measured with the pitch initially following the modulating frequency increase, then jumping abruptly to a lower value, and then continuing to increase again, repeating the entire process. De Boer, based on these effects, postulated that the perceived pitch does not actually correspond to the rate of amplitude modulation, rather, the pitch corresponds to those frequencies that are integral submultiples of the carrier frequency. When the modulating frequency is close to one of these submultiples, the pitch “locks in” on it, resulting in a jump to a lower value. Shouten [Shouten et al, 1962] termed these effects the “second effect of pitch shift”. These experiments demonstrated that the perception of pitch is not just based on the rate of amplitude modulation (i.e. the change in envelope of the signal), but depends on the temporal fine structure of the complex stimuli. According to Small, these effects also act as an argument against the “difference tone” theory of pitch perception. He argues that the spectral structure is uniquely determined by the modulating frequency, and if this remains constant, so should the difference tone. However, the perceived pitch increases when the spacing of spectral components does not change. Thus pitch perception is unrelated to difference tones.

Licklider [Licklider, 1951] hypothesized that the auditory system is able to calculate the autocorrelation function of a neural spike train, and to transform in this way temporal regularities into a place code for pitch. The neural scenario imagined by Licklider is depicted in figure 4.1. Nowadays, this specific neural scenario is often judged unrealistic [Kaernbach, Demany, 1998], [Strope et al, 2001], but Licklider’s basic proposal is still very influential [Lyon, 1983], [Slaney, Lyon, 1990], [Assmann, Summerfield, 1990], [Meddis, Hewitt, 1991] [de Cheveigné, 1998] [de Cheveigné, Kawahara, 2002], [Cariani, Delgutte, 1996].

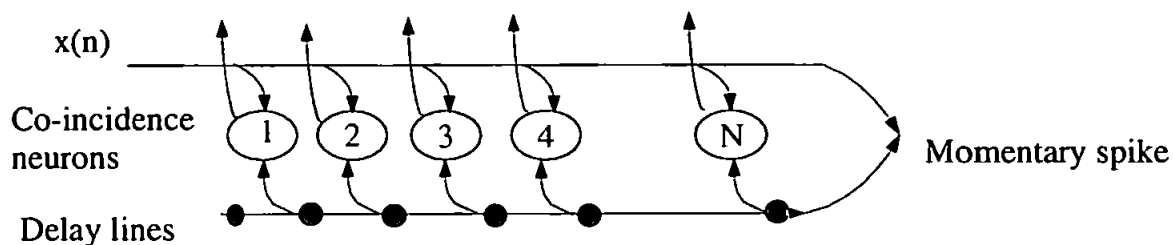


Figure 4.1. A neural autocorrelator by Licklider [Licklider, 1951]. A set of coincidence neurons is placed between a fast line and a delay line. The delay line is realized as a chain of neurons losing approximately 1 ms. per synaptic transmission.

The autocorrelation function and the summary autocorrelation calculations give a good indication of the dominant period in complex stimuli. However, Carlyon [Carlyon, 1996] and Kaernbach et al [Kaernbach, Demany, 1998] point out that relevant information about the pitch is primarily limited to first order differences or inter-spike intervals, and the method of autocorrelation does not distinguish between first order and higher order inter-spike intervals. They pointed out this inability of order distinction in autocorrelation based theory leads to the autocorrelation models giving rise to spurious peaks. Other problems with the autocorrelation method are those concerning multiple pitch tracking. Various schemes to estimate multiple periods have been proposed (see [de Cheveigné, 1998], [Wu et al, 2002] for a review). Single-period estimation models can be extended to estimate two periods by using secondary cues such as the second-largest peak in an autocorrelation pattern [Assmann, Summerfield, 1990]. However as de Cheveigne points out, this approach is not too effective. The "secondary cue" is often absent, or else not unique, or its position may not quite correspond to the secondary period. Also the primary period cues may themselves degrade in the presence of additional periods. He proposed a scheme which used a first period estimate to drive a harmonic segregation stage to suppress one voice, and then estimated the period of the other voice from the remainder. This "Joint Cancellation model" was based on a within channel "neural cancellation filter". After all the harmonics that belong to one period were removed, the remaining activity in the channels was used to derive the second period, and so on. This scheme has one fundamental theoretical problem. After cancellation, if the complex stimulus contains fundamentals that have common harmonics, the first cancellation stage will completely remove them and the detection of the second period may become erroneous. It has been shown in the literature that decreasing the

number of harmonics present in complex stimuli has a detrimental effect on the perception of pitch [Small, 1970]. Thus the removal of harmonics through the cancellation procedure leads to fewer harmonics for the secondary pitch, and therefore a reduced salience, leading to difficulty in estimation of the secondary pitch.

Moore [Moore, 1997] argued that most of the psychophysical data concerning the pitch of complex sounds can be understood on the basis of a simpler model. According to Moore, the pitch of a complex sound would simply correspond to the most frequent inter-spike interval (ISI) occurring in the responses of all the auditory nerve fibers excited by this sound. In a nerve fiber excited by a resolved spectral component with frequency f Hz, consecutive spikes will typically be separated by ISIs corresponding to $1/f$, $1/2f$, $1/3f$, ..., $1/nf$. In other nerve fibers excited by another resolved component, the ISIs will be partly different, but common ISIs will occur if the two components are harmonically related i.e., if the sound is periodic. The largest of the common ISIs will correspond to the period of the sound. As the corresponding ISI should also occur in fibers excited by the sum of several harmonics rather than by a single harmonic, this ISI should be overall the most frequent one. Note that although Moore's model posits that the pitch extraction process is the same for spectrally resolvable sounds and unresolvable sounds, it is possible in this conceptual framework to make sense of the fact that resolved harmonics provide more salient pitch cues than unresolved harmonics, due to the consistency of first order ISIs in the lower order harmonics. This forms one of the primary motivations for our choice of grouping of the harmonics in order to determine pitch period, and to carry out multiple pitch tracking.

From the study of different perceptual experiments and models, a picture of conflicting results emerges, a few which support one particular view, while other experimental data and models rule it out, at least in certain cases. This scenario has traditionally made it very difficult to propose a system that explains all the experimental data. In the end, the method of harmonic grouping was chosen because of its simplicity and plausibility as a pitch perception mechanism that naturally supports the estimation of multiple periodicities in the complex stimuli, as well as its low computational requirements.

4.2. Computation of Single Pitch Tracks through Harmonic Grouping

The sound signal that is presented to the system is first processed by a bank of damped harmonic oscillator (DHO) units. The output of this stage is further analysed to produce a representation that has high frequency resolution, and is robust to noise. This model of processing the auditory signal was presented in chapter 3, and its performance analysed in detail.

The high frequency resolution, which is a product of the temporal processing of the DHO unit outputs, is capable of resolving all the harmonics in the analysis range (lower frequency of 60 Hz, and the highest frequency of 1000 Hz). Information about these resolved harmonics is used in the system described below to produce pitch tracks for one voice present in the stimulus. In the next section, the system is further modified to handle the case of multiple pitch tracks, for stimuli which contain two simultaneous voices (with different fundamental frequency). An overview of the proposed system, with the various processing stages involved is presented in figure 4.2.

4.2.1 The Harmonic grouping and Periodicity Analysis System

The energy and frequency estimates calculated in the previous chapter provide estimates of sustained activity in the spectro-temporal domain for the resolved harmonics of the speech signal in the range of analysis (from $f_L = 60$ Hz to $f_H = 1000$ Hz). As pointed out previously, this range was chosen to reflect the range of resolved harmonics found in physiological data [Licklider, 1956]. It was also pointed out that the individual activity of the DHO is “locked in” to the harmonic closest to its characteristic frequency (see section 3.5 for detailed model of this processing). This output, produced by the energy-frequency analysis, is sampled at a rate of 100 Hz for the purpose of harmonic grouping.

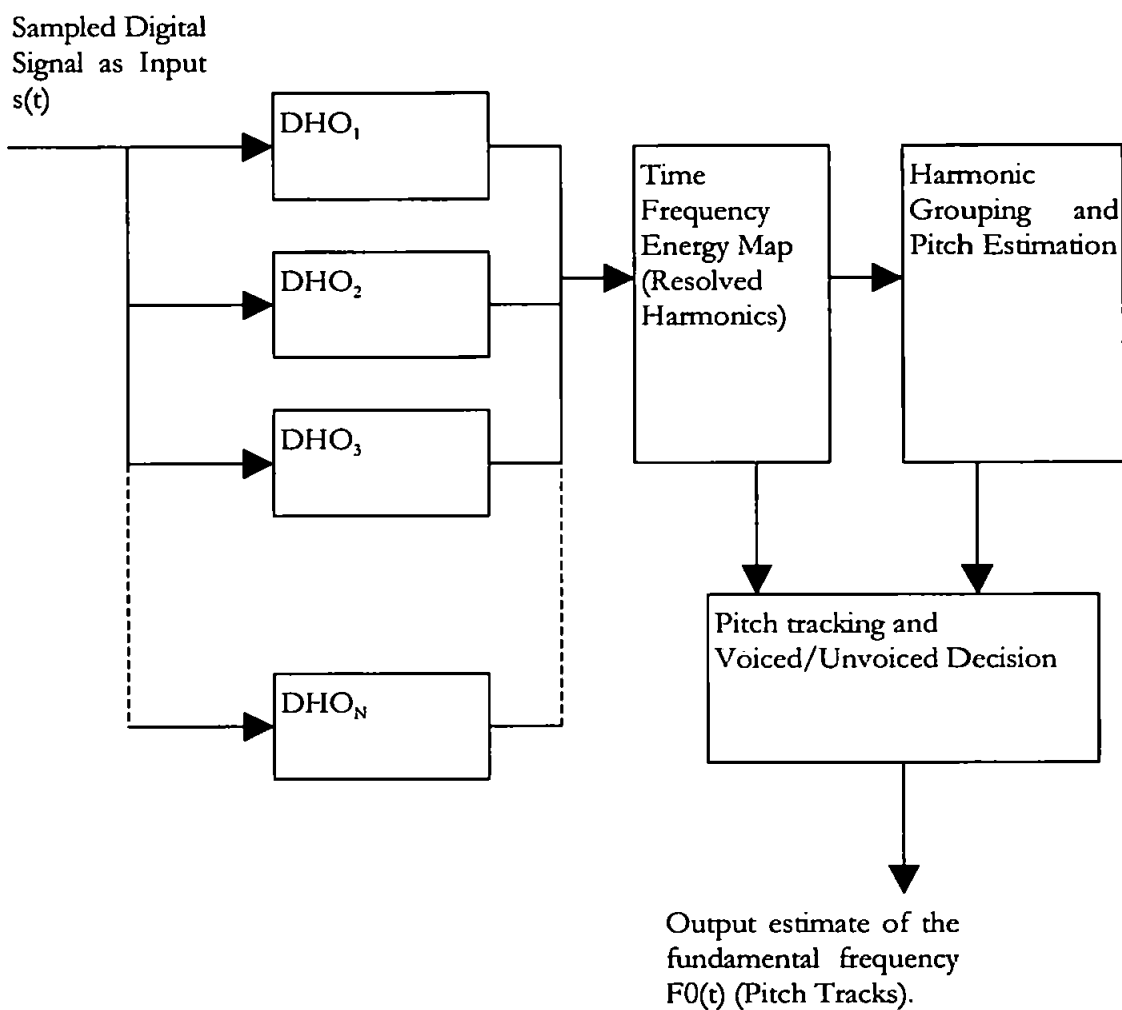


Figure 4.2. The system design and overview for single pitch track estimation. The Input signal is processed at several stages, giving the estimated pitch tracks of the speech signal as the system output.

The task of harmonic grouping is split into three stages of processing. In the first stage, sustained spectral peaks are found, which correspond to local maxima of the estimated energy from the amplitude of the bank of oscillators. The purpose of this stage is to pick out the harmonic frequencies and the corresponding energy in these harmonics from the full analysis range. The energy and the frequency of these locally maximum peaks in the estimate are treated as input for the next processing stage. The second stage is the estimation of possible candidates for fundamental frequency from the local maxima calculated in the first stage. The possible candidates are limited in this analysis by the range possible for normal speech, i.e. 65 Hz to

350 Hz. The third stage of processing then allocates a saliency figure to each of the possible candidates found, based on the energy of the harmonics that belong to that group. The membership of a group is calculated for each locally maximum energy-frequency pair, depending on whether the candidate fundamental frequency could have produced that harmonic. Apart from continuity constraints that are applied (described in the next section), the group with maximum saliency, is ranked as primary candidate for the foreground pitch estimation. This algorithm for harmonic grouping is described below in detail.

Stage 1. The input of this stage is the sustained activity in terms of the energy-frequency estimate calculated in section 3.5. This input is sampled at a rate of 100 Hz, irrespective of the sampling rate of the signal. Let the estimated energy content of the signal at this point be represented by $A(n)$, with n ranging from 1 to N (total number of DHO units = 40), and the frequency estimate of each unit (this is in general different from the characteristic frequency of the unit) as $F(n)$. These frequency estimates are tonotopically arranged, with $F(n)$ increasing from $n = 1$ to $n = N$. If a particular harmonic is present, more than one oscillator (depending on the amplitude) will have sustained activity at that frequency. However, the oscillator that has its characteristic frequency closest to the target harmonic has maximum sustained activity. Therefore, in an attempt to distinguish the output of other neighbouring oscillators from the maximum activity one, a simple peak picking algorithm is employed. These locally maximum energy frequency pairs are represented by $E(c)$ and $P(c)$. Physiologically, this situation could be likened with the 'lateral inhibition' in neural processing [McCabe, Denham, 1997][Denham, 2001]. Since the system here deals with populations of neurons but single units, a very simple and effective locally maximum principle is used. The Matlab code below describes this step algorithmically.

```

If (A(n) > A(n-1)) & (A(n) > A(n+1)),
    E(c) = A(n);
    P(c) = F(n);
    c=c+1;
End

```

In the code above, the estimated energy vector for the current time is stored in vector A (indexed by n), and the local peak energy and corresponding frequency are stored in vectors E and P respectively.

Stage 2. This state of processing provides the system with possible pitch frequency candidates that may explain the current activity pattern stored in the locally maximum energy-frequency pairs, designated above as $E(c)$ and $P(c)$. The initial estimation of possible pitch candidates is conditional on whether the system is currently processing voiced speech activity. This is indicated if the previous activity was assigned a pitch. If it is, then the initial pitch candidate is computed by the steps described below.

Let $f_{0,i}$ represent the previous pitch frequency estimate. Let the parameter for rate of change of pitch frequency (this parameter also allocates some margin for error in frequency estimation process) be a constant Γ . Then for each energy-frequency pair detected in stage 1, the algorithm described by the Matlab code below is applied to find the pitch hypothesis, given the prior knowledge of the previous pitch estimate.

```

M = round(P(k) / f0_{t-1});
If mod(P(k)/M, f0_{t-1}/f0_{t-1}) < Gamma,
    GF(1) = P(k);
    break;
end

```

In the Matlab code above, $GF(1)$ is the initial pitch estimate. The frequencies are stored in the variable P and k is the index which iterates through possibly all the local peak amplitude frequencies. The operation is intended to find the closest possible pitch estimate to the previous estimate. The term $f_{0,i}/M$ means that the $P(k)$ is the M^{th} harmonic of the possible current pitch estimate $GF(1)$. For example, if the previous estimate was $f_{0,i} = 100$ Hz, and $P(k) = 310$ Hz, then $M = 3$ ($P(k)$ is the 3rd harmonic), and $d = 10$, then $GF(1) = P(k)/M = 103.33$, because $d/f_{0,i} = 0.1$ which is less than $\Gamma (=0.2)$.

The above calculations ensure that if there are any harmonics in the current activity data that correspond to the previously estimated pitch, the corresponding pitch candidate is included in the list of all possible candidates first.

If the previous pitch estimate is zero (unvoiced), then an alternative method of finding the initial estimate is used. It involves generating the sub-harmonic series of the first peak (a *sub-harmonic series* is the inverse of the harmonic series, that is, it is comprised of frequencies of $P(k)$, $P(k)/2$,

$P(k)/3, P(k)/4, P(k)/5 \dots$), and selecting those elements of the series which lie within the pitch range as the initial pitch candidates, GF .

After the formation of initial estimates as described above, the task is to search for peaks in activity whose frequency, $P(k)$, is not explained by the initial pitch candidates, i.e., those harmonic frequencies in the set $P(k)$, which are not harmonics of any of the candidates in GF . This process is similar to finding the initial pitch candidate, given the previous pitch frequency $f_{0,1}$, where the initial estimate is replaced by $GF(1)$, and the constant Γ is replaced by θ , which is the tolerance factor for deviation of the harmonic due to inaccuracies in measurement, or noise. The value used for θ is 0.1 (10% deviation is allowed). For all these frequencies, the sub-harmonic series is constructed, and all candidates that fall within the normal human pitch range are appended to GF . Thus at the end of stage 2, GF contains all the possible pitch candidates which could explain the activity in the current estimate. The frequencies stored in GF are called *group frequencies* because they possibly explain a whole group of harmonics in the current frequency estimates $P(k)$.

Stage 3. This stage of analysis ascribes saliency to each of the pitch candidates discovered in the previous processing stage. This process is a simple double loop, in which all the pitch candidates of the previous stage (represented here by GF), are tested for all the current sustained activity data (represented in stage 1 by variable E). The process can be likened to assigning a membership value to each data point for each candidate, based on how much they explain each of the pitch estimates. In this scenario, a single data point (defined by a frequency and amplitude of activity, $P(k)$ and $E(k)$ for the k^{th} such element) can contribute to different pitch candidates, GF . The contribution of each data point to each pitch candidate is inversely proportional to its distance from the current pitch candidates, and directly proportional to the associated activity (the energy measure $E(k)$). Mathematically, if the contribution of data point k to candidate i is defined as $C_{k,i}$ and the distance of the frequency of the data point and the candidate as $D_{k,i}$ then, the algorithm in form of Matlab code below describes the computation of the saliency contributions. In the algorithm, the variable G contains the pitch hypothesis candidates, and P contains the frequency and E contains the energy of the local peaks. The computation of $D_{k,i}$ is represented by $D(k, i)$ and $C_{k,i}$ is represented by $C(k, I)$.

```

For I = 1 : NumberOfPeaks,
  For j = 1 : NumberOfCandidates,
    D(k, i) = mod(GF(i) / P(k);
    If (1 - D(k, i) <  $\theta$ , D(k, i) = 1-D(k, i);
    If D(k, i) <  $\theta$ , C(k, i) = E(k);
  End
End

```

Where θ is the tolerance factor defined in stage 2. The winning candidate is simply the candidate with maximum total contribution from all the current activity. This candidate is the primary or foreground pitch estimate. The second group (whose group frequency GF is not a harmonic or a subharmonic of the primary group), forms the background pitch estimate, and so on. The saliency of the pitch estimate is simply the sum of contributions for that candidate, i.e., we define saliency for pitch estimate $GF(i)$ as $GS(i)$, given by equation 4.1, where K is the total number of peaks.

$$GS(i) = \sum_{k=1}^K C_k \quad \dots 4.1.$$

The algorithm described above is used to form possible groups of harmonics. However, as may be noticed, the groups thus constructed may not always be “true” groups, and some groups may be sub-groups of others. This is because the sub-harmonic series that are constructed in step 2 give rise to a number of groups, which are harmonically related. Thus if two or more groups have the same saliency, as measured in equation 4.2, then the group with maximum group frequency is selected as the dominant group, and all the other groups that are harmonically related to the primary or foreground group are not considered when determining the background group. In the above algorithm, several constants are used for the formation of harmonic groups, and the calculation of group saliency. These parameters of the system are tabulated in table 4.1 below.

Parameter	Brief Description	Value
LowF0	Lowest possible pitch in the search range	60 Hz
HiF0	Highest possible pitch value in the search range	350 Hz
Gamma	The continuity constraint, as a percentage of current pitch	20%
θ	Tolerance factor for harmonic group membership as a percentage of the group frequency	10%

Table 4.1. The parameters of the harmonic grouping algorithm with typical values.

4.2.2. The Voiced Unvoiced Decision/ Pitch Tracking Algorithm

The voiced unvoiced decision is a very important part of the algorithm as it is critical for good performance in challenging noise conditions.

The input to the system is the primary group saliency, $GS(primary)$, associated with the primary group as described in section 4.2.1. The output is a flag which specifies if the input speech signal is voiced or unvoiced at a particular point in time. The aim of the algorithm is to produce continuous pitch tracks, and to avoid short duration voiced or unvoiced segments. These situations of short discontinuities may arise due to noise in the signal, or when the saliency of voicing $GS(primary)$ falls below a fixed threshold. The algorithm achieves this by the operation of a state machine, where the current state of the algorithm determines the action to be taken. The state variables that are tabulated in table 4.2, along with a brief description of their function.

State Variable	Description
<i>currentActivity</i>	The current saliency indicator, true if over a fixed threshold, otherwise false.
<i>tracking</i>	The current voicing indicator, if the algorithm has a recent history of voicing, then this is set to true, otherwise false.
<i>gapLength</i>	The state variable for the length of the unvoiced sections during unvoiced sections.
<i>wordLength</i>	The state variable for the length of the voiced sections during voiced sections.
<i>startTime</i>	The state variable for storing the starting of voicing.

Table 4.2. The state variables of the pitch tracking algorithm.

The operation of the algorithm also depends on some parameters that control the state transitions. These control the allowable length (in terms of frames) of the voiced and unvoiced sections, and also the decisions of whether a particular frame is voiced or unvoiced. These parameters were initially chosen heuristically, depending on the minimum word length, the possible gap length, and the general level of activity during voicing (the minimum values of pitch saliency for voiced sections), and later manipulated if found unsatisfactory, based on performance on test utterances. These parameters of the system are presented in table 4.3.

Parameter	Description	Value
θ_p	The voicing threshold, acting on the group saliency	3 (absolute threshold)
θ_v	The minimum track length that is a valid track	10 (corresponds to 100 ms)
θ_g	The minimum gap between voicing that may be a genuine gap in voicing	4 (corresponds to 40 ms)

Table 4.3. The parameters of the pitch tracking algorithm and their description.

The value of 100 ms for the minimum voiced sections as defined by the parameter θ_v has been found to be in accordance with the studies carried out by Greenberg and others in [Greenberg et al., 1996] on the properties of conversational speech derived from the Switchboard corpus.

The Voice Activity Detection Algorithm:

The algorithm for voice activity detection is presented here in the form of Matlab code. The various variables involved in the algorithm are explained in table 4.2, and parameter values in table 4.3. The basic feature of this algorithm is that it favours either continuous activity, or continuous silence. This algorithm enables the system to get rid of spurious peaks and valleys in noisy environments.

Let the group saliency be denoted by a variable called *gIntensity* at point in time *t*.

```
If gIntensity >  $\theta_{sp}$ ,
    currentActivity = 1;
else,
    currentActivity = 0;
end
If tracking == 1,
    If currentActivity == 1,
        trackLength = trackLength + 1;
        gapLength = 0;
    ElseIf gapLength >  $\theta_g$  ,
        tracking = 0;
        Output = 1;
    ElseIf trackLength >  $\theta_w$  ,
        gapLength = gapLength + 1;
        Output = 1;
    Else,
        Reset(startTime);
        Output = 0;
    End
Else,
    If CurrentActivity == 1,
        Tracking = 1;
        TrackLength = 1;
        GapLength = 0;
        StartTime = t;
        Output = 1;
    Else,
        Output = 0;
    End
End
```

The output of the algorithm, the variable *Output*, is a Boolean value which is 1 in the case of voiced sections and 0 otherwise. The function *Reset* in the code above rolls back the previous voiced decisions to unvoiced, starting from the time instant *startTime*.

4.2.3: Example output in response to complex stimuli

In this section the response of the system to two different complex stimuli is illustrated. The first stimulus is a speech like artificial signal, as described in equation 3.28. The second stimulus is speech stimulus from a male speaker. We present the analysis of these stimuli both as clean signals and white noise added at 0 decibels signal to noise ratio.

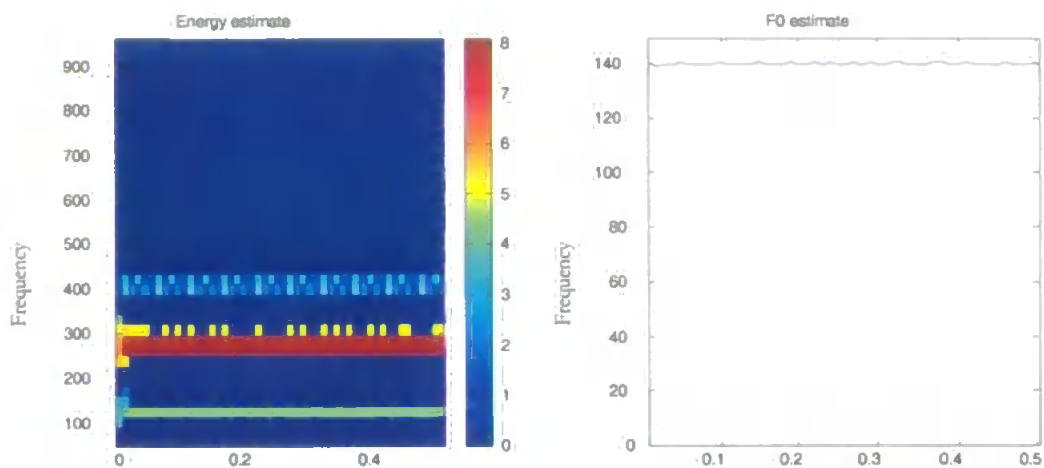


Figure 4.4. The system response to the system to the speech like synthetic stimuli. The y-axis is the frequency axis, and the x-axis is the time axis. It is interesting to note that there is some activity in the 420 Hz range, even though the signal component itself is not present. The periodicity estimate is quite accurate.

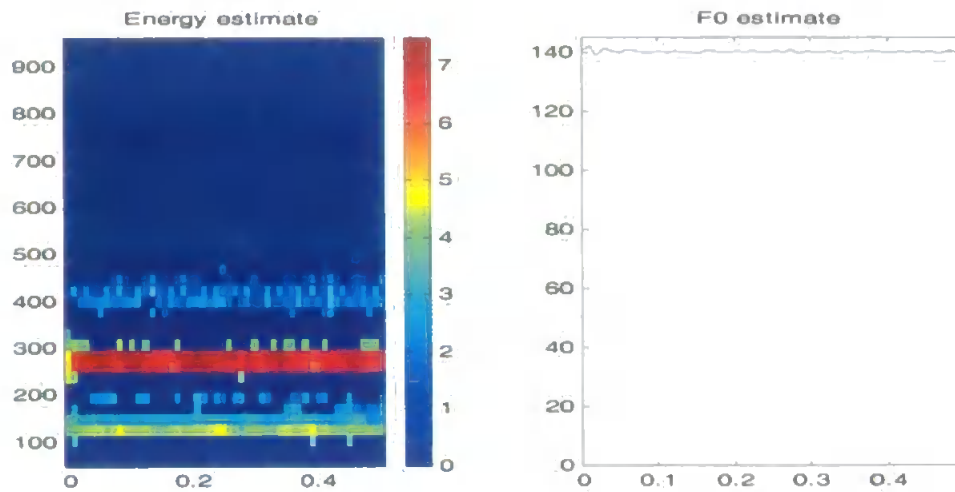


Figure 4.5. The same input stimulus as in figure 4.4, but with added white noise at 0 dB SNR. Although the channels with fundamental periodicity appear to be affected by noise in spectro-temporal view on the left, the periodicity estimate is still quite accurate.

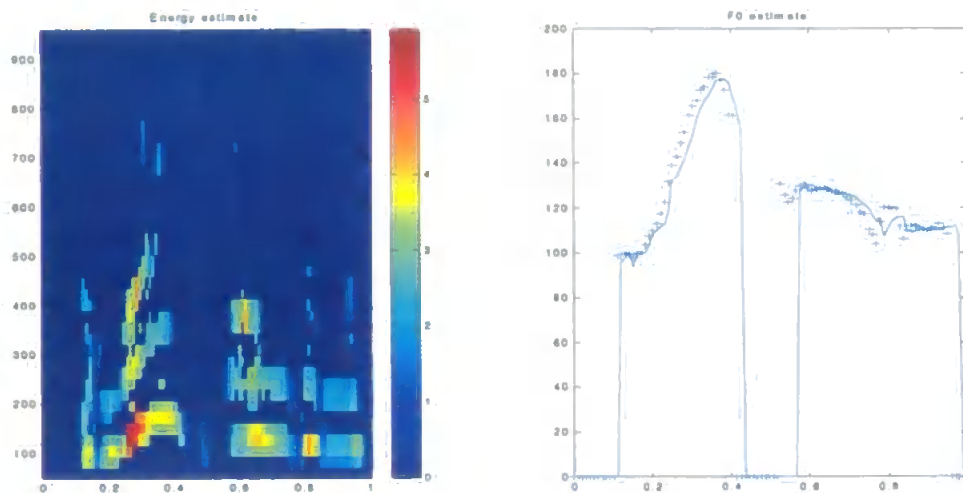


Figure 4.6. The stimulus is the speech from the first second of the m1 (male) speaker in the database collected by Plante et al [Plante et al, 1995], saying “the north wind”. The pitch estimate is quite accurate in clean speech conditions. There are some onset effects (around 0.58 s point on the x-axis), which make the reference track (marked by ‘+’) voiced, while the proposed system (continuous line), marks them as unvoiced. There is little activity for this region in the spectro-temporal view as well. The reference tracks were obtained by the laryngograph data collected during the recordings of the speech data.

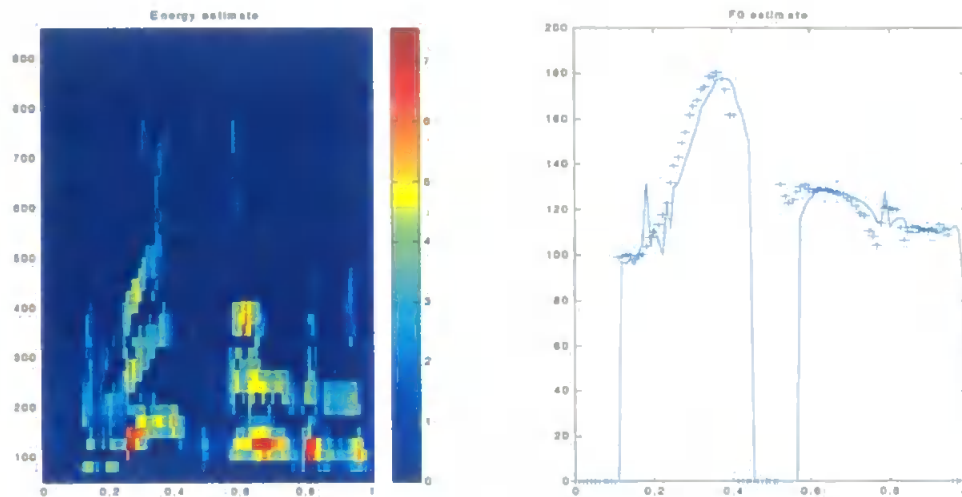


Figure 4.7. Response to the same stimulus as presented in figure 4.6, but with added 0 dB white noise. There are some estimation errors, but no pitch doubling or halving effects. The spectro-temporal view shows that the output at this stage is robust to high levels of noise.

4.3. Extension to the Estimation of Multiple Pitch Tracks for Simultaneous Speech

People are often presented with the perceptual task of segregating auditory signals occurring simultaneously, for example when listening to a target voice, in the presence of competing voices in the background*. Researchers have hypothesized that the auditory system is able to group these auditory stimuli into *perceptual auditory streams*, based on certain primitive features. One such feature is the pitch of the voiced sounds [Darwin, 1984], [Assmann, Summerfield, 1990], [Bregman, 1990]. Given the above assumptions about the mechanism of segregation, multiple pitch estimation for simultaneous speech must occur before the recognition of the perceived stream, and therefore, must be done at a lower processing level. This has been our motivation for the estimation of multiple pitch tracks in situations where speech from more than one speaker is presented to the listener (and our system) simultaneously.

Psychophysical experiments have shown that listeners use the fundamental frequency cue to group harmonic features together for simultaneous vowel recognition [Assmann, Summerfield, 1990], [Scheffers, 1983]. These experiments show that the relative level of the two vowels in a

* The foreground - background concept is referred to here as an attention correlate. That is, if the target of attention is stimulus A, then it is the foreground, irrespective of the stimulus strength of A, compared to the *background* stimulus B.

pair can be varied over a wide range, without affecting the recognition performance, as long as the vowels are present at different pitch. This difference is quantified in some experiments [Scheffers, 1983] as about 12% of the lower fundamental frequency, for pairs of vowels at least 200 ms in duration.

As presented in the previous chapter, there were a few modification made to the system for it to be able to extract multiple pitch tracks. These changes are not changes to the fundamental system of harmonic grouping, but some continuity constraints placed on the tracking and estimation system, which make up for the lack of the “attention” like feedback in order to form pitch tracks from simultaneous speech sounds. Although the system in theory is able to track multiple pitches, we have put it in the framework of foreground and background “streams”. This is because for more than two groups, not much experimental data is available, and also, it would require a finer frequency resolution than our current system, due to the increase in number of harmonics and decreasing spacing between the harmonics in the cases with more than two voices.

4.3.1. Two Pitch Track Considerations

Depending on the local saliency of a group of harmonics, there is a possibility that the estimates of pitch switch between foreground and background tracks. This was one of the biggest practical problems faced by this system. The second problem was that sometimes the pitch of one track becomes too close to the pitch track of the other because of the tolerance factor. In the original system, this tolerance factor was designed to take account of the maximum rate of change of pitch, and the error in estimating the frequency of the DHO output. These problems with the original system required some minor modifications, which are explained in this section.

All the problems are linked to the “attention” mechanism which acts as a feedback in the assignment of pitch estimates to pitch tracks. The modifications are implemented in terms of introducing an additional continuity heuristic to account for the “attention feedback”, and the tolerance factor is lowered.

The continuity constraint is applied once the harmonics have been grouped and their saliency calculated. At this point, a “look-back” heuristic is applied, which takes account of the modified tolerance factor for changes in pitch estimate from the previous section. If the previous pitch

estimates are available (the previous section was voiced), irrespective of the relative amplitude, the identity of the track is decided on the distance of the current estimates from the previous estimates. If the previous section is not voiced, but there has been a pitch estimate in the recent time (previous 0.5 seconds in this case), then these estimates are used as a reference. If however, no pitch tracks have been computed within this time, the foreground and background estimates are made on the relative saliency of the estimated groups.

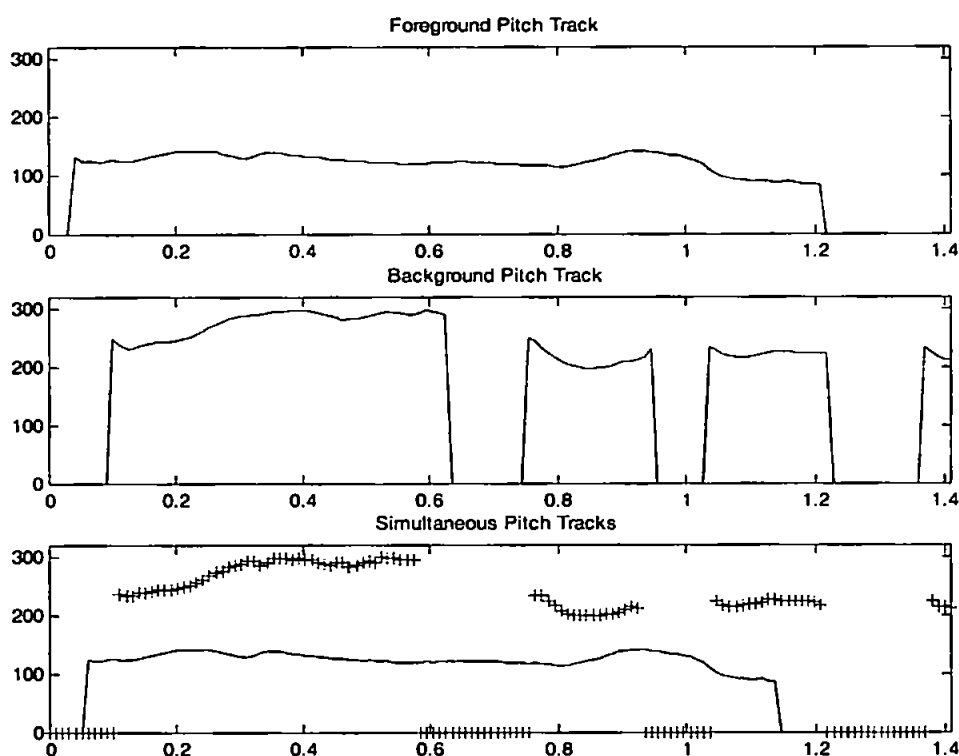


Figure 4.8. The pitch track on the top is the intended foreground pitch track. The one in the middle is the background pitch track. Both were computed independently with no noise. The bottom plot shows the multiple pitch tracking output of the proposed multiple pitch tracking system. The continuous line is for foreground pitch track, and the '+' line is for background pitch track. The signals in this case were added together at 0 dB SNR. The top (foreground) speaker is male, saying “Why were you all weary”, while the middle (background) speaker is female, saying “Why were we keen to use human”.

In the figure above, an example of processing that has two pitch tracks is presented, showing the pitch tracks separately, and then their simultaneous tracking using the proposed system. The isolated “reference” pitch tracks were computed using the proposed system prior to addition of the utterances, without any added noise (there was no laryngograph data available for these utterances, however, these utterances are very often used in computational auditory scene analysis research [Cooke, 1993]).

4.4. Modeling of Some Perceptual Pitch Phenomena

The primary purpose of the model presented in this chapter is to estimate pitch in speech signals and to be robust to noise and interfering speech. The model has been developed from the ground up for robustness to noise, in a biologically plausible way. An analysis of some psychophysical phenomena related to pitch perception is quite important in order to propose a pitch estimation model that aims to emulate the human performance in relation to pitch perception and its robustness to different kinds and levels of noise. In the next chapter, a detailed analysis pertaining to the noise robustness of the proposed pitch estimation system is presented. In this section, a few of the well researched perceptual pitch phenomena are analysed, and the behaviour of the proposed system for some controlled signal conditions is presented.

4.4.1. The Missing Fundamental Case

In the case of the *missing fundamental* phenomenon of pitch perception, the stimulus is composed of a set of successive harmonics, without the fundamental frequency. However, listeners perceive the pitch of the missing fundamental, even though it is not part of the stimulus [Boer, 1977].

To emulate this behaviour, the system was presented with a complex tone consisting of different successive harmonics of a fundamental that was missing from the complex. The system performance was found to be consistent with the perceptual data, and the pitch of the complex was correctly estimated as that of the missing fundamental. Figure 4.9 and 4.10 present two examples of the output of the system for these experiments.

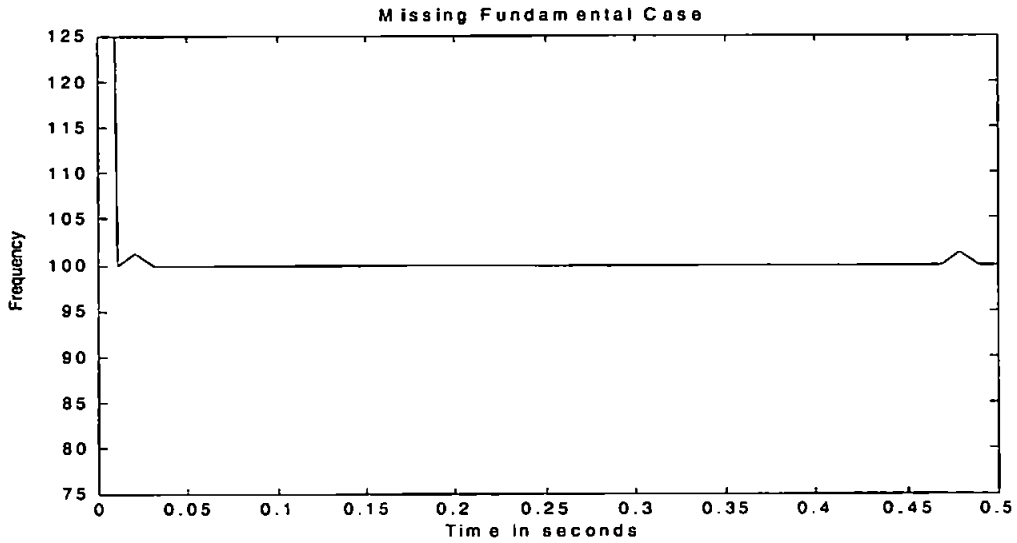


Figure 4.9. System output for the virtual pitch experiment. The input signal to the system consisted of the 3rd, 4th and 5th harmonic of the 100 Hz tone. After a brief transient period at the beginning, the system is able to track the pitch accurately.

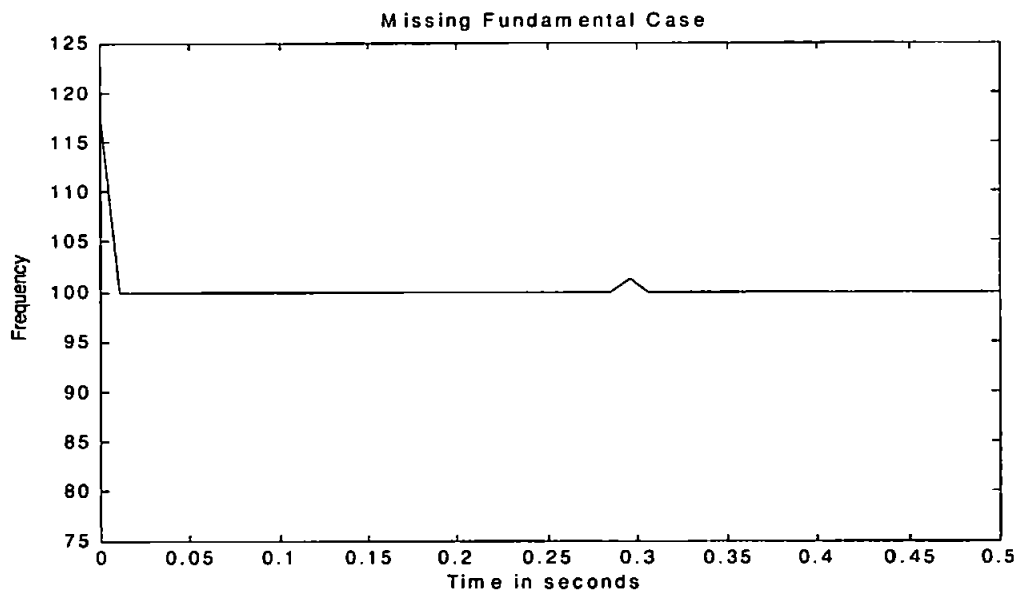


Figure 4.10. System output for the virtual pitch experiment. The input signal to the system consisted of the 6th, 7th and 8th harmonic of the 100 Hz tone. After a brief transient period at the beginning, the system is able to track the pitch accurately.

4.4.2. The Pitch of Iterated Ripple Noise Case

When a broadband noise stimulus is delayed and added to itself, the resulting stimulus has a pitch that increases as the delay is decreased. This phenomenon is called the *Repetition Pitch* or the

pitch of *Iterated Ripple Noise* [Bilsen, 1966], [Yost, 1978]. If the delayed noise is added to itself with a delay t , then the perceived pitch is $1/t$. Also, as the number of iterations is increased, the perceived pitch increases in saliency, and becomes more well defined. These phenomena were successfully emulated by the proposed system when presented with the *Iterated Ripple Noise* (IRN).

The perceptual results of IRN stimuli in terms of perception of a pitch in presence of such stimuli, have a pitch which varies with the reciprocal of the delay, [Bilsen, 1966], and the strength of the pitch increases with the number of iterations, and with the absolute value of the attenuation factor. IRN stimuli have been used to study pitch perception, principally, in time-domain auditory models which suggest that pitch is mediated by regularity in the fine-structure of the IRN stimulus. The temporal information is extracted by autocorrelation either directly from the waveform [Yost, 1997] or from the autocorrelogram [Meddis and Hewitt, 1991] [Patterson et al., 1996]. The transfer function of the IRN network has a spectral ripple, and, prior to the temporal hypothesis, the spectral peak spacing in the ripple was used to explain the perceived pitch [Bilsen, 1966]. The spectral ripple can probably not be resolved in the spectral region above about the eighth harmonic of the reciprocal of the delay, and yet IRN stimuli restricted to this spectral region still produce strong pitch perceptions [Patterson et al., 1996]. So the spectral hypothesis has been rejected for high-pass-filtered IRN stimuli at least.

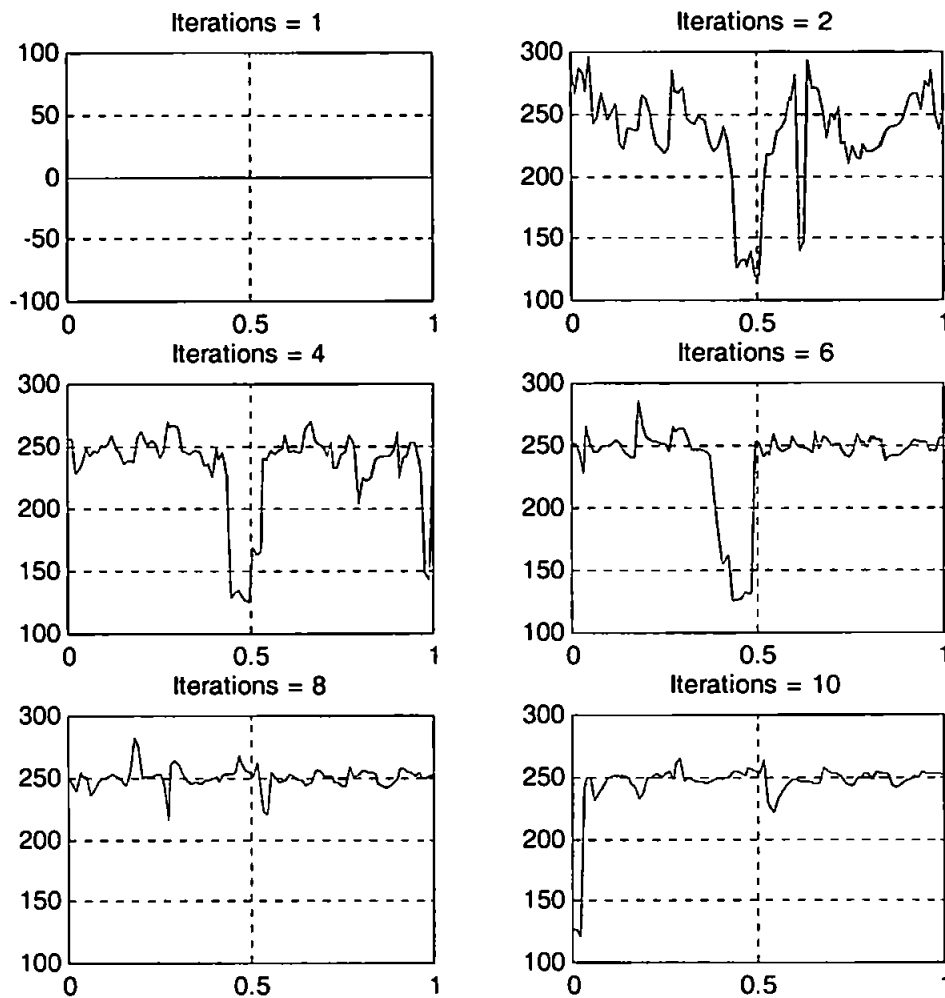


Figure 4.11. The Output of the system in response to iterated ripple noise input. As the number of iterations is increased, the output gets closer towards the frequency corresponding to the delay (delay was 4 ms, corresponding to a pitch of 250 Hz.).

In the experiments that were carried out on the proposed system, the stimulus included IRN. The system was presented with IRN with increasing number of iterations, starting with 1 (no addition, and therefore pure broadband noise in the input signal), and finishing with 10 (well defined envelope periodicity). It was observed that as the number of iterations were increased, the output of the system became increasingly accurate. These results are presented in figures 4.11 and 4.12. In figure 4.11, the actual output of the system corresponding to various iteration

counts is presented. In figure 4.12, the standard deviation of the outputs (around the 250 Hz mean) is calculated. The results are in general agreement with perceptual data.

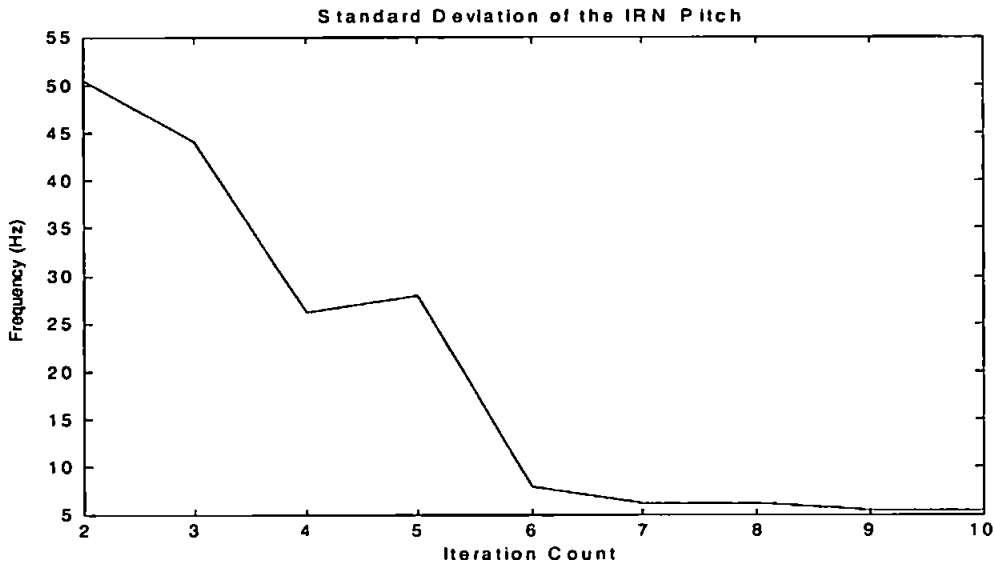


Figure 4.12. The standard deviation of the pitch estimates for IRN stimulus with increasing number of iterations, with iterations increasing from 1 to 10. The first point corresponding to the iteration 1 is excluded, as the system output was zero throughout for that case (see figure 4.11).

4.5. Key Aspects of the Proposed System

This chapter completes the presentation of the proposed system for pitch estimation. Chapter 3 provided the details of the DHO based processing of the input speech signal, while this chapter provided the details of a harmonic grouping algorithm for pitch estimation from the resulting representation.

The DHO based processing results in a broadly tuned tonotopic ordering of the frequency information in the input signal. This output is then used to develop a high resolution time-frequency representation that was shown to be robust to challenging noise conditions. The source of this noise robustness is the temporal processing of the output of the bank of DHO units, rather than the usual channel based processing. This processing results in DHO units “locking in” to the closest harmonic, and thus exhibiting a “noise masking” behavior.

The harmonic grouping based system is used to estimate the pitch of the voiced section of the input speech signal. One of the major differences with respect to other systems of grouping harmonics [Brown, Cooke, 1995], that the proposed system does not rely on any prior estimate of the pitch frequency, and the harmonic grouping itself leads to an estimate of the pitch frequency. Also, unlike the “Joint cancellation model” of de Cheveigné [de Cheveigné, 1998], the harmonic grouping system is based on “joint activation”, where one harmonic can contribute to several groups. One of the advantages of this approach of “joint activation” is that secondary or background harmonic grouping is more reliable, as more harmonics are available for their estimation (due to possible joint membership of the harmonic groups). This leads to more reliable background pitch estimation. Another source of robust performance is the use of continuity constraints in the algorithm, which reduces the likelihood of pitch halving and pitch doubling. It was also demonstrated that the system is able to emulate some perceptual pitch phenomena like pitch of the *missing fundamental*, and pitch of the *iterated ripple noise*. Because the pitch frequency is estimated through harmonic grouping, the process results in the establishment of foreground and background pitch estimates. This approach to pitch estimation makes the proposed system naturally suitable to multiple pitch estimation in the case of simultaneous speech from two speakers. The proposed system for harmonic grouping is different from systems which carry out explicit grouping of harmonics like those presented in [Duijfhuis et al, 1982], [Cooke, 1991]. While most of the systems would explicitly choose a fundamental to satisfy a template composed of a group of harmonics, the proposed system does not involve any template creation, and the operations are performed on spectral peaks, some of which may not be harmonics of the fundamental frequency.

The combined noise handling ability of the DHO output based representation and harmonic grouping makes the system robust to a large variety and levels of noises. The harmonic-grouping based pitch estimation system can handle the case of missing fundamental pitch, and is therefore suitable for pitch estimation of telephone quality speech, where the fundamental frequency component is usually missing. In the next chapter, we describe the experiments and the statistical study that was undertaken for the evaluation of the proposed system, and compare the performance with other pitch estimation systems.

Chapter 5

EXPERIMENTAL SETUP AND RESULTS

The Statistical Analysis

An appropriate evaluation of the various pitch estimation systems, and their performance on complex signals like noisy speech is far from being either simple or trivial. Many criteria for this evaluation have been provided in the literature [Rabiner et al, 1976], [Terhardt et al, 1982]. However, the performance of most modern systems has rarely been evaluated and compared in detail for clean speech signals and the evaluation of these systems in challenging noise conditions is completely non-existent. Even when a comparative analysis of performance is available for clean speech signals [de Cheveigné, Kawahara, 2002], it does not include all the different signal conditions, or the different error metrics. However, such an analysis is important for a meaningful evaluation for practical use of these systems.

Performance of these systems even for signals in quiet environments is difficult to compare, because of the fact that these evaluations are carried out in a variety of different conditions, which make it impossible to make quantitative assessment of their performance. The various sources of this variability in evaluation are enumerated below.

- a.) The use of different databases of signals.
- b.) In case of use of the same database(s), the use of different sub-sets of the same database (and selection of different parts of the target signal on which the evaluation is made), and organization of the test procedure.
- c.) Use of different methods to obtain and refine the reference pitch data to evaluate the performance of the target systems.
- d.) Use of different and sometimes only a sub-set of statistical error measures, which makes the assessment of the overall performance difficult.

Although a particular system may clearly demonstrate reproduction of a few perceptual effects, all the factors mentioned above conspire to make the task of overall performance assessment, comparison, and appreciation of relative strengths and shortcomings of the pitch estimation systems very difficult, and risky.

One of the primary aims of this dissertation is to evaluate the most prominent and modern pitch estimation systems, and compare them with the system proposed in this dissertation. This process of evaluation was carried out by means of detailed statistical analysis of the results of pitch estimation by these systems, with error measures that were recommended by Rabiner [Rabiner et al, 1976], and have become an accepted standard for such an evaluation. The data used was supplied by Plante et al [Plante et al, 1995] for clean speech, and the CASA [Cooke, 1993] dataset for multiple pitch track based evaluation. Noisy environments were simulated by adding various noises at various signal to noise ratios (SNRs) to the original data. These noises were also obtained from the CASA database of noises.

The evaluation process will be discussed in next section of the chapter. This is followed by a summary of the results for single pitch track evaluation, and then the multiple pitch track evaluation is described and a summary of results presented. The detailed results are presented in Appendices II (single pitch tracks, high resolution signals), III (single pitch tracks, low resolution signals), and IV (two pitch tracks, for both high and low resolution signals).

5.1. The Apparatus Used

In this section we describe the setup of the experiments that we used to evaluate the different systems, including the data, software and hardware used, and the software used to evaluate other reference pitch tracking algorithms, parameters, and other relevant details.

5.1.1. Description of the Data Used for Evaluation

The database used for evaluation of single pitch track estimation was prepared by Plante [Plante et al, 1995], especially for the purpose of evaluating of the pitch estimation and tracking systems. This database is commonly known as the “Keele Pitch Database”, and we shall refer to it by this name as well. The database is available on the internet via an anonymous ftp server at [ftp.cs.keele.ac.uk/pub/pitch](ftp://ftp.cs.keele.ac.uk/pub/pitch). The database is divided into modules for speech analysis, and psychophysical analysis. Only the module for speech analysis was used.

The speech analysis module of the database consists of ten speakers, five males, and five females. All the speakers are recorded speaking a phonetically balanced text, labeled the “north-wind story”. The original database also includes speech signals recorded by five children, which were not included in our analysis. During the recording, the original laryngograph signal was also recorded. After the recording, the signals were digitized at 20,000 Hz sampling rate, with a resolution of 16 bits for each sample, both for the speech signal and the laryngograph signals. The delay involved in the recording of the two signals is rectified as much as possible (the delay could not be fully removed because of slightly different vocal tract length of different speakers). The digitized laryngograph signal was then used to establish a reference periodicity signal, using an autocorrelation based algorithm. For those parts of the signal where there was observed periodicity in the laryngograph signal, but no clear periodicity in the speech signal, the corresponding frames were labeled with the negative of the period of the laryngograph activity. When periodicity was observed in the speech signal, but no laryngograph periodicity exists, the frames were assigned a value of -1 . These frames combined to produce a total of about 5% of the number of frames, depending on the speakers.

The reference signal of periodicity supplied with the database, was produced at 100 Hz, and used as such, with some modifications. The reference periodicity signal was pre-processed as follows. The signal was divided into three categories, i.e., the unvoiced or silence section, voiced or clearly periodic section, and the uncertain periodicity section. Those parts of the signal where there is clearly no activity for large durations (more than 5 frames), were labeled as unvoiced or silence. These frames were used for the estimation of the voiced/unvoiced errors (the different error criteria are described later in this chapter). Those parts of the signals that are clearly periodic, and have a duration of more than five frames, were labeled as periodic. The first and last two frames were removed from each such section to remove the effects of onset and offset activity. This was required because of the inherent assumptions of stationarity in the use of fixed frame-rate calculations, as the error introduced for these fast changing sections of the signals may give biased or inaccurate estimates for different systems. All the other sections (including the original negative valued frames), are labeled as uncertain. The total percentage of these uncertain frames varied from speaker to speaker, from 10% to 15%. No evaluation is made on these uncertain sections.

As mentioned above the original signals were sampled at the sampling rate of 20,000 Hz, with 16 bit resolution of each sample. We used this signal as it is for one set of the experiments, which were labeled as high resolution analysis. Another set was derived from this data, where the signal was down-sampled to 8000 Hz, 8 bit range. Also during the down-sampling process, the signal was band-pass filtered to the telephone signal bandwidth of 300 Hz to 3200 Hz, using a 2nd order band-pass filter. These telephone quality signals were labeled as low-resolution signals. It is interesting to note that for pitch frequencies below 300 Hz, the fundamental frequency components are not present in the low resolution signals.

Three different noise environments were considered for evaluation. These are the *white* noise, the *environmental* noise, and the *music* noise conditions. The noise data was obtained from the set of sound files provided at Cooke's website, [Cooke, 1993], [Cooke, 2002], and used to add noise at various signal to noise ratios. The noise signals were re-sampled according to the sampling rate of the original signal before adding. The *white* noise (labeled in the original database as n1) is a random sequence of samples with approximately unit variance and Gaussian distribution. The *environmental* noise, is noise as might be expected in a street or an office environment (labeled as n3 in the original database). The *music* noise is a piece of sampled tonal music (labeled as n4 in the original database). These different noise environments were chosen because they are known to occur in everyday experience. Moreover, they are quite different in temporal and spectral characteristics from each other and therefore present a variety of challenges to the evaluated systems. While the *white* noise is spectrally and temporally flat, the *environmental* noise is characterized by jumps in amplitude, and the *music* noise is clearly periodic in nature. The procedure for preparation of noisy signals is described later in this chapter. The spectra of the three types of noise described above is presented in figure 5.1.

The database used for simultaneous speech from two speakers was derived from the database prepared by Cooke [Cooke, 1993] (same database as that used to derive noise samples), consisting of ten different speakers, and three overlapping background speakers for each of the voices. Thus there were three sequences for each of the ten voices. The ten "foreground" voices were labeled from v0 to v9 in the original database, while the three "background" speech signals were labeled as n7 to n9. The "foreground" utterances have almost continuous voicing, while the "background" utterances do not. This database was used for the analysis of simultaneous speech

as it is quite often used in this context [Wu et al, 2002], and is suitable for the evaluation of multiple pitch tracking. However, unlike the Keele pitch database, there is no reference pitch track available. The reference pitch tracks were prepared using the PRAAT system [Boersma, 1993][Boersma, 2002] at 100 Hz rate, and visually checked for consistency. PRAAT was used as it was found to be the most accurate system in the evaluation of clean speech signals. These were prepared using the clean signals only, and used for all experiments as reference. The procedure for this is discussed later in this chapter. The original recordings for this database are available at 16,000 Hz sampling rate, with 16 bit resolution. These original recordings were used as such, and were labeled as the “high resolution” signals. As with the Keele Pitch Data, the signals were down-sampled to telephone quality speech at 8000 Hz sampling rate with 8 bit resolution and were labeled as “low resolution” signals.

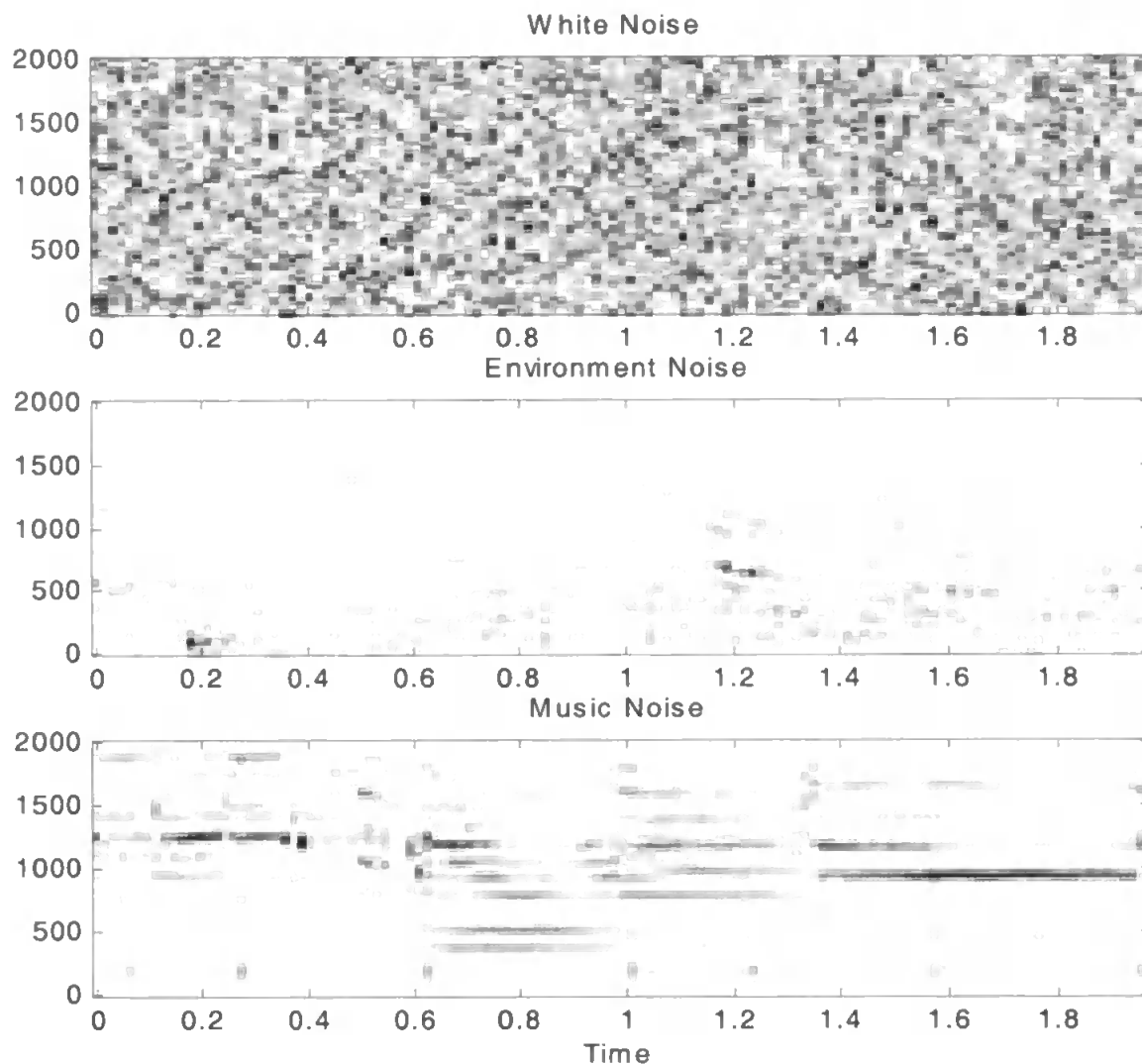


Figure 5.1. The spectrogram of the three different noise environments used in the analysis. The top plot is for the 'white noise'. The middle plot is for 'environmental noise', and the bottom plot is for 'music noise'. The spectrograms were computed using a 512 point FFT.

5.1.2. Description of the Software Used for Evaluation

The proposed system of Damped Harmonic Oscillators and Harmonic grouping for pitch estimation was developed using the Matlab[®] software. It provides a script based programming language which makes light work of implementing complex systems, although being an interpreted language, the programs developed run slower than possible under other programming languages like C++. The default parameters for the system were as described in

chapters 3 and 4. The frame rate for evaluation was 100 Hz, and the lowest and highest pitch allowed pitch estimates were restricted to 60 and 350 Hz respectively. The system is referred to in the tables, figures and charts as DHO.

The Correlogram system as described in section 2.6.3. was provided as part of the “Auditory Toolbox” for the Matlab system by Slaney [Slaney, 1998]. The function used to call the procedure which computes the pitch estimates was slightly modified from the original to give a simple voiced/unvoiced decision, as recommended in the documentation. All the default arguments are used. The system is referred to in the tables, figures and charts in shorthand as CORR.

The YIN system, as described in section 2.6.2. was provided by Cheveigne et al [de Cheveigné, Kawahara, 2002] on their web site [de Cheveigné, 2002]. This is also available as a Matlab package. The system was used with default parameters, and no changes made except for the change in output variable. The output variable is originally calculated in octaves, with the reference frequency of 440 Hz. The output was converted back into Hz for our evaluation. Also, we had to change the default frame rate for YIN, which was higher than the 100 Hz rate used for all other systems. Due to the comparison requirements, we had to change it to the default 100 Hz frame rate.

The PRAAT system, as described in section 2.6.1, was provided by Boersma [Boersma, 2002] as part of the PRAAT system. The system is available as an executable on various platforms (The Windows ® based system was used in this analysis). There are several pitch estimation algorithms available for this system, and we used the default algorithm. The system was used with the default parameters. A script file was prepared according to the instructions in the program. The main procedure call for the execution of the pitch tracking algorithm used was “To Pitch... 0.01 75 600” The parameter 0.01 sets the frame rate, and last two parameters determine the default lower and upper ranges of permitted frequencies in the pitch estimates. All these values are default values, and were not changed during the evaluation.

The probabilistic multiple pitch estimation system by Wu et al [Wu et al, 2002] is referred to as PMPT. The system was described in detail in section 2.6.4. The system was kindly made available by the authors upon request. The system supports only a sampling rate of 16000 Hz, therefore, it

was possible to use this system for the high resolution signal pitch estimation only, and it could not be used for the telephone quality 8000 Hz sampling rate signals. The system was made available in C programming language code, and was used without modifications to any parameters. Although, some changes were needed for the code to be used with a script to automate the evaluation.

5.1.3. Description of the Hardware Used for Evaluation

The experiments were run on a laptop computer with an Athlon 1500+ ® processor from AMD, and a desktop computer with a Pentium IV ® processor from Intel. Both the computers were running the Microsoft Windows XP Professional ® operating system.

5.2. The Evaluation Procedure and Error Metrics

One of the challenges of any comparative study, including this one, is the choice of meaningful error criteria. This choice is made difficult, as different systems are designed for different applications, and implement only selected functionality, based on the target research goals or practical applications. However, one can have a general list of desirable characteristics which we enumerate below.

1. The algorithm should accurately estimate the pitch period, within a suitable error margin. The error margin may change depending on the application area.
2. The algorithm should make robust estimates of the pitch period. I.e., the performance should not be affected by noise as far as possible, or that the fall in performance with increasing noise should be gradual, and it should also be robust to the signal conditioning (changes in resolution or sampling rate of the signal should not affect the performance of the algorithm too much). Needless to say, the system should be robust to change in speakers.
3. The algorithm should be efficient in terms of usage of computing resources. The resources considered are the number of operations, and the amount of memory required.
4. If possible, the algorithm should clearly indicate if a particular section of the signal has a well defined pitch or not (voiced / unvoiced decision).

Based on these desirable characteristics, and the recommendations made by other researchers, for example in [Rabiner et al, 1976], various error criteria were selected and used, as described in the next section.

5.2.1. The Criteria for Evaluation

In this section we define the various error criteria we used in our evaluation, and guidelines about how to interpret the results.

In order to calculate any error metric, we first need to consider the respective pitch frequency values of the reference signal and the estimated pitch value of the test signal by an algorithm. For example, if we assume that for a particular time, j , the reference pitch value, determined by the corresponding laryngograph signal is R_j . Further, the estimated pitch value by the pitch estimation system is P_j . Then the error metrics are defined on the corresponding values of R_j and P_j . Several different conditions arise when we compare these two values. These are:

- a.) $R_j > 0, P_j = 0$. This is the case where a voiced signal has not been assigned any pitch by the estimation system.
- b.) $R_j = 0, P_j > 0$. This is the case where the unvoiced or silence signal has been assigned a spurious pitch by the pitch estimation system.
- c.) $R_j = 0, P_j = 0$. This is the case where the unvoiced or silence signal has been recognized by the pitch estimation system, and assigned no periodicity.
- d.) $R_j > 0, P_j > 0$ And $|P_j - R_j| / R_j < \theta$. This is the case where the reference and estimated pitch values are close, as defined by the parameter θ .
- e.) $R_j > 0, P_j > 0$ And $|P_j - R_j| / R_j \geq \theta$. This is the case where the reference and estimated pitch values are quite different, as defined by the parameter θ .

Based on the possible conditions that may arise as described above, the opportunity to measure different characteristics is exploited by the error metrics defined below.

Gross Estimation Error Percentage:

This error metric is based on condition (c.) above. For this case, the pitch estimation algorithm has made a dramatic error in its estimation of the pitch. The parameter θ is the tolerance factor, which determines the allowed deviation from the reference pitch, exceeding which the estimate is labeled as a Gross Estimation Error (GEE). For a given pitch track, the total number of GEE counts are made for a given θ . The percentage of this count with respect to the length of the voiced pitch track determines the GEE percentage. Reasonable values of θ vary from 20% to 5%. The reason one would want to vary the value of θ is to test the suitability of an algorithm for an application, as different applications usually have different accuracy requirements. The calculations for the GEE percentages were carried out for all tests for the values of θ varying from 20, 10, and 5. The GEE percentage metric provides a measure of the “broad” or “rough” errors an algorithm makes in estimates of the pitch values over the whole track(s). The GEE percentage figures can be used as indicators of performance that is independent of the small local errors in calculations. It is clear that GEE percentage metric should be the most robust to noisy environments because (a.) it does not take into account the small pitch shifts that may occur due to noise, and (b.) it is a rough estimate of the performance of the system, for voiced section only, and does not involve a contribution of voiced/unvoiced errors (described later in this section). Thus, while analyzing noise robustness, the GEE percentage errors are most indicative of the underlying ability of an algorithm to handle noise in the voiced signal to produce accurate estimates of pitch period.

Fine Estimation Error Average:

This error metric is principally based on the converse of the gross estimation error described, and corresponds to the case of condition (d) as defined above. It provides a qualitative view of the deviations of the estimated pitch track from the reference pitch track. The Fine Estimate Error (FEE) average is mathematically described in equation 5.1.

$$FEE = \frac{1}{N} \sum_{j=1}^N (P_j - R_j) \quad \dots 5.1.$$

where N is the total number of samples for which the estimate P_j does not contribute to the GEE errors, i.e., for which the pitch estimates are not grossly inaccurate, based on parameter θ . Since the definition of FEE is based on θ as in the case of GEE, the calculations for the FEE

averages were carried out for all experiments for the values of θ varying from 20, 10, and 5. The FEE is a qualitative metric and indicates the bias inherent in the algorithms for pitch estimation. It tells us if the algorithm errs towards lower values, higher values, or equally towards higher and lower (making the FEE average closer to zero), when compared to the reference pitch estimate. One important use of this metric is for the evaluation of pitch algorithms used for speech compression for transmission over small bandwidth communication lines. In these cases, it is desirable, that the FEE averages for females be positive rather than negative, and for male voices, FEE averages may be lower rather than higher, in order to maintain naturalness and intelligibility.

Standard Deviation of Fine Error Estimates:

The standard deviation (STD) of the fine error estimates provides a related measure, which provides the metric for the amount of variation of the computed pitch estimate from the reference pitch estimate. Mathematically, it's calculation is dependent on the calculation of the FEE, and defined by equation 5.2. below.

$$STD = \sqrt{\frac{1}{N} \sum_{j=1}^N (P_j - R_j)^2 - FEE} \quad \dots 5.2.$$

Where the terms have their meaning as described for equation 5.1.

It is clear from the definition of equation 5.2. that the STD metric is a measure of accuracy of the pitch estimation algorithm in estimating pitch frequency during the portion of the signal where the estimate is not grossly incorrect.

Too-High / Too-Low Error Percentages:

This is a classification of those parts of the estimated pitch track that contribute to the GEE, as described above. This is often a useful estimate, as it provides an indication of the source of GEE percentages. Too-High Errors (THE) are defined as those gross errors (described above), which are higher (in frequency), compared to the reference pitch values. Too-Low Errors (TLE) are defined as those gross errors, which are lower in frequency compared to the reference pitch values. A higher percentage of THE indicates pitch frequency doubling, tripling, or similar phenomena. A higher percentage of TLE indicates pitch frequency halving or similar

phenomena. These are used in the analysis of the results as indicators to the possible source of GEE percentages.

Voiced-Unvoiced Errors:

Not all the algorithms that were evaluated implement a voiced/unvoiced decision system. These errors however were calculated for all systems, as they are indicative of the sources of the GEE percentages, as well as indicators of possible use of the inherent properties of the algorithm to make this decision. The voiced/unvoiced decision is quite important and has proven to be quite difficult in challenging noise conditions. Based on the conditions (a.) and (b.) as described above, the errors were split into two metrics, the Voiced-to-UnVoiced Error percentage (V_UVE), and the UnVoiced-to-Voiced Error percentage (UV_VE). The V_UVE percentage provides a measure of the inability of an algorithm to detect periodicity and voicing in the signal. The UV_VE percentage metric provides a measure of inability of an algorithm to distinguish unvoiced or noise sections of the signal from the voiced sections.

5.2.2. The Procedure for Evaluation of Performance for Single Pitch Tracks

This section describes the preparation of the reference pitch tracks, the noisy signal files, and the computation of the results for single pitch tracks estimation based experiments.

Preparation of the Reference Pitch Tracks:

The reference pitch tracks were computed from the estimate files provided with the Keele Pitch data. These files were originally prepared using autocorrelation method for periodicity estimation on the laryngograph signal, at a frame rate of 100 Hz (100 estimates are available per second). However, these reference pitch tracks were also partially corrected manually. We further refined these reference pitch tracks by eliminating any negative values (which were assigned for ambiguous cases in the original data), and selecting only those parts that were un-ambiguously voiced. Further refinement addressed the onset and offset effects, by not considering the first two and the last two frames. This was done to reduce the effects of fast changes in pitch during these phases. An example of the result is shown in figure 5.2, where the original pitch track and the reference pitch tracks are shown.

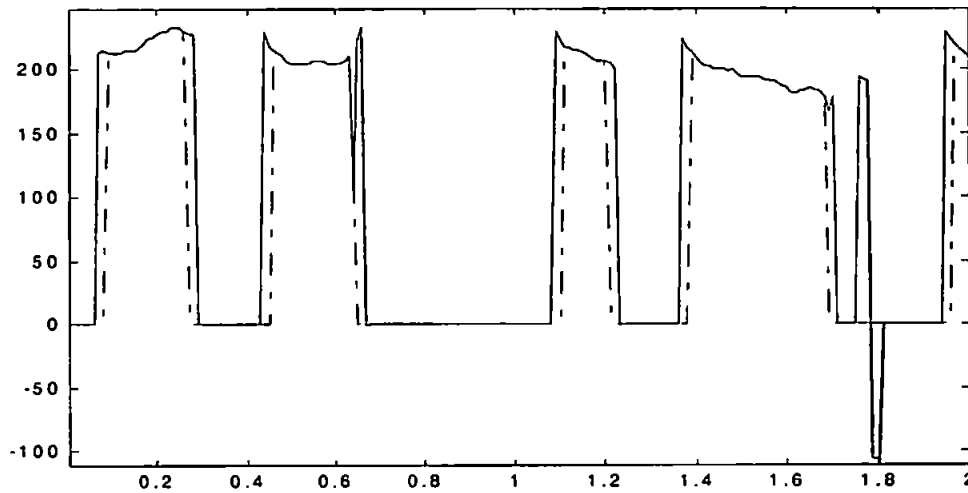


Figure 5.2. An example of the pre-processed reference signal. The original reference pitch track is shown in the solid line. The refined pitch estimate, after the onset and offset considerations, is shown as the broken line.

Preparation of the Noisy Signals:

The noisy environment conditions were simulated by adding different types of noises to the original speech signals, at different signal to noise ratios (SNRs). We recognize that this procedure does not take the Lombard effect into account [Pick et al, 1989], however, by adopting this procedure, we had a much more controlled noise environment. To produce realistic SNRs, the signal and noise power was computed locally before scaling and addition, thus the SNRs are controlled both locally and globally. The different noise types that were used are described in section 5.1.1. of this chapter. The equation for the calculation is given in chapter 3 (equation 3.22). The different SNRs used were 25 dB and 20 dB for “low” level noise conditions, 15 dB, 10 dB, and 5 dB for “medium to high” level noise conditions, and 0 dB and -5 dB for “very high” level noise conditions.

Computation of the Different Error Metrics:

Once the reference and test signals were available, control tests were performed first on the clean speech for all the different pitch estimation systems. This was followed by the experiment being repeated and different metrics computed for all the noise types (as described in section 5.1.1) and

noise levels. The resulting pitch tracks were stored in separate files on a computer for later reference and analysis, and the results were output to results tables. The different error metrics computed are described in section 5.2.1.

5.2.3. The Procedure for Evaluation of Performance for Two Pitch Tracks

This section describes the preparation of the reference pitch tracks, the noisy signal files, and the computation of the results for two pitch tracks estimation based experiments.

Preparation of the Reference Pitch Tracks:

The reference pitch tracks for the data that was used in the multiple pitch tracking experiments were not available as there was no accompanying laryngograph recordings. To obtain a reference system for this data, the PRAAT pitch tracking system (see section 2.6.1) was used. This system is quite accurate for clean speech pitch estimation (this will be justified when we present the results for the single pitch track experiments). As in the case of single pitch track reference data, the onset and offset sections of these tracks was excluded from any error analysis. The reference pitch tracks were calculated separately for the “foreground” and the “background” speech files.

Preparation of the “Noisy” or Multiple Speaker Data:

Of the systems that were evaluated, other than the proposed DHO and Harmonic Grouping based systems, most of them (with the exception of PMPT system) claim to track pitch for simultaneous speech, i.e. concurrent speech from two or more different speakers with different pitches. Therefore, a meaningful analysis of their performance can only be made by considering the “background” speech as “noise”, and testing their robustness to such noise. The test signals were therefore constructed with this in mind. The “clean” signal comprised of the foreground speakers’ speech only. The “noisy” signals were prepared by adding the “background” speakers’ speech at the SNRs that would allow significant presence of harmonic components of both the “foreground” and “background” signals. Thus, the “noisy” signals were created by adding each of the “background” speech files (described in section 5.1.1) to the “foreground” speech files at the SNRs of 5 dB, 0 dB and -5 dB. The addition process was similar to the one used for single pitch experiments. The PMPT system could only be evaluated for high resolution signals, because of it has a fixed sampling frequency, which is hard coded into the system.

Computation of the different Error Metrics:

As explained above, the “background” speech was considered as noise for the systems tested for the calculation of the error metrics. Additionally, for the DHO and Harmonic Grouping based system, as well as the PMPT system, the secondary or “background” pitch track was also evaluated, with additional error metric calculations with respect to the reference pitch tracks of the “background” speech. Thus the error analysis for this set of experiments was similar to the single pitch track experiments, apart from the analysis of the performance of the proposed system in estimation of the secondary pitch tracks.

5.3. The Summary of Results for Single Pitch Tracks Experiments

Detailed results for the single pitch track experiments, using the Keele data can be found in Appendix 2. Here, a summary of the results is presented. This summary was prepared in order to present concise results in the dissertation, and the details are included due to references to the details in the discussions and analysis sections. The summarized results were prepared by taking the average performance under various signal conditions for the whole database.

In this section, the results are presented in tabular format. For reasons of brevity and formatting we have used abbreviations in these tables, which are enumerated below.

The error measures are abbreviated as follows.

1. GEE_x – Gross estimation error percentage , where x is the value of the ‘percentage deviation allowed’ or constant θ .
2. FEE_x – Fine estimation error percentage, where x is the value of the ‘percentage deviation allowed’ or constant θ .
3. V_UVE – the voiced to unvoiced error estimate.
4. UV_VE – the unvoiced to voiced error estimate.
5. THE – The Too-High Error estimate.
6. TLE – The Too-Low Error estimate.

7. STD20 – The standard deviation error, computed using the FEE20 values.

The different pitch estimation systems have been enumerated as follows.

1. DHO – Damped Harmonic Oscillators and Harmonic grouping based system (proposed system).
2. CORR – The Correlogram computation based system (refer to section 2.6.3).
3. YIN – The very recent system proposed by de Cheveigne, based on the difference function calculation and parabolic interpolation.
4. PRAAT – The pitch estimation system based on autocorrelation, available in the PRAAT package [Boersma, 2002].

The different signal conditions are “Clean”, “Low”, “Medium to High”, and “Very High”, as discussed in section 5.2.2. The different noise “types” are as discussed in section 5.1.1.

5.3.1. High Resolution Speech Signals

The results presented below are for the original “high resolution” signals in the Keele database, i.e., sampling rate is 20000 Hz, and 16 bits are used to represent each signal value.

	GEE20	GEE10	GEE05	FEE20	FEE10	FEE05	V_UVE	UV_VE	THE	TLE	STD20
DHO	4.3	6.09	13.35	0.2	0.23	0.32	1.25	18.16	0.6	1.25	4.54
CORR	4.96	6.12	14.46	0.01	0.05	0.12	3.91	30.92	0.8	4.12	5.1
YIN	7.81	10.19	14.95	1.16	0.45	0.37	0	97.53	0.29	6.39	5.71
PRAAT	3.47	5.05	14.71	0.26	0.25	0.29	2.44	12.96	0.32	3.11	5.32

Table 5.1. Summary of results for “clean” high resolution speech signals.

System	Type	GEE20	GEE10	GEE05	FEE20	FEE10	FEE05	V_UVE	UV_VE	THE	TLE	STD20
DHO	White	4.9	6.64	13.75	0.15	0.2	0.32	1.22	19.27	0.52	1.44	4.62
	Env	3.5	5.32	12.41	0.22	0.23	0.31	1.28	18.16	0.44	1.3	4.58
	Music	3.73	5.51	12.85	0.21	0.23	0.3	0.89	19.75	0.3	0.89	4.57
CORR	White	6.31	7.32	15.34	0.01	0.01	0.11	5.47	19.15	0.64	5.65	5.04
	Env	5.8	7	15.29	0.01	0.05	0.15	4.48	28.6	0.87	4.89	5.12
	Music	5.43	6.86	15.58	-0.08	0.01	0.07	4.2	52.73	0.94	4.45	5.37
YIN	White	8.83	11.34	16.25	1.18	0.43	0.34	0	98.39	0.29	7.28	5.79
	Env	9.56	12.01	16.79	1.14	0.41	0.36	0	98.72	0.38	7.77	5.83
	Music	9.72	12.54	17.87	1.18	0.36	0.34	0	98.75	0.47	7.67	6.11
PRAAT	White	4.05	5.5	14.9	0.25	0.2	0.27	3.09	10.95	0.29	3.71	5.23
	Env	3.91	5.63	15.32	0.27	0.26	0.3	2.69	13.08	0.34	3.55	5.35
	Music	3.18	4.95	15.32	0.17	0.2	0.25	2.29	33.95	0.35	2.76	5.53

Table 5.2. Summary of results for “low noise” (25 dB and 20 dB SNR) conditions for high resolution signals.

System	Type	GEE20	GEE10	GEE05	FEE20	FEE10	FEE05	V_UVE	UV_VE	THE	TLE	STD20
DHO	White	4.35	6.23	13.4	0.08	0.18	0.3	1.11	27.11	0.32	1.14	4.86
	Env	6.56	9.55	16.89	0.16	0.26	0.33	0.97	47.62	0.42	1.59	5.47
	Music	4.34	6.16	13.78	0.13	0.19	0.28	1.46	30.5	0.66	1.6	4.7
CORR	White	27.83	28.43	34.57	-0.35	-0.36	-0.16	27.47	11.07	0.28	27.55	4.9
	Env	15.8	16.94	24.52	-0.01	-0.04	0.04	12.21	25.26	1.4	14.28	5.28
	Music	15.01	17.51	29.71	-0.29	-0.24	-0.19	7.98	67.76	3.7	9.17	6.68
YIN	White	18.14	22.88	29.97	2.32	0.71	0.24	0	99.04	0.66	14.08	7.88
	Env	19.73	24.54	32.02	2.08	0.55	0.21	0	99.46	0.98	14.99	8.05
	Music	25.57	31.62	40.55	1.93	0.36	0.21	0	99.17	4.06	16.1	9.1
PRAAT	White	24.35	25.09	31.91	-0.18	-0.19	0	23.93	5.07	0.06	24.27	4.95
	Env	15.01	16.39	25.42	0.2	0.12	0.17	9.83	10.89	0.18	14.81	5.48
	Music	11.15	13.93	27.72	0.08	0.03	-0.01	4.5	55.22	1.18	8.21	6.74

Table 5.3. Summary of results for “medium to high” (15 dB, 10 dB and 5 dB SNR) noise conditions for high resolution signals.

System	Type	GEE20	GEE10	GEE05	FEE20	FEE10	FEE05	V_UVE	UV_VE	THE	TLE	STD20
DHO	White	11.53	17.11	25.76	-0.06	0.18	0.34	0.3	89.53	0.52	1.3	7.31
	Env	18.18	28.14	38.65	0.29	0.37	0.43	0.23	96.01	0.98	0.87	9.87
	Music	7.82	12.46	21.89	0.05	0.03	0.21	1.69	78.48	1.2	1.85	6.75
CORR	White	87.94	87.99	89.21	-1.3	-1.34	-1	87.85	1.37	0.07	87.87	2.71
	Env	55.83	57.09	62.52	-0.19	-0.69	-0.45	47.96	22.9	2.26	52.29	6.83
	Music	50.57	58.4	69.74	-0.09	-0.46	-0.36	19.71	80.11	13.24	24.44	10.95
YIN	White	43.23	52.77	62.59	4.44	1.08	0.25	0	99.74	2.57	31.06	11.97
	Env	45.02	56.2	66.98	3.61	0.75	0.25	0	99.89	4.47	28.14	12.66
	Music	59.42	70.89	80.86	2.82	0.36	0.35	0	99.63	20.01	24.15	15.04
PRAAT	White	88.15	88.26	89.35	-0.56	-0.57	-0.36	87.49	0	0	88.15	2.36
	Env	57.01	57.95	63.26	-0.3	-0.68	-0.45	49.47	7.79	0.33	56.24	6.34
	Music	43.38	51.42	65.07	0.84	0.01	-0.14	16.26	71.24	4.81	26.33	10.38

Table 5.4. Summary of results for “very high” (0 dB and -5 dB SNR) noise conditions for high resolution signals.

5.3.2. Low Resolution Speech Signals

	GEE20	GEE10	GEE05	FEE20	FEE10	FEE05	V_UVE	UV_VE	THE	TLE	STD20
DHO	6.39	9.43	19.17	-0.27	0.13	0.22	2.91	15.95	0.43	2.91	5.94
CORR	6.59	7.64	16.07	0.06	0.1	0.19	4.39	0.4	1.92	4.65	5.17
YIN	50.67	53.88	59.3	3.32	0.36	-0.19	0	99.9	0.45	47.86	10.55
PRAAT	3.95	5.37	15.08	0.27	0.26	0.29	2.77	0.12	0.46	3.42	5.24

Table 5.5. Summary of results for “clean” low resolution speech signals.

System	Type	GEE20	GEE10	GEE05	FEE20	FEE10	FEE05	V_UVE	UV_VE	THE	TLE	STD20
DHO	White	7.03	10.25	19.99	-0.16	0.14	0.24	3.73	17.16	0.59	3.73	6.08
	Env	7.79	11	20.67	-0.28	0.12	0.24	3.21	20.5	0.47	3.21	6.02
	Music	7.06	10.16	19.83	-0.24	0.13	0.24	2.72	16.51	0.19	2.76	5.91
CORR	White	8.57	9.52	17.79	0.08	0.12	0.17	6.63	0.39	1.62	6.91	5.13
	Env	8.77	9.9	17.9	0.06	0.1	0.19	6.29	1.99	2.02	6.72	5.19
	Music	7.77	8.94	17.91	0.02	0.07	0.14	5.34	9.11	1.95	5.72	5.37
YIN	White	51.12	54.66	59.68	3.25	0.34	-0.04	0	99.98	0.47	48.18	10.62
	Env	51.47	55	60.47	3.17	0.37	-0.05	0	100	0.55	47.85	10.57
	Music	51.1	54.25	59.88	2.9	0.31	-0.03	0	99.97	0.46	48.11	10.35
PRAAT	White	5.19	6.27	16.02	0.26	0.25	0.27	4.11	0.12	0.38	4.78	5.15
	Env	5.63	6.87	16.51	0.27	0.26	0.3	3.99	0.37	0.59	5.03	5.25
	Music	4.29	5.75	16.28	0.19	0.23	0.26	3.03	6.55	0.35	3.8	5.47

Table 5.6. Summary of results for “low” (25 dB and 20 dB SNR) noise conditions for low resolution speech signals.

System	Type	GEE20	GEE10	GEE05	FEE20	FEE10	FEE05	V_UVE	UV_VE	THE	TLE	STD20
DHO	White	8.68	12.15	22.34	-0.11	0.22	0.27	2.67	33.71	0.59	2.91	6.31
	Env	12.09	17.06	27.2	-0.17	0.15	0.27	2.42	56.44	0.47	3.28	7.15
	Music	7.14	10.74	20.59	-0.19	0.19	0.29	2.91	23.77	0.48	2.98	6.13
CORR	White	32.05	32.6	38.22	-0.35	-0.35	-0.12	30.88	0.21	1	31.04	4.95
	Env	28.19	29.4	36.18	-0.02	-0.14	0.01	22.71	8.02	2.99	24.77	5.58
	Music	20.53	22.87	34.56	-0.28	-0.2	-0.09	10.04	30.97	6.71	11.35	6.78
YIN	White	48.23	53.71	61.15	2.77	0.2	0.12	0	99.99	1.36	42.23	11.14
	Env	50.87	57.8	65.84	3.32	0.12	-0.14	0	99.99	2.79	40.95	12.08
	Music	49.98	54.98	62.36	2.87	-0.04	0.34	0	99.97	2.03	44.11	11.36
PRAAT	White	27.6	28.21	34.94	-0.15	-0.18	0	27.12	0.04	0.14	27.47	4.95
	Env	27.29	28.36	36.36	0.15	-0.03	0	20.72	3.53	0.34	26.8	5.71
	Music	14.7	17.51	30.8	0.1	0.05	0.02	5.93	29.83	1.69	10.6	6.87

Table 5.7. Summary of results for “medium to high” (15 dB, 10 dB and 5 dB SNR) noise conditions for low resolution speech signals.

System	Type	GEE20	GEE10	GEE05	FEE20	FEE10	FEE05	V_UVE	UV_VE	THE	TLE	STD20
DHO	White	20.89	30.33	42.15	-0.73	0.27	0.37	0.29	93.53	0.85	0.95	10.02
	Env	29.25	44.86	57.34	0.66	0.25	0.43	0.14	99.73	1.62	2.04	12.34
	Music	12.45	19.07	30.16	0.25	0.16	0.19	4.1	60.74	1.62	4.1	7.93
CORR	White	88.51	88.57	89.51	-1.25	-1.28	-0.75	88.28	0	0.13	88.38	3.25
	Env	75.13	77.43	80.99	0.01	-0.95	-0.58	63.19	15.59	5.43	66.13	8.87
	Music	56.47	64.1	73.99	-0.58	-0.5	-0.23	22.74	60.69	17.16	27.23	11.17
YIN	White	47.72	58.85	70.5	3.19	0.46	0.22	0	99.99	6.46	27.9	13.26
	Env	60.44	72.58	82.1	3.66	1.34	0.23	0	100	11.28	30.31	15.08
	Music	47.16	57.85	69.32	1.45	0.28	0.01	0	100	11.89	24.57	12.86
PRAAT	White	89.27	89.34	90.15	-0.59	-0.64	-0.44	88.44	0	0	89.27	2.25
	Env	80.15	81.38	84.06	1.26	-0.61	-0.95	72.76	7.26	1.24	77.69	7.91
	Music	50.54	58.32	70.09	0.96	0	0.05	18.48	57.28	7.38	30.16	10.73

Table 5.8. Summary of results for “very high” (0 dB and -5 dB SNR) noise conditions for low resolution speech signals.

5.4. Summary of Results for Two Pitch Tracks Evaluation

The following results are for the speech signals with simultaneous speech from two speakers. The proposed Damped Harmonic Oscillators and Harmonic Grouping based system has been evaluated for its ability to track both the pitch estimates. Therefore, in the following results, the DHO_F is abbreviated for the foreground analysis, and DHO_B for the background analysis. All other systems are evaluated as usual. Similarly, for the output of the PMPT system, the foreground analysis is abbreviated as PMPT_F and the background track PMPT_B.

5.4.1. High Resolution Signal Analysis

The signals were sampled at 16000 Hz, at 16 bit resolution per sample. In this analysis, the PMPT system output is also considered (marked by PMPT_F for foreground pitch track based statistics, and PMPT_B for the background pitch track based statistics).

	GEE20	GEE10	GEE05	FEE20	FEE10	FEE05	V_UVE	UV_VE	THE	TLE	STD20
DHO_F	4.52	9.45	28.62	-0.77	-0.33	-0.52	2.84	10.42	0.06	2.84	5.71
DHO_B	8.70	12.70	16.46	0.12	0.12	0.04	7.70	0.00	0.00	6.70	2.17
PMPT_F	1.15	3.51	15.12	-0.53	-0.24	-0.26	0.51	53.24	0.54	0.51	4.20
PMPT_B	76.47	79.29	84.11	-2.39	0.66	1.54	71.26	21.41	0.00	71.82	10.32
CORR	19.76	23.04	29.34	-1.63	-0.98	-0.42	4.00	55.46	15.36	4.00	4.22
YIN	10.34	18.03	29.58	-0.62	-0.44	-0.23	0.00	80.00	2.43	1.73	6.26
PRAAT	5.33	13.95	23.13	-1.05	-0.63	-0.10	1.42	52.78	2.46	1.59	5.45

Table 5.9. Summary of results for 5 dB SNR for high resolution speech signals.

	GEE20	GEE10	GEE05	FEE20	FEE10	FEE05	V_UVE	UV_VE	THE	TLE	STD20
DHO_F	3.81	8.61	28.05	-0.78	-0.34	-0.51	2.23	8.96	0.02	2.23	5.70
DHO_B	9.84	12.49	16.30	-1.36	-0.43	-0.47	7.49	0.00	0.23	6.49	1.58
PMPT_F	1.68	3.97	15.48	-0.55	-0.25	-0.27	0.36	51.13	1.17	0.36	4.18
PMPT_B	48.15	48.39	51.81	0.30	0.38	0.23	47.47	0.36	0.00	48.15	4.63
CORR	7.55	9.69	12.80	0.01	-0.22	-0.25	5.17	39.65	0.13	5.81	2.85
YIN	5.21	9.83	18.41	0.65	0.18	0.03	0.00	80.00	0.00	2.49	4.93
PRAAT	4.45	7.77	11.43	0.57	0.16	0.08	1.12	43.01	0.00	2.18	3.13

Table 5.10. Summary of results for 0 dB SNR for high resolution speech signals.

	GEE20	GEE10	GEE05	FEE20	FEE10	FEE05	V_UVE	UV_VE	THE	TLE	STD20
DHO_F	4.70	9.59	29.00	-0.75	-0.32	-0.54	3.08	8.75	0.02	3.08	5.72
DHO_B	7.11	11.05	13.43	2.34	2.39	-0.44	6.05	0.00	0.00	5.05	2.84
PMPT_F	2.43	5.06	16.71	-0.49	-0.27	-0.29	0.45	53.34	1.29	0.94	4.24
PMPT_B	65.24	68.90	75.52	-2.03	0.81	1.09	54.40	33.41	0.00	55.83	10.90
CORR	14.54	19.77	27.48	-0.45	-0.40	-0.33	3.68	65.42	9.38	3.68	4.82
YIN	8.20	17.57	31.85	-0.44	-0.21	-0.22	0.00	77.78	1.09	2.15	6.38
PRAAT	4.72	14.08	24.65	-0.26	0.00	-0.01	1.54	65.21	0.97	1.62	5.69

Table 5.11. Summary of results for -5 dB SNR for high resolution speech signals.

5.4.2. Low Resolution Signal Analysis

The following results were obtained the same as before, but with the signals sampled at telephone quality speech, i.e. sampling rate of 8000 Hz, with a per sample resolution of 8 bits.

The PMPT system was not evaluated for this system due to restrictions on the sampling rate in design of the system.

System	GEE20	GEE10	GEE05	FEE20	FEE10	FEE05	V_UVE	UV_VE	THE	TLE	STD20
DHO_F	3.8	10.24	28.75	-0.91	-0.35	-0.61	2.38	8.83	0	2.38	6
DHO_B	8.09	12.09	15.86	1.27	1.27	1.18	35	0	0	7.91	1.7
CORR	23.57	25.14	28.68	-1.08	-0.74	-0.32	2.06	50.45	20.84	2.06	3.19
YIN	20.33	26.86	36.54	-0.12	-0.21	-0.26	0	80	2.34	11.89	6.36
PRAAT	10.1	14.81	21.65	-0.9	-0.56	-0.08	1.18	49.37	6.43	1.77	4.4

Table 5.11. Summary of results for 5 dB SNR for low resolution speech signals.

System	GEE20	GEE10	GEE05	FEE20	FEE10	FEE05	V_UVE	UV_VE	THE	TLE	STD20
DHO_F	3.75	9.84	28.43	-0.89	-0.41	-0.61	2.43	19.17	0	2.43	5.92
DHO_B	23.5	26.21	30.21	-0.04	-0.13	-0.13	76.34	0.11	0	20.34	0.92
CORR	10.03	11.83	14.01	0.08	-0.08	-0.16	6.38	37.71	0.95	7.3	2.55
YIN	13.77	19.09	28.29	0.78	0.31	-0.09	0	80	0.27	9.83	5.7
PRAAT	5.46	8.57	11.35	0.54	0.16	0.08	1.41	37.71	0	2.9	2.89

Table 5.12. Summary of results for 0 dB SNR for low resolution speech signals.

System	GEE20	GEE10	GEE05	FEE20	FEE10	FEE05	V_UVE	UV_VE	THE	TLE	STD20
DHO_F	3.54	9.55	28.3	-0.89	-0.41	-0.62	2.15	7.88	0	2.15	5.9
DHO_B	6.31	10.28	13.05	3.53	2.71	0.16	2.42	0	0	4.6	2.53
CORR	24.38	26.11	30.17	-0.13	-0.36	-0.23	3.71	62.58	19.67	3.71	3.32
YIN	22.53	30.5	42.46	0.5	-0.06	-0.15	0	80	5.66	8.69	7.03
PRAAT	11.68	16.69	23.98	0.16	0	0.02	2.9	62.17	6.79	3	4.51

Table 5.13. Summary of results for -5 dB SNR for low resolution speech signals.

In this chapter of the dissertation, the proposed system was evaluated and compared with some benchmark systems, and the experimental procedure and setup explained. In the next section, we analyse the results in detail, and make interpretations about the relative merits of each of the systems.

Chapter 6

DISCUSSION

The Interpretation and Analysis of Results

Different psychoacoustic and psychophysical experiments and their interpretations have spawned a large number of pitch estimation systems. In this dissertation, some of the most successful and recent pitch estimation systems were evaluated. An extensive study and comment on the performance of these systems in various realistic and challenging noise conditions is presented in this chapter. The performance of the proposed system under these conditions is also discussed, and compared with the other systems that were evaluated.

As discussed previously, the performance of most pitch estimation systems has not been evaluated extensively for noisy signals. This is quite surprising, because surely robustness should be one of the prime considerations for the development of any practical and useful system, as well as a critical test of biological plausibility*.

In this chapter, we shall discuss the performance of the various systems in the experiments described in the previous chapter. The performance for single pitch experiments will be discussed followed by the multiple pitch experiments. This will be followed by a discussion on the general conclusions that may be drawn from this study.

In the following discussions, reference to various systems and error parameters is made in their abbreviated form for the sake of brevity and flow. For a complete reference to their respective definitions and description, the reader is referred to the previous chapter (chapter 5).

6.1. The Single Pitch Estimation Performance for High Resolution Signals

In the experiments concerning the pitch estimation performance for high resolution signals, the signals were sampled at 20,000 Hz, with 16 bit resolution. This relatively high specification signal means more signal resolution for the lower frequencies, as well as a higher bandwidth (10,000

* The evolutionary requirements for the auditory system must argue for robustness to a large range of different types and levels of noises present in the environment.

Hz), leading to the possibility of much more accurate analysis and estimates for systems which use the full bandwidth of the signal for their computations. This is because the number of samples representing the same pitch frequency is larger for larger sampling periods, than for smaller sampling periods. A larger bit resolution on the other hand, means that the resulting digital representation of the signal has smaller quantisation errors. Therefore, under these signal conditions, one would expect more accurate pitch estimates. This was indeed the case, and the error statistics were generally better for these signals, although they contain the same speech information. In view of this fact however, more emphasis should be laid on the relative performance degradation in noise.

Let us start with the analysis of results for clean speech first, which forms a kind of control experiment. The 'clean' speech as discussed here actually contains some noise, but is mostly confined to very small values, mostly related to studio-related noise conditions. Figure 6.1 presents a graphical view of the experimental results for clean speech for these experiments.

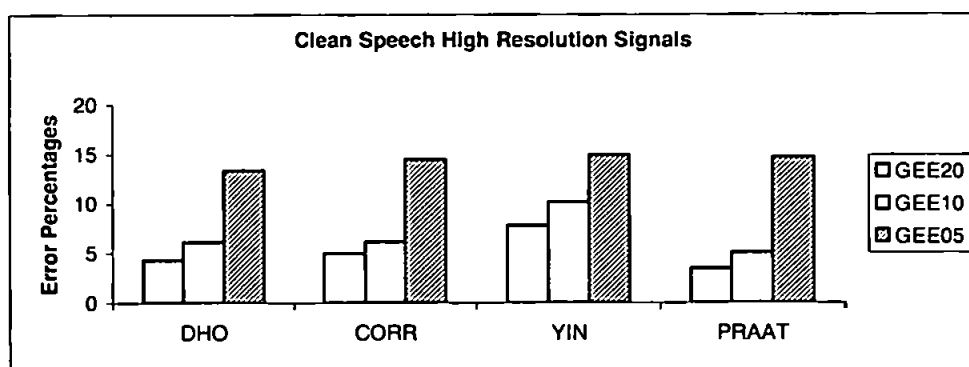


Figure 6.1. Average Gross Error Rates for clean speech for high resolution signals.

It is clear from the figure 6.1 above that the PRAAT system performs best when the error margins are large (GEE20, GEE10), but the DHO system has minimum errors for small error margins (GEE05). That is, although the PRAAT system makes fewer gross errors on the whole, the estimates are not very accurate when compared to the reference pitch tracks, while the DHO system makes a little more gross errors, a large percentage of the correct estimates are quite accurate (within the 5% margin of the reference pitch estimate). The CORR system and the YIN system are also quite accurate.

At this point, reference should be made to the source of the errors. For the DHO, PRAAT and CORR systems, the male speech based estimates contributed to much higher errors as compared to the female speech based estimates. Full results are presented in the appendix to this dissertation, to which the reader may refer for the speaker based breakdown of these experiments. A male-female error bar graph for clean speech is presented in figure 6.2. below to illustrate this.

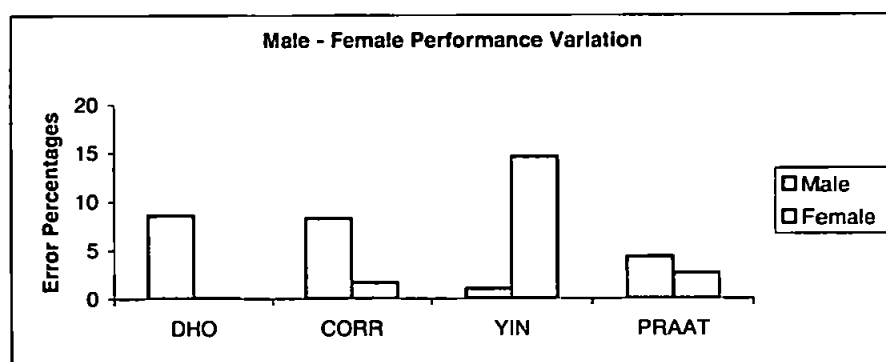


Figure 6.2. The GEE20 error contributions in clean speech conditions for various systems. The DHO female group errors are very small (0.04 %) and therefore not clearly visible here.

The PRAAT system appears to be the most balanced system. However, it should be noted that the majority of the problems for the male speaker group was caused by one speaker in particular, labeled in the database and our experiments as ‘m4’. On further analysis, it was observed that this particular speaker has reference pitch estimates which fall below 70 Hz for a large period. This seems to cause some problems to all the systems, except the YIN system, which estimates the pitch frequency quite accurately for this particular speaker. However, the YIN system consistently performs worse on most female speakers. Thus, although this system has best performance for males, overall, it has the worst average performance of all the systems that are compared here. On the other hand, the DHO system performs best on the female speaker group.

The THE error metric (Too -High Errors) is generally smaller for all systems than the TLE error metric (Too-Low Errors). That is, all the systems are more prone to “pitch halving” rather than “pitch-doubling”. The STD20 metric, that measures the standard deviation of the correct

estimates within 20% of the reference estimate, is smallest for the DHO system and largest for the YIN system, although the differences are small (DHO STD20 = 4.54, YIN STD20 = 5.71).

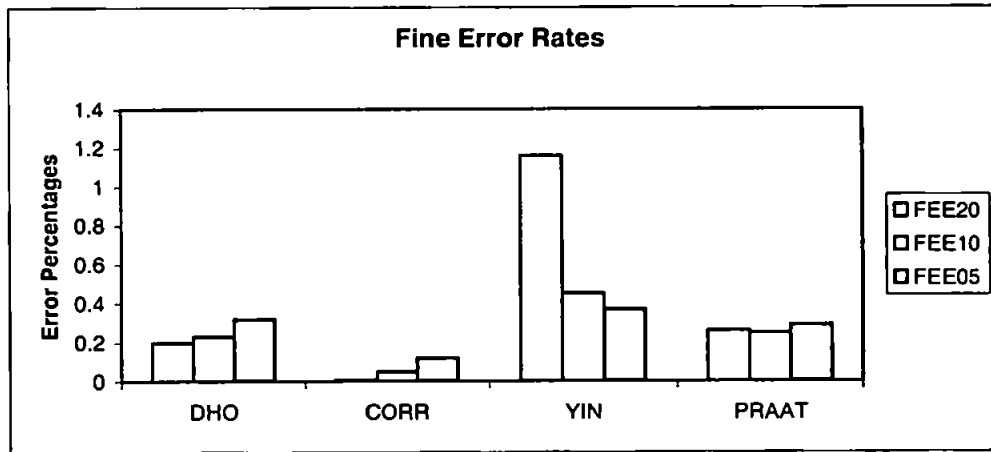


Figure 6.3. Fine error rates for clean speech signals. The FEE20 for CORR system is 0.01 and is not clearly visible in the graph above.

The Fine Estimation errors (FEE20, FEE10 and FEE05), are shown in figure 6.3. It is clear from the graph that for no noise conditions, the CORR system gives the most accurate estimates of the pitch. The case of the YIN system is quite interesting for this metric. The fine errors get smaller with smaller margins. This is not an error. It only suggests that for this system, when the estimate is correct within a smaller margin of error, i.e. GEE10 or GEE05, the estimates are more likely to be much closer to the reference values, as compared to the estimates where the system gives correct pitch values with the range of 20% of the reference estimates.

6.1.1. Analysis of Performance Degradation in Noise For High Resolution Signals

Lets us now turn our attention to the analysis of performance for noisy speech signals. In these experiments, the performance of the various systems was evaluated for increasing levels of various types of noises.

Figure 6.4 shows the GEE20 error metric performance for the various levels of white noise. From the figure, it is clear that that performance is more or less equal and quite low for all systems up to 20 dB of noise. However, for larger noise levels, the estimation errors keep increasing dramatically for the majority of the systems, apart from the DHO system, where the performance is very robust to even very high levels of noise. Even for moderate levels of noise

(10 dB for example), other systems have unacceptable error rates for most applications. The performance for GEE10 and GEE05 show similar trends.

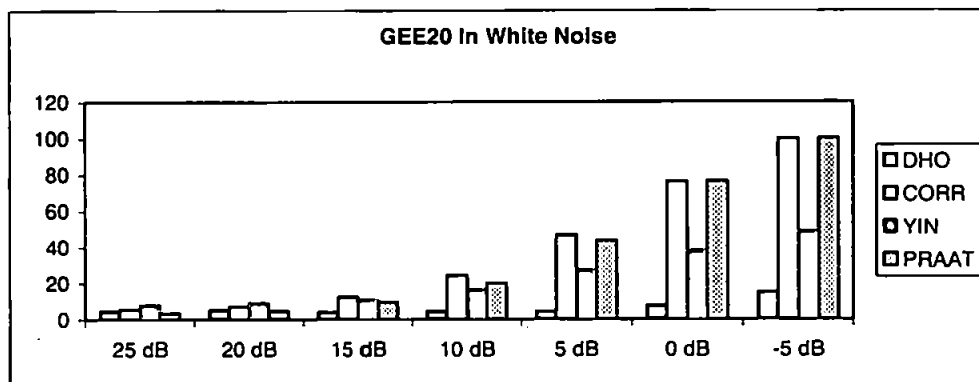


Figure 6.4. Gross error rates (GEE20), for different levels of white noise for high resolution signals.

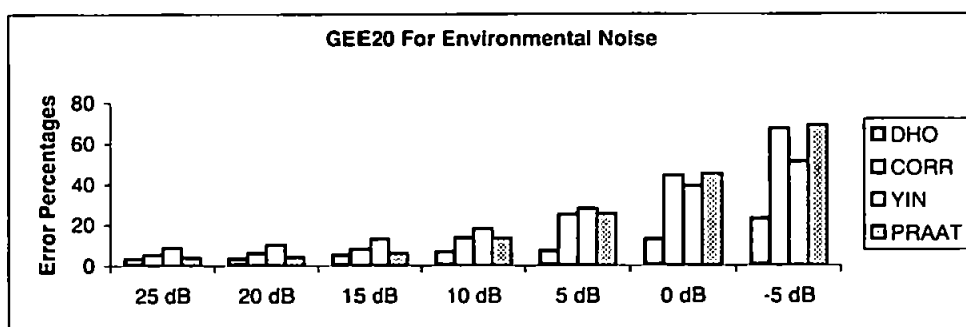


Figure 6.5. Gross error rates (GEE20), for different levels of environmental noise for high resolution signals.

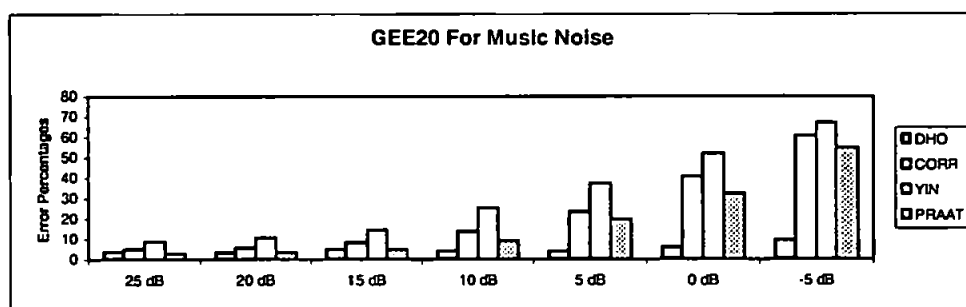


Figure 6.6. Gross error rates (GEE20), for different levels of music noise for high resolution signals.

For “Environmental” noise and “Music” noise, the GEE20 metric shows similar trends with some differences, as is apparent in figures 6.5. and 6.6. Overall, all the systems perform better in Music noise. One reason for this is the high frequency nature of the Music clip that was used for these experiments. Thus, the pitch range spectral properties should not change much. It is therefore surprising that systems other than DHO seem to deteriorate with increasing levels of Music noise. Analysis of the pitch track and V_UVE error show that the errors in case of the PRAAT and the CORR are partly due to non-detection of speech (i.e. labeling of voiced period as unvoiced), and pitch halving or pitch doubling errors.

Given such desperate performance in noise by the various systems, it is difficult to display the fine error estimates, as they can be misleading. This is because of the high error rates involved. When the gross errors are high, by definition, the fine errors are evaluated on a much smaller section of the data. Therefore, fine errors make sense only when compared against the gross error rates. For example, if the gross error rate GEE20 is 10%, the FEE20 can be expected to be high. However, if the GEE20 is 90%, the FEE20 is computed over the correct 10% of the values only, it is expected to decrease in absolute value, as the GEE20 increases.

The THE and TLE metrics show similar trends to the GEE20, as displayed in the figures above. That is apart from the DHO system, these metrics increase sharply with increasing noise levels. The voicing error measures (V_UVE and UV_VE) give an indication as to the reasons for high error rates in the case of CORR and PRAAT. For these systems, the majority of the gross errors at 0 and -5 dB levels are due to voiced portions of speech being classified as unvoiced, and therefore assigned a pitch of 0.

6.2. The Single Pitch Estimation Performance for Low Resolution Signals

The low resolution signals for these experiments were sampled at 8000 Hz, with 8 bit resolution. A lower sampling rate means lower bandwidth, and lower number of samples per pitch period. Small signal resolution of 8 bits means that there are likely to be larger quantisation errors. Thus the estimation of pitch frequency under these signal conditions can be expected to be more difficult than the high resolution signals. Addition of noise under these signal conditions should also affect the performance of pitch estimation systems more severely. The added problem is that of the *missing fundamental*. For the low resolution signals, the bandwidth was constrained to 200 to 3800 Hz, representing “telephone quality” speech. Therefore, this signal condition is quite

challenging, and systems that implicitly rely on the fundamental component being present in the signal are expected to fare badly.

The clean speech experiments, as in the case of high resolution signals, provide a kind of benchmark of performance for the various systems that were evaluated. Figure 6.7 provides the gross error estimation metrics for the various signals. Looking at the results, it is instantly apparent that the performance is worse than for the high resolution signals.

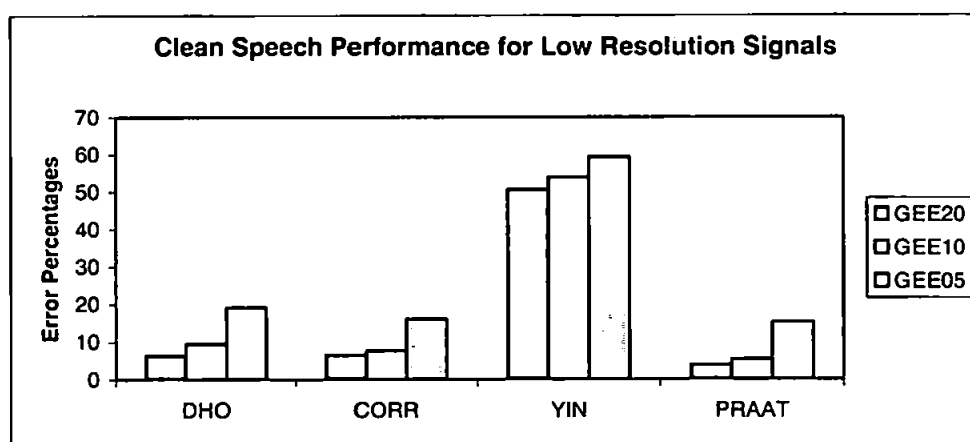


Figure 6.7. Average Gross Error Rates for Clean speech for low resolution signals.

The best performance on average for gross error metrics is shown by the PRAAT system for the clean speech signals. DHO and CORR performances are very similar, however, the YIN system shows signs of failure, with the average GEE20 metric crossing the 50% mark. Unlike the high resolution signals, the GEE05 is least for the PRAAT system as well. This may be explained by the fact that the system uses parabolic interpolation for estimation, and in the absence of enough signal resolution, this contributes to the accuracy of the estimated pitch tracks. The same reasoning may also explain the GEE20 and GEE10 results, as the PRAAT system provides the most accurate results for all these metrics.

The case of breakdown of the YIN system for these signal conditions even with no additive noise bears need for further analysis. It was observed that most of the errors were contributed by “pitch halving”, i.e., the system was found to be consistently finding lower estimates than the reference values, with estimates nearly half the reference values. This is also reflected in the TLE error metric. Figure 6.8 illustrates this with the help of an example.

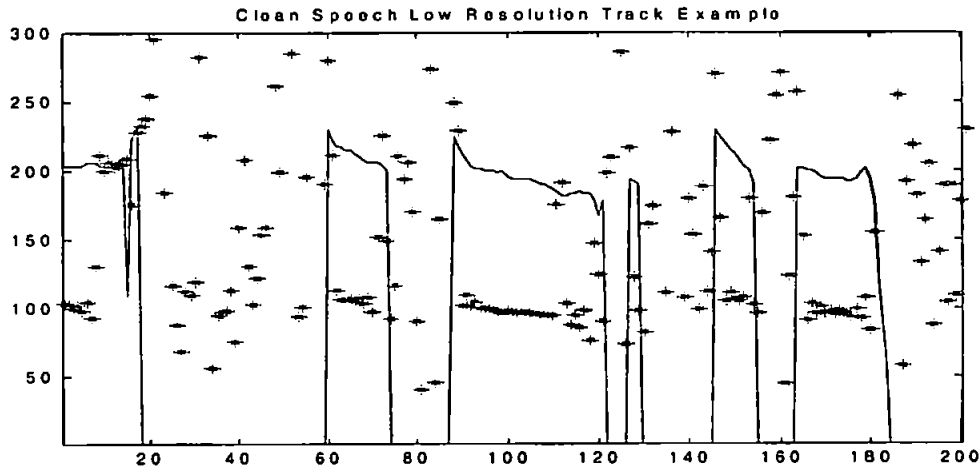


Figure 6.8. An illustration of the catastrophic failure of YIN system for low resolution clean speech signals. The YIN estimates are marked with a “*”, while the reference pitch tracks are shown as a continuous line. Since the YIN system does not employ a voicing detection system, those values for which the reference pitch values are zero should be ignored.

It is clear from the example in figure 6.8 that the system failure occurs in most instances due to the phenomenon of pitch halving. The YIN system uses a variant of the autocorrelation function, called the difference function (see section 2.6.2 for details). In the low resolution signal conditions with telephone quality speech, the lower frequency information is missing. This would lead a simple autocorrelation based system to give estimates higher than the reference estimates, i.e. to err on the side of pitch doubling. However, the YIN system, with its difference function, calculates the initial pitch estimates based on the equivalent minima of the autocorrelation function. These values are then normalised leading to the “cumulative mean normalized difference function”. During this process, the new function is obtained by dividing each value of the difference function with its average over short-lag values. These short-lag values would correspond to higher multiples of the pitch frequency for voiced speech, thus reducing the chances of “too-high” errors. However, in the absence of any low frequency energy in the signal, it is likely that this normalization produces spurious higher lag “valleys”, thus leading to the resulting pitch-halving for the signal conditions under discussion. This interpretation was further substantiated by a *missing fundamental* experiment on the system. When presented with a signal consisting of the 2nd, 3rd and the 4th harmonics of a 200 Hz tone (sampled at 8000 Hz), the system gives a pitch estimate of 100 Hz, thus exhibiting sub-harmonic errors.

However, if the fundamental is 100 Hz, the system produces correct pitch estimates (100 Hz), when presented with the 3rd, 4th and 5th harmonics (without the fundamental frequency component). Thus, although the normalization of the difference function used in the YIN system reduces the “too-high” errors, for telephone quality speech, this normalization seems to result in too many “too-low” errors. Figure 6.9 presents these “too-low” averaged errors for the various systems for clean, low resolution speech signals.

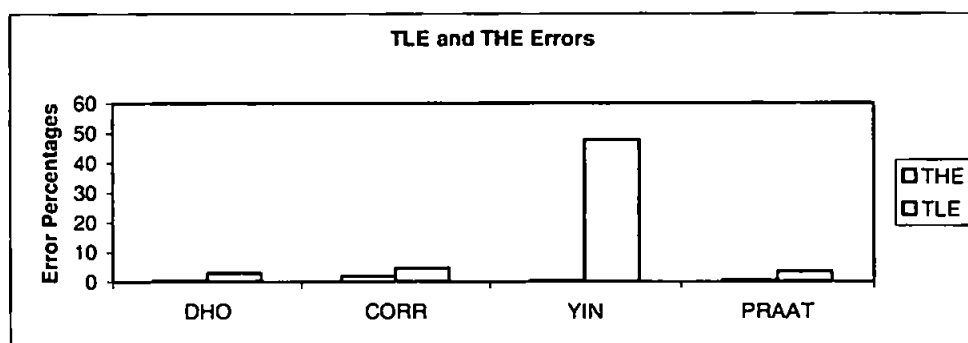


Figure 6.9. The TLE and THE error metrics for clean low resolution signals. Clear links between the TLE and the gross error rates can be seen (compared with figure 6.7). The THE errors are less than 2% for all systems.

Another interesting insight provided by the figure 6.9 is that although the PRAAT system is most accurate when gross error rates are compared, the DHO system has lower THE and TLE rates. This indicates that the DHO system is not making pitch-halving or pitch-doubling errors, but the majority of the contribution to errors comes from the V_UVE (voiced to unvoiced errors account for about 3% of the errors for DHO system), and from the general inaccuracy of the estimates.

The Fine errors are relative to the gross errors, and the PRAAT system has the lowest fine estimate errors as well. The standard deviation is the same for all systems, apart from the YIN system, which has double the standard deviation error (STD20), compared to the rest of the systems.

6.2.1. Analysis of Performance Degradation in Noise for Low Resolution Signals

Apart from the quantisation noise present in the original signal, experiments were carried out where various types of noise were added to the original signal at various SNRs. The noise types

and noise levels were same as those in the high resolution signal experiments, (with adjusted sampling rate and resolution). The preparation procedure for the noisy signals is described in the chapter 5. Here we shall discuss the results of these experiments.

The addition of different types of noise at different levels leads to progressive degradation in performance. The most indicative figures are those shown by the gross error rates. The figures are shown in figures 6.10, 6.11 and 6.12 for gross error rates GEE20.

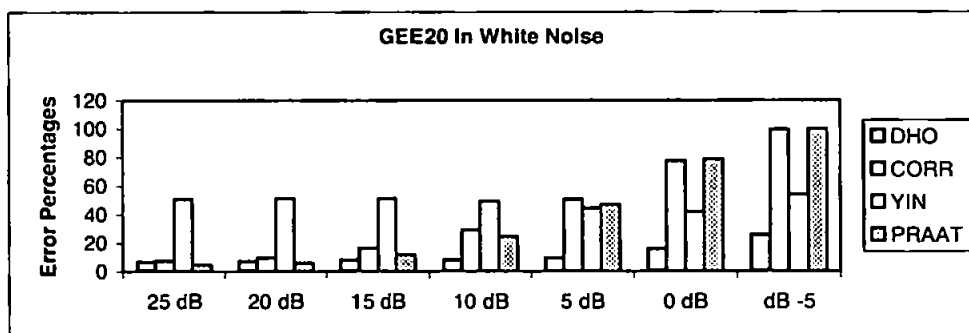


Figure 6.10. Gross error rates (GEE20), for different levels of White noise for low resolution signals.

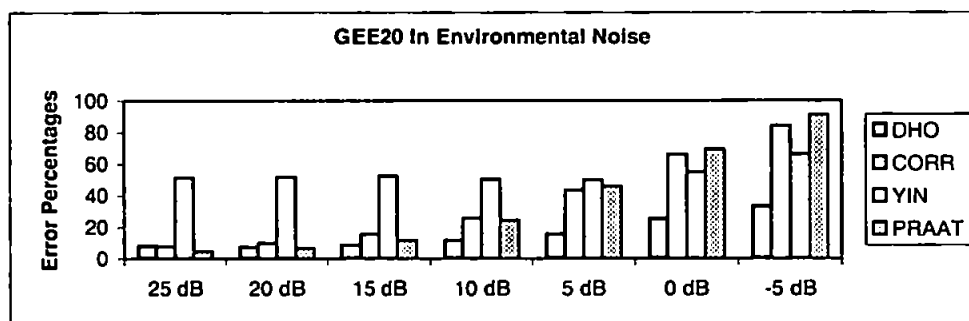


Figure 6.11. Gross error rates (GEE20), for different levels of Environmental noise for low resolution signals.

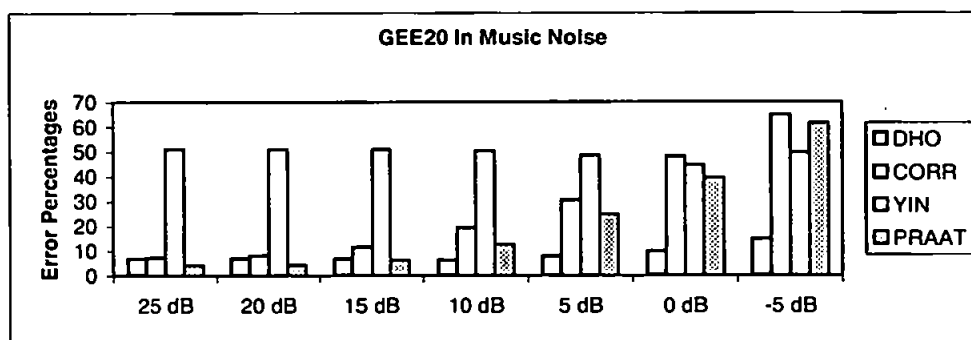


Figure 6.12. Gross error rates (GEE20), for different levels of Music noise for low resolution signals.

As can be seen from the figures above, the gradual reduction in performance according to the gross error measure, with the increasing level of noise, is present for all the systems, for all types of noises. The system that is most robust to high levels of noise is the DHO system. The YIN system shows a breakdown in pitch estimation for all types and levels of noise. The CORR and PRAAT system show good performance in low to medium levels of noise, but the systems break down under high to very high levels of noise.

The YIN system seems to improve a little in low levels of noise. This is unexpected, but explainable. As discussed in the previous section, the reason for the breakdown of the YIN system is due to its estimation equations and normalization. The system breaks down due to spurious “valleys” in the absence of low frequency information from the pitch frequencies. However, in the presence of noise, these effects may sometimes cancel out, leading to slightly increased probability of finding the correct pitch frequency.

GEE10 and GEE05 error metrics show similar results. However, it is very interesting to note that even in low signal to noise ratios i.e. very high levels of noise, the DHO system shows robustness to these error measures as well. For GEE10, for example, the 0 dB error measure is below 20%, while for all the other systems, it is more than 50%. This relatively high level of accuracy is achieved without any explicit interpolation mechanism. The method of lower order harmonic grouping makes it robust to high frequency noise on the one hand, and copes well with the missing fundamental on the other. The power density characteristics of speech make the lower frequency region quite robust to broadband noise in general, and this is exploited by the

DHO system in the grouping of harmonics of lower order, as these frequencies are not all simultaneously affected by noise.

The reasons for the breakdown of the PRAAT and CORR systems mainly lie in the decreasing pitch saliency with increasing levels of noise. The majority of voiced frames under high levels of noise are classified as unvoiced, leading to a dramatic breakdown in performance under high noise conditions. This is also reflected in the TLE measure, which is nearly equal to the gross error measures, indicating that these systems fail to identify any periodic information in the search range for these signals. This is also the major source of errors in high noise conditions for the DHO system, although the total percentage is comparatively much lower.

The DHO system performs very well in noisy conditions, for all types and levels of noise, even as other systems break down. Moreover, the degradation in performance as the noise levels increase is quite gradual. This ability of the system to handle noise was designed into the system from the ground-up, by the means of damped harmonic units, whose output produces effects similar to noise masking observed in the biological auditory systems of mammals. As discussed in chapter 3, the temporal processing of the output of DHO units provides a high frequency resolution, thus the system achieves good accuracy without the use of interpolation. The harmonic grouping system provides continuity constraints that prevent pitch halving and pitch doubling, and continuous estimation, rather than frame based analysis adds further robustness to the system.

6.3. The Multiple Pitch Estimation Performance for High Resolution Signals

The special characteristics of the DHO system allow it to estimate more than one pitch period simultaneously. The other systems that were studied did not have this capability. However, the performance of these systems is of interest if one considers the second voice as noise. Of the two simultaneous voices, the one with a more continuous spectrum is considered as foreground, and the other, more intermittent one is considered as background. Due to these continuity constraints that exist for a meaningful analysis of pitch for the two simultaneous voices, the Keele dataset (see section 5.1.1), could not be used, as although mixtures could be formed for various voices, there would be no logical argument for treating one voice as foreground, and the

other background. Instead, as discussed previously, the dataset of 30 mixtures, prepared by Cooke [Cooke, 1993], was used. This dataset was labeled as CASA (see section 5.1.1), and the analysis of the results based on this data is presented in this section.

An added complexity that occurs in the estimation of pitch for these signals is that of concurrent harmonics. If the pitch frequency of one voice is an exact multiple of the other, then all but one harmonic would nearly be coincidental. This additional challenge would contribute towards higher errors for mixtures of such voices.

In this category of analysis, the performance of the probabilistic multiple pitch tracking system (PMPT) [Wu et al, 2002] is also evaluated.

The perceptual experiments dealing with recognition of simultaneous vowels in [Darwin, Hukin 2000] show that the performance in terms of correct recognition of both the vowels (with different fundamental frequencies) is a function of the difference between the fundamental frequencies of the vowels, as well as the duration of the stimulus. For stimulus with duration of 200 ms, the correct recognition of the pairs of vowels is about 85 – 90%. For smaller durations, the listener performance is much reduced (about 65-70% correct recognition for 100 ms long or less duration). However, it is hard to draw straight inferences in terms of expected best performance of any biologically inspired pitch estimation system from these experiments. This is especially true for the database that was used to test the performance, with long durations of voicing being generally present in the stimulus presented to the systems that were evaluated.

Let us first discuss the performance of the different systems in the determination of the foreground pitch track estimation. Figure 6.13 shows the GEE20 measurements for the various systems.

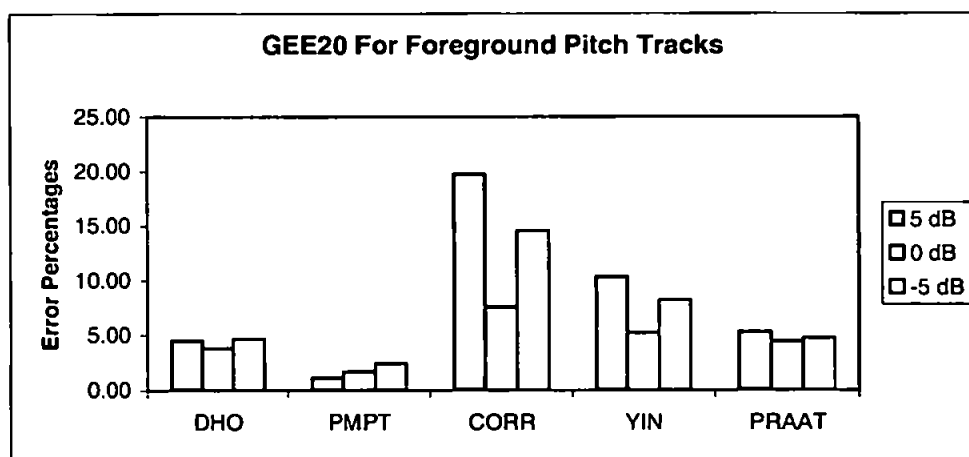


Figure 6.13. The GEE20 performance for the foreground pitch track estimation for different pitch estimation systems.

It can be seen in the figure 6.13. that the foreground pitch track estimation is most accurate for the PMPT system, followed by the DHO system, while that of the CORR system is worst. Interestingly, the 0 dB signal condition is most accurate individually for almost all the systems. When the SNR is 5 dB, the foreground speech signal is much higher in amplitude compared to the background speaker speech. At 5 dB SNR, the background signal power is much lower than at the 0 dB SNR. This makes the background signal more “noise-like” at 5 dB than at 0 dB, therefore the performance difference between 5 dB and 0 dB mixing levels. At -5 dB, the foreground signal power is much reduced as compared to the 5 dB and 0 dB conditions, thus the errors are expected to be higher as well. The PMPT system performs better under all signal conditions for the estimation of the foreground pitch track. This good performance could be explained by the fact that according to Wu et al [Wu et al, 2002], the system performance in terms of parameter estimation was tuned on the same dataset, while the rest of the systems were probably not.

Analysis of other error metrics reveals that most of the errors for the DHO foreground system are “Too-Low Errors” or TLE errors, i.e., the system failed to classify voiced sections and thus assigned a pitch of zero, combined with the pitch halving effects. For the CORR system, the majority of the errors are contributed by the THE error metrics, indicating pitch doubling errors. For the YIN and PRAAT systems, the error contributions are spread between TLE and THE errors.

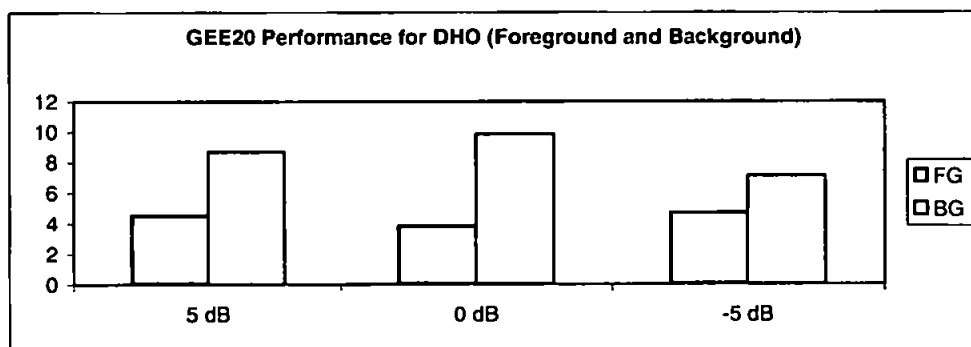


Figure 6.14. The foreground and background pitch track estimation errors (GEE20) for the DHO system (BG = background, FG = foreground).

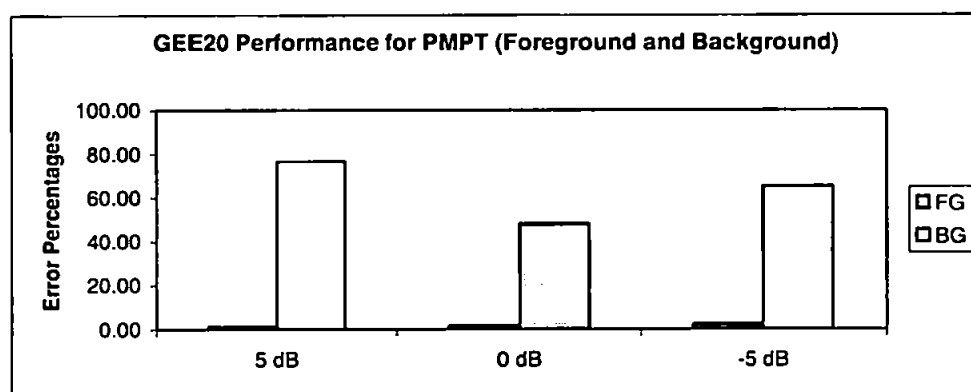


Figure 6.15. The foreground and background pitch track estimation errors (GEE20) for the PMPT system (BG = background, FG = foreground) The foreground GEE20 values are below 5%, and therefore not clearly visible in this plot. See figure 6.13. for a better resolution of these figures.

The DHO system finds the pitch tracks for the foreground and the background pitch tracks quite accurately. The errors are higher for the background pitch track as compared to the foreground pitch tracks because certain voice mixtures contain instances at which the background pitch track falls within a range which makes it nearly an integer multiple of the foreground pitch track. Under these conditions, the background pitch track is assigned a value of zero, thus increasing the background pitch track errors. It is quite interesting to note as well that the background pitch track errors are smaller than the foreground errors for the YIN and CORR systems for most conditions. As expected, the background pitch track errors are larger for the 5 and 0 dB conditions, as compared to the -5 dB condition.

The PMPT system performance for the background speaker are very poor, and does not improve as the signal to noise ratio improves in favour of background speakers. This performance is due to voiced to unvoiced errors, that is, the system fails to find the pitch of the background speaker in the majority of cases, leading to unreliable performance. However, comparing the TLE performance and the V_UVE performance points to the fact that this is not always the case, and sometimes, the estimated pitch track is probably making a sub-harmonic (pitch halving) error.

6.4. The Multiple Pitch Estimation Performance for Low Resolution Signals

For low resolution signal mixtures, it would be quite interesting to note how the performance is affected. It is to be expected for reasons similar to the single pitch track study, that the performance in general would be worse. However, there are additional factors which contribute under simultaneous speech mixtures. As compared to the high resolution signal with concurrent voiced with near integer multiple fundamentals, the low resolution signal conditions would exacerbate the accuracy problem, leading to worse estimates.

Another factor to be taken into account while discussing these results is the fact that for some voices, the fundamental may be missing. Therefore, in terms of resolvable harmonics, it would be quite difficult to assign different pitch values to tracks which are near integer multiples of each other.

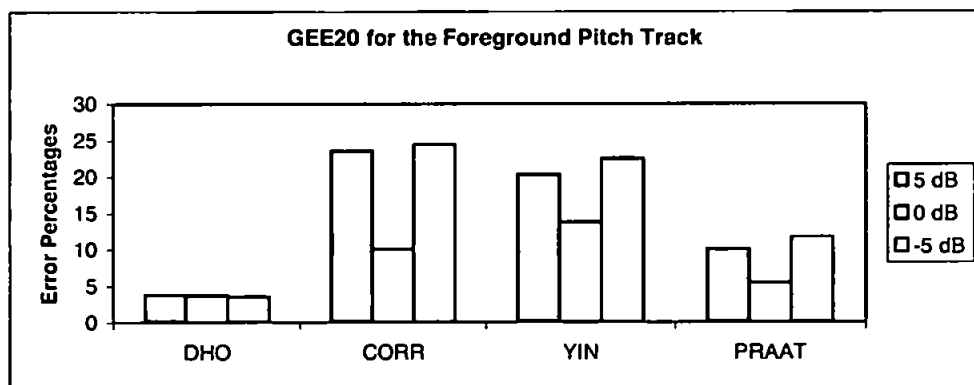


Figure 6.16. The foreground track average gross estimation error (GEE20) for the various systems at different SNRs.

The figure 6.16. shows the comparative gross error measure (GEE20) performance figures for the low resolution signal conditions with the mixtures prepared at different SNRs. It is clear that the DHO system is most accurate, while the YIN and CORR system are quite bad in terms of this error metric. The differences in error measures for the different levels are for the same reasons as discussed in the previous section (section 6.4). The fact that errors in the computation of the single pitch tracking systems is higher can be explained by the fact that when two competing pitch candidates are present, the algorithm responsible for assigning the “true” pitch frequency may sometimes err towards the background pitch track, thus the foreground pitch tracking errors are further increased. However, for a multiple pitch estimation capability system like DHO, the errors are comparatively low, as it considers both the foreground the background pitch estimates as valid. Continuity constraints, which keep a small history of previous estimates then make the task of assignment of the foreground and background tracks much easier. This is further illustrated by figure 6.17. below.

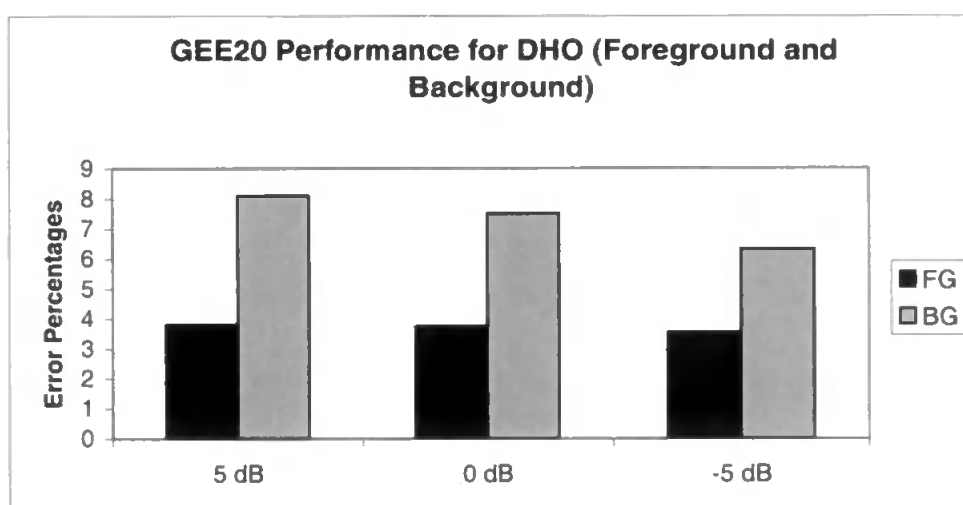


Figure 6.17. The foreground and background pitch track average error estimates, as computed using the GEE20 error metric.

The performance for the foreground pitch track remains nearly the same for all the three different power ratios in which the voices are mixed. The background pitch estimation errors decrease as the relative power of the background speech increases. Also, the overall performance

of the DHO system in this case is much better than the single pitch tracking systems. The PRAAT system is the best single pitch tracking system according to these figures, however the reader is pointed to the fact that the reference pitch tracks were not available in the form of laryngograph data and were computed using the PRAAT system, thus the performance of PRAAT should be expected to be better as in these experiments, the different error measures compared the performance of the system for clean signal, with the performance of the same system with the noisy signal.

6.5. Comparison of Computational Requirements

This section provides a summary of the computational requirements study that is presented in appendix 1.

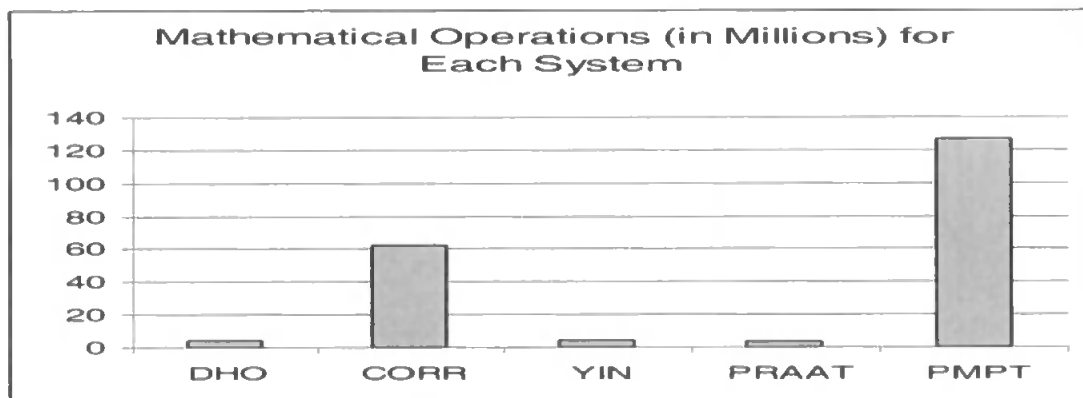


Figure 6.18. The number of mathematical operations requirement for the various systems analysed. The requirements are for processing one second of speech at 8000 Hz.

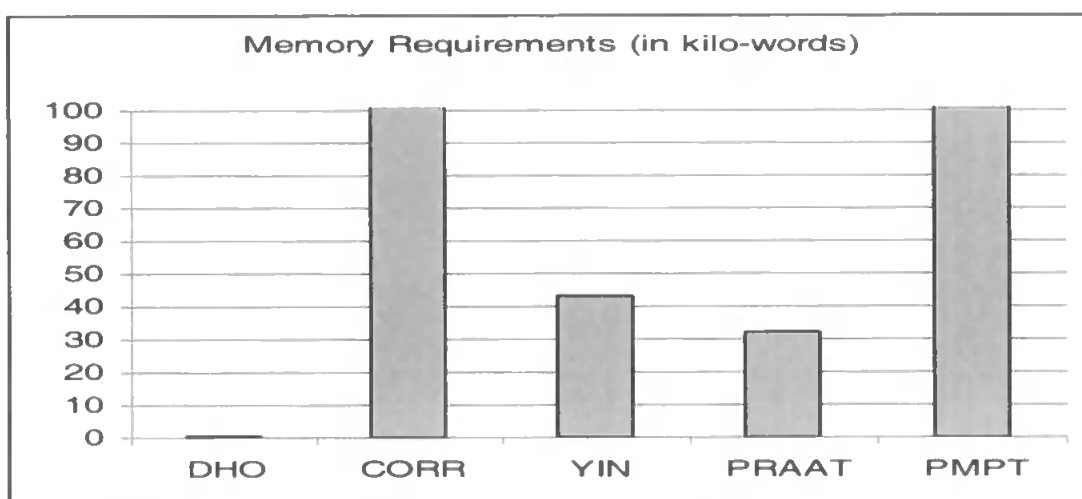


Figure 6.19. The memory requirements for processing one second of speech sampled at 8000 Hz, in kilo-words. The requirements for the CORR system is 1280 and for the PMPT system is 17856 words kilo-words (truncated for this graph).

The requirements for the PMPT system were calculated by halving the requirements, as the original processing is valid only for signals sampled at 16000 Hz. The requirements show that the PRAAT system is slightly better in terms of mathematical operations required than the DHO system. However, the memory requirements for the DHO system are much less (0.7 kilo-words) compared to all the other systems. The CORR and the PMPT systems have much higher computational and memory requirements compared to all the other systems. The DHO system can be implemented for 16 bit operation. However, it is not clear if other systems can be implemented for less than 32 bit resolution due to higher order of analysis and double precision requirements for their operation.

6.6. Comparison and Analysis of the Overall Pitch Estimation Performance

The experimental analysis presented above was aimed at demonstrating and analyzing the results of the experiments that were carried out, and detailed in chapter 5.

To compare the various systems, the different performance criteria and performance characteristics have been analysed. The PRAAT system is the most accurate under low noise conditions. The proposed DHO system is shown to be the most robust system under extreme and challenging noise conditions. The DHO system also emerged as the better system at tracking

the background pitch tracks, compared with the PMPT system. The DHO system also has far lower computational requirements compared to the other systems.

The PRAAT system is based on the window normalized autocorrelation and parabolic interpolation system, and was studied in detail in section 2.6.1. The performance of this system as calculated in the experiments described above is quite robust to signal conditions. The performance however, degrades under medium to high levels of noise. The performance of this system is most affected by white noise, and at and below 0 dB noise, the system fails to detect any periodicity and therefore one witnesses a catastrophic failure. This happens both with high resolution and low resolution signals. The performance is slightly worse than the DHO system for most noise types and noise conditions for low resolution signals, but is quite comparable in general for low-noise conditions. Unfortunately, the system does not support multiple pitch tracking, so full comparison with the DHO system could not be carried out. However, for the single pitch track study of simultaneous speech, the system shows a good degree of robustness. This could be partly attributed to the fact that the reference pitch tracks for the multiple pitch track analysis were prepared using clean signals as the input to this system. Also, as the nature of the background is intermittent, for those parts of the signal where the foreground signal is the only signal, the performance is guaranteed.

The CORR system is based on simulated peripheral filtering, followed by autocorrelation computation in each channel, followed by the summary autocorrelation. The peak of the summary autocorrelation function for all the channels gives the pitch period, and the model is detailed in section 2.6.3. In the experiments whose results were analysed in this chapter, the performance of the CORR system is for most parts in between the best system and the worst system. The system is very accurate for low noise and clean speech signals, and we get minimum fine error estimates for this system under these conditions. Like other systems, the CORR system also degrades in performance with signal quality. For most types of noises, the system makes a lot of gross errors for medium and high levels of all the different types of noises, although, the white noise affects its performance most. Comparatively, the performance is equal to the worst systems in high levels of noise, but worst for medium levels of noise. The CORR system does not implement a multiple pitch estimation algorithm. For simultaneous speech, the system was tested for its accuracy in tracking the foreground speech pitch frequency, and the

background speech was treated as noise. However, the performance for the foreground pitch estimation is worst for most cases for the CORR system. Thus we can conclude that the CORR system is quite sensitive to different noise conditions, although quite accurate for clean speech. Also, its performance is quite sensitive to background speech.

The YIN system, as studied in section 2.6.2., is based on a variation of autocorrelation, where the normalized difference function of the signal is computed. It also involves parabolic interpolation and local search for estimation of its pitch estimate. For clean speech, the system shows performance which is comparable to other systems, but with larger fine estimation errors. However, further analysis showed that the system was behaving differently to other systems, where it was making more errors for the female speakers, compared to other systems, which had a major part of their error measures contributed by the male speakers. For high resolution signals, the performance in noise for this system is quite accurate, but for medium and high levels of noise, the system shows large degradation in performance for all types of noises. Performance is most affected by the music noise. This may be due to the fact that the system was designed for both speech and musical pitch detection, and this causes confusion about the saliency of speech pitch, when music is present in the background. For low resolution signals, the system performance breaks down, and one sees large errors due to pitch halving or doubling in the output pitch track (see figure 6.8). As discussed before, this may be due to the normalization of the difference function used in the calculation of the pitch estimates. The performance does not degrade markedly further with increasing levels of noise in low resolution signals. This points to the fact that most errors in the high resolution cases may have been due to similar effects. The YIN system does not implement a multiple pitch detection system, so it cannot handle simultaneous speech and produce two different pitch candidates. However, as in the case of CORR and PRAAT systems, it was evaluated for the foreground pitch track accuracy only. Even for the foreground pitch estimates, the performance is much lower than the PRAAT and the DHO systems. Also, the system does not employ a speech detection system, so the evaluation of the voicing decision based errors cannot be carried out, and there is no way of telling if the errors contributed are due to the lack of pitch saliency.

The PMPT system (see section 2.6.4 for description) was only evaluated for the high resolution signals in the multiple pitch tracking case. The system has a fixed sampling rate requirements of

16000 Hz. For the input signal, and could not be tested for the low resolution, telephone quality signals. This was one of the reasons why it was not considered for the more extensive tests for the single pitch cases. Another factor is that the system is very slow and requires a very high performance computer for reasonable simulation times, and uses lot of memory (see appendix 1 for a computational requirements study). However, it was included in the analysis as it is quite recent [Wu et al, 2002], and was the only system available capable of multiple pitch tracking for speech sounds. The system performs well in the estimation of foreground pitch tracks for high resolution signals, in presence of simultaneous background speech from a different speaker. However, this good performance may be due to the fact that the system was originally designed using the same dataset (the CASA dataset, [Cooke, 2002]), in terms of estimation of its parameters. The authors mention in their presentation that half of the dataset was used for the estimation of the parameters. In spite of this advantage, the background pitch tracks are not reliable, and show a very large degree of errors, the system performance for this task is much worse than the proposed DHO system.

The DHO system was proposed in this dissertation in chapters 3 and 4. It is based on a bank of damped harmonic oscillators, whose characteristic frequencies vary from 80 Hz to 1000 Hz. The output of the damped harmonic oscillators is evaluated and integrated over time to produce a spectral representation of the input signal, and finally, the different harmonics are grouped together based on a common fundamental hypotheses. The dominant group, based on the continuity and saliency (determined by the group energy), is declared as the foreground group and the residual is treated as the background group, thus giving the foreground and background pitch estimates. The system was designed with noise robustness in mind from the beginning, and this is reflected in the performance figures discussed above. For clean speech, the system performs very well, giving error estimates that were comparable or better than other systems evaluated. For the noisy signals, especially medium and very high noise levels, for all the different types of noises the performance is most robust compared to all the other systems, and remains at acceptable levels under all noise and signal conditions. Noise robustness is highly desirable for the pitch estimates to be of any practical use in speech analysis and processing. Due to its multiple pitch tracking abilities, the DHO system performance compares most favorably when the noise is background speech. In this case, the second pitch track produced is the background pitch estimate. The performance of the system in estimating the background pitch track is better

than all other systems' performance in estimation of the foreground pitch track. For high resolution signals, the foreground estimate is comparable to the PRAAT system, and for the low resolution signals, the DHO system performance is much better than the PRAAT system. Another important factor favorable to this system is that the performance for more strict gross error measures follows similar noise robust trends, when compared to other systems, whose accuracy degrades markedly when the signal is noisy. The systems ability to estimate the background pitch tracks, compared with the PMPT system is also quite remarkable.

Chapter 7

CONCLUSIONS

This dissertation has presented a new and very robust pitch estimation and tracking system. Performance of the proposed system was extensively evaluated and compared with a variety of other systems, in different levels and types of noises and different signal conditions. The conclusions that can be drawn from this work, in practical and theoretical terms, are discussed in this chapter. The dissertation presents several new ideas and techniques both in terms of general auditory processing, and in terms of pitch perception. Therefore further implications and testing of these ideas and future direction of the research are also discussed.

7.1. Summary of Results

Earlier in this dissertation, a case was made for robust processing of speech in order to develop robust speech applications, and additionally to enhance our understanding of robustness that is observed in biological auditory systems. One of the most robust features of speech signals is its periodicity [Terhardt, 2002]. However, performance of various systems and models of pitch perception have not been evaluated systematically for noise robustness. Since the primary aim of the current research was to develop a biologically plausible and noise robust pitch estimation system, it was quite essential that this evaluation and comparison be carried out. In order to do this comparative analysis of performance in a meaningful way, publicly available data and standard metrics (as described in chapter 5) were used for the evaluation. A summary of this evaluation is presented in this section.

The detailed evaluation of the chosen pitch estimation systems revealed that performance of these systems abruptly breaks down in the presence of noise. It was also noted that some systems are able to handle certain noise conditions, and certain signal conditions better than others. For example, the performance of the PRAAT system [Boersma, 1993] reveals that it completely breaks down (100% errors) in speech signals with high white noise, but the performance was not so bad for other types of noises, and is very accurate for clean speech conditions. Similar trends were observed in the correlogram based system of pitch evaluation [Slaney, Lyon, 1990]. The performance of the YIN system [de Cheveigné, Kawahara, 2002]

severely degrades when presented with telephone quality speech signals. Further analysis of the performance of this system reveals that it produces sub-harmonic errors in the cases where the target pitch is high, and in these cases, is not able to handle very well signals from which the fundamental is missing (this occurs in most cases in telephone quality speech).

It was observed that the systems that are designed to estimate single pitch frequencies perform poorly when the signals are adulterated with interfering speech. This is an important finding because the most often encountered interference in practical systems is interfering speech from other talkers. The PMPT system [Wu et al, 2002] performs very well in estimation of the foreground speech pitch estimates in these cases. However, contrary to claims of ability to track multiple pitches from simultaneous speech from two talkers, the background pitch estimate for this system was unreliable and inaccurate.

The proposed DHO model of pitch estimation was found to be very reliable in the presence of challenging noise conditions for all types of noises, and demonstrates graceful degradation in performance, measured by slightly increasing error rates, for very high levels of noise. Although it was not found to be the most accurate system for clean speech, the major aim of noise robustness is met without doubt. For pitch tracking performance in the presence of interfering speech, the overall system performance was better than all the other systems evaluated, by a huge margin. The system was able to reliably estimate the pitch of both the foreground, as well as the background speaker.

In light of these results it may be concluded that the proposed system meets the design aim of noise robustness, with performance degradation with increasing noise levels that is gradual. The system performance had similar trends for all the different types of noises, and different signal conditions.

7.2. Contributions

This section briefly summarises the original contributions to the field of pitch estimation and auditory research made in this dissertation.

The Damped Harmonic Oscillator Based Frequency Analysis System

A new system for frequency analysis was presented. This system, based on the operation of very simple damped harmonic oscillators (DHO), implements a broad frequency analysis of the input signal and produces a tonotopically ordered output. It is based on the peripheral auditory system and provides a rough model of the mechanical selectivity of the basilar membrane. Some of the important properties of this model of signal analysis are enumerated below.

1. The damped harmonic oscillator is easy to visualise as a unit similar to a section of the basilar membrane, with corresponding damping and characteristic frequency. Computationally, it is much cheaper to simulate compared to either the traditionally used digital filters or the Fourier transform.
2. Unlike the most well established techniques, the frequency response to a signal by the proposed system is not fixed for a characteristic frequency. Each output can be analysed for frequency of response. This leads to several oscillators responding to the stimulus frequency, even when the characteristic design frequency is different from the stimulus frequency. This property provides the system with robustness to noise in the signal, with the individual units “locking on” to the periodic components in the neighbourhood of their characteristic frequency.
3. The DHO system is a dynamical system which does not have any delay lines and few design parameters. Compared to the Filter-Bank processing, which has to be designed with many more parameters and specifically designed delay lines, the system is capable of providing much more flexible operation, and a higher frequency resolution for further analysis.

Temporal Mode Processing and the Resulting Representation

The bank of damped harmonic oscillator units produces an output which has a tonotopic ordering. In this dissertation, the fine temporal processing of this output produces a representation which is far superior to channel based systems. Although such an encoding of sensory information has been suggested before, [Barlow, 1961], [Cariani, 1999], it has never been computationally modelled in this context. The following discussion contrasts this approach to the more traditional systems, and highlights the source of noise robustness in the proposed system.

The output of the DHO units is used for a fine temporal analysis of each of the outputs, based on the positive peaks and zero-crossings. This process can be likened to the temporal coding of the inter-spike intervals in the auditory system. The proposed temporal coding based analysis and representation differs from channel based representations in terms of the type of input required, and the qualitatively different roles that channels play. Inters-spike intervals are time intervals that describe temporal relations between pairs of jointly occurring spike events. Calculations similar to the calculation and detection of inter-spike intervals are done in our system, albeit in a manner more suitable to digital computers. Such time intervals constitute correlational information within them. In contrast, representations that are based on channels rely on probabilities or on rates of spike events over a time window to compute this information. Moreover in channel based systems, information about which particular channels are activated and by how much, are essential to the representation, and may not be the most robust way to code the information. In the DHO model, units are preferentially activated by stimulus components that are nearby in frequency, and these regions therefore contribute relatively more of their stimulus-related frequencies to the global representation (i.e. over the entire frequency range). Once the information about the intervals is combined, however, the representation does not rely on the particular channel identities of the DHO units to encode frequency because the intervals themselves bear this information, and in a much more precise and robust way. One could discard all information concerning characteristic frequency (or cochlear place) without affecting the representation. In contrast, in a channel based representation, such as a place frequency map, the identities of particular channels are absolutely critical for representational function. Consequently, stimulus representations are much less robust in these systems, and the

frequency resolution is a direct function of the number of channels. The basic informational constituents and the ways information is encoded in the proposed system, compared to the channel based representations are therefore very different. Moreover there is some evidence that in the biological auditory systems, representation of periodicity pitch appears to follow this pattern of fine temporal structure within more coarsely tuned frequency channels [Cariani, 1997].

The Harmonic Grouping Based Pitch Estimation System

The harmonic grouping based system for pitch frequency estimation that is proposed uses the representation described above. It takes advantage of higher frequency resolution afforded by the representation to find the pitch frequency that best explains the harmonically related groups of activation in the representation. The system is based on the Barlowian approach to perception for minimisation of representation [Barlow, 1959], and, unlike other systems of harmonic grouping, for example [Brown, Cooke, 1995], does not require any a-priori knowledge of the pitch period. The minimization of representation is achieved by establishing a common pitch frequency for the entire group (and all the frequencies in the group are thus explained by this common frequency). Moreover, since the grouping approach is used to separate sets of harmonically related frequencies, the system has an inherent ability to track multiple pitch frequencies present in simultaneous speech from speakers with different frequencies.

The Statistical Analysis of Performance

The conclusions from results of the statistical analysis of the proposed system's performance were presented earlier in this chapter. This analysis also included a variety of other systems which have been reported to be quite accurate.

The detailed statistical analysis performed is absolutely essential in order to clearly define the advantages for any system over other systems, and to identify sources of errors. However, the component of analysis in noisy signals has been missing in most systems, even when such an analysis is performed. A detailed analysis of this magnitude has not been published before. By presenting the analysis, it is hoped that the need for such an analysis will be highlighted, and that the performance figures as well as qualitative analysis presented in this dissertation be a reference to compare performance, as the analysis was done on publicly available data, with standard error

metrics. It is also intended to make the prepared data files and other materials available on the internet for this purpose (subject to permission from the original contributors).

7.3. Scope for Future Research Work and Extensions

The system based on damped harmonic oscillators exhibits some very interesting properties like noise masking and leads to natural grouping of activity around harmonic and formant frequencies. The harmonic grouping idea can be seen as an expression of minimisation of representation. One of the consequences of this minimisation of representation of auditory stimuli is the emergence of pitch, which was utilised for the work presented in this dissertation. However, it is quite a powerful concept which can be utilised further for the development of representations which lead to a better understanding of the auditory system.

In the current treatment of the proposed system, only the monaural case was considered. It will be quite interesting to see if the framework can be extended and used in the binaural case as well.

The current research was undertaken with the aim of producing a computationally efficient and noise robust pitch estimation system. However, the model of processing that is proposed in this dissertation can be extended as a more elaborate auditory processing model. This may be achieved by putting the research in the neural processing context. The aim of this research would be to extend the Barlowian approach to develop a sparse auditory information coding scheme in terms of temporal processing and inter-spike intervals based representation of the full frequency range using models of integrate and fire spiky neurons. This model of temporal processing would further analyse the phase relationships of different frequency components to discover other “binding” features like common onsets in order to group the components together. It would be very interesting to see if the emergent properties of such a system match the physiological data, and whether this model would provide more insights into biological auditory processing.

Other future work is to use the system in speech recognition applications for improved voice activity detection. Further research would also explore the potential of using the temporal analysis as the basic signal processing stage in speech recognition systems, and evaluate the effect on the performance in noise.

Appendix 1

IMPLEMENTATION DETAILS – COMPUTATIONAL COMPLEXITY AND MEMORY REQUIREMENT ANALYSIS

In this section, we analyse the computational costs of implementing and deploying the various systems considered in this dissertation for pitch period estimation. Due to the sparse availability of information on other systems, the analysis is limited to general calculation and analysis, and comparative figures. This should however, give a general idea of the merits of the various systems analysed in terms of practicality of their deployment in real-world applications and trade-offs in terms of performance versus speed of calculation and memory requirements.

Most of the systems that were considered for detailed analysis are based on autocorrelation computation. The multiple pitch estimation system by Wu [Wu et al, 2002], is based on the perceptual filter-bank, summary autocorrelation and hidden Markov Models. These systems will be discussed only in terms of theoretical requirements for implementation. The reason for this is that the systems under consideration are available as implementation on various platforms (Matlab, Windows executables, or a mixture of the two), making it very difficult to make direct comparisons.

A.1. The DHO and Harmonic Grouping Based System

The main computational task of this system is the simulation of the operation of the bank of Damped Harmonic Oscillators (DHOs). The harmonic grouping section of the code is operational at the rate 100 Hz, making it contribute a comparatively smaller amount to the total simulation time.

For each sample, and single damped harmonic oscillator, the total computational cost can be expressed in terms of number of arithmetic operations. All the operations can be done in the range of 16 bit integers, without the risk of overflow. The main functional operations for the whole process and their computational requirements are enumerated below.

1. The simulation of DHO operation takes 3 multiplications and 3 additions, provided the constants are evaluated and stored beforehand. Total cost for simulating 40 DHO units per sample is then 120 multiplications and 120 additions. Accumulated cost for this stage = **240 arithmetic operations**. Number of state variables and constants that need to be maintained = $40 \times 6 =$ **240 memory words**.

2. The next processing stage (results in an estimate of the spectral representation of the input signal) is conditionally called, when there is a local positive maxima, greater than a predefined constant. Given that the operation is performed, the total number of operations for the active DHO is 8 additions and 4 multiplications. The profiling information during a typical run was analysed, and this conditional logic was called on average 38% of the total time (i.e. for 38% of the samples). Therefore, the total contribution of this stage of processing is given by $\Rightarrow 0.38 \times 40 \times 12 =$ **182.4 arithmetic operations** per sample. The memory requirements for this stage, in terms of variables that are kept in memory during run-time = $8 \times 40 =$ **320 memory words**.

3. Harmonic Grouping and related operations are called 4% of the time on average according to the profiling analysis. The number of arithmetic operations per call to the harmonic grouping per channel is given by 20 additions and 5 multiplications. The preprocessing for the harmonic grouping (calculation of periods, handing of integration times, etc) are 5 additions and 6 multiplications. Therefore the total number of arithmetic operations per channel (per DHO) = 36 arithmetic operations, and total is = $0.04 \times 40 \times (5 + 10 + 20 + 12) =$ **57.6 arithmetic operations**. The memory requirements for the harmonic grouping (per call but not persistent) are dependent on the number of harmonics in the input data and the order of analysis (number of output streams or groups desired). The total on average of male and female voices was calculated to be **126 memory words**.

4. The overhead (total other computation time and memory requirements) also depend on the order of analysis (whether the input data consists of simultaneous voices or not) as well as the sampling rate. However, the profiling information gives us an indication that the total time for such processing is about 1% of the total simulation time. The memory requirements for this section are dependent on the number of samples in the input

string. Since this varies according to application and implementation, we shall not include this in our analysis further.

Considering the total of the various sections and stages of processing as described above, the total computation cost for this system (per sample) is given by adding all of these together. The result of this operation gives us the figure:

$$(240 + 182.4 + 57.6) * 101 / 100 = \mathbf{484.8 \text{ arithmetic operations per sample.}}$$

Therefore, at 8000 Hz sampling rate, the total arithmetic operations are equal to 3872000.60 per second, and at 16000 Hz sampling rate, the figure is 6831801.60 arithmetic operations per second. The total memory requirements at run-time are equal to: **686 memory words**.

A.2. The Autocorrelation based Systems

The computational requirements for the autocorrelation based systems vary depending on the method used for the estimation of the autocorrelation function. The different strategies could be FFT, Filter-bank, directly from discrete time input, windowed and normalized variants of the above methods etc. Another source of variability of requirements is the choice of pre-computed coefficient storage. If the coefficients are pre-computed for storage, the memory requirements increase considerably, while the computation load decreases. The computational requirements also depend on frame size for analysis, the lowest and highest frequencies considered in the pitch range, and if any interpolation is used during computation.

Due to the difficulty in comparing the various systems, and the wide choice of the methods available, let us first take the estimate of the operations, independent of the implementation. Considering the operation of autocorrelation function computation, we have equation A.1.

$$a(l) = \frac{1}{a(0)} \sum_{j=0}^{N-1-l} x_j x_{j+l} \quad \dots \text{ A.1.}$$

The number of operations for this operation depends on the search range in the lag domain. For typical sampling rate of 8000 Hz, the search range will be 95 lags (taking into account the lower frequency and upper frequency of analysis). Therefore the total number of operations for a single frame of 380 samples (three times the lowest period corresponding to 70 Hz frequency) is equal

to $380 \times 95 \times 3 = 98,325$ mathematical operations per frame. Considering the normal overlap of half the window range, we have the total number of computations per frame is 147,487 mathematical operations per frame, and **776.25 mathematical operations per sample**. The memory requirements are strongly dependent on the implementation details.

Let us consider the YIN system as available from the authors' web site, as an example of the computational requirements of a typical system using autocorrelation function based algorithm. The main difference function calculation in the system is done by a running summary autocorrelation function, thus reducing the memory requirements for a straight forward autocorrelation computation. For this calculation, each sample-lag pair requires 6 additions and 4 multiplications. Although the calculation is done on frame by frame basis, the total number of these operations per sample per computation is 8. Therefore, the total number of arithmetic operations in the difference function calculation is **290 arithmetic operations per sample**. The smoothing function takes another 18 arithmetic operations per sample. The parabolic interpolation and normalization take in total a further 65 arithmetic operations. This is followed by a search and further smoothing operations, for which it is difficult to estimate the number of operations. However, the profiling information shows the time taken in these tasks, is about 37%. The total estimated computational costs for the YIN system is **500.05 arithmetic operations per sample**. The memory requirements are difficult to establish because the majority of operations are done on the whole waveform, and not sample by sample. The memory requirements for the system, as implemented are **43520 memory words** for one second of speech signal, sampled at 8 kHz (the figure cannot be divided by the number of samples, because the memory is allocated for the whole waveform, and all operations are done in place). Due to logarithmic and other floating point operations, the algorithm is most likely to require a minimum of 32 bit operation and memory word size, after optimization.

Let us consider the PRAAT systems' implementation. While PRAAT uses the computation of the FFT of the window and the signal (the autocorrelation of the window can be computed once and stored). The FFT is computed for 400 samples windows (512 point FFT), with a frame size of 80 samples, for a signal sampled at 8000 Hz. The number of computations required for every frame is roughly given by $N \log(N)$, where N is the FFT analysis size. Therefore, for 2 FFT operations and one multiplication operation for power spectrum computation, we get 6,900

operations per frame or **193.9 arithmetic operations per sample**. However, profiling information collected showed that this operation took only 49% of the total execution time. Therefore, we can approximate that the number of other operations performed, for example the smoothing and parabolic interpolation included, equal to total computational requirements of about **387.7 arithmetic operations per sample**. The memory requirements are difficult to establish, as the operations are performed on the entire waveform as a whole and it is difficult to establish the optimised implementation. However, on analyzing the source code for the system, it is clear that the system needs at least **32000 memory words per second of speech analysed**. In our analysis, we have not considered the PRAAT system's calculation of the harmonic to noise ratio (HNR), which is used for the voiced/unvoiced decision.

The CORR system performs the same operations as the PRAAT system, but in each channel. Therefore, the total computational requirements are multiplied by the number of channels for the basic summary autocorrelation function. Therefore for a case of 40 channels, the number of computations per sample are 40 times greater than the PRAAT system i.e. in the range of **7756 arithmetic operations per sample**. The memory requirements are also quite large, and like the rest of the system, difficult to evaluate for an optimised evaluation.

A.3. The Probabilistic Multiple Pitch Estimation and Tracking System (PMPT)

The Probabilistic Multiple Pitch Estimation and Tracking system by Wu [Wu et al, 2002] was available as source code in C upon request from to the authors. The profiling was done on this code, and the results evaluated. As with the other system evaluated, the whole waveform is evaluated at once, therefore a true measure of the memory requirements is not available readily, but is only a rough approximation, based on profiling and observation of the source code.

The system operates on the 16 bit resolution signals, sampled at 16 kHz. The main part of the pitch tracking algorithm operates on the output of a 128 channel Gammatone bank of filters (4th order). The lower frequency channels (1 to 55) and the higher frequency channels (56 – 128), are treated separately. For simple 4th order filters, the number of arithmetic operations is 5 additions and 6 multiplications, i.e. for 128 channels, the filtering computation uses 1408 arithmetic operations per sample. The next stage of processing is envelope computation, which involves

low pass filtering of the high frequency channels, and high pass filtering of the low frequency channels. Based on the profiling information, this operations takes 1.08 times the front end filtering process. This process is followed by a realignment stage, where the different channels output is realigned to compensate for the filter delays. This process completes in about 0.08 times the front end filtering process. The memory required for this stage is 128 memory words per sample. The realigned envelope information is then used by the Correlogram computation algorithm. The Correlogram is computed twice for higher frequency channels and once for lower frequency channels, with different window sizes. The method used for the computation is straight autocorrelation function calculation, as presented in equation A.1. The total number of operations for the entire calculation is 7743 mathematical operations per sample. The correlogram computation also requires a large amount of memory. The total amount of memory used is 630 memory words per sample. The Correlogram computation is followed by a peak picking algorithm, for valid pitch candidates. The total computational time requirement for this operation is 0.08 times the filter bank computation. The memory requirements for this stage are 24 memory words per sample. The next stage of processing is the main probabilistic pitch tracking system, which has a memory requirement of 1450 memory words per sample. The profiling information calculation for this portion of processing indicates that it takes 3.5 times the front end filtering process computation. Therefore, the total requirements for the system are **15,825 arithmetic operations per sample**. The overall memory requirements are computed to be **2,232 memory words per sample**. For one second of speech data, the computational requirements are therefore **253,200,000 arithmetic operations**, and memory requirements of **35,712,000 memory words**, given a sampling rate of 16 kHz.

A.4. The Comparative Analysis

The computational requirements of an algorithm are very important for practical exploitation. A system which is computationally efficient is preferable over less efficient systems even when the computational resources are available, due to considerations of power consumption and scalability of the overall system. Therefore the analysis that is presented in this section of the dissertation is not very detailed and gives approximate figures, due to the difficulty in establishing the requirements; given an optimal implementation (most systems we have evaluated were probably not optimized). However, from the analysis, we can conclude that the PRAAT system has the least computational requirements. However, it is not very economical on the memory

requirements, due to limitations of the algorithm, as it can only operate on chunks of signal, which requires large amount of memory. However, the DHO system is quite economical on the both the computational requirements in terms of arithmetic operations per sample, and the memory required for the computation. This combination of low computational requirements and low memory requirements can be considered as the most optimal solution of all the systems analysed. The YIN and CORR systems require a much larger number of arithmetic operations, and also much large memory requirements. The PMPT system is the worst system both in terms of computational and memory requirements, and the difference with other systems is quite noticeable even when running on a very fast computer (we used a computer with 256 MB of RAM, and an Athlon 1500+ XP processor).

Another factor which is of vital importance for practical implementation purposes is the number of bits required for arithmetic operations and storage. All the systems that were analysed used a mixture of double (64 bit) and single precision (32 bit) floating point operations. Based on the understanding of the system and the algorithmic requirements, apart from the PRAAT and DHO, none of the other system can be implemented in less than 32 bit resolution. The DHO system can be easily implemented as a 16 bit fixed point operation algorithm, as it does not use FFT or high order filter banks, or non-linear interpolation.

The figures for computational requirements comparison in a graphical form are included in chapter 5 (figures 6.18 and 6.19).

Appendix 2

THE COMPLETE RESULTS FOR SINGLE PITCH TRACK EXPERIMENTS

This appendix presents the complete results for the single pitch track estimation experiments. For better readability, the various systems are referred to by brief names, which were also used throughout the main text. For reference to these, please see chapter 2 and chapter 5.

The various error measures and their names, as used in the tables were described in detail in chapter 5. These are abbreviated in the tables as follows.

GEE20 is Gross Error Rate within 20% of the reference pitch estimate.

GEE10 is Gross Error Rate within 10% of the reference pitch estimate.

GEE05 is Gross Error Rate within 5% of the reference pitch estimate.

FEE20 is Fine Error Rate within 20% of the reference pitch estimate.

FEE10 is Fine Error Rate within 10% of the reference pitch estimate.

FEE05 is Fine Error Rate within 5% of the reference pitch estimate.

V_UVE is percentage voiced to unvoiced error measure.

UV_VE is the percentage unvoiced to voiced error measure.

THE is the percentage of “too-high” errors.

TLE is the percentage of “too-low” errors.

STD20 is the standard deviation of FEE20 errors.

The Male Speakers average performance is represented by the name MA.

The Female Speakers average performance is represented by the name FA.

The term OA means Overall Average.

DHO	GER20	GER10	GER05	FER20	FER10	FER05	V_UVE	UV_VE	THE	TLE	STD20
f1	0	0.37	3.35	-0.45	-0.36	-0.28	0	9.05	0	0	3.82
f2	0.19	0.38	2.81	0.08	0.03	0.23	0	46.34	0	0	4.99
f3	0	0.26	3.6	-0.15	-0.1	0.18	0	22.06	0	0	3.73
f4	0	0.2	3.07	0.2	0.14	0.24	0	6.44	0	0	5.13
f5	0	0.29	3.22	0.33	0.41	0.55	0	23.83	0	0	5.34
m1	7.51	16.22	37.05	0.83	0.42	0	2.66	15.95	0.48	2.66	6.17
m2	0	1.24	18.63	1.61	1.55	1.35	0	23.13	0	0	5.7
m3	0.61	0.92	3.36	-0.52	-0.48	-0.35	0.61	17.7	0	0.61	2.84
m4	23.89	26	34.89	0.37	0.29	0.56	4.22	3.42	2.58	4.22	3.4
m5	10.82	15.04	23.48	-0.27	0.35	0.69	5.01	13.65	2.9	5.01	4.24
MA	8.56	11.88	23.48	0.4	0.43	0.45	2.5	14.77	1.19	2.5	4.47
FA	0.04	0.3	3.21	0	0.02	0.18	0	21.54	0	0	4.6
OA	4.3	6.09	13.35	0.2	0.23	0.32	1.25	18.16	0.6	1.25	4.54

CORR	GER20	GER10	GER05	FER20	FER10	FER05	V_UVE	UV_VE	THE	TLE	STD20
f1	0.74	3.35	4.46	0.86	0.97	0.93	0	16.52	0.74	0	4.66
f2	1.69	2.25	8.82	0.07	0.11	0.08	0.19	25	0.94	0.56	6.41
f3	4.37	4.88	9.77	0.31	0.42	0.6	1.8	44.12	1.03	3.34	4.96
f4	1.64	3.27	12.27	-0.63	-0.29	0.07	0.61	31.44	0.61	0.82	8.16
f5	0	1.17	8.77	-0.17	-0.41	-0.78	0	33.05	0	0	7.35
m1	15.5	16.95	22.52	0.1	0.23	0.34	12.59	38.04	2.91	12.59	3.35
m2	1.24	1.86	26.09	-1.45	-1.56	-0.69	0.62	33.7	0.62	0.62	5.73
m3	0.31	0.92	2.75	0.1	0.19	0.2	0.31	20.66	0	0.31	3.16
m4	16.86	18.27	27.87	0.27	0.24	0.13	16.39	35.76	0.47	16.39	3.29
m5	7.26	8.31	21.24	0.65	0.59	0.36	6.6	30.92	0.66	6.6	3.89
MA	8.23	9.26	20.09	-0.07	-0.06	0.07	7.3	31.82	0.93	7.3	3.88
FA	1.69	2.98	8.82	0.09	0.16	0.18	0.52	30.03	0.66	0.94	6.31
OA	4.96	6.12	14.46	0.01	0.05	0.12	3.91	30.92	0.8	4.12	5.1

YIN	GER20	GER10	GER05	FER20	FER10	FER05	V_UVE	UV_VE	THE	TLE	STD20
f1	6.69	9.67	14.5	1.01	0.08	0.1	0	97.51	0.37	4.83	6.95
f2	8.44	11.63	17.64	1.4	0.28	0.29	0	98.06	0	5.82	8
f3	4.63	7.2	11.83	1.12	0.29	0.49	0	96.56	0.26	3.6	6.54
f4	39.67	45.6	53.37	3.58	0.62	0.09	0	97.16	0.61	35.38	12.78
f5	13.45	17.25	21.93	2	0.52	0.41	0	96.81	0	13.16	8.84
m1	1.21	3.87	13.32	0.52	0.71	0.52	0	97.51	0	0	3.7
m2	0.62	0.62	1.86	0.81	0.81	0.78	0	98.02	0.62	0	2.53
m3	0.61	1.22	2.75	0.21	0.09	0	0	97.05	0	0.61	2.66
m4	1.17	2.11	6.79	0.61	0.7	0.58	0	98.86	0.47	0.23	2.37
m5	1.58	2.77	5.54	0.39	0.41	0.43	0	97.72	0.53	0.26	2.73
MA	1.04	2.12	6.05	0.51	0.54	0.46	0	97.83	0.32	0.22	2.8
FA	14.58	18.27	23.85	1.82	0.36	0.27	0	97.22	0.25	12.56	8.62
OA	7.81	10.19	14.95	1.16	0.45	0.37	0	97.53	0.29	6.39	5.71

PRAAT	GER20	GER10	GER05	FER20	FER10	FER05	V_UVE	UV_VE	THE	TLE	STD20
f1	4.09	6.32	8.55	1.04	1.08	1.1	0.37	9.28	0.74	3.35	4.7
f2	2.06	3.19	10.13	0.28	0.39	0.37	0	3.45	0.38	1.69	6.65
f3	3.86	4.11	10.54	0.58	0.65	0.83	1.29	26.22	0.77	3.08	5.19
f4	3.07	4.91	13.09	-0.11	-0.04	0.26	2.66	15.98	0	2.86	8.04
f5	0	1.17	8.48	0.07	-0.18	-0.66	0	24.33	0	0	7.35
m1	7.99	10.41	18.89	0.52	0.54	0.43	7.75	6.15	0	7.75	3.87
m2	0.62	2.48	30.43	-1.47	-1.63	-0.42	0	5.51	0.62	0	6.34
m3	0.61	1.22	5.5	0.26	0.36	0.28	0.61	11.48	0	0.61	3.54
m4	10.54	12.88	23.19	0.68	0.58	0.36	10.54	9.57	0	10.54	3.52
m5	1.85	3.83	18.34	0.76	0.75	0.38	1.19	17.67	0.66	1.19	3.96
MA	4.32	6.17	19.27	0.15	0.12	0.21	4.02	10.07	0.26	4.02	4.25
FA	2.62	3.94	10.16	0.37	0.38	0.38	0.86	15.85	0.38	2.2	6.39
OA	3.47	5.05	14.71	0.26	0.25	0.29	2.44	12.96	0.32	3.11	5.32

Table A.2.1. The clean speech high resolution signal results for evaluated systems.

DHO	GER20	GER10	GER05	FER20	FER10	FER05	V_UVE	UV_VE	THE	TLE	STD20
f1	0.74	1.12	4.09	-0.46	-0.36	-0.28	0	12.67	0	0	3.85
f2	0	0.38	2.63	0.21	0.09	0.27	0	43.75	0	0	5.21
f3	0	0.26	3.6	-0.14	-0.09	0.19	0	22.6	0	0	3.72
f4	1.43	1.64	4.5	0.11	0.05	0.14	1.43	7.22	0	1.43	5.11
f5	0	0.29	3.22	0.36	0.44	0.6	0	27.35	0	0	5.34
m1	7.02	15.01	36.08	0.76	0.47	0.02	3.39	24.92	1.45	3.39	6.09
m2	0	1.24	16.77	1.36	1.3	1.42	0	22.91	0	0	5.53
m3	0.61	0.92	3.06	-0.51	-0.48	-0.37	0.61	15.08	0	0.61	2.84
m4	28.81	29.74	39.11	0.35	0.36	0.58	5.85	6.61	1.17	5.85	3.23
m5	7.78	12.01	21.64	-0.29	0.32	0.65	1.85	13.25	3.43	1.85	4.25
MA	8.84	11.78	23.33	0.33	0.39	0.46	2.34	16.55	1.21	2.34	4.39
FA	0.43	0.74	3.61	0.02	0.03	0.18	0.29	22.72	0	0.29	4.65
OA	4.64	6.26	13.47	0.18	0.21	0.32	1.31	19.64	0.61	1.31	4.52

CORR	GER20	GER10	GER05	FER20	FER10	FER05	V_UVE	UV_VE	THE	TLE	STD20
f1	0.74	2.6	5.58	0.81	0.75	0.98	0	13.35	0.74	0	4.61
f2	1.88	2.44	9.57	0.06	0.1	0.03	0.38	15.09	0.94	0.75	6.46
f3	4.37	4.63	9.51	0.37	0.43	0.68	1.8	27.49	1.03	3.34	4.93
f4	3.07	4.09	12.88	-0.49	-0.32	0.16	2.45	23.71	0.41	2.45	7.92
f5	0	1.17	8.19	-0.21	-0.45	-0.88	0	25.34	0	0	7.35
m1	16.46	18.16	23.24	0.11	0.24	0.33	14.29	24.42	2.18	14.29	3.37
m2	1.24	1.86	25.47	-1.42	-1.53	-0.72	0.62	13.44	0.62	0.62	5.69
m3	1.22	1.83	3.67	0.05	0.15	0.16	1.22	16.39	0	1.22	3.17
m4	17.56	18.74	28.1	0.17	0.19	0.14	17.33	30.75	0.23	17.33	3.18
m5	8.71	9.76	22.16	0.66	0.61	0.35	7.78	20.88	0.92	7.78	3.88
MA	9.04	10.07	20.53	-0.09	-0.07	0.05	8.25	21.18	0.79	8.25	3.86
FA	2.01	2.99	9.15	0.11	0.1	0.19	0.93	20.99	0.62	1.31	6.26
OA	5.53	6.53	14.84	0.01	0.02	0.12	4.59	21.09	0.71	4.78	5.06

YIN	GER20	GER10	GER05	FER20	FER10	FER05	V_UVE	UV_VE	THE	TLE	STD20
f1	8.18	11.9	17.47	1	0.23	-0.02	0	98.64	0.37	6.69	7.07
f2	8.82	12.38	18.39	1.35	0.24	0.25	0	98.92	0	5.82	8.1
f3	4.88	6.94	10.54	0.97	0.33	0.47	0	97.47	0.26	3.6	6.02
f4	43.15	49.28	58.9	3.84	0.27	-0.15	0	97.68	0.41	38.85	13.81
f5	13.74	17.54	22.22	1.87	0.44	0.45	0	97.48	0	13.16	8.59
m1	1.21	3.87	13.08	0.54	0.73	0.56	0	98.34	0	0	3.7
m2	0.62	0.62	2.48	0.8	0.8	0.82	0	99.12	0.62	0	2.59
m3	0.61	1.22	2.45	0.21	0.06	0	0	98.36	0	0.61	2.82
m4	1.17	1.87	6.79	0.59	0.7	0.6	0	99.32	0.47	0.23	2.33
m5	1.32	2.77	5.94	0.45	0.46	0.46	0	97.86	0.53	0.13	2.84
MA	0.99	2.07	6.15	0.52	0.55	0.49	0	98.6	0.32	0.2	2.85
FA	15.75	19.61	25.5	1.81	0.3	0.2	0	98.04	0.21	13.62	8.72
OA	8.37	10.84	15.83	1.16	0.43	0.34	0	98.32	0.27	6.91	5.79

PRAAT	GER20	GER10	GER05	FER20	FER10	FER05	V_UVE	UV_VE	THE	TLE	STD20
f1	3.72	5.2	7.81	1.04	0.9	1.08	0	2.94	0.74	2.97	4.48
f2	2.25	2.81	9.57	0.3	0.26	0.34	0	2.37	0.38	1.69	6.38
f3	3.6	3.86	11.05	0.55	0.62	0.75	1.29	23.87	1.03	2.57	5.21
f4	4.09	5.73	12.88	-0.19	-0.04	0.25	3.89	18.3	0	3.89	7.78
f5	0	1.17	8.77	0.05	-0.2	-0.73	0	23.83	0	0	7.35
m1	7.26	9.93	18.64	0.56	0.55	0.45	7.02	5.65	0	7.02	3.93
m2	0.62	3.11	29.81	-1.41	-1.67	-0.41	0	4.85	0.62	0	6.41
m3	0.61	1.22	5.5	0.23	0.33	0.3	0.61	7.87	0	0.61	3.53
m4	10.3	12.65	23.19	0.68	0.59	0.38	10.3	7.06	0	10.3	3.54
m5	1.58	3.56	18.21	0.75	0.74	0.36	0.92	17.94	0.66	0.92	3.96
MA	4.08	6.09	19.07	0.16	0.11	0.21	3.77	8.67	0.26	3.77	4.28
FA	2.73	3.75	10.02	0.35	0.31	0.34	1.03	14.26	0.43	2.22	6.24
OA	3.4	4.92	14.54	0.26	0.21	0.28	2.4	11.47	0.34	3	5.26

Table A.2.2. Results for 25 dB White noise for high resolution speech signals.

DHO	GER20	GER10	GER05	FER20	FER10	FER05	V_UVE	UV_VE	THE	TLE	STD20
f1	0.74	1.12	4.09	-0.46	-0.36	-0.28	0	12.9	0	0	3.86
f2	4.32	4.5	6.57	0.02	-0.03	0.23	0	41.81	0	4.32	4.98
f3	0	0.26	3.86	-0.12	-0.07	0.23	0	19.89	0	0	3.71
f4	0.41	1.84	4.5	-0.36	0.18	0.32	0	7.47	0	0	7.38
f5	0.29	0.58	2.92	0.43	0.51	0.56	0	26.85	0	0	5.18
m1	5.81	15.25	35.11	0.65	0.32	-0.02	1.94	21.93	0.24	1.94	6.25
m2	1.86	3.11	19.25	1.43	1.37	1.43	0	24.01	0	0	5.61
m3	0.61	0.92	3.06	-0.53	-0.49	-0.39	0.61	15.08	0	0.61	2.84
m4	29.04	30.44	38.17	0.31	0.19	0.5	5.85	4.56	0.47	5.85	3.23
m5	8.44	12.27	22.82	-0.21	0.28	0.62	2.9	14.46	3.56	2.9	4.27
MA	9.15	12.4	23.68	0.33	0.33	0.43	2.26	16.01	0.85	2.26	4.44
FA	1.15	1.66	4.39	-0.1	0.05	0.21	0	21.78	0	0.86	5.02
OA	5.15	7.03	14.04	0.12	0.19	0.32	1.13	18.89	0.43	1.56	4.73

CORR	GER20	GER10	GER05	FER20	FER10	FER05	V_UVE	UV_VE	THE	TLE	STD20
f1	3.35	4.83	6.32	0.96	0.85	0.86	2.97	9.05	0.37	2.97	4.35
f2	3.19	3.94	10.13	0	0.03	0.14	2.06	15.52	0.75	2.44	6.56
f3	5.14	5.66	10.54	0.44	0.56	0.61	2.83	24.05	1.03	4.11	5.06
f4	5.93	6.95	15.13	-0.58	-0.4	0.14	5.32	23.45	0.41	5.32	7.75
f5	0	1.17	8.19	-0.22	-0.46	-0.9	0	22.99	0	0	7.34
m1	16.95	18.4	23	0.05	0.21	0.33	15.25	18.27	1.69	15.25	3.31
m2	3.11	4.35	27.33	-1.4	-1.62	-0.85	2.48	9.25	0.62	2.48	5.75
m3	1.53	2.14	4.28	0.05	0.14	0.17	1.53	11.8	0	1.53	3.22
m4	18.97	19.91	29.04	0.13	0.18	0.14	18.74	20.05	0.23	18.74	3.13
m5	12.8	13.85	24.54	0.66	0.6	0.31	12.27	17.67	0.53	12.27	3.82
MA	10.67	11.73	21.64	-0.1	-0.1	0.02	10.05	15.41	0.62	10.05	3.85
FA	3.52	4.51	10.06	0.12	0.12	0.17	2.64	19.01	0.51	2.97	6.21
OA	7.1	8.12	15.85	0.01	0.01	0.1	6.35	17.21	0.56	6.51	5.03

YIN	GER20	GER10	GER05	FER20	FER10	FER05	V_UVE	UV_VE	THE	TLE	STD20
f1	9.67	12.64	17.1	0.54	0.01	-0.07	0	99.1	0.74	7.43	6.63
f2	9.94	13.51	19.51	1.45	0.32	0.29	0	98.49	0	7.13	8.31
f3	5.14	6.68	10.03	0.79	0.37	0.64	0	97.47	0.26	3.6	5.23
f4	48.88	55.42	63.6	4.27	0.48	-0.35	0	98.71	0.41	44.17	13.84
f5	14.33	19.01	23.1	2.16	0.31	0.46	0	97.82	0	13.16	9.36
m1	1.21	4.36	14.29	0.58	0.77	0.58	0	98.34	0	0	3.78
m2	0.62	0.62	2.48	0.81	0.81	0.82	0	99.12	0.62	0	2.65
m3	0.61	1.22	2.75	0.29	0.14	0.04	0	99.34	0	0.61	2.86
m4	1.17	1.87	7.26	0.56	0.67	0.58	0	99.09	0.47	0.23	2.35
m5	1.32	3.03	6.6	0.46	0.47	0.47	0	97.05	0.66	0.13	2.99
MA	0.99	2.22	6.68	0.54	0.57	0.5	0	98.59	0.35	0.2	2.93
FA	17.59	21.45	26.67	1.84	0.3	0.19	0	98.32	0.28	15.1	8.68
OA	9.29	11.84	16.67	1.19	0.43	0.35	0	98.45	0.32	7.65	5.8

PRAAT	GER20	GER10	GER05	FER20	FER10	FER05	V_UVE	UV_VE	THE	TLE	STD20
f1	5.95	7.43	8.55	1.2	1.06	1.02	2.23	1.58	0.74	5.2	4.24
f2	3.56	4.32	11.82	0.25	0.29	0.35	1.5	1.94	0.38	3.19	6.73
f3	4.37	4.63	10.28	0.57	0.63	0.75	2.06	22.78	0.51	3.86	5.09
f4	5.73	7.16	13.7	-0.33	-0.23	0.16	5.52	17.53	0	5.52	7.62
f5	0	1.17	8.77	0.03	-0.22	-0.76	0	28.02	0	0	7.38
m1	8.96	11.14	19.61	0.48	0.54	0.45	8.72	5.81	0	8.72	3.82
m2	0.62	3.11	29.19	-1.38	-1.65	-0.44	0	3.52	0.62	0	6.38
m3	1.22	1.83	5.81	0.19	0.29	0.28	1.22	2.3	0	1.22	3.52
m4	11.48	13.35	23.89	0.57	0.55	0.4	11.48	3.87	0	11.48	3.38
m5	5.15	6.73	20.98	0.77	0.75	0.36	5.01	17	0.13	5.01	3.96
MA	5.48	7.23	19.9	0.12	0.09	0.21	5.29	6.5	0.15	5.29	4.21
FA	3.92	4.94	10.63	0.34	0.3	0.31	2.26	14.37	0.33	3.55	6.21
OA	4.7	6.09	15.26	0.23	0.2	0.26	3.77	10.44	0.24	4.42	5.21

Table A.2.3. Results for 20 dB White noise for high resolution speech signals.

DHO	GER20	GER10	GER05	FER20	FER10	FER05	V_UVE	UV_VE	THE	TLE	STD20
f1	0	0.37	3.72	-0.47	-0.37	-0.33	0	17.42	0	0	3.9
f2	1.13	1.5	3.38	0.19	0.08	0.3	0.75	45.26	0	0.75	5.06
f3	0.51	0.51	3.6	-0.07	-0.07	0.19	0	27.67	0	0	3.55
f4	1.43	3.27	6.34	-0.6	0.11	0.29	0	4.12	0	0	7.97
f5	0.29	0.58	3.51	0.36	0.44	0.59	0	28.52	0	0	5.36
m1	10.17	17.43	38.01	0.46	0.43	-0.03	0.97	18.27	0.24	0.97	6.17
m2	0.62	1.24	17.39	1.01	1.15	1.21	0	24.89	0	0	5.02
m3	0	0	3.06	-0.56	-0.56	-0.37	0	20.66	0	0	2.89
m4	20.14	21.55	29.98	0.31	0.3	0.58	5.85	2.96	0.23	5.85	3.21
m5	6.73	10.82	19.39	-0.28	0.24	0.64	2.51	14.99	2.11	2.51	4.25
MA	7.53	10.21	21.57	0.19	0.31	0.4	1.87	16.35	0.52	1.87	4.31
FA	0.67	1.25	4.11	-0.12	0.04	0.21	0.15	24.6	0	0.15	5.17
OA	4.1	5.73	12.84	0.03	0.18	0.31	1.01	20.48	0.26	1.01	4.74

CORR	GER20	GER10	GER05	FER20	FER10	FER05	V_UVE	UV_VE	THE	TLE	STD20
f1	4.83	5.2	7.06	1.12	1.04	0.97	4.83	4.75	0	4.83	3.69
f2	6.19	6.94	13.13	-0.07	-0.04	-0.02	5.07	14.87	0.75	5.44	6.64
f3	7.46	7.97	13.11	0.41	0.54	0.57	5.91	22.6	0.51	6.94	5.22
f4	18.61	19.02	26.18	-1	-0.86	-0.24	18	20.88	0.41	18	7.4
f5	0	1.17	9.65	-0.21	-0.46	-0.95	0	20.47	0	0	7.64
m1	21.07	22.28	26.88	0.04	0.18	0.31	19.61	14.95	1.45	19.61	3.35
m2	11.18	12.42	34.16	-1.63	-1.87	-0.9	11.18	4.85	0	11.18	5.79
m3	7.65	8.26	11.62	0.04	0.14	0.23	7.65	4.92	0	7.65	3.3
m4	21.55	22.48	31.62	0.13	0.19	0.15	21.55	9.34	0	21.55	3.14
m5	25.2	26.12	33.91	0.48	0.44	0.24	25.07	16.2	0.13	25.07	3.72
MA	17.33	18.31	27.64	-0.19	-0.18	0	17.01	10.05	0.32	17.01	3.86
FA	7.42	8.06	13.83	0.05	0.04	0.07	6.76	16.71	0.33	7.04	6.12
OA	12.37	13.19	20.73	-0.07	-0.07	0.04	11.89	13.38	0.33	12.03	4.99

YIN	GER20	GER10	GER05	FER20	FER10	FER05	V_UVE	UV_VE	THE	TLE	STD20
f1	9.67	14.13	18.59	1.11	0.18	-0.17	0	99.1	0.74	7.06	7.48
f2	11.82	16.7	23.26	1.83	0.52	0.15	0	98.92	0.19	8.44	8.93
f3	6.94	9.25	14.14	1.25	0.56	0.58	0	98.19	0.26	4.63	6.59
f4	54.19	61.35	71.57	5.9	2.13	0.2	0	99.23	0.41	48.88	15.49
f5	16.96	22.22	27.19	2.44	0.37	0.28	0	97.65	0	14.62	9.82
m1	1.69	4.6	13.8	0.62	0.79	0.59	0	98.84	0.48	0	3.76
m2	2.48	3.73	6.83	0.97	0.68	0.58	0	99.56	0.62	0.62	4.12
m3	2.75	3.98	6.42	0.23	0.08	0.02	0	100	0.31	1.22	3.25
m4	0.94	1.41	7.03	0.6	0.6	0.57	0	99.09	0.47	0.23	2.39
m5	2.51	4.49	7.78	0.43	0.42	0.46	0	97.86	1.32	0	2.99
MA	2.08	3.64	8.37	0.57	0.51	0.44	0	99.07	0.64	0.42	3.3
FA	19.92	24.73	30.95	2.51	0.75	0.21	0	98.62	0.32	16.73	9.66
OA	11	14.18	19.66	1.54	0.63	0.33	0	98.84	0.48	8.57	6.48

PRAAT	GER20	GER10	GER05	FER20	FER10	FER05	V_UVE	UV_VE	THE	TLE	STD20
f1	9.67	10.04	11.15	1.28	1.19	1.03	6.69	1.58	0	9.67	3.39
f2	6.57	7.13	13.88	0.24	0.2	0.28	4.5	1.08	0.38	6.19	6.49
f3	7.2	7.46	12.6	0.71	0.78	0.82	5.14	21.16	0.51	6.68	5.11
f4	23.11	23.31	30.06	-0.87	-0.82	-0.16	22.9	15.46	0	22.9	7.25
f5	0	1.17	9.06	-0.03	-0.28	-0.77	0	26.51	0	0	7.51
m1	10.9	12.35	21.31	0.34	0.5	0.41	10.9	4.32	0	10.9	3.72
m2	9.94	13.04	35.4	-1.58	-1.77	-0.63	9.94	1.98	0	9.94	6.4
m3	3.98	4.59	8.56	0.15	0.26	0.23	3.98	0.66	0	3.98	3.56
m4	13.82	14.99	26	0.48	0.5	0.37	13.82	2.96	0	13.82	3.3
m5	11.21	12.4	24.14	0.8	0.72	0.31	11.08	15.8	0.13	11.08	3.91
MA	9.97	11.47	23.08	0.04	0.04	0.14	9.94	5.14	0.03	9.94	4.18
FA	9.31	9.82	15.35	0.27	0.22	0.24	7.85	13.16	0.18	9.09	5.95
OA	9.64	10.65	19.22	0.15	0.13	0.19	8.89	9.15	0.1	9.52	5.07

Table A.2.4. Results for 15 dB White noise for high resolution speech signals.

DHO	GER20	GER10	GER05	FER20	FER10	FER05	V_UVE	UV_VE	THE	TLE	STD20
f1	1.49	1.49	4.09	-0.3	-0.3	-0.27	0	22.17	0	0	3.56
f2	0	0.38	2.44	0.3	0.18	0.32	0	47.41	0	0	5.26
f3	0.51	0.77	3.6	-0.08	-0.03	0.2	0	29.66	0	0	3.73
f4	0.61	2.45	6.13	-0.16	-0.11	0.28	0.61	9.28	0	0.61	6.94
f5	0.29	0.58	3.51	0.32	0.4	0.55	0	35.4	0	0	5.39
m1	8.23	16.22	35.35	0.54	0.17	0.02	0	25.42	0.48	0	6.36
m2	0.62	2.48	16.77	1.07	1.44	1.2	0	33.04	0	0	5.29
m3	0.31	0.61	5.2	-0.61	-0.56	-0.3	0	22.62	0.31	0	3.06
m4	23.19	25.53	34.19	0.25	0.32	0.45	4.45	4.33	0.47	4.45	3.41
m5	8.97	12.53	21.64	-0.42	0.12	0.58	5.28	15.93	2.24	5.28	4.27
MA	8.26	11.48	22.63	0.16	0.3	0.39	1.95	20.27	0.7	1.95	4.48
FA	0.58	1.13	3.95	0.01	0.03	0.22	0.12	28.78	0	0.12	4.98
OA	4.42	6.31	13.29	0.09	0.16	0.3	1.03	24.53	0.35	1.03	4.73

CORR	GER20	GER10	GER05	FER20	FER10	FER05	V_UVE	UV_VE	THE	TLE	STD20
f1	10.04	10.41	11.52	1.16	1.07	0.88	10.04	2.94	0	10.04	3.83
f2	13.13	13.7	19.7	0.01	-0.05	0.07	12.01	10.99	0.94	12.2	6.57
f3	15.94	16.45	19.54	0.66	0.67	0.56	14.91	23.51	0.51	15.42	4.84
f4	37.01	37.22	42.74	-1.57	-1.51	-0.8	36.4	18.04	0.61	36.4	7.39
f5	3.51	4.97	13.74	-0.26	-0.46	-1.04	3.51	26.51	0	3.51	7.89
m1	32.69	33.41	37.29	-0.05	0	0.12	31.72	9.47	0.97	31.72	3.23
m2	24.84	26.09	43.48	-2.77	-2.84	-1.22	24.84	2.42	0	24.84	5.4
m3	27.22	27.52	31.8	-0.08	-0.01	0.23	27.22	1.64	0	27.22	3.62
m4	33.49	34.19	42.15	0.24	0.27	0.15	33.49	5.47	0	33.49	3.09
m5	46.31	46.83	50.26	0.05	0.05	0.1	46.17	17.94	0.13	46.17	3.26
MA	32.91	33.61	41	-0.52	-0.51	-0.13	32.69	7.39	0.22	32.69	3.72
FA	15.93	16.55	21.45	0	-0.05	-0.07	15.37	16.4	0.41	15.51	6.11
OA	24.42	25.08	31.22	-0.26	-0.28	-0.1	24.03	11.89	0.32	24.1	4.91

YIN	GER20	GER10	GER05	FER20	FER10	FER05	V_UVE	UV_VE	THE	TLE	STD20
f1	18.96	25.28	33.46	2.66	0.88	0.29	0	99.55	0.37	15.24	8.95
f2	20.45	29.46	37.9	3.28	0.33	0.21	0	97.63	0.19	15.2	12.51
f3	16.97	22.11	30.85	2.16	0.63	0.53	0	98.73	0	12.6	8.34
f4	60.12	68.1	75.26	6.59	1.17	-0.87	0	98.2	0.41	54.19	17.09
f5	28.95	35.09	43.57	3.18	0.3	0.04	0	98.32	0	24.27	11.81
m1	2.42	5.33	15.5	0.64	0.72	0.62	0	99.5	1.45	0.24	3.8
m2	5.59	6.21	11.18	0.99	0.9	0.53	0	100	0.62	1.86	3.42
m3	4.89	8.26	11.01	0.73	0.15	0.01	0	100	0.61	3.98	4.6
m4	0.7	2.11	8.2	0.5	0.58	0.52	0	99.32	0	0	2.63
m5	4.49	7.52	11.35	0.34	0.38	0.47	0	98.53	1.98	0.13	3.66
MA	3.62	5.88	11.45	0.64	0.55	0.43	0	99.47	0.93	1.24	3.62
FA	29.09	36.01	44.21	3.57	0.66	0.04	0	98.49	0.19	24.3	11.74
OA	16.35	20.95	27.83	2.11	0.6	0.24	0	98.98	0.56	12.77	7.68

PRAAT	GER20	GER10	GER05	FER20	FER10	FER05	V_UVE	UV_VE	THE	TLE	STD20
f1	8.92	9.29	10.41	1.38	1.29	1.13	8.92	0.45	0	8.92	3.49
f2	14.26	14.63	20.26	0.13	0.15	0.23	12.95	0	0.38	13.88	6.46
f3	18.77	19.02	22.11	0.79	0.86	0.8	17.22	12.48	0.51	18.25	4.72
f4	41.31	41.51	47.24	-1.55	-1.48	-0.73	41.1	8.76	0	41.1	7.42
f5	5.26	6.43	15.79	0.04	-0.23	-0.86	5.26	19.46	0	5.26	7.91
m1	15.98	17.19	25.42	0.35	0.47	0.34	15.98	1.99	0	15.98	3.65
m2	23.6	26.71	43.48	-2.78	-2.78	-1.07	23.6	0.88	0	23.6	5.99
m3	19.57	19.88	23.55	0.07	0.15	0.22	19.57	0.66	0	19.57	3.73
m4	14.99	15.93	27.17	0.43	0.42	0.26	14.99	1.37	0	14.99	3.29
m5	37.2	37.86	43.14	0.39	0.36	0.24	37.2	6.83	0	37.2	3.32
MA	22.27	23.51	32.55	-0.31	-0.28	0	22.27	2.34	0	22.27	4
FA	17.7	18.18	23.16	0.16	0.12	0.11	17.09	8.23	0.18	17.48	6
OA	19.99	20.85	27.86	-0.08	-0.08	0.06	19.68	5.29	0.09	19.88	5

Table A.2.5. Results for 10 dB White noise for high resolution speech signals.

DHO	GER20	GER10	GER05	FER20	FER10	FER05	V_UVE	UV_VE	THE	TLE	STD20
f1	1.86	2.6	5.2	-0.47	-0.49	-0.41	0	40.27	0	0	3.85
f2	0.56	1.31	4.13	0.31	0.19	0.28	0	62.07	0	0	5.62
f3	1.29	1.54	4.11	0.12	0.05	0.22	0	45.21	0	0.77	3.78
f4	6.13	9.82	13.5	-0.41	0.15	0.38	2.66	31.7	0	2.66	9.44
f5	0	0.29	2.63	0.37	0.44	0.44	0	52.18	0	0	5.51
m1	9.2	17.68	36.8	1.14	0.54	-0.04	0	25.75	2.66	0	6.45
m2	0.62	2.48	15.53	0.92	1.31	1.31	0	40.09	0	0	5.44
m3	0.92	0.92	7.03	-0.65	-0.65	-0.38	0.92	23.28	0	0.92	3.1
m4	12.41	14.29	26	0.18	0.37	0.59	3.04	16.63	0.94	3.04	3.46
m5	12.14	15.57	25.73	-0.34	0.19	0.52	6.33	26.1	0	6.33	4.48
MA	7.06	10.19	22.22	0.25	0.35	0.4	2.06	26.37	0.72	2.06	4.59
FA	1.97	3.11	5.91	-0.02	0.07	0.18	0.53	46.29	0	0.69	5.64
OA	4.51	6.65	14.07	0.12	0.21	0.29	1.3	36.33	0.36	1.37	5.11

CORR	GER20	GER10	GER05	FER20	FER10	FER05	V_UVE	UV_VE	THE	TLE	STD20
f1	27.88	27.88	29.37	0.92	0.92	0.79	27.88	1.58	0	27.88	3.89
f2	31.71	32.08	38.27	-1.04	-1.03	-0.7	30.77	4.09	0.75	30.96	6.82
f3	40.36	40.36	43.19	0.43	0.43	0.52	40.1	17	0.26	40.1	4.37
f4	61.55	61.55	64.83	-2.09	-2.09	-1.08	60.94	13.4	0.61	60.94	7.06
f5	24.27	25.44	34.8	-0.56	-0.87	-1.36	24.27	21.98	0	24.27	8.3
m1	52.06	52.54	54.96	-0.2	-0.22	0	51.82	4.98	0.24	51.82	3.21
m2	47.2	47.83	60.87	-3.8	-3.68	-1.69	47.2	0.88	0	47.2	5.14
m3	59.02	59.02	60.55	-0.29	-0.29	-0.38	59.02	0	0	59.02	3.13
m4	55.74	55.97	62.06	0.04	0.1	0.1	55.74	2.51	0	55.74	3.14
m5	67.28	67.41	68.73	-0.52	-0.57	-0.28	67.15	12.85	0.13	67.15	2.83
MA	56.26	56.56	61.44	-0.95	-0.93	-0.45	56.19	4.24	0.07	56.19	3.49
FA	37.15	37.46	42.09	-0.47	-0.53	-0.37	36.79	11.61	0.32	36.83	6.09
OA	46.71	47.01	51.76	-0.71	-0.73	-0.41	46.49	7.93	0.2	46.51	4.79

YIN	GER20	GER10	GER05	FER20	FER10	FER05	V_UVE	UV_VE	THE	TLE	STD20
f1	38.29	47.58	61.71	5.13	1.41	-0.12	0	99.77	0.37	30.86	13.22
f2	47.47	57.04	65.85	6.14	1.29	-0.23	0	99.14	0	39.96	14.82
f3	35.73	44.73	56.81	3.77	0.57	0.01	0	99.28	0	28.79	11.94
f4	66.26	73.82	81.19	8.55	1.73	-0.24	0	97.94	0.82	58.69	17.3
f5	47.37	57.02	66.67	6.2	1.06	0.15	0	98.15	0	40.64	14.91
m1	4.6	7.51	19.13	0.88	0.79	0.58	0	99.5	1.94	0.97	4.35
m2	9.94	14.91	21.12	1.14	0.84	0.68	0	100	1.24	3.73	5.79
m3	8.26	12.54	17.43	0.92	0.14	-0.03	0	100	0.31	5.2	5.14
m4	1.87	3.51	11.01	0.49	0.59	0.52	0	100	0	0	2.71
m5	10.95	16.36	23.22	-0.13	0.42	0.35	0	99.06	4.75	0.26	4.7
MA	7.12	10.96	18.38	0.66	0.55	0.42	0	99.71	1.65	2.03	4.54
FA	47.02	56.04	66.45	5.96	1.21	-0.09	0	98.86	0.24	39.79	14.44
OA	27.07	33.5	42.41	3.31	0.88	0.17	0	99.28	0.94	20.91	9.49

PRAAT	GER20	GER10	GER05	FER20	FER10	FER05	V_UVE	UV_VE	THE	TLE	STD20
f1	28.62	28.62	29.37	1.22	1.22	1.12	28.62	0	0	28.62	3.31
f2	37.71	38.09	43.34	-0.81	-0.78	-0.64	35.83	0	0	37.71	6.89
f3	43.7	43.7	46.27	0.17	0.17	0.72	43.7	0.54	0	43.7	4.43
f4	65.24	65.24	67.48	-1.95	-1.95	-0.87	65.24	0.77	0	65.24	7.02
f5	26.9	27.78	34.8	-0.53	-0.99	-1.32	26.9	3.69	0	26.9	8.14
m1	33.41	34.14	39.47	0.04	0.1	0.22	33.41	0	0	33.41	3.54
m2	49.07	50.31	63.35	-4.41	-4.15	-1.78	49.07	0	0	49.07	5.07
m3	56.88	56.88	58.1	-0.16	-0.16	-0.14	56.88	0	0	56.88	3.31
m4	32.55	33.02	41.69	0.5	0.49	0.31	32.55	1.37	0	32.55	3.12
m5	60.03	60.03	62.66	-0.15	-0.15	-0.07	60.03	1.34	0	60.03	2.98
MA	46.39	46.88	53.06	-0.84	-0.77	-0.29	46.39	0.54	0	46.39	3.61
FA	40.43	40.69	44.25	-0.38	-0.47	-0.2	40.06	1	0	40.43	5.96
OA	43.41	43.78	48.65	-0.61	-0.62	-0.24	43.22	0.77	0	43.41	4.78

Table A.2.6. Results for 5 dB White noise for high resolution speech signals.

DHO	GER20	GER10	GER05	FER20	FER10	FER05	V_UVE	UV_VE	THE	TLE	STD20
f1	3.72	6.32	7.06	-1.11	-0.43	-0.29	0	96.15	0	0	5.43
f2	2.25	5.07	9.38	0.49	0.14	0.48	0	100	0	0.19	7.69
f3	0.77	0.77	3.86	-0.11	-0.11	0.17	0	78.48	0	0	3.87
f4	14.52	23.72	30.47	2.42	0.9	0.78	0	86.86	0	1.02	13.4
f5	0.29	1.17	5.56	0.54	0.34	0.44	0	94.8	0	0.29	6.76
m1	8.23	18.4	38.5	0.86	0.28	0.03	0	68.11	0.24	0	6.55
m2	6.83	12.42	28.57	0	1.15	1.16	0	79.74	0	0	6.64
m3	1.83	2.75	8.26	-0.81	-0.63	-0.46	0	87.87	0	0	3.82
m4	26.7	29.98	40.52	0.35	0.38	0.57	3.04	50.57	0	3.04	4.12
m5	13.19	19.39	30.08	-0.78	-0.03	0.41	2.77	65.19	1.58	2.77	5.26
MA	11.36	16.59	29.18	-0.08	0.23	0.34	1.16	70.29	0.37	1.16	5.28
FA	4.31	7.41	11.27	0.45	0.17	0.32	0	91.26	0	0.3	7.43
OA	7.83	12	20.22	0.18	0.2	0.33	0.58	80.78	0.18	0.73	6.35

CORR	GER20	GER10	GER05	FER20	FER10	FER05	V_UVE	UV_VE	THE	TLE	STD20
f1	61.34	61.34	61.34	0.62	0.62	0.62	61.34	0.68	0	61.34	3.17
f2	60.23	60.41	63.98	-2.72	-2.62	-1.39	60.04	0.86	0.19	60.04	6.26
f3	70.95	70.95	72.24	0.47	0.47	0.68	70.95	7.05	0	70.95	4.14
f4	77.71	77.71	79.75	-2.67	-2.67	-1.24	77.3	4.9	0.41	77.3	7.09
f5	57.6	58.48	64.62	-0.46	-1.19	-1.61	57.31	7.21	0	57.6	9.2
m1	87.41	87.41	87.89	-1.69	-1.69	-1.4	86.68	1.33	0.73	86.68	2.62
m2	74.53	74.53	82.61	-5.44	-5.44	-2.57	74.53	0	0	74.53	5.13
m3	87.46	87.46	88.07	-1.86	-1.86	-1.45	87.46	0	0	87.46	3.15
m4	92.27	92.27	94.15	-0.71	-0.71	-0.34	92.27	0.91	0	92.27	3.93
m5	91.03	91.03	91.29	-0.61	-0.61	-0.33	90.9	4.15	0.13	90.9	2.55
MA	86.54	86.54	88.8	-2.06	-2.06	-1.22	86.37	1.28	0.17	86.37	3.48
FA	65.57	65.78	68.39	-0.95	-1.08	-0.59	65.39	4.14	0.12	65.45	5.97
OA	76.05	76.16	78.59	-1.51	-1.57	-0.9	75.88	2.71	0.15	75.91	4.72

YIN	GER20	GER10	GER05	FER20	FER10	FER05	V_UVE	UV_VE	THE	TLE	STD20
f1	53.53	65.43	75.46	7.55	1.55	0.02	0	99.77	0	44.24	14.87
f2	62.85	74.11	83.11	8.93	2.2	-0.15	0	99.78	0.19	54.97	15.83
f3	52.19	66.58	76.35	7.11	1	0.43	0	99.82	0.26	41.39	15.44
f4	73.62	81.19	86.3	7.44	1.52	-0.52	0	98.97	1.02	62.58	19.6
f5	67.25	75.44	81.87	6.51	-1.12	-0.41	0	98.99	0	58.48	17.72
m1	8.72	13.8	25.67	0.82	0.89	0.61	0	99.67	2.66	2.42	4.79
m2	20.5	29.81	39.75	3	0.56	0.33	0	100	2.48	10.56	8.66
m3	16.51	26.61	37.92	0.77	0.18	0.23	0	100	0.61	8.56	8.42
m4	3.28	6.32	14.75	0.31	0.5	0.48	0	99.77	0.23	0	3.46
m5	20.32	30.87	40.11	0.02	0.5	0.58	0	99.46	8.44	1.32	6.34
MA	13.86	21.48	31.64	0.98	0.53	0.44	0	99.78	2.89	4.57	6.33
FA	61.89	72.55	80.62	7.51	1.03	-0.13	0	99.47	0.29	52.33	16.69
OA	37.88	47.02	56.13	4.25	0.78	0.16	0	99.62	1.59	28.45	11.51

PRAAT	GER20	GER10	GER05	FER20	FER10	FER05	V_UVE	UV_VE	THE	TLE	STD20
f1	69.52	69.52	69.52	0.78	0.78	0.78	66.17	0	0	69.52	3.1
f2	69.61	69.79	72.61	-3.4	-3.27	-2.04	66.04	0	0	69.61	6.29
f3	78.15	78.15	79.18	0.69	0.69	0.84	75.58	0	0	78.15	4.1
f4	84.25	84.25	85.28	-2.08	-2.08	-1.22	80.57	0	0	84.25	6.52
f5	66.08	67.25	72.51	-0.09	-0.72	-1.6	66.08	0	0	66.08	10.1
m1	74.09	74.33	75.06	-0.37	-0.53	-0.26	74.09	0	0	74.09	3.23
m2	80.12	80.75	86.34	-5.25	-4.72	-2.51	80.12	0	0	80.12	5.08
m3	87.77	87.77	88.38	-1.78	-1.78	-1.33	87.77	0	0	87.77	3.46
m4	67.21	67.21	71.66	0.43	0.43	0.23	67.21	0	0	67.21	2.99
m5	86.15	86.15	86.41	-0.19	-0.19	-0.02	86.15	0	0	86.15	2.35
MA	79.07	79.24	81.57	-1.43	-1.36	-0.78	79.07	0	0	79.07	3.42
FA	73.52	73.79	75.82	-0.82	-0.92	-0.65	70.89	0	0	73.52	6.02
OA	76.3	76.52	78.69	-1.13	-1.14	-0.71	74.98	0	0	76.3	4.72

Table A.2.7. Results for 0 dB White noise for high resolution speech signals.

DHO	GER20	GER10	GER05	FER20	FER10	FER05	V_UVE	UV_VE	THE	TLE	STD20
f1	7.81	12.64	15.61	-0.48	-0.59	-0.33	0	100	0	0	8.29
f2	8.82	13.7	18.01	0.22	0.35	0.6	0	100	0	0.75	9
f3	11.31	13.88	18.25	-0.62	-0.51	-0.12	0	93.85	0	1.29	6.14
f4	29.86	44.17	54.6	-2.3	-0.48	0.12	0	100	0	11.25	17.74
f5	13.74	17.54	23.68	0.51	1.02	0.69	0	95.13	0	5.26	9.65
m1	13.08	21.55	41.16	0.99	0.43	0.01	0	98.01	2.66	0	6.52
m2	7.45	16.77	31.68	0.19	1.66	1.75	0	100	0	0	8.38
m3	9.79	14.37	20.8	-0.7	-0.56	-0.03	0	100	0.92	0	5.34
m4	23.19	27.63	40.05	0.49	0.41	0.41	0.23	98.63	2.11	0.23	4.44
m5	27.18	39.97	49.21	-1.42	-0.07	0.34	0	97.19	2.9	0	7.24
MA	16.14	24.06	36.58	-0.09	0.38	0.49	0.05	98.77	1.72	0.05	6.38
FA	14.31	20.39	26.03	-0.54	-0.04	0.19	0	97.8	0	3.71	10.16
OA	15.22	22.22	31.31	-0.31	0.17	0.34	0.02	98.28	0.86	1.88	8.27

CORR	GER20	GER10	GER05	FER20	FER10	FER05	V_UVE	UV_VE	THE	TLE	STD20
f1	99.63	99.63	99.63	1.87	1.87	1.87	99.63	0	0	99.63	0
f2	99.81	99.81	99.81	-6.41	-6.41	-6.41	99.81	0	0	99.81	0
f3	100	100	100	0	0	0	100	0	0	100	0
f4	100	100	100	0	0	0	100	0	0	100	0
f5	98.83	98.83	98.83	-6.48	-6.48	-6.48	98.83	0	0	98.83	6.86
m1	100	100	100	0	0	0	100	0	0	100	0
m2	100	100	100	0	0	0	100	0	0	100	0
m3	100	100	100	0	0	0	100	0	0	100	0
m4	100	100	100	0	0	0	100	0.23	0	100	0
m5	100	100	100	0	0	0	100	0	0	100	0
MA	100	100	100	0	0	0	100	0.1	0	100	0
FA	100	100	100	-2	-2	-2	100	0	0	100	1.37
OA	100	100	100	-1	-1	-1	100	0	0	100	0.69

YIN	GER20	GER10	GER05	FER20	FER10	FER05	V_UVE	UV_VE	THE	TLE	STD20
f1	60.97	71	81.41	7.76	2.78	1.52	0	100	0.37	53.16	15.45
f2	71.29	79.92	86.87	10.24	1.5	-0.08	0	100	0.56	60.98	16.81
f3	60.93	72.49	83.29	7.52	0.78	0.05	0	100	1.03	48.59	15.89
f4	85.07	89.98	94.07	10.83	4.47	-0.47	0	99.23	1.64	74.03	18.2
f5	73.68	80.41	87.72	7.06	1.71	0.59	0	99.83	0	66.37	18.47
m1	19.61	29.3	43.58	0.64	0.61	0.72	0	99.83	5.57	2.42	6.42
m2	31.06	42.24	54.66	1.5	0.3	0.22	0	100	6.83	14.91	9.55
m3	29.66	44.04	59.33	1.12	0.65	-0.08	0	100	2.75	12.23	10.01
m4	16.16	23.65	36.07	-0.23	0.52	0.38	0	100	3.04	1.17	4.89
m5	37.47	52.11	63.59	-0.13	0.55	0.51	0	99.6	13.72	2.77	8.49
MA	27	38	51	0.6	0.5	0.4	0	100	6.4	6.7	7.87
FA	70	79	87	8.7	2.3	0.3	0	100	0.7	61	16.96
OA	49	59	69	4.6	1.4	0.3	0	100	3.6	34	12.42

PRAAT	GER20	GER10	GER05	FER20	FER10	FER05	V_UVE	UV_VE	THE	TLE	STD20
f1	100	100	100	0	0	0	100	0	0	100	0
f2	100	100	100	0	0	0	100	0	0	100	0
f3	100	100	100	0	0	0	100	0	0	100	0
f4	100	100	100	0	0	0	100	0	0	100	0
f5	100	100	100	0	0	0	100	0	0	100	0
m1	100	100	100	0	0	0	100	0	0	100	0
m2	100	100	100	0	0	0	100	0	0	100	0
m3	100	100	100	0	0	0	100	0	0	100	0
m4	100	100	100	0	0	0	100	0	0	100	0
m5	100	100	100	0	0	0	100	0	0	100	0
MA	100	100	100	0	0	0	100	0	0	100	0
FA	100	100	100	0	0	0	100	0	0	100	0
OA	100	100	100	0	0	0	100	0	0	100	0

Table A.2.8. Results for -5 dB White noise for high resolution speech signals.

DHO	GER20	GER10	GER05	FER20	FER10	FER05	V_UVE	UV_VE	THE	TLE	STD20
f1	0.74	1.49	4.09	-0.53	-0.34	-0.3	0	10.86	0	0	4.07
f2	0	0.19	2.25	0.06	0.02	0.16	0	43.75	0	0	4.89
f3	0	0.26	3.6	-0.16	-0.1	0.17	0	18.44	0	0	3.75
f4	2.86	3.27	6.13	0.27	0.13	0.24	0	5.67	0	0	5.48
f5	0.29	0.58	3.22	0.35	0.43	0.54	0	24.5	0	0	5.27
m1	8.72	16.95	36.56	1.01	0.46	-0.01	3.39	20.93	1.69	3.39	6.2
m2	0	1.24	18.63	1.61	1.56	1.46	0	22.25	0	0	5.71
m3	0	0.31	3.06	-0.55	-0.51	-0.36	0	14.75	0	0	2.91
m4	14.05	14.99	25.06	0.34	0.41	0.48	5.62	3.42	1.41	5.62	3.14
m5	8.84	12.8	20.84	-0.16	0.29	0.63	3.43	12.72	2.77	3.43	4.27
MA	6.3	9.3	21	0.5	0.4	0.4	2.5	15	1.2	2.5	4.44
FA	0.8	1.2	3.9	0	0	0.2	0	21	0	0	4.69
OA	3.6	5.2	12	0.2	0.2	0.3	1.2	18	0.6	1.2	4.57

CORR	GER20	GER10	GER05	FER20	FER10	FER05	V_UVE	UV_VE	THE	TLE	STD20
f1	0.74	2.97	4.46	0.88	0.91	0.93	0	24.43	0.74	0	4.59
f2	2.44	3.19	9.76	0.21	0.3	0.15	0.38	21.77	0.94	1.31	6.51
f3	4.11	5.14	9.25	0.3	0.52	0.66	1.8	37.07	1.03	3.08	5.12
f4	3.27	4.91	14.11	-0.57	-0.35	0.11	1.23	29.38	0.61	2.45	8.29
f5	0	1.17	8.77	-0.17	-0.42	-0.79	0	34.56	0	0	7.33
m1	16.95	18.16	23.24	0.08	0.2	0.34	13.32	35.38	3.63	13.32	3.3
m2	1.24	1.86	26.09	-1.48	-1.59	-0.59	0.62	29.96	0.62	0.62	5.76
m3	0.31	0.92	3.36	0.1	0.2	0.21	0.31	21.97	0	0.31	3.19
m4	16.86	18.03	27.87	0.2	0.22	0.13	16.63	31.89	0.23	16.63	3.22
m5	8.31	9.5	22.03	0.67	0.62	0.34	7.39	27.04	0.92	7.39	3.89
MA	8.7	9.7	21	-0	-0	0.1	7.7	29	1.1	7.7	3.87
FA	2.1	3.5	9.3	0.1	0.2	0.2	0.7	29	0.7	1.4	6.37
OA	5.4	6.6	15	0	0.1	0.2	4.2	29	0.9	4.5	5.12

YIN	GER20	GER10	GER05	FER20	FER10	FER05	V_UVE	UV_VE	THE	TLE	STD20
f1	7.81	10.78	15.99	0.92	0.16	0.08	0	99.32	0.37	4.83	6.87
f2	10.13	13.51	19.89	1.42	0.27	0.25	0	98.92	0	7.5	8.07
f3	5.4	6.94	10.28	0.75	0.29	0.53	0	97.11	0.26	3.86	5.37
f4	45.6	50.31	58.69	3.33	0.23	-0.04	0	98.2	0.61	41.31	13.13
f5	13.45	17.25	21.05	2.07	0.49	0.46	0	97.65	0	13.16	9.12
m1	1.45	3.87	12.83	0.54	0.71	0.53	0	99.83	0.48	0	3.61
m2	0.62	0.62	1.86	0.85	0.85	0.83	0	100	0.62	0	2.54
m3	0.92	1.53	3.36	0.26	0.13	0.01	0	98.69	0	0.61	2.71
m4	1.41	2.58	7.73	0.57	0.69	0.6	0	99.54	0.47	0.23	2.48
m5	1.45	2.77	5.8	0.41	0.44	0.45	0	97.99	0.53	0.26	2.7
MA	1.2	2.3	6.3	0.5	0.6	0.5	0	99	0.4	0.2	2.81
FA	16	20	25	1.7	0.3	0.3	0	98	0.3	14	8.51
OA	8.8	11	16	1.1	0.4	0.4	0	99	0.3	7.2	5.66

PRAAT	GER20	GER10	GER05	FER20	FER10	FER05	V_UVE	UV_VE	THE	TLE	STD20
f1	4.09	6.32	7.81	1.07	1.11	1.12	0.37	8.37	0.74	3.35	4.65
f2	2.81	3.94	11.44	0.37	0.49	0.39	0	3.66	0.38	2.44	6.75
f3	3.6	4.37	10.8	0.59	0.77	0.79	1.29	26.94	1.03	2.57	5.33
f4	4.5	6.13	14.52	-0.19	-0.17	0.24	3.27	16.49	0	4.5	8.02
f5	0.29	1.46	9.06	0.06	-0.19	-0.64	0	22.99	0	0.29	7.36
m1	8.47	10.9	19.13	0.51	0.53	0.44	7.99	6.98	0.24	7.99	3.87
m2	0.62	3.11	30.43	-1.43	-1.7	-0.4	0	4.85	0.62	0	6.42
m3	0.61	1.22	5.81	0.27	0.37	0.32	0.61	12.46	0	0.61	3.55
m4	10.54	12.88	23.65	0.69	0.59	0.37	10.54	7.97	0	10.54	3.55
m5	1.98	3.96	18.34	0.76	0.74	0.38	1.32	17.67	0.66	1.32	3.98
MA	4.4	6.4	19	0.2	0.1	0.2	4.1	10	0.3	4.1	4.27
FA	3.1	4.5	11	0.4	0.4	0.4	1	16	0.4	2.6	6.42
OA	3.8	5.4	15	0.3	0.3	0.3	2.5	13	0.4	3.4	5.35

Table A.2.9. Results for 25 dB Environment noise for high resolution speech signals.

DHO	GER20	GER10	GER05	FER20	FER10	FER05	V_UVE	UV_VE	THE	TLE	STD20
f1	0	0.37	3.35	-0.47	-0.37	-0.29	0	9.28	0	0	3.83
f2	2.06	2.44	4.32	0.14	0.14	0.26	0	43.97	0	0.38	4.94
f3	0	0.26	3.86	-0.21	-0.15	0.17	0	23.15	0	0	3.87
f4	1.23	1.64	3.89	0.16	0.05	0.13	1.23	4.9	0	1.23	5.19
f5	0.29	0.58	3.51	0.37	0.45	0.61	0	27.35	0	0	5.27
m1	8.72	18.89	36.32	1	0.43	0.01	2.91	22.09	0.24	2.91	6.48
m2	0	1.86	18.63	1.6	1.68	1.52	0	24.01	0	0	5.85
m3	0.61	0.92	3.67	-0.51	-0.47	-0.33	0.61	15.08	0	0.61	2.91
m4	14.29	15.93	27.4	0.4	0.32	0.47	4.68	2.05	1.41	4.68	3.23
m5	7.39	11.48	19.79	-0.34	0.28	0.66	3.83	14.06	1.32	3.83	4.32
MA	6.2	9.8	21	0.4	0.5	0.5	2.4	15	0.6	2.4	4.56
FA	0.7	1.1	3.8	0	0	0.2	0.3	22	0	0.3	4.62
OA	3.5	5.4	12	0.2	0.2	0.3	1.3	19	0.3	1.4	4.59

CORR	GER20	GER10	GER05	FER20	FER10	FER05	V_UVE	UV_VE	THE	TLE	STD20
f1	0.74	2.97	4.46	0.89	0.92	0.92	0	23.3	0.74	0	4.48
f2	3.38	4.13	10.69	0.17	0.25	0.1	0.56	21.34	0.94	2.25	6.54
f3	4.11	5.4	9.51	0.31	0.5	0.66	1.8	34.54	1.03	3.08	5.21
f4	5.32	6.75	15.95	-0.8	-0.5	0.04	2.66	25.77	0.82	4.29	8.17
f5	0.29	1.46	9.06	-0.18	-0.43	-0.8	0	31.88	0	0.29	7.35
m1	16.95	18.16	23.24	0.09	0.21	0.35	13.56	33.55	3.39	13.56	3.31
m2	3.11	4.35	27.95	-1.49	-1.52	-0.56	2.48	27.09	0.62	2.48	5.85
m3	1.22	1.83	4.89	0.1	0.19	0.25	1.22	23.93	0	1.22	3.2
m4	17.33	18.5	28.1	0.21	0.23	0.15	16.86	29.84	0.47	16.86	3.24
m5	9.37	10.55	22.96	0.68	0.63	0.31	8.71	27.31	0.66	8.71	3.88
MA	9.6	11	21	-0	-0	0.1	8.6	28	1	8.6	3.9
FA	2.8	4.1	9.9	0.1	0.2	0.2	1	27	0.7	2	6.35
OA	6.2	7.4	16	0	0.1	0.1	4.8	28	0.9	5.3	5.12

YIN	GER20	GER10	GER05	FER20	FER10	FER05	V_UVE	UV_VE	THE	TLE	STD20
f1	9.67	14.13	17.47	1.26	0.01	-0.01	0	99.55	0.74	7.06	7.65
f2	11.63	15.57	22.89	1.77	0.46	0.23	0	98.92	0	9.01	8.46
f3	6.94	9	12.85	0.75	0.22	0.34	0	97.29	0.26	5.14	5.87
f4	52.56	57.67	65.44	3.12	-0.18	-0.15	0	98.2	0.41	46.01	13.9
f5	15.2	19.3	23.68	2.14	0.51	0.49	0	97.48	0	14.62	9.15
m1	1.69	4.12	13.08	0.55	0.72	0.54	0	99.83	0.73	0	3.59
m2	0.62	1.86	3.73	0.78	0.93	0.86	0	99.78	0.62	0	2.9
m3	1.22	2.14	4.28	0.41	0.22	0.08	0	98.69	0	1.22	3
m4	1.64	3.04	7.73	0.56	0.71	0.6	0	99.54	0.94	0.23	2.52
m5	1.85	3.3	7.26	0.32	0.41	0.44	0	97.86	0.53	0.26	2.89
MA	1.4	2.9	7.2	0.5	0.6	0.5	0	99	0.6	0.3	2.98
FA	19	23	28	1.8	0.2	0.2	0	98	0.3	16	9.01
OA	10	13	18	1.2	0.4	0.3	0	99	0.4	8.4	5.99

PRAAT	GER20	GER10	GER05	FER20	FER10	FER05	V_UVE	UV_VE	THE	TLE	STD20
f1	4.09	6.32	7.81	1.1	1.14	1.14	0.37	6.33	0.74	3.35	4.6
f2	3.94	4.88	12.2	0.49	0.54	0.35	0.38	5.6	0.38	3.56	6.64
f3	3.6	4.63	10.54	0.57	0.71	0.79	1.29	29.48	1.03	2.57	5.38
f4	5.52	7.36	15.75	-0.36	-0.15	0.26	4.29	14.43	0	5.52	8.01
f5	0.29	1.46	9.06	0.08	-0.17	-0.62	0	22.65	0	0.29	7.37
m1	8.23	10.65	18.89	0.5	0.52	0.43	7.75	9.14	0.24	7.75	3.87
m2	0.62	3.73	31.68	-1.38	-1.58	-0.47	0	8.15	0.62	0	6.5
m3	0.61	1.53	6.12	0.32	0.38	0.32	0.61	11.8	0	0.61	3.64
m4	10.3	12.65	24.12	0.68	0.58	0.33	10.3	7.06	0	10.3	3.56
m5	3.56	5.15	19.26	0.76	0.74	0.38	3.43	18.61	0.13	3.43	3.97
MA	4.7	6.7	20	0.2	0.1	0.2	4.4	11	0.2	4.4	4.31
FA	3.5	4.9	11	0.4	0.4	0.4	1.3	16	0.4	3.1	6.4
OA	4.1	5.8	16	0.3	0.3	0.3	2.8	13	0.3	3.7	5.35

Table A.2.10. Results for 20 dB Environment noise for high resolution speech signals.

DHO	GER20	GER10	GER05	FER20	FER10	FER05	V_UVE	UV_VE	THE	TLE	STD20
f1	0.74	1.12	4.09	-0.45	-0.35	-0.28	0	16.97	0	0	3.82
f2	6.57	6.57	8.44	0.2	0.2	0.43	0	49.57	0	6	4.74
f3	0	0.77	4.63	-0.38	-0.18	0.21	0	26.58	0	0	4.3
f4	5.93	6.75	9.41	0.45	0.35	0.34	0	12.89	0	0	6.06
f5	0	0.29	3.51	0.36	0.43	0.62	0	34.73	0	0	5.25
m1	7.02	16.46	36.08	0.77	0.4	-0.02	0.24	21.43	1.45	0.24	6.24
m2	0	1.86	17.39	1.6	1.68	1.54	0	28.85	0	0	5.8
m3	0.61	0.92	3.98	-0.45	-0.41	-0.34	0.61	22.95	0	0.61	3.05
m4	18.74	20.14	31.15	0.51	0.46	0.57	4.92	7.06	0.23	4.92	3.32
m5	14.38	19.13	27.44	-0.44	0.28	0.63	6.07	12.45	2.9	6.07	4.53
MA	8.2	12	23	0.4	0.5	0.5	2.4	19	0.9	2.4	4.59
FA	2.7	3.1	6	0	0.1	0.3	0	28	0	1.2	4.83
OA	5.4	7.4	15	0.2	0.3	0.4	1.2	23	0.5	1.8	4.71

CORR	GER20	GER10	GER05	FER20	FER10	FER05	V_UVE	UV_VE	THE	TLE	STD20
f1	2.6	4.09	5.95	1.09	0.94	0.92	1.12	22.4	0.37	2.23	4.16
f2	4.69	5.44	11.82	0.1	0.18	-0.01	1.31	20.47	0.94	3.56	6.61
f3	4.88	6.17	10.8	0.37	0.57	0.6	2.06	31.46	1.03	3.86	5.37
f4	9	11.25	18.61	-1	-0.57	-0.14	5.73	26.03	0.61	8.18	8.33
f5	1.17	2.34	10.23	-0.18	-0.42	-0.84	0	30.37	0	1.17	7.39
m1	17.68	18.89	23.73	0.08	0.21	0.37	13.56	30.9	4.12	13.56	3.31
m2	8.7	9.94	32.3	-1.66	-1.67	-0.67	5.59	22.69	3.11	5.59	5.93
m3	3.36	3.98	6.42	0.15	0.25	0.26	3.36	24.26	0	3.36	3.25
m4	18.03	18.97	28.57	0.2	0.25	0.16	17.56	24.83	0.47	17.56	3.18
m5	12.8	13.85	24.67	0.67	0.61	0.28	12.27	27.31	0.53	12.27	3.86
MA	12	13	23	-0	-0	0.1	10	26	1.6	10	3.91
FA	4.5	5.9	11	0.1	0.1	0.1	2	26	0.6	3.8	6.37
OA	8.3	9.5	17	-0	0	0.1	6.3	26	1.1	7.1	5.14

YIN	GER20	GER10	GER05	FER20	FER10	FER05	V_UVE	UV_VE	THE	TLE	STD20
f1	13.75	18.59	23.79	1.48	0.04	-0.11	0	99.77	0.74	10.41	8.31
f2	16.14	20.08	27.58	1.72	0.54	0.29	0	99.57	0	11.82	8.97
f3	9.77	12.85	16.97	1	0.46	0.37	0	98.73	0.51	7.71	6.59
f4	58.9	64.42	74.23	4.46	0.12	-0.38	0	98.45	0.41	53.17	15.58
f5	19.3	25.73	30.41	3.26	0.65	0.38	0	98.83	0	17.54	10.69
m1	2.91	5.08	15.5	0.54	0.67	0.57	0	100	0.97	0	3.75
m2	1.86	3.11	7.45	1.15	0.96	0.76	0	99.56	0.62	0	3.47
m3	3.06	4.89	10.09	0.53	0.16	0.01	0	100	0	1.53	3.93
m4	1.87	3.28	7.96	0.58	0.73	0.61	0	99.09	0.94	0.23	2.52
m5	3.43	6.2	10.82	0.21	0.39	0.47	0	98.8	1.58	0.53	3.43
MA	2.6	4.5	10	0.6	0.6	0.5	0	99	0.8	0.5	3.42
FA	24	28	35	2.4	0.4	0.1	0	99	0.3	20	10.03
OA	13	16	22	1.5	0.5	0.3	0	99	0.6	10	6.72

PRAAT	GER20	GER10	GER05	FER20	FER10	FER05	V_UVE	UV_VE	THE	TLE	STD20
f1	4.46	6.69	8.18	1.24	1.27	1.17	0.74	4.3	0.74	3.72	4.55
f2	6.57	7.5	14.82	0.35	0.4	0.24	1.5	4.31	0	6.57	6.67
f3	4.88	5.91	12.08	0.63	0.77	0.74	0.77	25.68	1.03	3.86	5.48
f4	10.84	13.09	20.65	-0.38	-0.22	0.13	7.36	15.21	0	10.84	8.24
f5	2.63	3.8	11.4	0.07	-0.19	-0.65	0	21.48	0	2.63	7.45
m1	9.93	11.62	20.1	0.41	0.53	0.43	9.69	8.97	0.24	9.69	3.72
m2	1.86	4.97	33.54	-1.27	-1.63	-0.45	1.24	9.69	0.62	1.24	6.76
m3	1.53	2.14	7.34	0.31	0.41	0.36	1.53	6.56	0	1.53	3.6
m4	12.18	14.05	24.82	0.65	0.56	0.35	12.18	6.15	0	12.18	3.46
m5	5.41	6.99	20.71	0.74	0.71	0.38	5.28	19.54	0.13	5.28	3.97
MA	6.2	8	21	0.2	0.1	0.2	6	10	0.2	6	4.3
FA	5.9	7.4	13	0.4	0.4	0.3	2.1	14	0.4	5.5	6.48
OA	6	7.7	17	0.3	0.3	0.3	4	12	0.3	5.8	5.39

Table A.2.11. Results for 15 dB Environment noise for high resolution speech signals.

DHO	GER20	GER10	GER05	FER20	FER10	FER05	V_UVE	UV_VE	THE	TLE	STD20
f1	0	1.12	3.72	-0.72	-0.36	-0.32	0	58.14	0	0	4.86
f2	0.38	0.56	2.81	0.05	0.1	0.23	0	73.28	0	0	4.94
f3	0.26	0.51	4.11	-0.17	-0.11	0.26	0	43.76	0	0	3.87
f4	11.45	14.31	18.61	0.76	0.3	0.2	1.64	25.52	0	6.75	9.23
f5	6.14	6.43	8.77	0.63	0.5	0.55	0	62.25	0	4.97	5.34
m1	11.38	20.82	40.44	0.65	0.47	-0.01	1.21	35.22	0.24	1.21	6.52
m2	4.35	6.21	19.88	1.33	1.41	1.27	0	55.51	0	0	5.68
m3	1.53	2.14	7.03	-0.51	-0.54	-0.34	0.61	46.56	0	0.61	3.51
m4	24.36	26.7	35.36	0.56	0.44	0.42	2.34	29.84	0	2.34	3.57
m5	8.31	12.53	20.98	-0.33	0.29	0.64	3.83	27.71	2.37	3.83	4.27
MA	10	14	25	0.3	0.4	0.4	1.6	39	0.5	1.6	4.71
FA	3.6	4.6	7.6	0.1	0.1	0.2	0.3	53	0	2.3	5.65
OA	6.8	9.1	16	0.2	0.3	0.3	1	46	0.3	2	5.18

CORR	GER20	GER10	GER05	FER20	FER10	FER05	V_UVE	UV_VE	THE	TLE	STD20
f1	6.69	8.18	9.29	1.39	1.21	1.05	3.35	21.72	0.37	6.32	4.07
f2	9.38	10.13	17.07	-0.16	-0.07	-0.11	4.13	20.26	0.94	8.44	6.66
f3	7.46	9	14.14	0.57	0.72	0.6	3.34	32.91	1.03	6.43	5.56
f4	19.63	21.68	28.43	-0.95	-0.7	-0.32	11.45	27.84	1.43	17.79	8.83
f5	2.92	4.09	13.16	-0.02	-0.27	-0.76	0.88	32.55	0	2.92	7.58
m1	22.76	23.49	28.09	0.05	0.11	0.3	18.16	29.9	4.6	18.16	3.22
m2	18.63	19.88	39.75	-2.18	-2.44	-1.14	14.91	19.6	3.73	14.91	5.67
m3	8.26	9.17	13.15	0.39	0.44	0.38	8.26	21.97	0	8.26	3.66
m4	22.25	23.19	32.32	0.22	0.28	0.16	21.08	21.18	1.17	21.08	3.14
m5	20.45	21.77	30.74	0.54	0.48	0.25	19.26	26.51	1.06	19.26	3.88
MA	18	20	29	-0	-0	-0	16	24	2.1	16	3.91
FA	9.2	11	16	0.2	0.2	0.1	4.6	27	0.8	8.4	6.54
OA	14	15	23	-0	-0	0	10	25	1.4	12	5.23

YIN	GER20	GER10	GER05	FER20	FER10	FER05	V_UVE	UV_VE	THE	TLE	STD20
f1	23.79	28.62	37.92	2.3	0.4	-0.22	0	100	0.74	19.7	9.56
f2	25.14	32.65	38.84	2.71	0.32	0.07	0	100	0	20.83	11.71
f3	16.71	19.79	29.31	1.75	0.94	0.14	0	99.1	0.26	13.11	7.55
f4	66.46	73.01	82.41	5.47	0.57	-0.24	0	97.94	0.82	60.33	17.44
f5	27.49	34.5	42.11	3.62	0.57	0.16	0	99.16	0	24.85	11.85
m1	5.33	7.99	18.16	0.54	0.66	0.56	0	100	2.18	0	3.98
m2	5.59	8.07	15.53	1.75	1.34	0.81	0	100	1.86	1.24	4.65
m3	2.45	4.89	11.93	0.86	0.47	0.09	0	99.67	0	1.83	4.52
m4	1.87	3.51	9.6	0.58	0.66	0.63	0	99.32	0.94	0	2.74
m5	6.46	9.76	15.17	0.14	0.39	0.47	0	99.33	2.37	0.26	3.6
MA	4.3	6.9	14	0.8	0.7	0.5	0	100	1.5	0.7	3.9
FA	32	38	46	3.2	0.6	-0	0	99	0.4	28	11.62
OA	18	22	30	2	0.6	0.3	0	99	0.9	14	7.76

PRAAT	GER20	GER10	GER05	FER20	FER10	FER05	V_UVE	UV_VE	THE	TLE	STD20
f1	14.13	15.24	16.36	1.76	1.48	1.3	4.46	4.3	0	14.13	3.98
f2	16.14	16.89	23.26	0.01	0.03	0.18	4.69	3.02	0	16.14	6.56
f3	9.25	10.54	16.2	0.79	0.75	0.75	2.57	21.34	1.03	8.23	5.63
f4	29.45	30.88	37.01	-0.48	-0.47	-0.2	14.11	12.37	0	29.45	8.26
f5	10.53	11.7	20.47	0.14	-0.15	-0.58	0	16.44	0	10.53	7.79
m1	10.17	11.86	20.34	0.4	0.52	0.42	9.93	9.8	0.24	9.93	3.74
m2	16.77	18.63	40.37	-1.94	-2.41	-0.72	16.77	9.25	0	16.77	6.48
m3	6.12	7.03	13.76	0.51	0.56	0.52	5.81	4.92	0	5.81	3.98
m4	12.88	14.05	24.12	0.53	0.49	0.32	12.88	4.78	0	12.88	3.21
m5	8.71	10.55	22.69	0.78	0.72	0.34	8.58	19.28	0.13	8.58	4.03
MA	11	12	24	0.1	-0	0.2	11	9.6	0.1	11	4.29
FA	16	17	23	0.4	0.3	0.3	5.2	11	0.2	16	6.44
OA	13	15	23	0.3	0.2	0.2	8	11	0.1	13	5.36

Table A.2.12. Results for 10 dB Environment noise for high resolution speech signals.

DHO	GER20	GER10	GER05	FER20	FER10	FER05	V_UVE	UV_VE	THE	TLE	STD20
f1	1.49	3.72	5.58	-0.55	-0.25	-0.23	0	89.37	0	0.74	5.46
f2	0.94	2.44	5.25	0.24	0.15	0.37	0	83.41	0	0	5.82
f3	4.63	8.23	12.34	-0.83	-0.3	0.15	0	74.86	0	0	6.6
f4	15.54	24.13	30.27	0.03	0.33	0.38	2.04	71.65	0	3.68	13.54
f5	5.56	7.31	9.65	0.34	0.49	0.57	0	79.36	0	0	7.24
m1	9.69	21.07	37.29	0.51	0.47	0.06	0.97	63.46	0.24	0.97	6.4
m2	2.48	8.07	22.36	1.6	1.49	1.37	0	71.81	0	0	6.66
m3	0.92	2.45	10.7	-0.29	-0.49	-0.34	0	81.31	0	0	4.39
m4	21.08	24.36	33.96	0.36	0.45	0.43	3.28	59	4.22	3.28	3.73
m5	12.27	19.39	31.4	-0.94	0	0.45	1.45	63.05	0.79	1.45	5.41
MA	9.3	15	27	0.3	0.4	0.4	1.1	68	1.1	1.1	5.32
FA	5.6	9.2	13	-0	0.1	0.3	0.4	80	0	0.9	7.73
OA	7.5	12	20	0.1	0.2	0.3	0.8	74	0.5	1	6.53

CORR	GER20	GER10	GER05	FER20	FER10	FER05	V_UVE	UV_VE	THE	TLE	STD20
f1	13.75	14.5	17.84	1.65	1.45	1.14	5.58	21.04	0	13.38	4.52
f2	18.2	19.7	27.02	-0.04	-0.11	-0.25	9.19	19.4	0.75	17.26	7.45
f3	16.97	18.25	23.91	0.56	0.65	0.56	12.85	33.63	0.51	16.45	5.67
f4	42.13	43.35	48.88	-0.48	-0.88	-0.73	30.88	24.48	1.64	40.29	8.95
f5	7.89	8.77	19.01	0.28	-0.04	-0.56	3.22	33.89	0	7.89	8.02
m1	34.14	34.62	38.5	-0.01	-0.01	0.24	29.54	26.08	4.36	29.78	3.19
m2	26.09	27.95	45.96	-2.38	-2.89	-1.35	21.12	19.16	3.73	21.74	6.07
m3	22.94	22.94	29.05	0.33	0.33	0.33	21.1	18.69	1.22	21.1	3.69
m4	30.91	31.85	40.05	-0.06	0.13	0.13	29.51	18.91	1.17	29.51	3.33
m5	39.58	40.9	46.31	0.17	0.09	0.25	36.02	27.44	3.03	36.02	3.87
MA	31	32	40	-0	-0	-0	27	22	2.7	28	4.03
FA	20	21	27	0.4	0.2	0	12	26	0.6	19	6.92
OA	25	26	34	0	-0	-0	20	24	1.6	23	5.48

YIN	GER20	GER10	GER05	FER20	FER10	FER05	V_UVE	UV_VE	THE	TLE	STD20
f1	39.78	46.1	55.39	3.8	1	0.26	0	100	0	33.46	10.48
f2	40.9	49.16	57.79	4.31	0.7	-0.42	0	99.78	0	34.9	13.36
f3	30.33	39.07	48.84	2.61	0.29	0.2	0	99.64	0	22.11	10.68
f4	71.57	81.8	88.34	6.89	-1.05	-1.59	0	98.2	1.02	61.55	21.64
f5	47.95	56.43	64.91	6.71	2.11	0.69	0	99.5	0	42.4	14.46
m1	7.51	12.35	24.7	0.58	0.54	0.55	0	99.83	3.15	0.24	4.81
m2	14.29	19.25	23.6	1.38	0.85	0.5	0	100	3.73	4.35	5.75
m3	9.17	17.13	26.61	0.94	0.01	-0.14	0	99.67	0	5.2	6.68
m4	5.15	8.2	14.99	0.39	0.57	0.57	0	100	0.7	0	3.5
m5	13.06	19.79	29.68	0.23	0.34	0.38	0	99.73	5.8	0.26	5.43
MA	9.8	15	24	0.7	0.5	0.4	0	100	2.7	2	5.23
FA	46	55	63	4.9	0.6	-0	0	99	0.2	39	14.12
OA	28	35	43	2.8	0.5	0.1	0	100	1.4	20	9.68

PRAAT	GER20	GER10	GER05	FER20	FER10	FER05	V_UVE	UV_VE	THE	TLE	STD20
f1	17.47	18.59	21.93	2.04	1.7	1.38	5.58	5.43	0	17.47	5.03
f2	31.52	32.65	38.65	-0.27	-0.29	-0.18	9.57	2.59	0	31.52	7.18
f3	17.48	18.51	24.16	0.93	0.94	0.81	10.54	16.46	1.03	16.45	5.58
f4	60.94	61.96	66.46	-1.67	-1.97	-1.27	39.26	9.54	0	60.94	8.92
f5	19.3	20.18	30.41	0.27	-0.11	-0.86	2.34	15.94	0	19.3	8.33
m1	15.98	16.95	25.18	0.37	0.39	0.33	15.74	9.97	0.24	15.74	3.69
m2	27.33	28.57	49.69	-2.54	-2.86	-1.03	26.09	10.13	0	27.33	6.1
m3	19.88	20.49	28.44	0.76	0.62	0.45	19.57	7.87	0	19.57	4.32
m4	19.2	20.84	31.85	0.48	0.55	0.28	19.2	7.97	0	19.2	3.57
m5	26.91	28.89	37.47	0.47	0.39	0.27	26.91	13.39	0	26.91	4.04
MA	22	23	35	-0	-0	0.1	22	9.9	0.1	22	4.34
FA	29	30	36	0.3	0.1	-0	13	10	0.2	29	7.01
OA	26	27	35	0.1	-0	0	17	9.9	0.1	25	5.67

Table A.2.13. Results for 5 dB Environment noise for high resolution speech signals.

DHO	GER20	GER10	GER05	FER20	FER10	FER05	V_UVE	UV_VE	THE	TLE	STD20
f1	1.12	4.83	7.06	-0.44	0.05	-0.1	0	96.83	0	0	6.96
f2	4.88	11.63	18.2	1.7	0.6	0.74	0	97.84	0	0	10.54
f3	9.25	14.91	20.31	-0.95	0.01	0.28	0	92.77	0	0	9.08
f4	27.4	44.58	52.35	0.34	0.32	0.5	0	96.39	0	0.41	20.29
f5	7.31	10.53	13.74	0.37	0.29	0.3	0	96.31	0	5.26	8.11
m1	20.58	32.69	46.73	0.56	0.59	0.16	0	85.38	0.24	0	6.89
m2	6.21	11.8	26.09	0.81	1.43	1.43	0	91.85	0.62	0	6.4
m3	9.17	12.54	26.91	0.05	-0.3	-0.33	0	96.39	0.31	0	5.54
m4	26.46	33.26	44.5	0.18	0.55	0.6	1.64	88.61	4.68	1.64	4.71
m5	19.53	32.06	44.59	-1.28	-0.06	0.29	2.24	87.28	2.24	2.24	6.81
MA	16	24	38	0.1	0.4	0.4	0.8	90	1.6	0.8	6.07
FA	10	17	22	0.2	0.3	0.3	0	96	0	1.1	11
OA	13	21	30	0.1	0.4	0.4	0.4	93	0.8	1	8.53

CORR	GER20	GER10	GER05	FER20	FER10	FER05	V_UVE	UV_VE	THE	TLE	STD20
f1	32.71	33.09	37.17	1.76	1.65	1.01	23.79	21.27	0	32.34	4.54
f2	34.9	36.77	43.15	-0.73	-0.7	-0.47	22.89	18.32	0.75	33.96	8.13
f3	32.39	34.45	39.59	0.7	1.04	0.75	27.25	31.83	0.77	31.36	6.3
f4	61.96	63.8	67.28	-0.93	-2.36	-2.07	49.08	24.23	1.43	59.71	10.78
f5	22.81	23.98	37.13	0.54	0.06	-0.57	14.33	35.07	0	22.81	9.46
m1	53.75	54.24	57.38	-0.1	-0.25	0.07	48.67	23.09	4.84	48.91	3.31
m2	39.75	40.37	56.52	-3.02	-3.27	-1.33	33.54	19.16	2.48	36.02	5.81
m3	51.68	52.6	55.66	0.56	0.14	0.15	46.18	16.72	2.75	46.48	4.69
m4	51.29	51.76	58.08	-0.24	-0.09	0.04	48.95	19.13	1.87	48.95	3.45
m5	62.53	63.19	65.83	-0.07	-0.25	0.08	54.88	26.37	6.6	54.88	3.72
MA	52	52	59	-1	-1	-0	46	21	3.7	47	4.2
FA	37	38	45	0.3	-0	-0	27	26	0.6	36	7.84
OA	44	45	52	-0	-0	-0	37	24	2.2	42	6.02

YIN	GER20	GER10	GER05	FER20	FER10	FER05	V_UVE	UV_VE	THE	TLE	STD20
f1	48.7	55.76	66.54	4.5	1.14	0.29	0	100	0	42.75	11.86
f2	57.79	68.11	76.74	5.41	-0.11	-0.38	0	99.35	0.56	49.72	16.73
f3	45.5	54.76	63.24	4.32	0.23	0.01	0	100	0.77	34.45	13.02
f4	72.39	83.23	88.96	9.2	-0.16	-0.46	0	99.23	1.02	61.15	21.64
f5	59.94	69.3	80.41	8.37	1.92	0.93	0	100	0	51.75	17.34
m1	18.64	28.81	39.95	0.16	0.51	0.55	0	100	4.6	1.69	6.71
m2	21.74	32.3	43.48	1	-0.09	-0.45	0	100	5.59	9.32	8.8
m3	25.69	39.76	51.99	0.79	0.32	-0.11	0	99.34	3.67	11.62	9.8
m4	9.6	15.93	26.46	0.07	0.53	0.38	0	100	2.34	0	4.57
m5	29.68	43.27	56.2	0.06	0.33	0.37	0	99.87	10.69	1.72	7.71
MA	21	32	44	0.4	0.3	0.2	0	100	5.4	4.9	7.52
FA	57	66	75	6.4	0.6	0.1	0	100	0.5	48	16.12
OA	39	49	59	3.4	0.5	0.1	0	100	2.9	26	11.82

PRAAT	GER20	GER10	GER05	FER20	FER10	FER05	V_UVE	UV_VE	THE	TLE	STD20
f1	40.52	40.89	43.87	1.53	1.4	0.97	25.65	6.56	0	40.52	4.25
f2	47.28	48.59	54.78	-0.11	0.02	-0.12	24.58	3.66	0	47.28	8.4
f3	41.13	42.42	47.3	1.12	1.46	0.97	30.59	9.22	0.51	40.36	6.41
f4	70.55	71.98	74.44	-1.15	-2.3	-2.22	56.85	7.47	0	70.35	10.1
f5	39.18	40.64	50.58	0.3	-0.5	-1.17	17.25	10.07	0	39.18	9.93
m1	34.38	35.11	42.13	0.13	0.09	0.1	33.66	8.14	0	34.38	3.76
m2	40.37	41.61	57.14	-3.77	-3.88	-1.52	38.51	10.79	0	39.75	5.37
m3	52.29	52.91	57.8	0.72	0.42	0.06	49.54	10.49	0.92	49.54	4.5
m4	32.55	33.26	41.92	0.56	0.52	0.31	32.08	7.97	0	32.08	3.27
m5	53.69	54.49	58.58	0.3	0.1	0.24	52.77	7.76	0.79	52.77	3.77
MA	43	43	52	-0	-1	-0	41	9	0.3	42	4.14
FA	48	49	54	0.3	0	-0	31	7.4	0.1	48	7.82
OA	45	46	53	-0	-0	-0	36	8.2	0.2	45	5.98

Table A.2.14. Results for 0 dB Environment noise for high resolution speech signals.

DHO	GER20	GER10	GER05	FER20	FER10	FER05	V_UVE	UV_VE	THE	TLE	STD20
f1	18.59	21.93	27.51	-0.93	-0.12	0.02	0	98.42	0	0	7.7
f2	12.76	19.89	29.08	0.55	0.7	0.86	0	99.78	0	0.75	12.27
f3	26.48	37.28	43.7	-1.65	-0.14	0.29	0	100	0	0	12.08
f4	32.31	57.06	70.76	2.53	1.15	1.13	0	98.97	0	3.07	23.66
f5	14.04	25.15	35.67	2.69	0.72	1.42	0	98.99	0	3.22	15.14
m1	32.45	46.73	61.99	0.42	0.71	-0.03	0	98.01	2.18	0	8.13
m2	17.39	32.3	45.34	-0.05	0.97	0.91	0	98.9	0.62	0	9.87
m3	19.27	32.72	47.09	2.55	-0.21	-0.5	0	100	0.61	0	9.88
m4	31.85	40.75	53.63	-0.24	0.33	0.29	0	98.86	4.92	0	5.65
m5	26.52	40.11	57.78	-1.38	-0.21	0.31	0.79	98.53	3.17	0.79	7.75
MA	25	39	53	0.3	0.3	0.2	0.2	99	2.3	0.2	8.26
FA	21	32	41	0.6	0.5	0.7	0	99	0	1.4	14.17
OA	23	35	47	0.5	0.4	0.5	0.1	99	1.2	0.8	11.21

CORR	GER20	GER10	GER05	FER20	FER10	FER05	V_UVE	UV_VE	THE	TLE	STD20
f1	52.04	53.53	56.51	1.79	1.22	0.84	44.98	22.62	0.74	50.19	7.02
f2	59.1	60.98	66.6	-0.67	-1.3	-0.79	48.22	18.75	0.38	57.6	10.5
f3	57.33	58.61	62.47	0.3	1.03	0.6	50.9	26.76	1.03	55.01	7.17
f4	74.85	77.91	81.39	0.42	-3	-2.62	60.74	21.91	0.61	71.98	14.67
f5	52.34	55.56	64.04	1.14	-0.73	-1.25	43.57	31.54	0.88	51.46	11.09
m1	81.11	81.36	82.81	-0.82	-1	-0.57	76.03	19.93	4.36	76.03	3.68
m2	62.11	62.73	75.16	-4.38	-4.81	-2.06	54.66	19.82	1.86	56.52	6.09
m3	74.92	76.15	77.98	0.76	-0.37	-0.46	65.75	17.05	3.36	66.67	6.31
m4	76.35	77.05	80.33	-0.84	-0.4	-0.36	72.6	19.82	2.34	72.6	4.27
m5	82.59	83.77	85.22	-0.08	-0.44	-0.08	72.3	24.63	8.18	72.3	5.7
MA	75	76	80	-1	-1	-1	68	20	4	69	5.21
FA	59	61	66	0.6	-1	-1	50	24	0.7	57	10.09
OA	67	69	73	-0	-1	-1	59	22	2.4	63	7.65

YIN	GER20	GER10	GER05	FER20	FER10	FER05	V_UVE	UV_VE	THE	TLE	STD20
f1	51.67	60.22	69.14	4.34	0.96	0.35	0	100	0	44.24	14.03
f2	65.67	75.8	83.49	4.07	0.04	0.07	0	100	1.13	54.22	17.76
f3	57.33	69.41	82.52	5.82	1.19	0.72	0	100	2.31	44.22	15.78
f4	72.39	82.62	89.98	12.39	3.07	1.07	0	100	0.82	59.92	21.05
f5	70.47	79.53	88.3	8.09	3.18	1.26	0	100	0	62.28	18.84
m1	33.17	47.94	62.23	-0.05	0.98	0.63	0	100	10.41	3.15	8.06
m2	41.61	57.76	68.94	1.71	0.88	-0.47	0	100	8.07	11.8	11.88
m3	45.26	62.69	74.92	1.44	-0.65	-0.62	0	100	9.17	13.76	12.08
m4	27.87	37.47	53.16	0.02	0.55	0.38	0	100	9.84	1.41	6.28
m5	45.25	59.37	72.96	0.4	0.26	0.39	0	100	18.47	3.56	9.24
MA	39	53	66	0.7	0.4	0.1	0	100	11	6.7	9.51
FA	64	74	83	6.9	1.7	0.7	0	100	0.9	53	17.49
OA	51	63	75	3.8	1.1	0.4	0	100	6	30	13.5

PRAAT	GER20	GER10	GER05	FER20	FER10	FER05	V_UVE	UV_VE	THE	TLE	STD20
f1	58.36	58.74	59.85	1.18	0.83	0.96	51.67	5.66	0	57.99	5.13
f2	68.29	69.04	72.8	-2.26	-2.25	-1.65	57.97	5.17	0	68.29	8.65
f3	70.95	71.47	72.49	0.65	1.09	1.17	64.27	7.41	0	70.95	5.42
f4	80.78	83.03	85.48	-0.02	-2.89	-2.51	70.96	3.87	0	80.37	13.92
f5	66.08	67.54	73.1	-1.37	-2.11	-1.81	50.88	7.05	0	66.08	9.69
m1	67.8	68.28	69.73	-0.12	-0.15	-0.04	67.55	8.31	0	67.55	3.55
m2	60.87	60.87	75.16	-4.98	-4.98	-2.2	59.01	10.13	0	60.87	5.34
m3	75.23	76.76	78.29	1.26	-0.2	-0.66	70.34	11.48	0.92	71.25	6.78
m4	59.48	60.19	67.45	-0.35	-0.14	0.01	58.31	8.2	0.47	58.31	3.79
m5	80.34	81.13	82.32	0.48	-0.07	0.18	76.91	6.29	2.9	76.91	4.76
MA	69	69	75	-1	-1	-1	66	8.9	0.9	67	4.84
FA	69	70	73	-0	-1	-1	59	5.8	0	69	8.56
OA	69	70	74	-1	-1	-1	63	7.4	0.4	68	6.7

Table A.2.15. Results for -5 dB Environment noise for high resolution speech signals.

DHO	GER20	GER10	GER05	FER20	FER10	FER05	V_UVE	UV_VE	THE	TLE	STD20
f1	0	0.74	3.72	-0.56	-0.36	-0.27	0	8.6	0	0	4.16
f2	0	0.38	3	0.16	0.04	0.25	0	45.69	0	0	5.19
f3	0	0.26	3.6	-0.21	-0.16	0.15	0	23.33	0	0	3.68
f4	0	0.2	2.86	0.24	0.18	0.25	0	4.9	0	0	5.1
f5	0	0.29	2.63	0.34	0.42	0.5	0	29.03	0	0	5.33
m1	7.02	15.5	36.32	0.78	0.46	0	2.66	18.6	0.48	2.66	6.16
m2	0	1.24	19.25	1.62	1.57	1.27	0	24.67	0	0	5.76
m3	0	0.31	2.45	-0.52	-0.48	-0.38	0	13.77	0	0	2.86
m4	25.29	27.4	36.77	0.34	0.26	0.52	4.22	4.56	2.58	4.22	3.43
m5	6.2	10.16	18.87	-0.15	0.38	0.67	2.24	12.99	0.53	2.24	4.11
MA	7.7	10.92	22.73	0.41	0.44	0.42	1.82	14.92	0.72	1.82	4.46
FA	0	0.37	3.16	-0.01	0.03	0.18	0	22.31	0	0	4.69
OA	3.85	5.65	12.95	0.2	0.23	0.3	0.91	18.61	0.36	0.91	4.58

CORR	GER20	GER10	GER05	FER20	FER10	FER05	V_UVE	UV_VE	THE	TLE	STD20
f1	0.74	2.97	5.2	0.7	0.79	0.94	0	52.49	0.74	0	5.15
f2	1.69	2.81	11.07	0	0.09	0.1	0.19	37.72	0.94	0.56	6.85
f3	4.37	4.88	9.51	0.38	0.49	0.67	1.8	63.29	1.03	3.34	4.97
f4	2.25	5.11	13.09	-0.93	-0.43	-0.05	0.82	40.46	0.61	1.43	8.86
f5	0	1.17	9.06	-0.31	-0.55	-0.99	0	49.16	0	0	7.3
m1	15.74	16.95	22.03	0.12	0.24	0.33	13.08	59.14	2.66	13.08	3.34
m2	1.24	3.11	25.47	-1.39	-1.33	-0.62	0.62	54.63	0.62	0.62	5.85
m3	0.61	1.22	3.67	0.01	0.1	0.12	0.61	53.11	0	0.61	3.3
m4	16.86	18.03	27.87	0.22	0.24	0.16	16.63	47.15	0.23	16.63	3.2
m5	7.92	8.84	22.16	0.66	0.62	0.33	6.46	47.39	1.45	6.46	3.9
MA	8.47	9.63	20.24	-0.08	-0.02	0.06	7.48	52.28	0.99	7.48	3.92
FA	1.81	3.39	9.59	-0.03	0.08	0.13	0.56	48.62	0.66	1.07	6.63
OA	5.14	6.51	14.91	-0.05	0.03	0.1	4.02	50.45	0.83	4.27	5.27

YIN	GER20	GER10	GER05	FER20	FER10	FER05	V_UVE	UV_VE	THE	TLE	STD20
f1	7.81	12.27	15.99	1.07	0.08	0.14	0	99.1	0.37	4.83	7.47
f2	9.76	12.38	19.14	1.24	0.44	0.14	0	98.71	0	6.75	7.85
f3	5.91	7.97	12.08	0.75	0.28	0.57	0	97.83	0.26	3.86	5.74
f4	45.6	52.97	61.76	3.89	-0.26	0.06	0	97.94	0.61	42.33	15.22
f5	13.45	16.96	21.93	2.01	0.65	0.34	0	97.65	0	12.87	8.72
m1	0.97	2.91	13.08	0.5	0.61	0.52	0	99.34	0	0	3.55
m2	0.62	0.62	2.48	0.85	0.85	0.88	0	99.78	0.62	0	2.62
m3	0.61	1.53	3.06	0.28	0.07	-0.02	0	99.67	0	0.61	3.07
m4	1.41	2.11	7.26	0.64	0.68	0.58	0	99.09	0.47	0.47	2.28
m5	1.85	3.3	6.07	0.4	0.42	0.43	0	97.59	0.53	0.26	2.81
MA	1.09	2.09	6.39	0.53	0.53	0.48	0	99.09	0.32	0.27	2.87
FA	16.51	20.51	26.18	1.79	0.24	0.25	0	98.24	0.25	14.13	9
OA	8.8	11.3	16.28	1.16	0.38	0.36	0	98.67	0.29	7.2	5.93

PRAAT	GER20	GER10	GER05	FER20	FER10	FER05	V_UVE	UV_VE	THE	TLE	STD20
f1	0.74	2.6	5.2	1.1	1.07	1.18	0	35.97	0.74	0	4.77
f2	2.25	3.56	12.38	0.27	0.32	0.38	0	13.15	0.38	1.69	6.89
f3	3.86	4.37	10.54	0.65	0.66	0.85	1.54	47.74	1.03	2.83	5.2
f4	2.04	4.09	13.09	-0.23	-0.07	0.06	1.64	24.74	0	1.84	8.22
f5	0	0.88	9.06	-0.1	-0.41	-0.8	0	39.6	0	0	7.36
m1	7.02	9.44	19.13	0.54	0.55	0.42	6.54	18.6	0.24	6.54	3.9
m2	0.62	3.73	30.43	-1.4	-1.39	-0.42	0	24.67	0.62	0	6.47
m3	0.92	1.53	6.42	0.22	0.31	0.28	0.92	25.25	0	0.92	3.7
m4	10.3	12.65	23.65	0.69	0.59	0.38	10.3	12.76	0	10.3	3.54
m5	2.11	4.22	18.21	0.71	0.7	0.35	1.58	24.23	0.53	1.58	4.02
MA	4.2	6.31	19.57	0.15	0.15	0.2	3.87	21.1	0.28	3.87	4.33
FA	1.78	3.1	10.06	0.34	0.31	0.34	0.64	32.24	0.43	1.27	6.49
OA	2.99	4.71	14.81	0.24	0.23	0.27	2.25	26.67	0.35	2.57	5.41

Table A.2.16. Results for 25 dB Music noise for high resolution speech signals.

DHO	GER20	GER10	GER05	FER20	FER10	FER05	V_UVE	UV_VE	THE	TLE	STD20
f1	0.74	1.49	4.46	-0.56	-0.36	-0.27	0	14.03	0	0	4.19
f2	0	0.19	2.81	0.09	0.04	0.26	0	44.4	0	0	4.92
f3	0.51	0.51	3.6	-0.2	-0.2	0.1	0	21.34	0	0	3.61
f4	0	0.2	2.86	0.26	0.2	0.27	0	8.51	0	0	5.1
f5	0	0.29	2.63	0.34	0.42	0.5	0	30.54	0	0	5.32
m1	6.05	14.29	35.35	0.72	0.44	0.01	2.66	19.44	0.73	2.66	6.07
m2	0	1.24	19.25	1.64	1.59	1.28	0	25.55	0	0	5.78
m3	0	0.31	2.45	-0.52	-0.48	-0.37	0	25.25	0	0	2.87
m4	23.42	26	35.6	0.51	0.37	0.63	3.75	5.47	1.17	3.75	3.59
m5	5.41	9.23	18.6	-0.19	0.32	0.64	2.24	14.32	0.53	2.24	4.17
MA	6.98	10.21	22.25	0.43	0.45	0.44	1.73	18	0.49	1.73	4.49
FA	0.25	0.54	3.27	-0.02	0.02	0.17	0	23.76	0	0	4.63
OA	3.61	5.37	12.76	0.21	0.23	0.3	0.87	20.88	0.24	0.87	4.56

CORR	GER20	GER10	GER05	FER20	FER10	FER05	V_UVE	UV_VE	THE	TLE	STD20
f1	0.74	2.6	7.06	0.66	0.83	0.96	0	59.5	0.74	0	5.32
f2	1.88	3	11.26	-0.03	0.07	0.11	0.19	43.32	0.94	0.75	6.98
f3	4.63	5.66	11.83	0.43	0.58	0.67	2.06	67.81	1.03	3.6	5.69
f4	2.45	5.52	15.34	-0.73	-0.29	-0.22	1.23	41.75	0.61	1.64	9.46
f5	0	1.17	9.36	-0.57	-0.81	-1.25	0	54.36	0	0	7.42
m1	15.98	17.19	22.52	0.12	0.24	0.31	12.83	62.13	3.15	12.83	3.3
m2	1.86	4.35	26.09	-1.62	-1.47	-0.68	1.24	57.27	0.62	1.24	5.94
m3	1.83	2.45	6.12	-0.15	-0.06	0.05	1.83	60.66	0	1.83	3.4
m4	17.33	18.5	28.34	0.24	0.26	0.17	17.1	51.25	0.23	17.1	3.22
m5	10.55	11.74	24.54	0.63	0.59	0.31	7.26	51.94	3.17	7.26	3.9
MA	9.51	10.85	21.52	-0.16	-0.09	0.03	8.05	56.65	1.43	8.05	3.95
FA	1.94	3.59	10.97	-0.05	0.07	0.05	0.69	53.35	0.66	1.2	6.97
OA	5.73	7.22	16.24	-0.1	-0.01	0.04	4.37	55	1.05	4.62	5.46

YIN	GER20	GER10	GER05	FER20	FER10	FER05	V_UVE	UV_VE	THE	TLE	STD20
f1	9.67	14.13	20.07	0.97	0.04	0.17	0	98.87	1.12	5.95	7.64
f2	11.44	15.76	23.45	1.48	0.68	0.27	0	99.14	0	7.88	8.74
f3	6.17	8.23	11.83	0.78	0.22	0.51	0	98.19	0.51	4.37	5.66
f4	54.19	61.96	70.76	4.06	-0.65	-0.43	0	97.94	1.02	46.83	15.34
f5	14.91	18.42	24.85	2.03	0.65	0.16	0	98.83	0	13.74	9.09
m1	3.87	6.3	17.43	0.63	0.7	0.52	0	99.34	1.69	0.73	3.83
m2	0.62	1.86	3.73	0.69	0.87	0.91	0	99.78	0.62	0	2.98
m3	1.22	3.06	4.28	0.4	-0.05	-0.04	0	99.34	0	0.92	3.96
m4	1.87	3.04	9.13	0.61	0.67	0.58	0	99.09	0.47	0.7	2.58
m5	2.51	5.01	8.97	0.36	0.34	0.45	0	97.86	1.19	0.26	3.13
MA	2.02	3.85	8.71	0.54	0.51	0.48	0	99.08	0.79	0.52	3.3
FA	19.28	23.7	30.19	1.86	0.19	0.14	0	98.59	0.53	15.75	9.29
OA	10.65	13.78	19.45	1.2	0.35	0.31	0	98.84	0.66	8.14	6.3

PRAAT	GER20	GER10	GER05	FER20	FER10	FER05	V_UVE	UV_VE	THE	TLE	STD20
f1	0.74	2.6	6.69	0.83	1	1.18	0	43.89	0.74	0	5.35
f2	1.69	3.19	12.57	0.3	0.3	0.37	0	22.63	0.19	1.31	7.18
f3	4.37	4.88	12.85	0.66	0.67	0.87	1.54	64.01	1.03	3.34	5.53
f4	5.11	8.59	16.36	-1.09	-0.35	-0.08	1.64	30.93	0.2	4.7	9.34
f5	0	0.88	9.94	-0.35	-0.66	-0.98	0	56.04	0	0	7.47
m1	7.26	9.69	18.89	0.54	0.55	0.47	6.78	38.87	0.24	6.78	3.88
m2	0.62	3.11	29.81	-1.3	-1.19	-0.43	0	46.04	0.62	0	6.39
m3	0.92	1.53	8.26	0.06	0.15	0.22	0.92	41.97	0	0.92	3.88
m4	10.77	13.11	24.36	0.67	0.58	0.38	10.77	28.47	0	10.77	3.57
m5	2.24	4.35	18.47	0.69	0.69	0.32	1.72	39.36	0.53	1.72	4.03
MA	4.36	6.36	19.96	0.13	0.15	0.19	4.04	38.94	0.28	4.04	4.35
FA	2.38	4.03	11.68	0.07	0.19	0.27	0.64	43.5	0.43	1.87	6.97
OA	3.37	5.19	15.82	0.1	0.17	0.23	2.34	41.22	0.36	2.95	5.66

Table A.2.17. Results for 20 dB Music noise for high resolution speech signals.

DHO	GER20	GER10	GER05	FER20	FER10	FER05	V_UVE	UV_VE	THE	TLE	STD20
f1	0.74	1.49	4.83	-0.58	-0.38	-0.33	0	17.87	0	0	4.19
f2	0	0.56	3.38	0.14	0.08	0.32	0	46.98	0	0	5.26
f3	0.51	0.51	3.34	-0.13	-0.13	0.12	0	18.44	0	0	3.49
f4	4.91	5.11	7.98	0.01	-0.05	0.05	1.64	3.09	0	2.45	5.15
f5	1.75	2.05	4.97	0.42	0.5	0.67	0	28.19	0	0	5.39
m1	7.51	15.98	36.08	0.65	0.34	0.03	2.66	22.09	0.97	2.66	6.04
m2	0	1.24	18.63	1.7	1.65	1.26	0	25.11	0	0	5.81
m3	0	0.31	2.14	-0.51	-0.47	-0.34	0	20.98	0	0	2.87
m4	25.29	27.87	36.53	0.42	0.28	0.55	3.98	3.87	4.92	3.98	3.56
m5	8.05	11.87	20.45	-0.2	0.32	0.66	4.88	14.19	0.66	4.88	4.22
MA	8.17	11.45	22.77	0.41	0.42	0.43	2.31	17.25	1.31	2.31	4.5
FA	1.58	1.94	4.9	-0.03	0	0.17	0.33	22.92	0	0.49	4.7
OA	4.88	6.7	13.83	0.19	0.21	0.3	1.32	20.08	0.65	1.4	4.6

CORR	GER20	GER10	GER05	FER20	FER10	FER05	V_UVE	UV_VE	THE	TLE	STD20
f1	1.49	3.35	7.81	0.46	0.62	0.79	0.74	66.97	0.37	0.74	5.72
f2	2.81	4.5	15.76	0.12	0.06	0.08	0.19	49.14	0.94	1.13	7.74
f3	4.88	6.43	15.68	0.54	0.83	0.52	2.31	73.6	1.03	3.86	6.47
f4	5.73	7.57	20.04	-0.55	-0.39	-0.61	1.84	45.88	1.23	4.09	9.05
f5	0	1.17	13.45	-0.78	-1.03	-1.58	0	58.89	0	0	7.93
m1	18.4	19.37	24.21	0.22	0.23	0.36	15.25	67.94	3.15	15.25	3.21
m2	7.45	9.94	32.92	-1.82	-1.67	-0.84	3.73	61.67	2.48	3.73	6.16
m3	3.98	4.28	15.29	-0.36	-0.31	0.06	2.75	66.89	0.31	2.75	4.04
m4	18.27	18.97	29.27	0.25	0.27	0.14	17.8	56.26	0.47	17.8	3.21
m5	18.47	20.05	30.61	0.58	0.51	0.22	10.82	59.57	6.2	10.82	4
MA	13.31	14.52	26.46	-0.22	-0.19	-0.01	10.07	62.47	2.52	10.07	4.12
FA	2.98	4.6	14.55	-0.04	0.02	-0.16	1.02	58.89	0.71	1.96	7.38
OA	8.15	9.56	20.5	-0.13	-0.09	-0.09	5.54	60.68	1.62	6.02	5.75

YIN	GER20	GER10	GER05	FER20	FER10	FER05	V_UVE	UV_VE	THE	TLE	STD20
f1	12.64	16.73	21.56	0.74	0.36	0.38	0	98.42	3.35	6.32	7.34
f2	19.51	23.08	31.71	1.5	0.93	0.51	0	98.92	1.31	12.95	8.71
f3	12.85	17.99	23.65	1.22	0.18	0.35	0	98.73	0.77	9	7.74
f4	54.81	62.78	71.17	5.28	0.57	0	0	97.94	2.04	48.26	16.62
f5	21.05	26.9	34.21	2.19	-0.19	-0.31	0	99.16	0	19.59	10.45
m1	5.81	8.96	21.55	0.67	0.93	0.64	0	99.17	2.91	0.73	4.13
m2	1.86	4.35	9.32	0.49	0.78	0.77	0	99.34	0.62	0.62	4.86
m3	5.81	8.56	12.84	0.18	-0.12	-0.08	0	99.67	0.61	3.67	4.53
m4	2.11	2.58	7.73	0.62	0.69	0.64	0	99.54	1.41	0.47	2.3
m5	7.78	9.89	14.12	0.3	0.3	0.4	0	98.13	4.09	1.19	3.25
MA	4.68	6.87	13.11	0.45	0.52	0.47	0	99.17	1.93	1.33	3.81
FA	24.17	29.5	36.46	2.18	0.37	0.18	0	98.63	1.5	19.22	10.17
OA	14.42	18.18	24.79	1.32	0.44	0.33	0	98.9	1.71	10.28	6.99

PRAAT	GER20	GER10	GER05	FER20	FER10	FER05	V_UVE	UV_VE	THE	TLE	STD20
f1	0.74	2.6	7.81	0.73	0.89	1.04	0	58.6	0.74	0	5.73
f2	1.69	3	15.2	0.35	0.31	0.16	0	31.25	0.19	1.31	7.55
f3	6.43	7.97	16.45	0.53	0.8	0.57	1.29	70.34	1.03	5.14	6.47
f4	7.98	9.41	20.86	-0.46	-0.36	-0.43	2.25	35.57	0.2	7.57	8.35
f5	3.22	4.09	15.79	-0.71	-1.03	-1.43	0	60.4	0	3.22	7.87
m1	8.47	11.14	21.07	0.57	0.6	0.49	7.75	46.01	0.48	7.75	4
m2	3.73	6.83	32.92	-1.68	-1.62	-0.63	1.86	51.32	0.62	1.86	6.47
m3	1.83	2.45	14.07	-0.12	-0.03	0.15	1.22	54.1	0	1.22	4.23
m4	9.13	11.48	24.82	0.85	0.62	0.32	9.13	41	0	9.13	3.77
m5	5.41	7.12	21.5	0.64	0.59	0.3	3.3	49.67	1.45	3.3	4.07
MA	5.72	7.8	22.88	0.05	0.03	0.13	4.65	48.42	0.51	4.65	4.51
FA	4.01	5.41	15.22	0.09	0.12	-0.02	0.71	51.23	0.43	3.45	7.2
OA	4.86	6.61	19.05	0.07	0.08	0.05	2.68	49.83	0.47	4.05	5.85

Table A.2.18. Results for 15 dB Music noise for high resolution speech signals.

DHO	GER20	GER10	GER05	FER20	FER10	FER05	V_UVE	UV_VE	THE	TLE	STD20
f1	0	0.74	3.72	-0.6	-0.39	-0.31	0	19.46	0	0	4.16
f2	1.13	1.69	5.07	-0.05	0.02	0.35	0.75	46.77	0	0.75	5.41
f3	0	1.03	3.6	-0.33	-0.13	0.1	0	28.03	0	0	4
f4	5.32	5.52	7.98	0.17	0.1	0.13	2.66	12.89	0	2.66	5.18
f5	0	0.29	3.51	0.32	0.4	0.6	0	37.92	0	0	5.35
m1	8.23	16.22	36.32	0.72	0.46	0.06	0.73	23.26	2.91	0.73	6.12
m2	0	1.24	19.88	1.74	1.69	1.11	0	34.14	0	0	5.93
m3	0.61	0.92	2.75	-0.58	-0.54	-0.42	0.61	29.84	0	0.61	2.9
m4	18.74	19.67	29.51	0.26	0.2	0.42	8.43	7.52	4.45	8.43	3.21
m5	7.26	10.95	20.45	-0.35	0.14	0.62	5.41	15.26	0	5.41	4.2
MA	6.97	9.8	21.78	0.36	0.39	0.36	3.04	22	1.47	3.04	4.47
FA	1.29	1.85	4.77	-0.1	0	0.17	0.68	29.01	0	0.68	4.82
OA	4.13	5.83	13.28	0.13	0.19	0.27	1.86	25.51	0.74	1.86	4.65

CORR	GER20	GER10	GER05	FER20	FER10	FER05	V_UVE	UV_VE	THE	TLE	STD20
f1	2.23	3.72	13.38	0.24	0.27	0.81	1.12	73.98	0.74	1.12	6.5
f2	5.07	7.88	22.7	0.35	-0.21	0.01	0.56	54.74	0.94	2.06	9.46
f3	4.88	10.54	25.19	0.68	0.79	0.15	2.06	80.47	1.03	3.86	9.16
f4	14.72	16.16	29.24	-0.25	-0.4	-0.73	4.5	56.19	3.27	10.43	9.21
f5	0.29	1.46	19.59	-1.16	-1.42	-1.79	0	73.32	0	0.29	8.71
m1	23	23.97	28.33	0.25	0.34	0.41	15.5	70.76	7.02	15.5	3.36
m2	18.63	21.74	41.61	-2.43	-2.17	-1.39	4.97	66.08	9.32	4.97	6.44
m3	10.7	12.54	25.69	-0.93	-0.53	-0.11	5.81	69.84	0.61	5.81	5.21
m4	24.59	25.53	33.96	0.21	0.2	0.06	20.61	67.43	3.04	20.61	3.22
m5	31.66	33.38	41.42	0.16	0.2	0.18	17.02	69.48	10.42	17.02	4.2
MA	21.72	23.43	34.2	-0.55	-0.39	-0.17	12.78	68.72	6.08	12.78	4.49
FA	5.44	7.95	22.02	-0.03	-0.19	-0.31	1.65	67.74	1.2	3.55	8.61
OA	13.58	15.69	28.11	-0.29	-0.29	-0.24	7.21	68.23	3.64	8.17	6.55

YIN	GER20	GER10	GER05	FER20	FER10	FER05	V_UVE	UV_VE	THE	TLE	STD20
f1	35.32	40.89	50.56	2.07	0.62	0.01	0	99.55	3.72	24.16	9.99
f2	32.65	40.34	50.28	3.57	0.23	0.19	0	98.92	1.13	25.33	13.01
f3	29.31	39.33	50.13	2.01	-0.55	0.25	0	99.1	2.31	18.77	11.76
f4	61.96	70.55	81.39	5.64	0.08	-0.01	0	98.71	5.32	49.28	18.96
f5	34.21	41.81	51.46	3.11	-0.6	-0.19	0	98.32	0	30.7	13.48
m1	7.51	10.17	22.76	0.77	0.94	0.62	0	99.67	3.63	0.97	4.28
m2	9.94	18.01	29.19	1.23	0.24	0.18	0	99.78	1.86	1.86	7.62
m3	14.37	18.65	28.75	-0.08	-0.1	-0.08	0	100	2.14	6.73	5.94
m4	7.26	9.84	16.63	0.67	0.78	0.53	0	100	3.28	1.41	3.07
m5	18.6	22.3	27.57	0.31	0.41	0.5	0	98.66	9.23	3.3	4.19
MA	11.54	15.79	24.98	0.58	0.45	0.35	0	99.62	4.03	2.85	5.02
FA	38.69	46.59	56.76	3.28	-0.05	0.05	0	98.92	2.49	29.65	13.44
OA	25.11	31.19	40.87	1.93	0.2	0.2	0	99.27	3.26	16.25	9.23

PRAAT	GER20	GER10	GER05	FER20	FER10	FER05	V_UVE	UV_VE	THE	TLE	STD20
f1	8.92	10.78	17.84	1	1.19	1.02	1.12	64.03	0.37	8.18	6.18
f2	5.25	8.82	23.26	0.78	0.11	0.24	0	37.5	0.19	3.75	9.76
f3	6.68	10.8	25.71	1.01	1.38	0.44	1.29	71.61	1.03	5.4	8.61
f4	15.75	16.77	29.65	-0.08	-0.06	-0.26	4.91	38.14	0.2	14.93	8.9
f5	4.97	5.85	23.68	-0.95	-1.28	-1.72	0.29	64.09	0	4.97	8.72
m1	10.9	13.32	22.76	0.65	0.72	0.55	8.96	49.67	1.45	8.96	3.94
m2	8.7	12.42	37.27	-1.86	-1.71	-0.82	4.97	58.37	0.62	4.97	6.72
m3	6.42	7.95	23.85	-0.55	-0.31	0.08	2.14	61.97	0	2.14	5.07
m4	10.07	12.88	26.46	0.81	0.53	0.31	10.07	54.67	0	10.07	3.79
m5	12.93	14.64	27.31	0.47	0.4	0.24	4.88	55.29	5.01	4.88	4.21
MA	9.8	12.24	27.53	-0.09	-0.07	0.07	6.2	55.99	1.42	6.2	4.75
FA	8.32	10.6	24.03	0.35	0.27	-0.06	1.52	55.07	0.36	7.45	8.43
OA	9.06	11.42	25.78	0.13	0.1	0.01	3.86	55.53	0.89	6.82	6.59

Table A.2.19. Results for 10 dB Music noise for high resolution speech signals.

DHO	GER20	GER10	GER05	FER20	FER10	FER05	V_UVE	UV_VE	THE	TLE	STD20
f1	0	0.74	4.09	-0.78	-0.57	-0.36	0	58.14	0	0	4.27
f2	3.38	3.94	6.38	-0.06	0.05	0.32	0	67.67	0	3.38	5.52
f3	0	0.26	3.34	-0.16	-0.11	0.06	0	52.98	0	0	3.74
f4	6.54	8.18	10.84	0.29	0.27	0.21	3.68	31.96	0	3.68	6.23
f5	0	0.29	4.09	0.19	0.27	0.54	0	66.11	0	0	5.42
m1	4.84	12.11	34.87	0.6	0.48	-0.04	1.21	39.37	0.97	1.21	6.12
m2	1.24	3.11	22.98	1.52	1.56	1.37	0	54.85	0	0	6.16
m3	0	0.31	3.67	-0.62	-0.57	-0.38	0	32.79	0	0	3.07
m4	16.86	19.44	29.74	0.22	0.24	0.42	3.98	29.38	3.28	3.98	3.48
m5	7.26	11.08	22.3	-0.38	0.07	0.63	3.03	25.97	1.72	3.03	4.52
MA	6.04	9.21	22.71	0.27	0.36	0.4	1.65	36.47	1.19	1.65	4.67
FA	1.98	2.68	5.75	-0.1	-0.02	0.16	0.74	55.37	0	1.41	5.03
OA	4.01	5.95	14.23	0.08	0.17	0.28	1.19	45.92	0.6	1.53	4.85

CORR	GER20	GER10	GER05	FER20	FER10	FER05	V_UVE	UV_VE	THE	TLE	STD20
f1	7.06	10.78	26.02	0.06	0.19	1.08	2.6	80.77	1.12	5.2	8.74
f2	12.01	18.39	36.02	1.09	0.37	-0.05	3.19	59.05	0.94	8.26	11.26
f3	7.46	19.54	39.59	1.46	1.55	0.3	5.4	83	1.03	6.43	12.31
f4	32.72	34.56	46.22	-0.13	-0.33	-0.36	13.09	65.98	5.11	23.72	10.29
f5	6.14	9.06	28.65	-1.12	-1.89	-1.96	2.34	81.88	1.17	4.68	10.19
m1	35.84	36.8	41.4	-0.02	0.08	0.3	18.16	71.76	15.5	18.16	3.73
m2	21.12	27.95	45.96	-3.88	-2.66	-1.52	4.97	71.37	8.07	4.97	7.39
m3	22.63	25.99	38.84	-1.34	-0.48	0.02	10.4	75.08	2.75	10.4	6.28
m4	37.94	38.88	46.84	0.04	0.11	-0.07	28.1	76.99	6.56	28.1	3.5
m5	50	50.79	55.67	-0.56	-0.4	-0.05	23.48	77.91	16.23	23.48	3.77
MA	33.5	36.08	45.74	-1.15	-0.67	-0.26	17.02	74.62	9.82	17.02	4.93
FA	13.08	18.47	35.3	0.27	-0.02	-0.2	5.32	74.14	1.87	9.66	10.56
OA	23.29	27.27	40.52	-0.44	-0.35	-0.23	11.17	74.38	5.85	13.34	7.74

YIN	GER20	GER10	GER05	FER20	FER10	FER05	V_UVE	UV_VE	THE	TLE	STD20
f1	49.07	60.59	68.4	4.72	0.18	-0.24	0	99.32	5.2	36.8	14.68
f2	47.28	56.29	68.67	4.1	0.4	-0.69	0	98.71	2.63	38.27	15.29
f3	38.82	48.07	64.78	3.23	1.22	0.25	0	99.28	4.63	26.48	13.15
f4	66.46	76.69	83.44	8.58	0.49	0.12	0	99.23	10.43	46.63	19.34
f5	49.12	57.02	69.01	3.34	-0.14	0.12	0	98.99	2.34	38.6	16.28
m1	14.53	19.37	29.54	0.76	0.48	0.54	0	99.5	5.81	3.39	5.1
m2	25.47	32.92	45.34	1.47	0.78	0.06	0	99.78	7.45	9.32	7.96
m3	27.52	35.78	45.26	-1.04	-0.28	-0.05	0	99.34	7.65	10.09	7.81
m4	16.86	20.61	30.21	0.81	0.72	0.57	0	99.54	7.73	3.75	4.03
m5	36.68	47.63	55.28	-0.45	0.39	0.43	0	99.6	18.21	4.49	7.11
MA	24.21	31.26	41.13	0.31	0.42	0.31	0	99.55	9.37	6.21	6.4
FA	50.15	59.73	70.86	4.79	0.43	-0.09	0	99.1	5.05	37.36	15.75
OA	37.18	45.5	55.99	2.55	0.42	0.11	0	99.33	7.21	21.78	11.08

PRAAT	GER20	GER10	GER05	FER20	FER10	FER05	V_UVE	UV_VE	THE	TLE	STD20
f1	16.36	20.82	29.37	0.65	0.89	1.1	2.23	72.17	0.37	15.61	8.53
f2	17.64	23.08	40.71	1.3	0.33	-0.13	1.5	40.09	0.38	14.63	10.83
f3	11.31	22.37	40.1	2.88	1.76	0.32	3.86	71.97	1.03	9.25	11.77
f4	35.17	36.61	47.44	-0.31	-0.27	-0.39	12.88	47.94	0	31.7	9.71
f5	16.37	19.01	39.47	-0.89	-1.86	-1.94	0.29	64.26	0	16.37	10.43
m1	14.29	16.71	26.39	0.85	0.86	0.65	10.17	53.32	3.15	10.17	4.08
m2	21.74	27.95	47.83	-3.66	-2.69	-0.99	6.21	61.89	7.45	7.45	7.43
m3	15.29	18.96	35.47	-1.01	-0.15	0.18	5.81	66.23	0.31	5.81	6.52
m4	13.35	16.39	32.08	0.69	0.3	0.18	10.77	65.83	0.23	10.77	4.1
m5	33.91	35.62	44.46	-0.03	-0.01	0.11	15.83	59.17	8.84	15.83	4.27
MA	19.71	23.13	37.25	-0.63	-0.34	0.03	9.76	61.29	4	10.01	5.28
FA	19.37	24.37	39.42	0.73	0.17	-0.21	4.15	59.29	0.36	17.51	10.26
OA	19.54	23.75	38.33	0.05	-0.08	-0.09	6.96	60.29	2.18	13.76	7.77

Table A.2.20. Results for 5 dB Music noise for high resolution speech signals.

DHO	GER20	GER10	GER05	FER20	FER10	FER05	V_UVE	UV_VE	THE	TLE	STD20
f1	0	1.86	5.58	-0.14	-0.24	-0.23	0	85.07	0	0	5.06
f2	1.5	2.25	5.63	-0.11	-0.05	0.24	0	78.23	0	0	5.67
f3	0	2.06	5.91	-0.42	-0.28	-0.01	0	63.65	0	0	5.45
f4	11.25	16.36	22.29	2.11	0.34	0.19	5.32	67.01	0	5.73	10.94
f5	1.75	2.92	7.6	0.43	0.36	0.78	0	72.82	0	0.88	6.32
m1	8.96	17.43	36.8	0.49	0.36	-0.05	0.97	67.44	3.39	0.97	6.36
m2	7.45	11.18	27.95	1.7	1.15	1.38	0	79.96	0	0	6.25
m3	2.14	3.06	7.34	-0.61	-0.75	-0.45	0	77.05	0	0	3.78
m4	18.03	21.78	33.96	0.04	0.24	0.48	6.09	82.46	3.04	6.09	4.01
m5	9.76	16.49	28.1	-0.69	0.11	0.52	0.13	65.06	4.09	0.13	5.24
MA	9.27	13.99	26.83	0.19	0.22	0.38	1.44	74.39	2.1	1.44	5.13
FA	2.9	5.09	9.4	0.37	0.02	0.19	1.06	73.36	0	1.32	6.69
OA	6.09	9.54	18.12	0.28	0.12	0.29	1.25	73.87	1.05	1.38	5.91

CORR	GER20	GER10	GER05	FER20	FER10	FER05	V_UVE	UV_VE	THE	TLE	STD20
f1	17.84	24.91	44.24	1.1	0.89	1	8.92	83.26	2.23	13.38	11.1
f2	23.45	33.58	51.41	1.15	-0.04	-0.69	9.01	67.67	1.31	14.82	13.96
f3	15.68	36.25	57.84	4.38	2.48	0.87	10.8	84.09	1.29	12.34	15.71
f4	55.83	59.51	68.71	0.38	-0.16	-0.72	15.95	72.94	7.77	36.61	13.82
f5	18.42	24.27	44.44	-0.11	-1.52	-1.37	7.31	85.23	2.92	11.7	12.84
m1	63.68	64.16	67.55	-0.92	-0.92	-0.32	26.39	75.75	32.2	26.39	4.4
m2	35.4	41.61	59.01	-5.04	-3.33	-1.62	15.53	75.77	7.45	15.53	7.82
m3	44.95	50.76	59.94	-2.94	-0.75	0.01	18.04	78.03	7.65	18.04	8.18
m4	59.25	59.95	68.15	-0.4	-0.35	-0.44	36.53	79.27	18.5	36.53	3.95
m5	71.77	72.56	75.2	-1.32	-0.81	-0.2	26.78	80.59	22.56	26.78	4.64
MA	55.01	57.81	65.97	-2.12	-1.23	-0.51	24.66	77.88	17.67	24.66	5.8
FA	26.25	35.7	53.33	1.38	0.33	-0.18	10.4	78.64	3.1	17.77	13.48
OA	40.63	46.76	59.65	-0.37	-0.45	-0.35	17.53	78.26	10.39	21.21	9.64

YIN	GER20	GER10	GER05	FER20	FER10	FER05	V_UVE	UV_VE	THE	TLE	STD20
f1	54.28	66.54	76.95	3.48	1.21	1.13	0	99.55	8.55	36.06	15.91
f2	57.79	69.42	81.99	3.48	0.95	1.39	0	99.35	6	41.28	18.02
f3	44.47	55.27	70.95	4.9	1.13	0.54	0	99.28	8.74	26.22	13.75
f4	72.8	80.78	87.73	7.39	0.34	-0.4	0	99.23	16.56	44.38	20.61
f5	58.19	71.64	82.16	8.12	0.94	0.33	0	99.66	7.02	43.27	20.12
m1	37.53	45.52	57.63	0.6	0.45	0.36	0	99.67	17.19	7.51	7.2
m2	45.34	57.14	70.81	-1.72	-0.26	0.12	0	99.78	11.8	16.77	11.54
m3	56.57	69.42	81.04	0.46	-1.19	-0.43	0	100	21.41	14.98	12.71
m4	35.83	43.33	51.52	0.66	0.62	0.62	0	99.32	18.97	7.03	5.38
m5	56.73	66.36	73.88	-0.05	0.15	0.28	0	99.46	30.74	7.26	8.87
MA	46.4	56.35	66.98	-0.01	-0.05	0.19	0	99.65	20.02	10.71	9.14
FA	57.5	68.73	79.96	5.47	0.91	0.6	0	99.41	9.38	38.24	17.68
OA	51.95	62.54	73.47	2.73	0.43	0.39	0	99.53	14.7	24.48	13.41

PRAAT	GER20	GER10	GER05	FER20	FER10	FER05	V_UVE	UV_VE	THE	TLE	STD20
f1	20.07	27.51	46.1	0.37	0.51	0.57	5.58	77.38	0.37	17.47	11.05
f2	34.15	42.4	57.22	3.65	1.01	-0.05	9.01	53.88	0.19	27.77	13.04
f3	10.54	30.33	53.98	5.59	2.48	0.61	4.88	74.5	1.03	8.48	14.39
f4	59.51	62.58	70.35	1.3	0.76	-0.09	17.79	55.67	0.2	46.63	12.5
f5	31.29	34.5	54.39	-0.63	-2.14	-0.97	5.26	68.96	0	28.36	11.69
m1	32.45	33.66	44.79	0.1	0.14	0.22	14.04	60.96	13.32	14.04	4.53
m2	26.71	35.4	55.9	-5.62	-3.56	-1.55	8.7	66.74	4.97	9.32	8.07
m3	30.89	37	50.46	-2.21	-0.33	0.18	10.4	71.48	0.61	10.4	8.04
m4	19.67	21.55	39.81	0.35	0.16	0.06	14.52	75.17	1.41	14.52	4.04
m5	56.86	58.44	64.64	-0.86	-0.66	-0.12	21.9	67.34	13.19	21.9	4.99
MA	33.31	37.21	51.12	-1.64	-0.85	-0.24	13.91	68.34	6.7	14.04	5.93
FA	31.11	39.46	56.41	2.06	0.53	0.01	8.5	66.08	0.36	25.74	12.53
OA	32.21	38.34	53.76	0.21	-0.16	-0.11	11.21	67.21	3.53	19.89	9.23

Table A.2.21. Results for 0 dB Music noise for high resolution speech signals.

DHO	GER20	GER10	GER05	FER20	FER10	FER05	V_UVE	UV_VE	THE	TLE	STD20
f1	10.41	15.61	18.22	-1.92	-0.84	-0.46	3.72	85.07	0	3.72	7.47
f2	3	7.13	11.82	-0.04	-0.07	0.08	0	79.74	0.19	0.19	8.78
f3	4.88	9	13.88	-1.75	-0.85	-0.62	0	83	0	0	6.53
f4	19.63	30.06	39.26	2.98	0.58	0.21	8.79	86.6	0	10.22	15.82
f5	4.97	9.65	15.79	-0.43	-0.15	0.31	1.75	83.05	0	2.05	9.45
m1	17.43	24.21	47.94	0.19	0.21	0.02	1.45	71.93	4.36	1.45	6.24
m2	7.45	17.39	31.06	0.44	0.61	1.2	0	85.24	0	0	7.54
m3	3.67	4.89	15.29	-0.83	-0.62	-0.26	0	82.3	0.31	0	4.52
m4	13.82	16.63	32.79	0.06	0.37	0.42	4.22	85.65	5.62	4.22	3.92
m5	10.29	19.26	30.61	-0.51	0.18	0.41	1.45	88.35	3.03	1.45	5.56
MA	10.53	16.48	31.54	-0.13	0.15	0.36	1.42	82.69	2.66	1.42	5.56
FA	8.58	14.29	19.79	-0.23	-0.27	-0.1	2.85	83.49	0.04	3.24	9.61
OA	9.56	15.38	25.67	-0.18	-0.06	0.13	2.14	83.09	1.35	2.33	7.58

CORR	GER20	GER10	GER05	FER20	FER10	FER05	V_UVE	UV_VE	THE	TLE	STD20
f1	34.2	53.53	71.75	5.6	1.39	0.38	17.47	84.84	4.09	21.93	15.71
f2	45.03	58.54	69.79	4.03	0.07	-0.81	16.7	77.16	3.19	27.95	16.72
f3	26.99	52.44	73.01	7.19	3.29	0.59	14.91	83.73	4.11	16.97	17.56
f4	73.21	78.94	85.07	5.14	2.29	0.71	19.63	80.41	7.98	48.26	18.45
f5	42.98	52.63	67.25	2.81	-0.66	-0.97	15.79	83.89	4.68	27.19	16.42
m1	84.99	85.96	87.41	-2.14	-0.84	0.04	25.42	80.73	45.52	25.42	6.84
m2	59.01	67.7	80.75	-9.18	-5.59	-2.58	24.22	79.52	11.8	24.22	8.98
m3	59.94	70.34	78.9	-7.14	-2.13	-0.29	19.57	83.28	13.46	19.57	10.15
m4	86.89	87.82	90.87	-0.78	-0.7	-0.2	39.58	82.46	36.07	39.58	4.94
m5	91.82	92.61	93.54	-3.65	-1.8	-0.61	25.59	83.53	30.08	25.59	6.85
MA	76.53	80.89	86.29	-4.58	-2.21	-0.73	26.88	81.9	27.38	26.88	7.55
FA	44.48	59.22	73.37	4.95	1.27	-0.02	16.9	82.01	4.81	28.46	16.97
OA	60.51	70.05	79.83	0.19	-0.47	-0.37	21.89	81.95	16.1	27.67	12.26

YIN	GER20	GER10	GER05	FER20	FER10	FER05	V_UVE	UV_VE	THE	TLE	STD20
f1	60.59	76.21	85.5	4.7	-0.13	0.7	0	99.55	15.24	32.34	19.27
f2	64.92	77.11	86.68	5.1	0.54	-0.36	0	99.57	13.32	38.65	19.17
f3	60.15	72.75	86.89	4.34	2	0.52	0	100	19.54	27.51	17.9
f4	80.37	88.14	94.68	6.98	0.97	1.3	0	99.48	21.68	43.15	24.32
f5	64.91	78.65	88.01	6.42	-0.23	0.63	0	99.5	14.91	37.72	23.23
m1	67.55	78.45	86.68	1.8	0.53	0.42	0	99.83	35.11	10.17	11.5
m2	55.9	71.43	83.23	-0.15	0.53	-0.06	0	100	19.88	13.04	13.85
m3	65.14	81.35	89.6	-0.83	-2.33	-0.21	0	100	29.36	14.68	15.93
m4	69.56	79.86	87.59	0.67	0.65	-0.15	0	99.54	40.05	11.94	9.28
m5	79.68	88.39	93.67	0.02	0.42	0.3	0	99.87	44.2	9.1	12.27
MA	67.57	79.89	88.15	0.3	-0.04	0.06	0	99.85	33.72	11.79	12.56
FA	66.19	78.57	88.35	5.51	0.63	0.56	0	99.62	16.94	35.87	20.78
OA	66.88	79.23	88.25	2.9	0.29	0.31	0	99.73	25.33	23.83	16.67

PRAAT	GER20	GER10	GER05	FER20	FER10	FER05	V_UVE	UV_VE	THE	TLE	STD20
f1	37.55	57.99	73.23	6.66	2.52	0.67	13.01	78.05	0.37	30.48	16.27
f2	52.91	65.1	75.42	7.73	1.44	-0.07	18.01	69.4	0.19	41.65	15.48
f3	21.08	46.79	69.67	8.13	3.45	0.61	10.8	79.75	0.77	16.97	16.37
f4	77.1	80.78	86.3	8.35	4.81	1.09	24.54	71.91	0.2	60.74	15.37
f5	48.83	55.85	73.39	2.22	-1.18	-0.39	16.96	76.34	0	42.4	15.46
m1	69.01	70.94	75.3	-1.1	-0.28	-0.2	25.18	71.93	27.6	28.09	6.14
m2	49.07	61.49	79.5	-10	-6.26	-2.97	22.36	74.23	3.11	22.98	8.54
m3	52.91	63.3	73.7	-6.09	-1.89	-0.08	18.65	77.38	3.98	19.88	9.67
m4	51.05	55.97	68.38	0.76	-0.27	-0.32	34.89	77.9	8.43	35.13	5.21
m5	85.88	86.81	88.92	-1.88	-0.52	-0.09	28.63	75.77	16.23	29.42	6.76
MA	62	68	77	-4	-2	-1	26	75	12	27	7.26
FA	47	61	76	6.6	2.2	0.4	17	75	0.3	38	15.79
OA	55	65	76	1.5	0.2	-0	21	75	6.1	33	11.53

Table A.2.22. Results for 0 dB Music noise for high resolution speech signals.

DHO	GER20	GER10	GER05	FER20	FER10	FER05	V_UVE	UV_VE	THE	TLE	STD20
f1	0.74	1.86	3.35	-0.63	-0.37	-0.5	0	18.55	0	0	4.32
f2	0.75	2.63	10.51	-0.97	-0.49	0.24	0	26.29	0	0	7.21
f3	0	3.08	7.97	-0.91	-0.27	-0.07	0	9.04	0	0	6.24
f4	5.52	7.98	14.72	-0.74	0.03	-0.01	5.32	1.8	0	5.32	8.33
f5	4.09	5.56	15.2	-0.05	0.37	0.74	0.58	51.85	0	0.58	7.54
m1	12.59	22.76	41.89	0.33	0.24	-0.15	4.6	9.47	1.21	4.6	7.05
m2	10.56	14.29	32.3	1.7	2.14	1.3	0	30.18	0	0	6.65
m3	3.67	4.28	11.01	-0.92	-0.83	-0.45	3.67	9.51	0	3.67	3.61
m4	9.84	11.01	24.12	0.15	0.31	0.43	2.11	1.37	3.04	2.11	3.49
m5	16.09	20.84	30.61	-0.66	0.18	0.65	12.8	1.47	0	12.8	4.94
MA	10.55	14.64	27.98	0.12	0.41	0.36	4.63	10.4	0.85	4.63	5.15
FA	2.22	4.22	10.35	-0.66	-0.15	0.08	1.18	21.51	0	1.18	6.73
OA	6.39	9.43	19.17	-0.27	0.13	0.22	2.91	15.95	0.43	2.91	5.94

CORR	GER20	GER10	GER05	FER20	FER10	FER05	V_UVE	UV_VE	THE	TLE	STD20
f1	0.74	2.23	4.09	0.85	0.89	0.95	0	0.23	0.74	0	4.82
f2	1.88	3	9.38	0.02	0.14	0.18	0.56	0.22	0.75	1.13	6.85
f3	3.08	4.11	9.25	0.35	0.49	0.68	0.26	0	1.03	2.06	5.29
f4	3.07	4.91	14.72	-0.48	-0.12	0.13	2.04	0.26	0.61	2.25	8.34
f5	0	1.17	10.23	-0.33	-0.57	-0.94	0	0.84	0	0	7.54
m1	16.22	17.19	22.03	0.2	0.28	0.43	10.9	0	5.33	10.9	3.15
m2	6.21	6.83	31.68	-1.49	-1.6	-0.68	1.86	1.98	4.35	1.86	5.91
m3	1.83	2.45	4.89	0.1	0.19	0.28	1.22	0	0.61	1.22	3.1
m4	20.84	21.31	30.91	0.65	0.66	0.44	15.46	0.23	5.39	15.46	2.91
m5	12.01	13.19	23.48	0.69	0.66	0.4	11.61	0.27	0.4	11.61	3.76
MA	11.42	12.19	22.6	0.03	0.04	0.17	8.21	0.5	3.21	8.21	3.76
FA	1.75	3.08	9.54	0.08	0.17	0.2	0.57	0.31	0.63	1.09	6.57
OA	6.59	7.64	16.07	0.06	0.1	0.19	4.39	0.4	1.92	4.65	5.17

YIN	GER20	GER10	GER05	FER20	FER10	FER05	V_UVE	UV_VE	THE	TLE	STD20
f1	85.87	89.96	94.05	6.61	1.24	-0.7	0	100	0.37	84.76	15.91
f2	89.12	92.5	95.5	4.22	-0.66	-0.79	0	99.78	0	87.62	18.77
f3	79.69	84.58	90.23	5.74	0.71	0.66	0	99.82	0.26	76.09	14.29
f4	87.73	91.41	93.25	7.83	-0.17	-1.78	0	100	0.41	83.44	20.27
f5	89.77	92.11	93.27	5.29	0.37	-0.96	0	100	0	86.84	14.32
m1	11.14	15.25	27.12	0.68	0.53	0.28	0	99.83	2.18	6.78	4.73
m2	36.02	37.89	44.1	0.83	0.54	0.6	0	100	0.62	31.68	5.16
m3	21.1	24.77	30.28	0.89	0.03	-0.19	0	99.67	0	18.65	5.28
m4	1.41	2.58	11.71	0.63	0.63	0.5	0	100	0	0	2.72
m5	4.88	7.78	13.46	0.52	0.41	0.46	0	99.87	0.66	2.77	4.02
MA	14.91	17.65	25.33	0.71	0.43	0.33	0	99.87	0.69	11.98	4.38
FA	86.44	90.11	93.26	5.94	0.3	-0.71	0	99.92	0.21	83.75	16.71
OA	50.67	53.88	59.3	3.32	0.36	-0.19	0	99.9	0.45	47.86	10.55

PRAAT	GER20	GER10	GER05	FER20	FER10	FER05	V_UVE	UV_VE	THE	TLE	STD20
f1	4.46	6.69	8.18	1.17	1.22	1.13	0.74	0	0.74	3.72	4.57
f2	2.44	3.38	10.32	0.23	0.3	0.3	0.38	0	0.38	2.06	6.61
f3	2.57	3.08	9.51	0.61	0.61	0.71	0	0	0.77	1.8	5.12
f4	2.25	3.89	12.47	-0.06	0.09	0.29	2.04	0	0	2.04	7.92
f5	0	1.17	7.6	0.05	-0.19	-0.64	0	0.5	0	0	7.24
m1	8.96	10.9	19.85	0.47	0.56	0.4	8.47	0	0.24	8.47	3.8
m2	0.62	3.11	31.06	-1.38	-1.66	-0.4	0	0.66	0.62	0	6.44
m3	0.92	1.53	6.73	0.2	0.31	0.31	0.92	0	0	0.92	3.55
m4	14.52	15.22	26.46	0.63	0.59	0.38	12.41	0	1.87	12.41	3.18
m5	2.77	4.75	18.6	0.81	0.77	0.43	2.77	0	0	2.77	3.97
MA	5.56	7.1	20.54	0.15	0.11	0.22	4.91	0.13	0.55	4.91	4.19
FA	2.34	3.64	9.62	0.4	0.41	0.36	0.63	0.1	0.38	1.93	6.29
OA	3.95	5.37	15.08	0.27	0.26	0.29	2.77	0.12	0.46	3.42	5.24

Table A.2.23. Results for clean speech pitch estimation for low resolution signals.

DHO	GER20	GER10	GER05	FER20	FER10	FER05	V_UVE	UV_VE	THE	TLE	STD20
f1	0.74	1.86	3.35	-0.68	-0.42	-0.55	0	20.81	0	0	4.29
f2	0	1.88	9.76	-0.94	-0.46	0.18	0	25	0	0	7.18
f3	0	2.83	7.97	-0.82	-0.27	-0.02	0	15.73	0	0	6.1
f4	12.07	14.93	20.45	-0.76	0.08	0.07	10.02	1.29	0	10.02	8.6
f5	0.58	2.05	11.7	-0.13	0.28	0.62	0.58	42.62	0	0.58	7.45
m1	13.8	25.18	41.65	0.61	0.07	-0.21	4.36	9.8	1.94	4.36	7.46
m2	9.32	13.04	30.43	1.85	2.28	1.65	4.97	22.47	0	4.97	6.64
m3	3.67	4.28	11.31	-0.96	-0.87	-0.47	3.67	9.18	0	3.67	3.63
m4	9.37	11.01	25.06	0.12	0.35	0.46	3.28	6.61	3.51	3.28	3.67
m5	19.39	24.01	33.77	-0.51	0.24	0.69	14.51	1.34	0.92	14.51	4.92
MA	11.11	15.5	28.45	0.22	0.41	0.42	6.16	9.88	1.27	6.16	5.27
FA	2.68	4.71	10.64	-0.67	-0.16	0.06	2.12	21.09	0	2.12	6.72
OA	6.89	10.11	19.54	-0.22	0.13	0.24	4.14	15.48	0.64	4.14	6

CORR	GER20	GER10	GER05	FER20	FER10	FER05	V_UVE	UV_VE	THE	TLE	STD20
f1	2.23	3.35	5.2	0.94	1.07	0.97	1.86	0.23	0.37	1.86	4.53
f2	2.81	3.56	10.13	0.09	0.1	0.08	1.31	0	0.75	1.88	6.6
f3	3.34	4.63	10.03	0.23	0.44	0.58	0.51	0	1.03	2.31	5.41
f4	4.91	6.75	16.36	-0.28	-0.01	0.25	3.89	0.26	0.41	4.29	8.4
f5	0	1.17	9.94	-0.34	-0.59	-0.9	0	0.84	0	0	7.58
m1	16.95	17.92	22.76	0.23	0.3	0.43	12.35	0	4.6	12.35	3.16
m2	6.21	6.83	32.92	-1.44	-1.55	-0.77	1.86	1.98	4.35	1.86	5.93
m3	2.75	3.36	6.12	0.08	0.18	0.29	2.75	0	0	2.75	3.21
m4	21.55	22.01	31.62	0.63	0.64	0.44	16.86	0.46	4.68	16.86	2.94
m5	14.78	15.83	25.86	0.67	0.66	0.35	14.51	0.27	0.26	14.51	3.75
MA	12.45	13.19	23.85	0.03	0.05	0.15	9.67	0.54	2.78	9.67	3.8
FA	2.66	3.89	10.33	0.13	0.2	0.19	1.51	0.26	0.51	2.07	6.5
OA	7.55	8.54	17.09	0.08	0.12	0.17	5.59	0.4	1.65	5.87	5.15

YIN	GER20	GER10	GER05	FER20	FER10	FER05	V_UVE	UV_VE	THE	TLE	STD20
f1	86.62	89.96	93.68	4.7	0.51	-0.72	0	100	0.37	85.13	15.62
f2	89.68	93.25	95.31	4.85	-2.26	-2.91	0	99.78	0	88.18	19.86
f3	80.46	86.89	91.52	7.34	0.98	-0.04	0	100	0.26	76.35	15.31
f4	87.93	91	93.05	4.57	-0.9	-0.3	0	100	0.2	84.25	20.27
f5	90.35	91.81	93.86	4.72	0.69	-1	0	100	0	88.01	12.62
m1	11.14	15.5	27.12	0.7	0.48	0.28	0	100	2.18	6.3	4.76
m2	34.16	37.89	42.86	1.43	0.6	0.62	0	100	0.62	30.43	6.01
m3	21.71	25.69	31.19	0.87	-0.17	-0.26	0	100	0	18.96	5.71
m4	1.41	3.04	11.48	0.61	0.6	0.51	0	100	0	0	2.82
m5	5.8	8.84	15.04	0.57	0.41	0.49	0	100	0.92	2.77	4.11
MA	14.84	18.19	25.54	0.83	0.38	0.33	0	100	0.74	11.69	4.68
FA	87.01	90.58	93.48	5.24	-0.2	-0.99	0	99.96	0.17	84.39	16.74
OA	50.93	54.39	59.51	3.03	0.09	-0.33	0	99.98	0.46	48.04	10.71

PRAAT	GER20	GER10	GER05	FER20	FER10	FER05	V_UVE	UV_VE	THE	TLE	STD20
f1	5.2	7.06	8.92	1.13	1.26	1.17	1.49	0	0.74	4.46	4.58
f2	3.19	3.94	11.07	0.26	0.3	0.34	1.13	0	0.38	2.81	6.67
f3	3.08	3.08	9	0.54	0.54	0.66	0.51	0	0.51	2.57	4.85
f4	4.09	5.32	13.5	-0.11	-0.03	0.22	3.89	0	0	3.89	7.63
f5	0	0.88	8.48	0.04	-0.26	-0.76	0	0.5	0	0	7.27
m1	9.2	11.14	20.1	0.48	0.58	0.45	8.96	0	0	8.96	3.79
m2	0.62	2.48	31.06	-1.35	-1.51	-0.34	0	0.66	0.62	0	6.44
m3	1.22	1.83	6.42	0.21	0.32	0.31	1.22	0	0	1.22	3.53
m4	14.52	15.22	27.17	0.66	0.61	0.42	12.65	0	1.87	12.65	3.2
m5	4.09	5.94	20.05	0.83	0.79	0.37	4.09	0	0	4.09	4.04
MA	5.93	7.32	20.96	0.17	0.16	0.24	5.38	0.13	0.5	5.38	4.2
FA	3.11	4.06	10.19	0.37	0.36	0.33	1.4	0.1	0.33	2.75	6.2
OA	4.52	5.69	15.58	0.27	0.26	0.28	3.39	0.12	0.41	4.06	5.2

Table A.2.24. Results for 25 dB White noise for low resolution speech signals.

DHO	GER20	GER10	GER05	FER20	FER10	FER05	V_UVE	UV_VE	THE	TLE	STD20
f1	1.49	2.23	3.72	-0.1	-0.37	-0.5	0	16.97	0	0	4.65
f2	1.13	2.25	10.32	-0.74	-0.48	0.19	0	32.33	0	0	6.77
f3	0.26	2.83	9	-0.85	-0.23	0.06	0	11.75	0	0	6.02
f4	10.43	13.7	20.25	-0.59	-0.04	-0.07	6.34	5.93	0	6.34	9.35
f5	0.58	2.63	12.28	-0.21	0.35	0.71	0.58	47.99	0	0.58	7.68
m1	13.32	23.97	42.37	0.69	0.24	-0.14	4.84	9.14	3.63	4.84	7.44
m2	6.21	9.32	29.81	2.1	2.42	1.43	2.48	23.13	0	2.48	6.53
m3	3.67	5.2	11.93	-0.73	-0.81	-0.43	3.67	19.02	0	3.67	4
m4	21.08	23.65	36.3	0.09	0.33	0.39	7.03	18.91	1.17	7.03	4.18
m5	13.46	18.21	28.36	-0.64	0.2	0.65	8.18	3.21	0.66	8.18	4.97
MA	11.55	16.07	29.76	0.3	0.48	0.38	5.24	14.68	1.09	5.24	5.42
FA	2.78	4.73	11.11	-0.5	-0.15	0.08	1.38	22.99	0	1.38	6.89
OA	7.16	10.4	20.43	-0.1	0.16	0.23	3.31	18.84	0.55	3.31	6.16

CORR	GER20	GER10	GER05	FER20	FER10	FER05	V_UVE	UV_VE	THE	TLE	STD20
f1	3.72	4.46	5.95	1.11	1.14	1	3.35	0.23	0.37	3.35	4.35
f2	4.69	5.25	11.44	0.04	0	0.08	3.19	0	0.75	3.75	6.5
f3	4.88	6.17	11.31	0.33	0.51	0.62	2.31	0	0.77	4.11	5.28
f4	7.77	9.41	18	-0.42	-0.06	0.22	6.75	0.26	0.41	7.16	8.3
f5	0	1.17	9.94	-0.3	-0.55	-0.85	0	0.5	0	0	7.56
m1	17.92	18.89	23.97	0.22	0.3	0.4	13.8	0	4.12	13.8	3.2
m2	9.32	10.56	33.54	-1.53	-1.57	-0.77	4.97	1.98	4.35	4.97	5.96
m3	4.89	5.5	8.56	0.12	0.22	0.31	4.89	0	0	4.89	3.3
m4	22.95	23.19	33.02	0.57	0.61	0.43	18.27	0.46	4.68	18.27	2.92
m5	19.66	20.45	29.16	0.61	0.6	0.32	19.26	0.27	0.4	19.26	3.7
MA	14.95	15.72	25.65	0	0.03	0.14	12.24	0.54	2.71	12.24	3.82
FA	4.21	5.29	11.33	0.15	0.21	0.21	3.12	0.2	0.46	3.67	6.4
OA	9.58	10.5	18.49	0.08	0.12	0.18	7.68	0.37	1.58	7.96	5.11

YIN	GER20	GER10	GER05	FER20	FER10	FER05	V_UVE	UV_VE	THE	TLE	STD20
f1	87.36	90.71	92.94	3.36	1.84	0.85	0	100	0.37	85.87	15.67
f2	90.06	93.81	95.12	8.16	1.28	1.56	0	100	0	89.12	18.36
f3	83.03	90.49	93.06	9.67	0.12	-0.09	0	99.82	0.26	78.92	18.16
f4	85.07	88.75	92.43	7.02	0	-0.92	0	100	0.2	80.98	18.6
f5	91.52	92.11	94.44	2.59	0.84	-0.56	0	100	0	89.18	10.99
m1	12.11	17.19	26.63	0.67	0.5	0.33	0	100	2.42	6.78	4.8
m2	34.16	38.51	44.72	1.61	0.56	0.64	0	100	0.62	30.43	6.48
m3	21.41	25.08	30.28	0.42	-0.29	-0.16	0	100	0	18.96	5.23
m4	2.34	3.75	12.65	0.62	0.57	0.48	0	100	0	0	2.91
m5	6.07	8.97	16.23	0.54	0.39	0.51	0	100	1.06	3.03	4.09
MA	15.22	18.7	26.1	0.77	0.35	0.36	0	100	0.82	11.84	4.7
FA	87.41	91.17	93.6	6.16	0.82	0.17	0	99.96	0.17	84.81	16.36
OA	51.31	54.94	59.85	3.47	0.58	0.26	0	99.98	0.49	48.33	10.53

PRAAT	GER20	GER10	GER05	FER20	FER10	FER05	V_UVE	UV_VE	THE	TLE	STD20
f1	7.43	8.18	10.41	1.44	1.28	1.19	4.46	0	0	7.43	3.82
f2	4.13	4.88	12.95	0.17	0.21	0.33	2.06	0	0.38	3.75	6.73
f3	3.6	3.6	9.77	0.57	0.57	0.61	1.03	0	0.51	3.08	4.89
f4	6.13	6.95	14.93	-0.29	-0.21	0.07	5.93	0	0	5.93	7.5
f5	0	0.88	8.19	0.05	-0.26	-0.8	0	0.5	0	0	7.31
m1	10.17	11.62	19.85	0.38	0.53	0.43	10.17	0	0	10.17	3.66
m2	2.48	4.97	31.68	-1.55	-1.44	-0.32	1.86	0.66	0.62	1.86	6.22
m3	1.83	2.45	7.03	0.18	0.3	0.25	1.83	0	0	1.83	3.61
m4	15.22	15.93	27.4	0.64	0.6	0.42	13.35	0	1.87	13.35	3.17
m5	7.52	9.1	22.43	0.85	0.82	0.39	7.52	0	0	7.52	4.04
MA	7.45	8.81	21.68	0.1	0.16	0.23	6.95	0.13	0.5	6.95	4.14
FA	4.26	4.9	11.25	0.39	0.32	0.28	2.7	0.1	0.18	4.04	6.05
OA	5.85	6.86	16.46	0.24	0.24	0.26	4.82	0.12	0.34	5.49	5.09

Table A.2.25. Results for 20 dB White noise for low resolution speech signals.

DHO	GER20	GER10	GER05	FER20	FER10	FER05	V_UVE	UV_VE	THE	TLE	STD20
f1	1.86	2.97	4.09	-0.75	-0.48	-0.56	0	29.41	0	0	4.3
f2	2.81	4.88	12.38	-0.96	-0.39	0.17	0	25.86	0	0.94	7.36
f3	2.31	5.14	10.8	-0.89	-0.17	0.13	0	19.71	0	0	6.2
f4	6.54	10.02	17.59	-0.59	-0.1	-0.05	4.5	5.67	0	4.5	9.19
f5	4.68	6.14	15.5	0.06	0.48	0.82	0.58	52.01	0	1.17	7.52
m1	12.59	22.76	41.89	0.52	0.22	-0.06	4.12	9.14	3.87	4.12	7.08
m2	6.21	10.56	29.19	1.59	2.19	1.21	1.86	29.96	0	1.86	6.93
m3	3.67	4.59	12.23	-0.93	-0.81	-0.39	3.67	15.74	0	3.67	3.79
m4	23.65	24.82	37.7	0.14	0.35	0.35	6.32	10.25	1.41	6.32	3.73
m5	16.09	21.37	31.4	-0.7	0.21	0.66	8.31	1.47	1.58	8.31	5.02
MA	12.44	16.82	30.48	0.12	0.43	0.35	4.86	13.31	1.37	4.86	5.31
FA	3.64	5.83	12.07	-0.62	-0.13	0.1	1.02	26.53	0	1.32	6.91
OA	8.04	11.33	21.28	-0.25	0.15	0.23	2.94	19.92	0.69	3.09	6.11

CORR	GER20	GER10	GER05	FER20	FER10	FER05	V_UVE	UV_VE	THE	TLE	STD20
f1	5.58	5.95	7.43	1.23	1.14	0.91	5.58	0	0	5.58	4.08
f2	7.5	8.26	14.07	-0.02	-0.12	-0.05	6	0	0.75	6.57	6.55
f3	8.48	9.25	14.14	0.55	0.61	0.61	6.43	0	0.51	7.97	5.18
f4	22.49	23.52	30.88	-1.13	-0.9	-0.25	21.88	0.26	0.41	22.09	7.92
f5	0.29	1.75	10.82	-0.44	-0.62	-0.99	0.29	0.34	0	0.29	7.83
m1	23	23.49	27.6	0.22	0.21	0.34	19.61	0	3.39	19.61	2.99
m2	17.39	18.63	38.51	-1.89	-1.92	-0.78	14.29	1.76	3.11	14.29	5.89
m3	15.9	16.21	18.96	0.23	0.3	0.37	15.9	0	0	15.9	3.3
m4	29.51	29.74	37.47	0.37	0.4	0.35	24.59	0.46	4.92	24.59	2.83
m5	34.43	34.83	40.11	0.23	0.25	0.21	34.17	0.27	0.26	34.17	3.19
MA	24.05	24.58	32.53	-0.17	-0.15	0.1	21.71	0.5	2.34	21.71	3.64
FA	8.87	9.75	15.47	0.04	0.02	0.05	8.04	0.12	0.33	8.5	6.31
OA	16.46	17.16	24	-0.07	-0.07	0.07	14.87	0.31	1.34	15.1	4.98

YIN	GER20	GER10	GER05	FER20	FER10	FER05	V_UVE	UV_VE	THE	TLE	STD20
f1	88.1	91.08	95.17	5.03	2.53	0.3	0	100	0.37	86.62	17.91
f2	89.31	91.74	94.93	5.58	1.39	0.73	0	100	0	86.3	17.55
f3	83.55	88.95	93.57	7.7	1.43	0.8	0	100	0	76.35	16.89
f4	77.71	84.46	89.16	7.3	0.11	-0.79	0	100	0	73.82	19.28
f5	91.23	92.4	94.74	2.79	-1.33	-0.06	0	100	0	88.01	13.55
m1	12.59	17.19	28.33	0.72	0.5	0.38	0	99.83	3.15	7.02	4.82
m2	34.78	38.51	49.07	1.53	0.63	0.64	0	99.78	0.62	30.43	6.68
m3	24.16	27.83	34.86	0.37	-0.25	-0.21	0	100	0	20.49	5.39
m4	2.58	4.92	13.35	0.58	0.58	0.47	0	100	0.47	0	3.06
m5	8.31	12.8	19.39	0.49	0.44	0.52	0	100	2.51	2.64	4.5
MA	16.48	20.25	29	0.74	0.38	0.36	0	99.92	1.35	12.12	4.89
FA	85.98	89.72	93.51	5.68	0.83	0.19	0	100	0.07	82.22	17.03
OA	51.23	54.99	61.26	3.21	0.61	0.28	0	99.96	0.71	47.17	10.96

PRAAT	GER20	GER10	GER05	FER20	FER10	FER05	V_UVE	UV_VE	THE	TLE	STD20
f1	7.06	7.81	8.92	1.45	1.29	1.16	7.06	0	0	7.06	3.62
f2	7.5	8.26	14.82	0.17	0.17	0.28	4.88	0	0.38	7.13	6.56
f3	8.23	8.74	13.62	0.84	0.74	0.68	6.17	0	0.51	7.71	4.96
f4	24.95	25.36	31.9	-0.74	-0.75	-0.29	24.74	0	0	24.95	7.31
f5	0	0.88	8.77	0	-0.31	-0.87	0	0.5	0	0	7.4
m1	13.56	15.01	23.49	0.4	0.57	0.43	13.56	0	0	13.56	3.75
m2	12.42	14.91	37.89	-1.93	-1.96	-0.47	12.42	0.66	0	12.42	6.21
m3	10.4	10.7	15.6	0.28	0.35	0.29	10.4	0	0	10.4	3.64
m4	16.39	17.1	28.34	0.65	0.6	0.42	15.22	0	1.17	15.22	3.21
m5	18.07	18.73	28.5	0.62	0.63	0.35	18.07	0	0	18.07	3.62
MA	14.17	15.29	26.76	0	0.04	0.2	13.94	0.13	0.23	13.94	4.09
FA	9.55	10.21	15.61	0.35	0.23	0.19	8.57	0.1	0.18	9.37	5.97
OA	11.86	12.75	21.18	0.18	0.13	0.2	11.25	0.12	0.21	11.65	5.03

Table A.2.26. Results for 15 dB White noise for low resolution speech signals.

DHO	GER20	GER10	GER05	FER20	FER10	FER05	V_UVE	UV_VE	THE	TLE	STD20
f1	1.12	1.86	3.72	-0.6	-0.42	-0.5	0	40.05	0	0	4.18
f2	4.5	6.57	13.51	-0.74	-0.15	0.38	0	37.07	0	1.5	7.42
f3	2.83	5.66	11.57	-0.71	-0.17	0.16	0	50.81	0	0	6.33
f4	13.09	18.4	24.13	0.43	0.49	0.28	5.73	20.88	0	9.61	10.83
f5	7.6	9.06	17.84	-0.01	0.43	0.67	0	46.64	0	0	7.61
m1	10.9	22.76	41.89	0.76	0.26	-0.08	1.45	14.12	4.84	1.45	7.23
m2	4.35	5.59	27.95	1.69	1.89	1.15	1.86	38.55	0	1.86	5.99
m3	5.5	7.03	13.46	-0.55	-0.64	-0.25	3.98	34.1	0	3.98	4.14
m4	14.75	16.16	29.98	0.22	0.43	0.39	3.28	34.17	0.23	3.28	3.67
m5	19.79	24.54	34.17	-0.32	0.45	0.72	9.37	11.38	1.32	9.37	4.99
MA	11.06	15.22	29.49	0.36	0.48	0.38	3.99	26.46	1.28	3.99	5.2
FA	5.83	8.31	14.15	-0.33	0.04	0.2	1.15	39.09	0	2.22	7.27
OA	8.44	11.76	21.82	0.02	0.26	0.29	2.57	32.78	0.64	3.1	6.24

CORR	GER20	GER10	GER05	FER20	FER10	FER05	V_UVE	UV_VE	THE	TLE	STD20
f1	12.27	12.64	14.13	1.34	1.26	1.05	12.27	0	0	12.27	4.17
f2	14.26	15.2	21.76	-0.03	-0.18	-0.1	12.95	0	0.94	13.32	6.89
f3	16.97	17.48	21.59	0.72	0.86	0.97	16.2	0	0.51	16.45	5
f4	39.06	39.67	45.6	-1.5	-1.23	-0.5	38.04	0.26	0.41	38.65	7.93
f5	5.85	7.31	17.84	-0.22	-0.41	-1.06	5.56	0.17	0	5.85	8.17
m1	35.59	36.08	39.23	0.13	0.12	0.33	33.41	0	2.18	33.41	3.1
m2	31.68	32.92	48.45	-2.94	-3	-1.56	29.19	1.32	2.48	29.19	5.33
m3	38.84	38.84	41.9	0.1	0.1	0.17	38.84	0	0	38.84	3.3
m4	44.5	44.73	50.82	0.3	0.34	0.34	40.52	0.23	3.98	40.52	2.93
m5	51.72	52.11	55.01	-0.09	-0.06	0	51.19	0.13	0.53	51.19	3.08
MA	40.46	40.94	47.08	-0.5	-0.5	-0.14	38.63	0.34	1.83	38.63	3.55
FA	17.68	18.46	24.18	0.06	0.06	0.07	17	0.09	0.37	17.31	6.43
OA	29.07	29.7	35.63	-0.22	-0.22	-0.04	27.81	0.21	1.1	27.97	4.99

YIN	GER20	GER10	GER05	FER20	FER10	FER05	V_UVE	UV_VE	THE	TLE	STD20
f1	86.99	90.33	94.42	0.58	-4.04	0.5	0	100	0.37	83.27	16.14
f2	85.93	88.93	93.06	4.05	0.37	0.01	0	100	0	82.36	15.53
f3	76.86	84.58	89.46	6.63	1.14	-0.45	0	100	0.26	72.24	15.94
f4	70.76	76.89	85.07	6.2	0.77	-1.35	0	100	0.2	64.42	18.35
f5	87.13	90.64	93.27	6.37	-1.23	-0.88	0	100	0	83.04	16.64
m1	14.77	21.07	30.75	0.71	0.68	0.29	0	100	4.36	7.99	4.89
m2	34.78	40.37	48.45	1.35	0.56	0.8	0	100	0.62	30.43	6.66
m3	20.49	26.3	35.47	0.49	-0.34	-0.42	0	100	0	15.9	6.81
m4	3.51	5.62	13.82	0.58	0.65	0.44	0	100	0.47	0	3.06
m5	10.95	16.36	25.59	0.26	0.3	0.45	0	100	3.56	2.51	4.94
MA	16.9	21.94	30.82	0.68	0.37	0.31	0	100	1.8	11.37	5.27
FA	81.53	86.28	91.06	4.76	-0.6	-0.44	0	100	0.17	77.07	16.52
OA	49.22	54.11	60.94	2.72	-0.11	-0.06	0	100	0.98	44.22	10.89

PRAAT	GER20	GER10	GER05	FER20	FER10	FER05	V_UVE	UV_VE	THE	TLE	STD20
f1	12.64	13.38	14.87	1.23	1.05	0.87	12.64	0	0	12.64	4.13
f2	14.45	15.2	21.58	0.16	0.07	0.19	13.13	0	0.38	14.07	6.64
f3	19.28	19.28	23.14	0.76	0.76	0.87	18.25	0	0.51	18.77	4.76
f4	43.15	43.35	48.06	-1.28	-1.22	-0.67	41.92	0	0	43.15	7.19
f5	7.6	8.48	16.96	-0.15	-0.47	-1.09	7.6	0	0	7.6	7.64
m1	21.55	22.52	29.3	0.27	0.37	0.31	21.55	0	0	21.55	3.64
m2	26.09	27.95	46.58	-3	-2.94	-0.95	26.09	0	0	26.09	5.83
m3	34.56	34.56	38.23	0.28	0.28	0.28	34.56	0	0	34.56	3.54
m4	22.72	23.19	33.96	0.62	0.62	0.34	21.55	0	1.17	21.55	3.24
m5	40.77	41.03	46.04	0.3	0.3	0.17	40.77	0	0	40.77	3.27
MA	29.13	29.85	38.82	-0.3	-0.27	0.03	28.9	0	0.23	28.9	3.9
FA	19.42	19.94	24.92	0.14	0.04	0.03	18.71	0	0.18	19.25	6.07
OA	24.28	24.89	31.87	-0.08	-0.12	0.03	23.81	0	0.21	24.07	4.99

Table A.2.27. Results for 10 dB White noise for low resolution speech signals.

DHO	GER20	GER10	GER05	FER20	FER10	FER05	V_UVE	UV_VE	THE	TLE	STD20
f1	5.2	6.32	7.43	-0.66	-0.55	-0.52	0	61.09	0	0	4.3
f2	3.75	5.82	13.7	-0.07	0.03	0.32	0.38	58.19	0	0.38	7.82
f3	2.06	4.11	11.05	-0.9	-0.42	0.11	1.03	55.15	0	1.03	5.96
f4	20.25	27.4	34.15	-1.63	-0.12	0.32	6.34	29.64	0	6.54	12.58
f5	2.92	4.68	13.74	0.04	0.61	0.7	0.29	65.27	0	0.29	7.79
m1	14.04	26.39	45.28	0.81	0.54	-0.07	0	24.92	0.48	0	7.53
m2	4.97	8.07	32.3	1.82	2.06	1.13	1.86	48.9	0	1.86	6.75
m3	3.98	4.28	13.15	-0.73	-0.68	-0.32	3.98	59.67	0	3.98	3.92
m4	18.97	20.61	31.62	0.58	0.66	0.63	7.73	48.29	2.11	7.73	3.68
m5	19.26	25.99	36.68	-0.36	0.25	0.61	3.43	33.07	1.85	3.43	5.56
MA	12.24	17.07	31.8	0.42	0.57	0.4	3.4	42.97	0.89	3.4	5.49
FA	6.84	9.67	16.02	-0.64	-0.09	0.19	1.61	53.87	0	1.65	7.69
OA	9.54	13.37	23.91	-0.11	0.24	0.29	2.5	48.42	0.44	2.52	6.59

CORR	GER20	GER10	GER05	FER20	FER10	FER05	V_UVE	UV_VE	THE	TLE	STD20
f1	31.6	31.6	32.34	0.77	0.77	0.77	31.6	0	0	31.6	3.77
f2	34.33	35.08	41.84	-1.02	-1.03	-0.79	33.4	0	0.56	33.77	7.15
f3	43.96	44.47	46.79	0.6	0.8	0.87	43.7	0	0.26	43.7	4.77
f4	61.55	61.96	65.03	-2.22	-1.96	-0.98	60.74	0	0.41	61.15	7.94
f5	23.1	24.27	34.21	-0.62	-0.93	-1.3	22.81	0	0	23.1	8.63
m1	57.14	57.63	59.32	-0.05	-0.07	0.1	55.69	0	1.45	55.69	3.25
m2	53.42	53.42	63.98	-4.08	-4.08	-1.95	53.42	0.88	0	53.42	4.84
m3	64.53	64.53	65.14	-0.54	-0.54	-0.4	64.53	0	0	64.53	2.92
m4	68.15	68.15	72.13	0.01	0.01	0.15	65.81	0.23	2.34	65.81	2.9
m5	68.34	68.34	69.66	-0.61	-0.61	-0.39	67.81	0.13	0.53	67.81	2.53
MA	62.31	62.41	66.04	-1.05	-1.06	-0.5	61.45	0.25	0.86	61.45	3.29
FA	38.91	39.48	44.04	-0.5	-0.47	-0.29	38.45	0	0.25	38.66	6.45
OA	50.61	50.94	55.04	-0.78	-0.76	-0.39	49.95	0.12	0.56	50.06	4.87

YIN	GER20	GER10	GER05	FER20	FER10	FER05	V_UVE	UV_VE	THE	TLE	STD20
f1	72.86	77.32	84.01	3.04	0.44	-0.58	0	100	0.74	65.43	13.55
f2	75.61	82.36	88.37	4.81	0.47	1.36	0	100	0.19	69.98	17.28
f3	59.9	70.95	80.46	4.03	-0.77	0.12	0	100	0.51	50.9	16.11
f4	63.8	73.42	81.8	6.36	1.84	-0.1	0	100	1.02	55.83	19.49
f5	76.32	80.99	85.96	2.33	-2.25	-0.85	0	100	0	71.35	16
m1	16.22	22.28	34.38	0.71	0.53	0.46	0	100	6.78	5.81	5.41
m2	31.68	41.61	50.93	0.95	0.16	0.32	0	100	2.48	22.36	8.84
m3	16.51	25.99	40.98	1.01	-0.26	-0.19	0	100	0.61	8.87	8.04
m4	6.79	13.11	22.95	0.35	0.58	0.41	0	100	2.58	0	4.36
m5	22.69	32.32	42.74	0.28	0.35	0.5	0	100	8.97	2.64	6.52
MA	18.78	27.06	38.4	0.66	0.27	0.3	0	100	4.28	7.94	6.63
FA	69.7	77.01	84.12	4.11	-0.06	-0.01	0	100	0.49	62.7	16.48
OA	44.24	52.04	61.26	2.39	0.11	0.14	0	100	2.39	35.32	11.56

PRAAT	GER20	GER10	GER05	FER20	FER10	FER05	V_UVE	UV_VE	THE	TLE	STD20
f1	32.71	32.71	33.83	0.9	0.9	0.71	32.71	0	0	32.71	3.55
f2	38.27	38.46	44.47	-0.83	-0.73	-0.27	36.21	0	0	38.27	6.85
f3	45.5	45.76	47.04	0.55	0.66	0.8	45.24	0	0	45.5	4.57
f4	64.42	64.42	66.87	-1.81	-1.81	-0.7	64.42	0	0	64.42	6.84
f5	28.65	29.82	37.72	-0.31	-0.64	-1.49	27.19	0	0	28.65	8.34
m1	40.19	41.16	45.04	0.1	0.19	0.32	40.19	0	0	40.19	3.47
m2	53.42	53.42	66.46	-4.24	-4.24	-1.51	53.42	0	0	53.42	5.16
m3	59.94	59.94	61.47	-0.27	-0.27	-0.35	59.94	0	0	59.94	3.35
m4	43.09	43.56	51.29	0.56	0.55	0.32	43.09	0	0	43.09	3.27
m5	60.55	60.55	63.32	-0.18	-0.18	-0.11	60.55	0	0	60.55	2.95
MA	51.44	51.73	57.52	-0.81	-0.79	-0.26	51.44	0	0	51.44	3.64
FA	41.91	42.24	45.99	-0.3	-0.33	-0.19	41.16	0	0	41.91	6.03
OA	46.68	46.98	51.75	-0.55	-0.56	-0.23	46.3	0	0	46.68	4.83

Table A.2.28. Results for 5 dB White noise for low resolution speech signals.

DHO	GER20	GER10	GER05	FER20	FER10	FER05	V_UVE	UV_VE	THE	TLE	STD20
f1	10.78	14.87	16.36	-0.59	-0.23	-0.24	0	93.21	0	0.74	7.23
f2	7.13	11.44	20.45	-0.36	-0.42	0.43	0	92.46	0	0	9.67
f3	10.54	14.65	20.31	-0.53	-0.21	0.2	0	83.54	0	0	7.21
f4	32.92	47.85	55.42	-4.14	-0.39	-0.44	0.2	87.11	0	5.93	18.58
f5	13.16	18.42	28.07	-1.37	0.14	0.52	0	96.14	0	5.56	10.15
m1	21.55	31.72	53.27	1.2	0.73	-0.18	0	88.54	4.6	0	7.63
m2	16.15	22.36	46.58	0.73	1.6	1.07	2.48	83.04	0	2.48	7.94
m3	1.22	6.42	21.1	-0.95	-0.67	-0.43	0.31	85.57	0	0.31	5.87
m4	20.14	26	40.28	0.43	0.76	0.68	1.41	77.45	1.17	1.41	5.19
m5	26.65	37.34	50.92	-1.29	0.15	0.69	1.32	93.71	1.72	1.32	6.8
MA	17.14	24.77	42.43	0.02	0.52	0.37	1.1	85.66	1.5	1.1	6.68
FA	14.91	21.45	28.12	-1.4	-0.22	0.1	0.04	90.49	0	2.45	10.57
OA	16.02	23.11	35.28	-0.69	0.15	0.23	0.57	88.08	0.75	1.77	8.63

CORR	GER20	GER10	GER05	FER20	FER10	FER05	V_UVE	UV_VE	THE	TLE	STD20
f1	62.45	62.45	62.45	0.38	0.38	0.38	62.45	0	0	62.45	3.3
f2	62.48	62.85	66.79	-3.04	-2.84	-1.71	61.35	0	0.38	62.1	6.64
f3	72.75	72.75	74.04	-0.05	-0.05	0.59	72.75	0	0	72.75	4.64
f4	78.12	78.12	79.55	-2.27	-2.27	-1.26	77.51	0	0.2	77.91	7
f5	59.06	59.94	64.62	-1.09	-1.87	-1.94	57.6	0	0.58	58.48	9.04
m1	88.86	88.86	88.86	-1.27	-1.27	-1.27	88.14	0	0.73	88.14	2.08
m2	75.78	75.78	81.37	-5.05	-5.05	-3.13	75.78	0	0	75.78	4.67
m3	87.77	87.77	88.38	-1.72	-1.72	-1.35	87.77	0	0	87.77	3.01
m4	95.55	95.55	95.55	-0.44	-0.44	-0.44	95.55	0	0	95.55	2.1
m5	92.48	92.48	92.61	-0.8	-0.8	-0.68	91.82	0	0.66	91.82	2.19
MA	88.09	88.09	89.35	-1.86	-1.86	-1.37	87.81	0	0.28	87.81	2.81
FA	66.97	67.22	69.49	-1.22	-1.33	-0.79	66.33	0	0.23	66.74	6.12
OA	77.53	77.66	79.42	-1.54	-1.59	-1.08	77.07	0	0.26	77.27	4.47

YIN	GER20	GER10	GER05	FER20	FER10	FER05	V_UVE	UV_VE	THE	TLE	STD20
f1	54.65	65.43	76.95	5.04	-0.45	-0.24	0	100	0	46.84	15.48
f2	59.47	71.86	81.99	6.85	0.49	-0.4	0	100	0	52.16	18.18
f3	49.36	63.75	73.78	4.78	-0.1	0.65	0	100	0.51	35.99	16.08
f4	65.24	73.21	84.87	6.26	-0.21	0.6	0	100	1.23	53.99	20.22
f5	59.65	72.51	80.41	6.7	-0.1	-1.06	0	100	0	55.26	19.45
m1	21.31	30.51	44.79	1.35	0.98	0.54	0	100	9.2	2.91	6.59
m2	31.06	40.37	52.8	1.98	0.4	0.19	0	100	5.59	13.04	8.92
m3	23.85	36.7	52.29	0.44	-0.09	-0.21	0	100	3.98	8.56	9.72
m4	17.1	22.25	34.66	0.21	0.58	0.46	0	99.77	6.56	0.7	4.28
m5	35.62	46.97	59.63	0.57	0.65	0.51	0	100	16.89	2.11	7.71
MA	25.79	35.36	48.83	0.91	0.5	0.3	0	99.95	8.44	5.47	7.44
FA	57.67	69.35	79.6	5.93	-0.07	-0.09	0	100	0.35	48.85	17.88
OA	41.73	52.36	64.22	3.42	0.22	0.1	0	99.98	4.4	27.16	12.66

PRAAT	GER20	GER10	GER05	FER20	FER10	FER05	V_UVE	UV_VE	THE	TLE	STD20
f1	69.14	69.14	69.14	0.52	0.52	0.52	64.31	0	0	69.14	2.75
f2	70.92	70.92	74.11	-3.83	-3.83	-2.45	67.54	0	0	70.92	5.82
f3	79.43	79.43	79.43	0.58	0.58	0.58	77.89	0	0	79.43	3.62
f4	82.21	82.21	83.64	-2.71	-2.71	-1.47	80.37	0	0	82.21	7.12
f5	70.18	71.05	74.27	-0.08	-1.11	-1.88	65.2	0	0	70.18	9.95
m1	77	77.48	78.45	0.03	-0.01	-0.1	77	0	0	77	3.1
m2	77.64	77.64	83.23	-4.7	-4.7	-2.8	77.64	0	0	77.64	4.21
m3	88.99	88.99	88.99	-1.32	-1.32	-1.32	88.99	0	0	88.99	3.12
m4	83.37	83.37	84.78	0.01	0.01	0.1	83.37	0	0	83.37	2.78
m5	86.54	86.54	86.94	-0.29	-0.29	-0.05	86.54	0	0	86.54	2.54
MA	82.71	82.81	84.48	-1.26	-1.26	-0.83	82.71	0	0	82.71	3.15
FA	74.38	74.55	76.12	-1.1	-1.31	-0.94	71.06	0	0	74.38	5.85
OA	78.54	78.68	80.3	-1.18	-1.29	-0.89	76.89	0	0	78.54	4.5

Table A.2.29. Results for 0 dB White noise for low resolution speech signals.

DHO	GER20	GER10	GER05	FER20	FER10	FER05	V_UVE	UV_VE	THE	TLE	STD20
f1	20.07	25.28	30.11	-1.11	-0.11	-0.09	0	99.1	0	0	8.44
f2	12.95	18.95	28.71	-0.44	0.06	0.37	0	98.92	0	0	10.74
f3	24.68	36.5	44.22	0.87	0.12	0.43	0	96.38	0	0	12.83
f4	47.65	69.73	76.89	-3.89	2.5	1.84	0	100	0	0.82	27.33
f5	22.51	33.04	43.57	1.09	1.02	1.17	0	100	0	0.29	14.9
m1	30.99	43.58	60.53	0.49	0.29	-0.04	0	96.68	2.91	0	7.65
m2	19.88	33.54	52.17	-0.58	0.42	0.88	0	100	0	0	10.46
m3	7.34	22.02	35.17	-2.2	-0.62	-0.36	0	100	0	0	8.29
m4	34.89	42.39	54.8	0.32	0.34	0.38	0	100	5.15	0	5.47
m5	36.68	50.4	63.98	-2.25	-0.07	0.58	0.13	98.8	1.45	0.13	7.97
MA	25.96	38.39	53.33	-0.84	0.07	0.29	0.03	99.09	1.9	0.03	7.97
FA	25.57	36.7	44.7	-0.7	0.72	0.74	0	98.88	0	0.22	14.85
OA	25.76	37.54	49.02	-0.77	0.39	0.52	0.01	98.99	0.95	0.12	11.41

CORR	GER20	GER10	GER05	FER20	FER10	FER05	V_UVE	UV_VE	THE	TLE	STD20
f1	99.26	99.26	99.26	4.83	4.83	4.83	99.26	0	0	99.26	5.69
f2	99.25	99.25	99.25	-0.16	-0.16	-0.16	99.25	0	0	99.25	3.79
f3	99.23	99.23	99.23	1.04	1.04	1.04	99.23	0	0	99.23	4.1
f4	99.8	99.8	99.8	0	0	0	99.8	0	0	99.8	0
f5	97.66	97.66	98.83	-12.42	-12.42	-7.02	97.66	0	0	97.66	6.82
m1	100	100	100	0	0	0	100	0	0	100	0
m2	100	100	100	0	0	0	100	0	0	100	0
m3	99.69	99.69	99.69	-2.94	-2.94	-2.94	99.69	0	0	99.69	0
m4	100	100	100	0	0	0	100	0	0	100	0
m5	100	100	100	0	0	0	100	0	0	100	0
MA	99.94	99.94	99.94	-0.59	-0.59	-0.59	99.94	0	0	99.94	0
FA	99.04	99.04	99.27	-1.34	-1.34	-0.26	99.04	0	0	99.04	4.08
OA	99.49	99.49	99.61	-0.96	-0.96	-0.42	99.49	0	0	99.49	2.04

YIN	GER20	GER10	GER05	FER20	FER10	FER05	V_UVE	UV_VE	THE	TLE	STD20
f1	56.13	67.66	78.44	4.51	0.81	0.43	0	100	0.74	44.98	16.1
f2	65.67	74.48	84.05	6.55	1.93	0.83	0	100	1.31	49.53	17.06
f3	55.78	69.67	80.46	2.46	1.14	0.01	0	100	2.06	38.05	17.19
f4	71.78	80.78	86.71	8.82	2.05	-0.74	0	100	1.23	58.49	19.96
f5	62.28	72.22	83.33	4.5	-0.14	1.82	0	100	0	55.26	19.67
m1	46	55.69	68.28	0.31	0.71	0.36	0	100	20.82	4.36	8.2
m2	45.34	57.14	69.57	0.31	-1.02	-0.53	0	100	8.7	20.5	11.23
m3	40.37	56.27	70.95	2.44	0.6	0.34	0	100	8.26	9.79	11.78
m4	39.11	50.12	63.93	-0.32	0.75	0.25	0	100	18.03	0.7	6.97
m5	54.62	69.39	82.19	0.13	0.29	0.53	0	100	24.01	4.75	10.44
MA	45.09	57.72	70.98	0.57	0.27	0.19	0	100	15.96	8.02	9.73
FA	62.33	72.96	82.6	5.37	1.16	0.47	0	100	1.07	49.26	18
OA	53.71	65.34	76.79	2.97	0.71	0.33	0	100	8.52	28.64	13.86

PRAAT	GER20	GER10	GER05	FER20	FER10	FER05	V_UVE	UV_VE	THE	TLE	STD20
f1	100	100	100	0	0	0	100	0	0	100	0
f2	100	100	100	0	0	0	100	0	0	100	0
f3	100	100	100	0	0	0	100	0	0	100	0
f4	100	100	100	0	0	0	100	0	0	100	0
f5	100	100	100	0	0	0	100	0	0	100	0
m1	100	100	100	0	0	0	100	0	0	100	0
m2	100	100	100	0	0	0	100	0	0	100	0
m3	100	100	100	0	0	0	100	0	0	100	0
m4	100	100	100	0	0	0	100	0	0	100	0
m5	100	100	100	0	0	0	100	0	0	100	0
MA	100	100	100	0	0	0	100	0	0	100	0
FA	100	100	100	0	0	0	100	0	0	100	0
OA	100	100	100	0	0	0	100	0	0	100	0

Table A.2.30. Results for -5 dB White noise for low resolution speech signals.

DHO	GER20	GER10	GER05	FER20	FER10	FER05	V_UVE	UV_VE	THE	TLE	STD20
f1	0.74	1.86	3.35	-0.69	-0.43	-0.56	0	23.98	0	0	4.3
f2	1.69	3.75	11.26	-0.67	-0.27	0.35	0	31.25	0	0	7.34
f3	0.77	3.86	9.25	-0.96	-0.36	-0.1	0	8.14	0	0	6.15
f4	9.61	12.07	19.43	-0.97	-0.17	-0.08	4.29	1.03	0	4.29	8.5
f5	0.58	2.05	11.99	-0.11	0.29	0.59	0.58	50.5	0	0.58	7.43
m1	14.29	23	41.65	-0.04	0.22	-0.12	5.57	6.81	2.18	5.57	6.67
m2	3.73	6.83	25.47	1.78	2.08	1.52	0	30.18	0	0	6.39
m3	8.26	9.17	15.9	-0.8	-0.77	-0.33	7.65	18.69	0	7.65	3.85
m4	19.44	21.08	33.02	-0.04	0.22	0.48	4.45	16.4	1.17	4.45	3.76
m5	23.75	30.61	40.63	-0.92	0.2	0.68	11.08	3.61	0.4	11.08	5.75
MA	13.89	18.14	31.33	0	0.39	0.45	5.75	15.14	0.75	5.75	5.28
FA	2.68	4.72	11.05	-0.68	-0.19	0.04	0.98	22.98	0	0.98	6.74
OA	8.29	11.43	21.19	-0.34	0.1	0.24	3.36	19.06	0.37	3.36	6.01

CORR	GER20	GER10	GER05	FER20	FER10	FER05	V_UVE	UV_VE	THE	TLE	STD20
f1	1.86	3.35	4.83	1	1.04	1.05	1.49	1.81	0.37	1.49	4.77
f2	3.56	4.5	11.26	0.14	0.19	0.24	1.13	0.65	0.94	2.44	6.78
f3	3.08	4.63	10.28	0.3	0.57	0.59	0.26	0.72	1.03	2.06	5.57
f4	5.11	7.57	17.38	-0.5	-0.18	0.08	2.86	1.03	0.61	4.5	8.63
f5	0	1.17	10.23	-0.22	-0.47	-0.84	0	2.35	0	0	7.63
m1	16.22	17.19	21.79	0.19	0.26	0.44	11.86	0.83	4.36	11.86	3.14
m2	8.7	9.32	32.92	-1.69	-1.81	-0.67	3.11	3.3	5.59	3.11	5.74
m3	2.75	3.36	5.81	0.09	0.2	0.3	2.75	0.33	0	2.75	3.34
m4	22.01	22.48	31.85	0.64	0.65	0.43	15.93	0.23	6.09	15.93	2.93
m5	13.72	14.91	24.41	0.71	0.67	0.41	13.46	0.67	0.26	13.46	3.73
MA	12.68	13.45	23.36	-0.01	0	0.18	9.42	1.07	3.26	9.42	3.78
FA	2.72	4.24	10.8	0.14	0.23	0.23	1.15	1.31	0.59	2.1	6.68
OA	7.7	8.85	17.08	0.07	0.11	0.2	5.28	1.19	1.93	5.76	5.23

YIN	GER20	GER10	GER05	FER20	FER10	FER05	V_UVE	UV_VE	THE	TLE	STD20
f1	87.73	90.71	94.05	3.93	-0.51	1.63	0	100	0.37	85.87	14.78
f2	90.24	93.81	95.31	5.42	0.84	-1.54	0	100	0	87.62	17.93
f3	80.21	87.15	91.26	8.3	-0.09	-0.29	0	100	0.26	76.61	15.7
f4	86.3	90.18	93.05	6.84	-0.11	-1.07	0	100	0	82.62	19.11
f5	90.64	92.4	93.86	4.04	0.25	-1.83	0	100	0	88.01	14.01
m1	11.14	16.71	27.12	0.6	0.56	0.37	0	100	2.18	6.3	4.93
m2	35.4	38.51	45.96	1.3	0.59	0.82	0	100	0.62	31.06	6.15
m3	22.94	27.83	34.86	0.93	-0.03	-0.09	0	100	0	18.04	6.22
m4	2.58	4.45	14.29	0.58	0.59	0.49	0	100	0.47	0	2.97
m5	5.54	8.84	14.64	0.55	0.43	0.46	0	100	0.92	2.24	4.13
MA	15.52	19.27	27.37	0.79	0.43	0.41	0	100	0.84	11.53	4.88
FA	87.02	90.85	93.51	5.71	0.07	-0.62	0	100	0.13	84.15	16.31
OA	51.27	55.06	60.44	3.25	0.25	-0.1	0	100	0.48	47.84	10.59

PRAAT	GER20	GER10	GER05	FER20	FER10	FER05	V_UVE	UV_VE	THE	TLE	STD20
f1	4.09	6.32	8.18	1.12	1.16	1.23	0.37	0	0.74	3.35	4.69
f2	3.38	4.13	10.88	0.38	0.39	0.36	1.31	0	0.38	3	6.5
f3	2.57	3.86	9.51	0.58	0.67	0.68	0.26	0	0.51	2.06	5.49
f4	4.5	6.34	13.91	-0.14	-0.02	0.31	4.29	0	0	4.29	7.88
f5	0	1.17	7.89	0.08	-0.16	-0.64	0	1.01	0	0	7.25
m1	9.93	11.38	20.1	0.38	0.53	0.4	9.44	0	0.48	9.44	3.67
m2	1.86	3.11	31.68	-1.55	-1.6	-0.39	1.24	0.66	0.62	1.24	6.24
m3	1.22	1.83	7.65	0.19	0.31	0.3	1.22	0	0	1.22	3.59
m4	14.75	15.46	26.23	0.63	0.59	0.37	11.24	0	3.51	11.24	3.18
m5	4.09	5.94	19.26	0.78	0.72	0.41	4.09	0	0	4.09	3.96
MA	6.37	7.54	20.98	0.09	0.11	0.22	5.45	0.13	0.92	5.45	4.13
FA	2.91	4.36	10.07	0.41	0.41	0.39	1.25	0.2	0.33	2.54	6.36
OA	4.64	5.95	15.53	0.25	0.26	0.3	3.35	0.17	0.63	3.99	5.25

Table A.2.31. Results for 25 dB Environment noise for low resolution speech signals.

DHO	GER20	GER10	GER05	FER20	FER10	FER05	V_UVE	UV_VE	THE	TLE	STD20
f1	0.74	1.86	3.35	-0.68	-0.42	-0.55	0	22.17	0	0	4.33
f2	0.56	2.44	10.69	-0.82	-0.34	0.31	0	22.63	0	0	7.14
f3	0.26	3.34	8.23	-0.83	-0.26	0.02	0	26.58	0	0	6.18
f4	5.73	8.18	15.34	-0.88	-0.1	-0.11	3.27	15.72	0	3.27	8.38
f5	5.26	7.02	17.25	-0.09	0.39	0.76	0.58	48.15	0	0.58	7.61
m1	15.98	28.33	45.52	0.84	0.22	-0.18	5.81	4.98	4.36	5.81	7.65
m2	4.97	8.07	26.09	1.94	2.26	1.54	0	30.4	0	0	6.39
m3	5.2	5.2	12.23	-0.75	-0.75	-0.42	5.2	28.52	0	5.2	3.56
m4	11.71	13.82	24.82	-0.02	0.27	0.49	1.41	16.17	0.7	1.41	3.67
m5	22.56	27.44	37.86	-0.9	0.08	0.49	14.38	4.02	0.66	14.38	5.31
MA	12.08	16.57	29.31	0.22	0.41	0.38	5.36	16.82	1.14	5.36	5.31
FA	2.51	4.57	10.97	-0.66	-0.15	0.09	0.77	27.05	0	0.77	6.73
OA	7.3	10.57	20.14	-0.22	0.13	0.24	3.07	21.94	0.57	3.07	6.02

CORR	GER20	GER10	GER05	FER20	FER10	FER05	V_UVE	UV_VE	THE	TLE	STD20
f1	2.6	3.72	5.2	1.15	1.07	0.98	1.86	4.3	0.37	2.23	4.36
f2	4.13	5.25	12.01	0.03	0.03	0.08	2.25	1.72	0.94	3	6.83
f3	3.08	4.37	10.03	0.57	0.67	0.59	0.77	2.17	1.03	2.06	5.39
f4	9.82	12.27	20.65	-0.82	-0.32	-0.04	7.98	3.61	0.61	9	8.62
f5	0.29	1.46	10.53	-0.12	-0.37	-0.84	0	4.19	0	0.29	7.55
m1	17.43	18.64	23	0.12	0.24	0.44	12.59	1.99	4.84	12.59	3.24
m2	13.04	13.66	35.4	-1.8	-1.92	-0.67	7.45	4.85	5.59	7.45	5.7
m3	5.81	6.42	8.87	0.18	0.3	0.36	5.81	2.3	0	5.81	3.3
m4	23.89	24.36	33.26	0.57	0.58	0.4	16.63	1.59	7.03	16.63	2.9
m5	18.34	19.39	28.36	0.64	0.64	0.36	17.68	1.2	0.66	17.68	3.65
MA	15.7	16.5	25.78	-0.06	-0.03	0.18	12.03	2.39	3.62	12.03	3.76
FA	3.98	5.41	11.68	0.16	0.22	0.16	2.57	3.2	0.59	3.32	6.55
OA	9.84	10.96	18.73	0.05	0.09	0.17	7.3	2.79	2.11	7.67	5.16

YIN	GER20	GER10	GER05	FER20	FER10	FER05	V_UVE	UV_VE	THE	TLE	STD20
f1	89.96	92.57	95.17	6.65	1.65	0.14	0	100	0.37	88.1	15.57
f2	89.12	93.06	94.93	8.03	0.8	-0.91	0	100	0	86.87	17.75
f3	81.49	85.09	91	4.07	-0.97	-0.99	0	100	0.26	77.38	14.43
f4	84.25	88.75	92.23	3.81	-1.48	-0.51	0	100	0.41	78.73	20.08
f5	89.18	90.64	91.81	4.96	2.1	0.2	0	100	0	87.13	13.72
m1	11.38	15.74	25.67	0.57	0.71	0.42	0	100	2.18	6.05	4.68
m2	37.27	40.99	47.2	1.26	0.75	0.7	0	100	0.62	31.68	6.45
m3	23.85	26.3	34.86	0.77	0.25	0.01	0	100	0	20.49	5.48
m4	3.51	6.32	14.52	0.52	0.63	0.49	0	100	0.94	0	3.22
m5	6.73	10.03	17.68	0.35	0.42	0.43	0	100	1.32	2.24	4.01
MA	16.55	19.88	27.99	0.69	0.55	0.41	0	100	1.01	12.09	4.77
FA	86.8	90.02	93.03	5.51	0.42	-0.41	0	100	0.21	83.64	16.31
OA	51.67	54.95	60.51	3.1	0.49	0	0	100	0.61	47.87	10.54

PRAAT	GER20	GER10	GER05	FER20	FER10	FER05	V_UVE	UV_VE	THE	TLE	STD20
f1	6.32	7.43	8.92	1.51	1.26	1.2	3.35	1.36	0	6.32	3.81
f2	4.32	5.07	12.38	0.33	0.33	0.31	2.25	0	0.38	3.94	6.6
f3	2.31	4.11	10.03	0.59	0.57	0.66	0	0.18	1.03	1.29	5.62
f4	13.91	15.95	23.72	-0.36	-0.13	0.06	7.77	0	0	13.91	8.53
f5	2.34	3.51	10.82	0.1	-0.15	-0.63	0	2.01	0	2.34	7.33
m1	10.65	11.86	20.1	0.44	0.55	0.47	10.17	0.33	0.48	10.17	3.56
m2	3.11	3.73	32.92	-1.44	-1.55	-0.26	2.48	1.54	0.62	2.48	6.31
m3	2.45	3.06	8.87	0.22	0.34	0.34	2.45	0	0	2.45	3.67
m4	14.52	14.99	26.23	0.67	0.6	0.44	11.48	0	3.04	11.48	3.16
m5	6.33	8.18	20.98	0.78	0.72	0.41	6.33	0.27	0	6.33	3.95
MA	7.41	8.36	21.82	0.14	0.13	0.28	6.58	0.43	0.83	6.58	4.13
FA	5.84	7.21	13.17	0.43	0.38	0.32	2.67	0.71	0.28	5.56	6.38
OA	6.63	7.79	17.5	0.28	0.26	0.3	4.63	0.57	0.56	6.07	5.25

Table A.2.32. Results for 20 dB Environment noise for low resolution speech signals.

DHO	GER20	GER10	GER05	FER20	FER10	FER05	V_UVE	UV_VE	THE	TLE	STD20
f1	3.35	4.46	5.95	-0.43	-0.32	-0.44	0	43.44	0	0	4.33
f2	3.94	5.63	12.01	-0.24	-0.05	0.47	0	39.01	0	1.31	6.99
f3	0.51	3.86	9.25	-0.63	-0.09	0.11	0	30.38	0	0	6.21
f4	9.61	13.5	20.65	-0.46	0.45	0.06	5.11	20.62	0	5.11	9.77
f5	2.05	3.51	12.87	0.04	0.45	0.67	0	61.58	0	0	7.35
m1	22.03	32.2	47.94	0.06	0.23	-0.18	9.93	30.73	1.94	9.93	6.97
m2	13.04	15.53	31.06	1.72	2.26	1.53	4.97	29.3	0	4.97	6.23
m3	3.06	3.36	11.01	-0.7	-0.66	-0.18	0.61	16.72	0	0.61	3.85
m4	14.52	16.86	28.81	-0.05	0.23	0.43	2.81	7.74	0	2.81	3.82
m5	15.3	21.77	34.17	-0.95	0.13	0.55	9.76	10.04	0	9.76	5.57
MA	13.59	17.94	30.6	0.02	0.44	0.43	5.62	18.91	0.39	5.62	5.29
FA	3.89	6.19	12.15	-0.34	0.09	0.17	1.02	39	0	1.29	6.93
OA	8.74	12.07	21.37	-0.16	0.26	0.3	3.32	28.96	0.19	3.45	6.11

CORR	GER20	GER10	GER05	FER20	FER10	FER05	V_UVE	UV_VE	THE	TLE	STD20
f1	4.83	5.58	7.06	1.35	1.18	1.05	3.35	8.37	0.37	4.46	4.09
f2	7.32	8.44	14.63	-0.12	-0.04	0.1	5.07	3.02	0.75	6.57	6.66
f3	5.66	6.94	11.83	0.7	0.79	0.7	3.6	6.15	0.51	5.14	5.44
f4	20.04	22.49	30.27	-0.53	-0.43	-0.07	13.5	5.15	0.61	18.81	9
f5	1.17	2.34	12.87	0.01	-0.24	-0.77	0.29	6.04	0	1.17	7.79
m1	22.52	23.24	27.85	0.16	0.21	0.41	15.98	4.32	6.54	15.98	3.18
m2	18.63	19.88	39.75	-2	-2.27	-0.91	13.04	8.59	5.59	13.04	5.78
m3	14.98	15.6	18.96	0.27	0.25	0.37	14.68	6.23	0.31	14.68	3.59
m4	29.27	29.27	37.7	0.48	0.48	0.39	21.78	2.96	7.49	21.78	2.82
m5	29.68	30.61	36.28	0.27	0.28	0.26	27.97	2.01	1.72	27.97	3.49
MA	23.02	23.72	32.11	-0.17	-0.21	0.1	18.69	4.82	4.33	18.69	3.77
FA	7.8	9.16	15.33	0.28	0.25	0.2	5.16	5.75	0.45	7.23	6.6
OA	15.41	16.44	23.72	0.06	0.02	0.15	11.93	5.28	2.39	12.96	5.18

YIN	GER20	GER10	GER05	FER20	FER10	FER05	V_UVE	UV_VE	THE	TLE	STD20
f1	89.59	92.57	94.8	7.53	1.43	0.24	0	100	0.74	86.99	14.89
f2	89.12	92.5	94.75	3.5	-1.47	0.91	0	100	0	85.55	19.67
f3	84.32	87.66	92.8	5.01	0.78	0.13	0	100	0.26	79.95	14.95
f4	82.82	87.32	91.21	1.44	-3.53	-2.24	0	100	0.61	76.89	19.87
f5	86.84	91.81	93.86	10.75	0.25	-1.72	0	100	0	83.63	18.61
m1	12.11	17.19	26.63	0.48	0.55	0.32	0	100	1.69	6.78	4.61
m2	39.75	44.72	54.66	1.52	0.42	0.21	0	100	3.73	30.43	6.99
m3	22.32	27.83	36.7	1.44	0.28	-0.13	0	100	0.31	18.35	6.7
m4	6.32	10.54	18.97	0.59	0.67	0.47	0	100	1.64	0	3.51
m5	11.74	16.89	25.2	0.2	0.37	0.43	0	100	3.69	2.77	4.86
MA	18.45	23.43	32.43	0.85	0.46	0.26	0	100	2.21	11.67	5.33
FA	86.54	90.37	93.48	5.65	-0.51	-0.53	0	100	0.32	82.6	17.6
OA	52.49	56.9	62.96	3.25	-0.03	-0.14	0	100	1.27	47.13	11.47

PRAAT	GER20	GER10	GER05	FER20	FER10	FER05	V_UVE	UV_VE	THE	TLE	STD20
f1	7.81	8.55	10.04	1.7	1.53	1.28	4.83	2.94	0	7.81	3.89
f2	7.69	8.63	16.51	0.27	0.23	0.19	4.13	1.08	0	7.69	6.7
f3	6.17	7.71	13.88	0.83	0.74	0.71	1.29	2.71	1.03	5.14	5.77
f4	30.27	31.9	38.04	-0.74	-0.7	-0.4	14.72	1.55	0	30.27	8.84
f5	2.34	3.51	11.99	0.26	0.01	-0.41	0.29	3.69	0	2.34	7.45
m1	10.9	12.11	20.58	0.46	0.57	0.52	10.41	1	0.48	10.41	3.56
m2	11.8	13.66	37.89	-1.94	-2.08	-0.58	11.18	2.42	0.62	11.18	6.32
m3	7.95	8.56	14.37	0.47	0.45	0.42	7.95	0.66	0	7.95	3.76
m4	14.99	15.46	26.93	0.69	0.62	0.44	13.58	0.46	0	13.58	3.17
m5	15.83	17.41	27.84	0.62	0.59	0.28	15.83	0.67	0	15.83	3.83
MA	12.29	13.44	25.52	0.06	0.03	0.22	11.79	1.04	0.22	11.79	4.13
FA	10.85	12.06	18.09	0.46	0.36	0.27	5.05	2.39	0.21	10.65	6.53
OA	11.57	12.75	21.81	0.26	0.2	0.24	8.42	1.72	0.21	11.22	5.33

Table A.2.33. Results for 15 dB Environment noise for low resolution speech signals.

DHO	GER20	GER10	GER05	FER20	FER10	FER05	V_UVE	UV_VE	THE	TLE	STD20
f1	4.46	5.95	7.43	-0.6	-0.36	-0.38	0	56.56	0	0	4.86
f2	3.75	4.88	14.26	-0.44	-0.26	0.45	0	63.15	0	0	7.13
f3	2.06	5.66	11.83	-0.92	-0.38	0.26	0	52.8	0	0	6.85
f4	19.63	28.83	40.08	-2.36	-0.94	-0.54	3.48	60.05	0.2	5.93	13.84
f5	15.2	16.67	25.44	0.22	0.49	0.7	0	75.84	0	2.92	7.77
m1	17.43	27.6	43.1	0.27	0.12	-0.23	4.36	52.99	1.21	4.36	6.79
m2	11.8	13.66	31.68	2	2.08	1.52	4.97	28.63	0.62	4.97	6.06
m3	2.45	3.06	14.07	-0.15	-0.18	-0.03	0.61	46.56	0	0.61	4.5
m4	17.8	21.08	31.62	0.16	0.5	0.51	7.49	27.33	3.04	7.49	4.13
m5	22.43	29.16	40.77	-0.71	0.15	0.47	5.67	47.26	2.24	5.67	5.79
MA	14.38	18.91	32.24	0.31	0.53	0.45	4.62	40.55	1.42	4.62	5.45
FA	9.02	12.4	19.81	-0.82	-0.29	0.1	0.7	61.68	0.04	1.77	8.09
OA	11.7	15.65	26.03	-0.25	0.12	0.27	2.66	51.12	0.73	3.2	6.77

CORR	GER20	GER10	GER05	FER20	FER10	FER05	V_UVE	UV_VE	THE	TLE	STD20
f1	11.52	12.64	14.87	1.65	1.38	1.03	6.32	10.86	0.37	10.78	4.56
f2	12.76	14.45	22.33	-0.02	0	-0.2	7.88	6.47	0.75	11.63	7.58
f3	10.8	12.34	18.25	0.64	0.8	0.62	9.25	8.32	0.51	10.28	5.73
f4	38.65	41.51	47.24	-0.6	-0.94	-0.45	30.27	6.44	0.41	37.42	9.36
f5	4.09	5.26	17.84	0.18	-0.08	-0.62	2.34	8.89	0	4.09	8.22
m1	34.14	34.38	37.77	0.23	0.17	0.36	26.15	7.81	7.99	26.15	2.92
m2	25.47	26.71	45.34	-2.74	-2.81	-1.23	22.98	13.44	2.48	22.98	5.5
m3	33.33	33.33	36.7	0.21	0.21	0.18	30.28	9.51	2.14	30.28	3.51
m4	41.45	41.69	47.78	0.25	0.33	0.32	32.79	6.15	8.67	32.79	3.16
m5	45.51	46.04	49.87	-0.01	0.04	0.13	40.11	4.28	4.62	40.11	3.18
MA	35.98	36.43	43.49	-0.41	-0.41	-0.05	30.46	8.24	5.18	30.46	3.65
FA	15.56	17.24	24.1	0.37	0.23	0.08	11.21	8.2	0.41	14.84	7.09
OA	25.77	26.84	33.8	-0.02	-0.09	0.01	20.84	8.22	2.79	22.65	5.37

YIN	GER20	GER10	GER05	FER20	FER10	FER05	V_UVE	UV_VE	THE	TLE	STD20
f1	84.39	90.33	94.05	9.72	0.58	-2.38	0	100	0.74	81.41	16.83
f2	83.3	88.74	92.5	6.04	0.05	-0.62	0	100	0.19	78.42	19.37
f3	68.64	76.35	82.01	3.82	0.78	-0.35	0	100	0.26	63.24	14.19
f4	76.28	85.48	91.21	6.52	1.36	-1.57	0	100	1.43	67.48	23.02
f5	83.92	87.72	91.81	5.29	-2.5	-1.2	0	100	0	79.82	16.57
m1	18.64	26.63	38.5	0.25	0.63	0.48	0	100	4.12	6.05	5.76
m2	34.16	43.48	52.17	1.43	0.19	0.37	0	100	3.73	23.6	9.51
m3	24.77	33.03	48.93	1	0.09	-0.55	0	100	3.06	14.37	8.55
m4	9.84	13.82	22.95	0.46	0.55	0.47	0	100	2.81	0	3.76
m5	19.13	27.31	37.99	0.02	0.23	0.49	0	100	6.86	2.77	6.11
MA	21.31	28.85	40.11	0.63	0.34	0.25	0	100	4.11	9.36	6.74
FA	79.3	85.73	90.31	6.28	0.05	-1.22	0	100	0.52	74.08	18
OA	50.31	57.29	65.21	3.46	0.2	-0.49	0	100	2.32	41.72	12.37

PRAAT	GER20	GER10	GER05	FER20	FER10	FER05	V_UVE	UV_VE	THE	TLE	STD20
f1	17.1	17.84	21.56	1.67	1.47	1.21	6.32	5.88	0	17.1	4.74
f2	23.64	25.33	32.46	0.49	0.49	0.13	8.44	3.23	0	23.64	7.78
f3	11.57	12.6	18.25	0.63	0.65	0.55	7.71	4.16	1.03	10.54	5.67
f4	53.99	55.01	59.71	-0.87	-1.31	-1.02	32.52	2.58	0	53.99	8.58
f5	18.71	19.59	30.7	0.43	0.06	-0.88	3.22	5.2	0	18.71	8.46
m1	16.71	17.68	26.63	0.33	0.41	0.35	16.22	3.32	0.48	16.22	3.67
m2	22.36	23.6	47.2	-2.58	-2.88	-1.06	22.36	5.07	0	22.36	6.29
m3	27.22	27.52	33.03	0.37	0.42	0.23	26.3	4.26	0.92	26.3	4.06
m4	20.61	21.31	31.15	0.68	0.58	0.37	19.44	1.82	0	19.44	3.14
m5	31.66	32.45	39.84	0.21	0.29	0.28	31.27	1.47	0.4	31.27	3.45
MA	23.71	24.51	35.57	-0.2	-0.24	0.03	23.12	3.19	0.36	23.12	4.12
FA	25	26.07	32.54	0.47	0.27	0	11.64	4.21	0.21	24.8	7.05
OA	24.36	25.29	34.05	0.13	0.02	0.02	17.38	3.7	0.28	23.96	5.58

Table A.2.34. Results for 10 dB Environment noise for low resolution speech signals.

DHO	GER20	GER10	GER05	FER20	FER10	FER05	V_UVE	UV_VE	THE	TLE	STD20
f1	5.58	8.92	10.78	-0.85	-0.41	-0.52	0	91.86	0	1.86	5.91
f2	9.01	13.51	22.14	-1.29	-0.23	0.53	0	94.83	0	0	9.62
f3	5.91	9.25	16.45	-1.07	-0.35	0.18	0	87.88	0	0	6.93
f4	34.97	51.53	60.12	3.06	0.19	-0.04	1.84	94.07	0	7.16	19.08
f5	21.93	26.32	37.13	-0.63	-0.68	-0.07	0.58	97.48	0	12.57	10.38
m1	14.53	26.39	44.31	0.02	0.02	-0.21	0.24	85.22	1.45	0.24	7.12
m2	13.04	22.98	34.16	0.93	1.68	1.51	0	80.18	0	0	8.79
m3	3.36	9.17	24.46	-1.34	-1.17	-0.35	0	88.52	0	0	6.17
m4	24.12	30.68	43.79	0.17	0.84	0.74	5.62	87.93	2.11	5.62	5.06
m5	25.86	35.75	48.55	0.1	0.69	0.66	4.49	84.47	1.32	4.49	6.69
MA	16.18	25	39.06	-0.02	0.41	0.47	2.07	85.26	0.98	2.07	6.77
FA	15.48	21.91	29.33	-0.16	-0.3	0.02	0.49	93.22	0	4.32	10.38
OA	15.83	23.45	34.19	-0.09	0.06	0.24	1.28	89.24	0.49	3.19	8.57

CORR	GER20	GER10	GER05	FER20	FER10	FER05	V_UVE	UV_VE	THE	TLE	STD20
f1	27.51	29.74	33.46	2.07	1.38	0.85	20.82	13.57	0.74	26.02	6.36
f2	33.02	35.65	42.21	-0.55	-0.52	-0.36	21.95	9.27	0.75	31.71	8.95
f3	28.53	30.85	36.5	1.12	1.4	0.97	24.42	9.58	0.51	27.25	6.62
f4	55.01	58.69	63.39	-0.03	-1.78	-0.97	45.81	6.7	0.41	53.17	11.27
f5	19.59	22.22	35.09	0.51	-0.22	-0.75	12.57	11.07	0	19.59	9.74
m1	54.96	54.96	57.87	-0.15	-0.15	0.08	45.04	10.96	9.44	45.28	3.07
m2	39.75	40.99	55.9	-3.41	-3.51	-1.54	35.4	15.64	3.11	36.02	5.45
m3	54.13	54.43	56.57	0	0.17	0.17	47.09	11.15	4.28	47.09	4.09
m4	57.61	57.85	62.76	-0.07	0.05	0.23	47.54	10.71	9.37	47.54	3.43
m5	63.72	63.98	66.36	-0.39	-0.3	-0.08	53.17	6.83	9.37	53.17	2.94
MA	54.04	54.44	59.89	-0.8	-0.75	-0.23	45.65	11.06	7.11	45.82	3.79
FA	32.73	35.43	42.13	0.63	0.05	-0.05	25.11	10.04	0.48	31.55	8.59
OA	43.38	44.94	51.01	-0.09	-0.35	-0.14	35.38	10.55	3.8	38.68	6.19

YIN	GER20	GER10	GER05	FER20	FER10	FER05	V_UVE	UV_VE	THE	TLE	STD20
f1	71.38	74.72	82.53	3.95	1.27	-0.3	0	100	0.37	61.34	11.88
f2	71.11	78.24	84.8	4.41	-0.02	0.64	0	100	0.56	64.73	16.67
f3	59.13	69.41	80.21	4.21	0.34	0.43	0	99.82	0.77	49.36	15.36
f4	74.85	84.46	88.55	6.93	-1.78	-1.24	0	100	2.25	63.19	22.61
f5	71.93	81.58	88.6	9.69	0.73	1.15	0	100	0.58	66.37	18.34
m1	25.42	36.08	48.43	0.53	0.44	0.45	0	100	7.26	4.84	6.74
m2	36.65	44.1	59.01	1.87	-0.31	-0.06	0	100	8.07	15.53	9.07
m3	35.17	51.07	66.36	0.89	0.41	-0.05	0	100	6.42	11.93	10.73
m4	16.63	24.36	36.07	0.15	0.54	0.55	0	100	6.79	0	4.83
m5	35.88	48.15	58.97	0	0.39	0.42	0	100	14.64	2.64	7.9
MA	29.95	40.75	53.77	0.69	0.3	0.26	0	100	8.64	6.99	7.85
FA	69.68	77.68	84.94	5.84	0.11	0.14	0	99.96	0.91	61	16.97
OA	49.81	59.22	69.35	3.26	0.2	0.2	0	99.98	4.77	33.99	12.41

PRAAT	GER20	GER10	GER05	FER20	FER10	FER05	V_UVE	UV_VE	THE	TLE	STD20
f1	38.66	40.89	44.98	2.68	1.6	1.21	23.42	7.47	0	38.66	7.33
f2	47.09	48.78	54.6	-0.61	-0.36	-0.34	24.58	3.88	0	47.09	8.59
f3	39.07	40.36	43.96	1.18	1.3	0.97	26.99	5.79	0.51	38.56	6.14
f4	70.96	72.6	75.26	-0.51	-1.96	-2.2	56.24	2.58	0	70.76	10.81
f5	37.43	40.35	50	0.21	-0.65	-1.58	13.74	5.03	0	37.43	10.35
m1	38.5	38.74	43.83	0.26	0.2	0.24	36.32	4.82	1.21	37.29	3.31
m2	39.75	40.37	58.39	-3.88	-4.09	-1.66	39.75	8.59	0	39.75	5.61
m3	55.66	55.66	58.1	0.36	0.36	0.19	53.21	5.57	0.92	53.21	3.65
m4	36.53	36.77	44.03	0.6	0.54	0.4	36.53	4.33	0	36.53	3.18
m5	55.8	55.94	59.1	0.11	0.07	0.12	52.9	3.75	2.51	52.9	3.1
MA	45.25	45.5	52.69	-0.51	-0.58	-0.14	43.74	5.41	0.93	43.94	3.77
FA	46.64	48.6	53.76	0.59	-0.02	-0.39	28.99	4.95	0.1	46.5	8.64
OA	45.95	47.05	53.22	0.04	-0.3	-0.27	36.37	5.18	0.51	45.22	6.21

Table A.2.35. Results for 5 dB Environment noise for low resolution speech signals.

DHO	GER20	GER10	GER05	FER20	FER10	FER05	V_UVE	UV_VE	THE	TLE	STD20
f1	26.77	37.92	40.89	-1.91	-1.06	-0.52	0	100	0	1.86	11.18
f2	24.39	30.96	40.34	2.36	1.81	1.46	0	100	0	2.44	10.94
f3	21.59	28.28	35.99	-2.07	-0.22	0.06	0	98.92	0	0.26	9.97
f4	28.22	58.9	70.35	4.39	0.25	0.65	0	100	0	3.68	24.14
f5	14.62	23.1	36.84	-0.99	-0.51	0.26	0	100	0	6.43	12.46
m1	26.15	38.74	54.96	-0.5	0.18	-0.1	0	98.01	6.78	0	7.67
m2	17.39	26.09	41.61	0.83	1.06	1.47	0	98.68	0.62	0	7.81
m3	25.99	43.73	56.57	-0.48	-0.9	-0.6	0	100	0.31	0	11.2
m4	40.28	50.59	60.89	0.34	0.93	0.98	2.34	99.09	4.45	2.34	6.15
m5	27.44	42.22	59.5	-0.57	0.54	0.5	0.53	100	8.58	0.53	8.23
MA	27.45	40.27	54.71	-0.07	0.36	0.45	0.57	99.15	4.15	0.57	8.21
FA	23.12	35.83	44.88	0.36	0.05	0.38	0	99.78	0	2.93	13.73
OA	25.28	38.05	49.8	0.14	0.21	0.42	0.29	99.47	2.07	1.75	10.97

CORR	GER20	GER10	GER05	FER20	FER10	FER05	V_UVE	UV_VE	THE	TLE	STD20
f1	47.21	49.81	53.53	1.68	1.18	0.89	41.26	16.06	0.74	44.61	7.76
f2	56.85	59.85	66.79	-0.13	-1.29	-0.62	45.03	11.21	0.56	54.41	11.68
f3	55.27	57.58	62.21	0.68	0.89	0.87	47.04	12.12	1.03	51.67	8.02
f4	71.17	76.89	80.16	1.51	-2.25	-1.96	57.67	10.57	0	68.3	15.8
f5	49.71	54.09	62.87	0.27	-1.43	-1.26	41.52	13.09	0	48.83	11.91
m1	81.84	81.84	82.81	-0.94	-0.94	-0.47	69.98	13.46	9.93	70.22	3.02
m2	62.11	63.98	74.53	-4.18	-3.63	-1.69	52.17	18.72	5.59	52.8	5.5
m3	74.92	74.92	77.98	-0.76	-0.76	-0.36	63	14.43	6.73	63.3	4.14
m4	78.69	78.92	81.5	-0.6	-0.37	-0.09	65.34	14.35	12.18	65.34	3.77
m5	82.32	82.72	84.04	-0.56	-0.56	-0.21	66.36	10.31	13.46	66.36	4.02
MA	75.98	76.48	80.17	-1.41	-1.25	-0.56	63.37	14.25	9.58	63.6	4.09
FA	56.04	59.65	65.11	0.8	-0.58	-0.42	46.5	12.61	0.47	53.56	11.03
OA	66.01	68.06	72.64	-0.3	-0.91	-0.49	54.94	13.43	5.02	58.58	7.56

YIN	GER20	GER10	GER05	FER20	FER10	FER05	V_UVE	UV_VE	THE	TLE	STD20
f1	52.04	66.91	73.23	6.1	0.77	0.19	0	100	0.74	43.49	16.26
f2	64.35	72.98	83.3	4.03	1.16	-0.24	0	100	1.5	54.6	17.23
f3	62.21	75.32	85.6	5.86	1.78	-0.81	0	100	2.83	48.59	17.48
f4	76.28	84.66	91.62	4.21	1.25	-0.02	0	100	3.07	60.53	22.57
f5	71.35	79.82	88.3	5.13	1.85	-0.27	0	100	0.58	61.4	19.94
m1	43.34	55.69	69.73	0.64	1.05	0.65	0	100	17.43	3.63	8.92
m2	42.86	52.8	65.22	0.36	-0.02	0.38	0	100	9.32	13.66	10.87
m3	51.07	64.83	75.23	2.85	0.36	-0.24	0	100	11.31	10.09	11.88
m4	35.6	48.24	63.23	0.47	0.91	0.4	0	100	16.86	0.23	7.06
m5	50.26	62.93	73.75	-0.75	-0.13	0.08	0	100	23.88	3.83	9.25
MA	44.63	56.9	69.43	0.72	0.44	0.25	0	100	15.76	6.29	9.6
FA	65.25	75.94	84.41	5.07	1.36	-0.23	0	100	1.74	53.72	18.69
OA	54.94	66.42	76.92	2.89	0.9	0.01	0	100	8.75	30.01	14.14

PRAAT	GER20	GER10	GER05	FER20	FER10	FER05	V_UVE	UV_VE	THE	TLE	STD20
f1	54.65	56.13	59.85	2.55	1.53	1	46.1	8.37	0	54.28	7.17
f2	68.29	69.42	73.17	-0.48	-1.13	-1.01	54.22	4.53	0	68.11	9.49
f3	70.18	71.47	74.04	1.71	1.56	0.69	56.56	6.33	0.26	69.41	7.21
f4	80.16	83.64	86.3	3.87	-0.81	-1.01	65.85	2.58	0	79.35	15.18
f5	67.25	69.59	74.85	-0.72	-2.21	-2.1	49.42	6.21	0	67.25	11.01
m1	67.31	67.55	69.73	-0.13	-0.25	0.08	64.65	6.64	2.66	64.65	3.31
m2	65.22	65.84	77.02	-4.78	-4.6	-2.17	63.35	9.69	0	64.6	5.79
m3	77.68	77.98	78.59	-0.55	-0.8	-0.57	73.39	7.54	0.92	73.39	3.99
m4	62.76	62.76	68.62	0.27	0.27	0.11	61.12	7.29	1.64	61.12	3.3
m5	78.63	78.63	79.55	-0.15	-0.15	0.06	71.9	4.95	5.41	72.03	2.56
MA	70.32	70.55	74.7	-1.07	-1.11	-0.5	66.88	7.22	2.13	67.16	3.79
FA	68.11	70.05	73.64	1.39	-0.21	-0.49	54.43	5.6	0.05	67.68	10.01
OA	69.21	70.3	74.17	0.16	-0.66	-0.49	60.66	6.41	1.09	67.42	6.9

Table A.2.36. Results for 0 dB Environment noise for low resolution speech signals.

DHO	GER20	GER10	GER05	FER20	FER10	FER05	V_UVE	UV_VE	THE	TLE	STD20
f1	23.42	37.55	45.35	0.83	-1.38	-0.37	0	100	0	0	13.12
f2	20.64	40.53	53.66	4.29	1.37	0.67	0	100	0	0.94	18.04
f3	32.65	50.9	59.38	-2.38	0.05	0.34	0	100	0	1.54	15.43
f4	38.04	66.87	81.6	12.45	2.87	1.36	0	100	0	10.63	23.06
f5	40.94	56.43	71.64	0.29	-0.01	1.08	0	100	0	10.23	18.29
m1	30.75	49.64	66.1	-0.21	-0.21	-0.53	0	100	4.84	0	9.32
m2	31.06	44.1	58.39	-0.46	0.66	1.15	0	100	0	0	9.89
m3	39.76	58.41	70.03	-0.46	-0.8	-0.54	0	100	0	0	12.2
m4	38.17	55.74	70.73	-0.4	0.32	0.95	0	100	2.58	0	7.68
m5	36.68	56.6	72.03	-2.16	-0.02	0.28	0	100	4.22	0	10.06
MA	35.28	52.9	67.46	-0.74	-0.01	0.26	0	100	2.33	0	9.83
FA	31.14	50.46	62.33	3.1	0.58	0.62	0	100	0	4.67	17.59
OA	33.21	51.68	64.89	1.18	0.28	0.44	0	100	1.16	2.33	13.71

CORR	GER20	GER10	GER05	FER20	FER10	FER05	V_UVE	UV_VE	THE	TLE	STD20
f1	68.4	72.86	76.21	2.42	1.17	0.47	61.71	19.91	0.74	63.57	11.88
f2	76.74	80.11	83.86	2.05	-1.71	-0.24	68.48	15.73	0.38	72.8	14.65
f3	75.06	77.12	80.46	-0.14	0.47	0.57	67.35	15.91	1.03	68.89	9.44
f4	82.41	89.57	91.82	7.19	-0.74	-0.73	70.14	15.46	0	77.71	19.75
f5	73.98	79.53	85.09	3.84	-0.29	-2.39	63.74	15.6	0	70.47	15.85
m1	94.92	94.92	95.64	-1.35	-1.35	-0.78	80.63	18.6	10.41	80.63	4.35
m2	86.96	88.82	93.79	-4.82	-3.12	-1.04	75.16	23.79	4.97	75.16	8.3
m3	91.74	92.05	92.35	-0.9	-1.83	-1.59	75.84	18.03	7.95	76.15	5.96
m4	95.55	96.02	96.49	-3.26	-1.69	-0.71	76.11	18.91	16.16	76.11	5.85
m5	96.7	96.97	97.63	-1.76	-0.72	-0.23	75.33	15.53	16.75	75.33	5.84
MA	93.17	93.75	95.18	-2.42	-1.74	-0.87	76.61	18.97	11.25	76.67	6.06
FA	75.32	79.84	83.49	3.07	-0.22	-0.46	66.29	16.52	0.43	70.69	14.31
OA	84.25	86.8	89.33	0.33	-0.98	-0.67	71.45	17.75	5.84	73.68	10.19

YIN	GER20	GER10	GER05	FER20	FER10	FER05	V_UVE	UV_VE	THE	TLE	STD20
f1	60.22	71.38	79.55	6.15	2.05	0.6	0	100	1.86	43.49	16.37
f2	69.04	81.43	88.56	4.57	0.27	0.13	0	100	1.69	53.1	21.33
f3	67.87	80.98	91.26	8.89	2.18	0.86	0	100	3.86	46.79	17.86
f4	80.57	89.98	93.25	8.15	2.61	0.38	0	100	3.07	64.01	24.82
f5	76.9	85.67	92.69	8.96	4.89	1.13	0	100	2.34	63.45	20.88
m1	61.26	74.58	85.47	0.53	0.04	0.58	0	100	29.54	3.87	11.06
m2	55.9	68.94	80.75	4.29	1.68	0.47	0	100	18.01	12.42	12.3
m3	59.94	78.29	87.46	2.05	2.28	0.11	0	100	16.51	11.93	14.81
m4	63.93	78.92	86.42	-0.78	0.77	0.19	0	100	30.91	1.87	9.72
m5	63.85	77.31	87.34	1.43	1.04	0.09	0	100	30.21	5.15	10.93
MA	60.98	75.61	85.49	1.51	1.16	0.29	0	100	25.04	7.05	11.76
FA	70.92	81.89	89.06	7.34	2.4	0.62	0	100	2.56	54.17	20.25
OA	65.95	78.75	87.27	4.42	1.78	0.45	0	100	13.8	30.61	16.01

PRAAT	GER20	GER10	GER05	FER20	FER10	FER05	V_UVE	UV_VE	THE	TLE	STD20
f1	79.55	81.41	84.39	3.11	2.22	0.79	76.21	8.82	0	79.18	10.76
f2	87.05	89.49	91.93	3.93	-2.3	-2.15	80.11	6.68	0	86.49	15.6
f3	87.66	89.46	91.26	1.1	1.68	0.23	79.43	7.23	0.26	87.4	11.22
f4	89.98	93.87	94.89	13.43	1.47	1.15	84.46	5.15	0	88.75	17.5
f5	89.47	93.27	95.91	11.27	0.51	-6.73	80.12	7.21	0	89.47	18.04
m1	95.64	95.64	96.13	-0.49	-0.49	-0.07	90.8	7.97	3.63	90.8	3.44
m2	96.89	96.89	98.14	-6.41	-6.41	-5.09	91.93	11.89	0	91.93	2.77
m3	94.8	94.8	95.11	-2.03	-2.03	-1.71	86.85	9.84	1.22	86.85	3.05
m4	91.57	91.57	93.68	-0.63	-0.63	-0.87	88.29	9.34	3.04	88.29	4.05
m5	98.15	98.15	98.15	0.35	0.35	0.35	90.37	6.83	5.67	90.37	2.74
MA	95.41	95.41	96.24	-1.84	-1.84	-1.48	89.65	9.17	2.71	89.65	3.21
FA	86.74	89.5	91.67	6.57	0.72	-1.34	80.07	7.02	0.05	86.26	14.62
OA	91.08	92.46	93.96	2.36	-0.56	-1.41	84.86	8.1	1.38	87.95	8.92

Table A.2.37. Results for -5 dB Environment noise for low resolution speech signals.

DHO	GER20	GER10	GER05	FER20	FER10	FER05	V_UVE	UV_VE	THE	TLE	STD20
f1	0.74	1.86	3.35	-0.67	-0.41	-0.54	0	19.91	0	0	4.29
f2	0.38	2.25	10.51	-0.95	-0.47	0.31	0	23.71	0	0	7.21
f3	0	2.83	7.97	-0.82	-0.27	-0.02	0	9.95	0	0	6.07
f4	12.07	14.52	20.65	-1.02	-0.19	-0.15	8.59	9.54	0	8.59	8.42
f5	2.34	3.8	13.45	-0.02	0.39	0.75	0.58	44.3	0	0.58	7.43
m1	13.32	23.73	42.37	0.24	0.2	-0.21	6.3	4.82	0.24	6.3	6.98
m2	6.83	10.56	28.57	1.9	2.33	1.5	0	30.62	0	0	6.65
m3	3.36	3.98	10.7	-0.91	-0.83	-0.44	3.36	9.84	0	3.36	3.64
m4	18.74	19.91	32.32	0.28	0.45	0.51	2.11	0.23	0.94	2.11	3.6
m5	12.93	17.28	28.36	-0.47	0.17	0.55	8.31	0.8	0.26	8.31	4.71
MA	11.04	15.09	28.47	0.21	0.47	0.38	4.02	9.26	0.29	4.02	5.11
FA	3.1	5.05	11.19	-0.7	-0.19	0.07	1.83	21.48	0	1.83	6.69
OA	7.07	10.07	19.83	-0.24	0.14	0.23	2.93	15.37	0.14	2.93	5.9

CORR	GER20	GER10	GER05	FER20	FER10	FER05	V_UVE	UV_VE	THE	TLE	STD20
f1	1.12	2.23	5.2	0.87	0.98	1.07	0.37	10.18	0.74	0.37	4.84
f2	1.88	3.19	11.07	0.02	0.07	0.15	0.19	3.88	0.94	0.75	7.01
f3	3.6	4.37	9.51	0.4	0.47	0.71	0.77	3.25	1.03	2.57	5.16
f4	4.5	7.36	17.79	-0.61	-0.19	-0.03	1.84	6.19	0.61	3.89	9.07
f5	0	1.17	10.53	-0.48	-0.72	-1.07	0	8.22	0	0	7.55
m1	16.95	17.92	22.76	0.21	0.29	0.41	12.11	2.49	4.84	12.11	3.15
m2	6.21	7.45	30.43	-1.46	-1.5	-0.59	1.86	6.61	4.35	1.86	5.9
m3	3.06	3.67	7.34	-0.02	0.08	0.28	3.06	4.92	0	3.06	3.34
m4	21.31	21.78	31.15	0.64	0.65	0.44	16.16	2.73	5.15	16.16	2.91
m5	14.91	16.09	25.73	0.65	0.62	0.35	14.38	0.67	0.53	14.38	3.77
MA	12.49	13.38	23.48	0	0.03	0.18	9.51	3.48	2.97	9.51	3.81
FA	2.22	3.66	10.82	0.04	0.12	0.16	0.63	6.34	0.66	1.52	6.73
OA	7.35	8.52	17.15	0.02	0.08	0.17	5.07	4.91	1.82	5.51	5.27

YIN	GER20	GER10	GER05	FER20	FER10	FER05	V_UVE	UV_VE	THE	TLE	STD20
f1	84.76	88.48	92.57	4.72	0.35	-0.26	0	100	0.37	82.53	15.05
f2	90.06	93.06	96.25	2.98	-1.68	-1.45	0	100	0	87.8	19.39
f3	81.75	86.63	91.77	5.62	0.83	-0.86	0	100	0.26	77.89	14.88
f4	89.78	92.02	94.48	5.71	0.44	-0.47	0	100	0.2	86.71	18.29
f5	90.06	91.81	93.57	3.22	0.51	-0.63	0	100	0	87.13	11.87
m1	10.9	14.77	26.39	0.74	0.6	0.29	0	100	2.18	6.3	4.75
m2	35.4	37.89	43.48	1.24	0.63	0.6	0	100	0.62	32.3	6.02
m3	21.41	25.99	32.72	1.09	0.09	-0.17	0	100	0	18.96	5.95
m4	2.34	4.22	13.35	0.68	0.63	0.45	0	100	0.23	0	2.91
m5	5.28	7.78	13.59	0.6	0.43	0.48	0	100	0.66	2.51	3.81
MA	15.07	18.13	25.91	0.87	0.48	0.33	0	100	0.74	12.01	4.69
FA	87.28	90.4	93.73	4.45	0.09	-0.74	0	100	0.17	84.41	15.9
OA	51.17	54.27	59.82	2.66	0.28	-0.2	0	100	0.45	48.21	10.29

PRAAT	GER20	GER10	GER05	FER20	FER10	FER05	V_UVE	UV_VE	THE	TLE	STD20
f1	4.09	5.95	8.55	1.2	1.15	1.24	0.37	7.69	0.74	3.35	4.67
f2	2.25	3.38	11.63	0.28	0.29	0.38	0	3.66	0.38	1.69	6.91
f3	3.08	3.6	9.77	0.6	0.61	0.79	0.26	0	0.77	2.31	5.16
f4	4.7	6.54	15.34	-0.23	-0.12	0.08	1.84	3.87	0	4.5	8.21
f5	0	1.17	8.77	-0.15	-0.4	-0.88	0	3.02	0	0	7.26
m1	9.44	11.14	20.1	0.5	0.55	0.4	8.96	1.16	0.24	8.96	3.65
m2	0.62	3.11	30.43	-1.31	-1.21	-0.22	0	5.29	0.62	0	6.48
m3	0.92	1.53	7.03	0.26	0.37	0.32	0.92	1.97	0	0.92	3.68
m4	14.99	15.46	26.7	0.66	0.59	0.42	13.11	1.59	1.87	13.11	3.16
m5	2.24	4.35	18.87	0.78	0.75	0.39	2.24	0.27	0	2.24	4.04
MA	5.64	7.12	20.63	0.18	0.21	0.26	5.05	2.06	0.55	5.05	4.2
FA	2.83	4.13	10.81	0.34	0.31	0.32	0.49	3.65	0.38	2.37	6.44
OA	4.23	5.62	15.72	0.26	0.26	0.29	2.77	2.85	0.46	3.71	5.32

Table A.2.38. Results for 25 dB Music noise for low resolution speech signals.

DHO	GER20	GER10	GER05	FER20	FER10	FER05	V_UVE	UV_VE	THE	TLE	STD20
f1	1.86	2.97	4.09	-0.75	-0.49	-0.57	0	20.36	0	0	4.22
f2	0.38	2.25	10.13	-0.9	-0.42	0.3	0	32.97	0	0	7.17
f3	0.51	3.34	8.23	-0.8	-0.24	0.03	0	11.03	0	0	6.1
f4	8.38	10.84	17.18	-0.94	-0.14	-0.13	5.73	2.58	0	6.54	8.28
f5	4.09	5.56	15.5	-0.05	0.37	0.79	0.58	50.34	0	0.58	7.53
m1	12.83	24.46	41.89	0.41	0.09	-0.18	5.33	5.32	0.73	5.33	7.09
m2	6.83	11.18	29.19	1.83	2.37	1.54	0	29.52	0	0	6.75
m3	3.36	3.98	11.01	-0.89	-0.8	-0.34	3.36	11.15	0	3.36	3.65
m4	18.74	19.91	32.32	0.26	0.43	0.52	2.11	10.48	0.94	2.11	3.62
m5	13.59	18.07	28.76	-0.59	0.14	0.55	8.05	2.81	0.66	8.05	4.87
MA	11.07	15.52	28.63	0.21	0.45	0.42	3.77	11.85	0.46	3.77	5.19
FA	3.05	4.99	11.02	-0.69	-0.19	0.08	1.26	23.46	0	1.43	6.66
OA	7.06	10.26	19.83	-0.24	0.13	0.25	2.52	17.65	0.23	2.6	5.93

CORR	GER20	GER10	GER05	FER20	FER10	FER05	V_UVE	UV_VE	THE	TLE	STD20
f1	1.49	2.97	6.32	0.83	1.02	1.05	0.37	24.21	0.37	0.37	5.11
f2	1.88	3	12.76	-0.01	-0.01	-0.01	0.38	10.13	0.94	0.75	7.2
f3	4.11	4.88	12.08	0.6	0.67	0.66	1.29	13.74	1.03	3.08	5.62
f4	3.48	6.34	18.81	-0.77	-0.28	-0.08	1.84	11.86	0.61	2.86	9.74
f5	0	1.17	10.82	-0.59	-0.82	-1.25	0	18.79	0	0	7.59
m1	17.92	18.89	23.73	0.19	0.27	0.39	12.83	10.96	5.08	12.83	3.14
m2	7.45	8.7	30.43	-1.36	-1.4	-0.65	3.11	16.74	4.35	3.11	5.92
m3	4.89	5.2	10.7	0.06	0.12	0.23	4.59	16.39	0	4.59	3.56
m4	22.72	23.19	32.55	0.6	0.61	0.47	17.1	7.06	5.62	17.1	2.91
m5	17.94	19.13	28.5	0.58	0.54	0.29	14.51	3.21	2.77	14.51	3.82
MA	14.18	15.02	25.18	0.01	0.03	0.14	10.43	10.87	3.56	10.43	3.87
FA	2.19	3.67	12.16	0.01	0.12	0.08	0.77	15.75	0.59	1.41	7.05
OA	8.19	9.35	18.67	0.01	0.07	0.11	5.6	13.31	2.08	5.92	5.46

YIN	GER20	GER10	GER05	FER20	FER10	FER05	V_UVE	UV_VE	THE	TLE	STD20
f1	84.01	87.36	91.82	3.42	0.77	0.3	0	99.77	0.37	81.41	15.47
f2	88.74	92.31	95.68	6.43	-1.86	-2.36	0	99.78	0	86.49	19.45
f3	82.01	86.12	90.75	6.16	1.67	0.99	0	100	0.26	78.66	13.96
f4	87.93	91.62	93.87	7.51	-0.13	0.48	0	100	0.2	84.87	20.04
f5	90.64	92.4	92.98	3.7	0.97	0.92	0	100	0	88.89	11.14
m1	11.38	15.5	27.12	0.73	0.53	0.28	0	100	2.18	6.3	4.79
m2	36.65	39.13	45.96	0.97	0.42	0.17	0	100	0.62	32.3	5.88
m3	21.71	26.91	34.25	1.29	-0.12	-0.19	0	100	0	18.65	6.63
m4	2.34	3.98	13.35	0.68	0.6	0.44	0	100	0.47	0	2.9
m5	4.88	7.12	13.59	0.48	0.42	0.47	0	99.87	0.53	2.51	3.73
MA	15.39	18.53	26.85	0.83	0.37	0.24	0	99.97	0.76	11.95	4.79
FA	86.67	89.96	93.02	5.44	0.29	0.07	0	99.91	0.17	84.06	16.01
OA	51.03	54.24	59.94	3.14	0.33	0.15	0	99.94	0.46	48.01	10.4

PRAAT	GER20	GER10	GER05	FER20	FER10	FER05	V_UVE	UV_VE	THE	TLE	STD20
f1	1.12	2.97	7.43	0.93	1.1	1.24	0.37	19.23	0.74	0.37	5.41
f2	2.25	3.38	13.51	0.25	0.26	0.24	0	6.47	0	2.06	7.09
f3	3.6	4.11	12.08	0.78	0.79	0.76	0.26	10.85	0.77	2.83	5.57
f4	3.27	5.93	16.77	-1.02	-0.41	-0.1	1.64	11.34	0	3.07	9.42
f5	0	1.17	9.94	-0.34	-0.58	-1.04	0	17.79	0	0	7.41
m1	10.17	11.86	20.82	0.54	0.58	0.49	9.69	8.31	0.24	9.69	3.66
m2	1.86	4.97	31.06	-1.48	-1.26	-0.37	1.24	11.01	0.62	1.24	6.44
m3	2.75	3.36	10.4	0.11	0.21	0.25	2.75	11.48	0	2.75	3.82
m4	15.22	15.69	26.93	0.66	0.59	0.42	13.82	4.33	0	13.82	3.17
m5	3.17	5.41	19.39	0.72	0.69	0.36	3.17	1.74	0	3.17	4.12
MA	6.63	8.26	21.72	0.11	0.16	0.23	6.13	7.37	0.17	6.13	4.24
FA	2.05	3.51	11.95	0.12	0.23	0.22	0.45	13.13	0.3	1.67	6.98
OA	4.34	5.89	16.83	0.12	0.2	0.23	3.29	10.25	0.24	3.9	5.61

Table A.2.39. Results for 20 dB Music noise for low resolution speech signals.

DHO	GER20	GER10	GER05	FER20	FER10	FER05	V_UVE	UV_VE	THE	TLE	STD20
f1	0.74	1.86	3.35	-0.69	-0.43	-0.56	0	23.53	0	0	4.3
f2	1.13	2.25	10.51	-0.69	-0.43	0.27	0	29.96	0	0	6.73
f3	0	2.83	7.97	-0.82	-0.26	0.04	0	11.93	0	0	6.08
f4	8.18	10.22	16.56	-0.51	0.16	-0.02	3.48	2.84	0	3.48	8.04
f5	4.97	6.73	17.25	-0.15	0.35	0.88	0.58	42.11	0	0.58	7.74
m1	12.35	23.97	41.89	0.49	0.34	-0.15	6.3	6.48	0.73	6.3	7.07
m2	8.7	14.29	31.06	1.89	2.47	1.64	1.86	26.87	0	1.86	6.86
m3	3.36	3.98	11.01	-0.87	-0.78	-0.32	3.36	12.13	0	3.36	3.66
m4	21.55	22.48	36.07	0.09	0.23	0.44	3.28	9.79	2.58	3.28	3.76
m5	8.84	13.46	24.8	-0.46	0.16	0.56	4.62	12.99	1.19	4.62	4.78
MA	10.96	15.63	28.96	0.23	0.48	0.43	3.88	13.65	0.9	3.88	5.23
FA	3	4.78	11.13	-0.57	-0.12	0.12	0.81	22.07	0	0.81	6.58
OA	6.98	10.21	20.05	-0.17	0.18	0.28	2.35	17.86	0.45	2.35	5.9

CORR	GER20	GER10	GER05	FER20	FER10	FER05	V_UVE	UV_VE	THE	TLE	STD20
f1	2.23	4.09	8.55	0.52	0.69	0.93	0.74	31.22	0.37	0.74	6.05
f2	3.19	5.07	16.51	0.23	0.01	0.05	0.75	18.97	0.94	1.5	8.06
f3	4.63	6.43	16.2	0.49	0.84	0.4	2.06	23.69	1.03	3.34	6.57
f4	6.75	10.02	23.93	-0.74	-0.4	-0.38	3.07	17.01	1.43	5.11	10.26
f5	0	1.17	13.45	-0.89	-1.13	-1.7	0	24.66	0	0	8.02
m1	22.76	23.49	28.57	0.22	0.26	0.42	13.56	22.76	9.2	13.56	3.16
m2	16.77	18.63	38.51	-2.15	-2.08	-1.06	6.21	31.28	10.56	6.21	6.11
m3	8.26	8.56	18.04	-0.2	-0.15	0.21	6.12	28.85	0.31	6.12	4.14
m4	25.76	26.23	35.36	0.52	0.53	0.36	18.74	16.63	6.56	18.74	2.92
m5	26.39	27.57	35.75	0.44	0.44	0.22	17.68	10.04	7.26	17.68	3.85
MA	19.99	20.9	31.25	-0.23	-0.2	0.03	12.46	21.91	6.78	12.46	4.04
FA	3.36	5.35	15.73	-0.08	0	-0.14	1.32	23.11	0.75	2.14	7.79
OA	11.67	13.13	23.49	-0.16	-0.1	-0.05	6.89	22.51	3.76	7.3	5.91

YIN	GER20	GER10	GER05	FER20	FER10	FER05	V_UVE	UV_VE	THE	TLE	STD20
f1	85.87	89.22	93.31	3.21	-1.1	1.5	0	99.77	0.37	82.9	15.58
f2	88.93	92.12	94.93	3.31	-2.72	-1.21	0	100	0	86.3	18.4
f3	80.72	85.6	90.75	5.25	-0.26	0.01	0	100	0	78.15	14.97
f4	83.44	87.73	91.21	6.84	0.18	0.95	0	99.74	0.41	78.53	16.91
f5	89.47	91.23	92.69	3.8	1.26	0.68	0	100	0	87.72	11.3
m1	11.62	15.98	28.09	0.82	0.59	0.32	0	100	2.91	6.3	4.98
m2	35.4	38.51	47.83	1.28	0.69	0.56	0	100	0.62	31.68	5.73
m3	23.85	26.91	34.56	0.71	-0.12	0.06	0	100	0	18.65	5.59
m4	3.28	5.15	13.82	0.54	0.63	0.46	0	100	1.41	0	3.04
m5	6.99	10.03	16.89	0.45	0.37	0.5	0	100	2.37	2.77	4.01
MA	16.23	19.32	28.23	0.76	0.43	0.38	0	100	1.46	11.88	4.67
FA	85.69	89.18	92.58	4.48	-0.53	0.38	0	99.9	0.16	82.72	15.43
OA	50.96	54.25	60.41	2.62	-0.05	0.38	0	99.95	0.81	47.3	10.05

PRAAT	GER20	GER10	GER05	FER20	FER10	FER05	V_UVE	UV_VE	THE	TLE	STD20
f1	5.95	7.81	12.27	1.1	1.29	1.23	1.12	28.51	0.37	5.2	5.69
f2	2.44	4.13	14.63	0.61	0.42	0.15	0	18.1	0.19	1.31	7.87
f3	5.4	8.23	18.77	0.51	0.79	0.61	0	25.14	0.77	4.37	7.49
f4	9	10.22	21.27	-0.46	-0.38	-0.5	2.66	17.78	0	8.79	8.41
f5	2.92	4.09	14.33	-0.66	-0.92	-1.43	0	24.83	0	2.92	7.76
m1	11.14	13.08	22.52	0.63	0.63	0.52	9.69	18.77	1.21	9.69	3.84
m2	1.86	5.59	32.3	-1.83	-1.69	-0.51	1.24	27.53	0.62	1.24	6.53
m3	4.59	4.89	16.82	0	0.06	0.11	3.36	26.89	0	3.36	4.25
m4	13.35	14.99	26.46	0.86	0.6	0.45	11.94	13.9	0	11.94	3.59
m5	8.31	10.16	23.09	0.71	0.57	0.29	6.86	9.91	0.79	6.86	4.03
MA	7.85	9.74	24.24	0.07	0.04	0.17	6.62	19.4	0.52	6.62	4.45
FA	5.14	6.9	16.25	0.22	0.24	0.01	0.75	22.87	0.27	4.52	7.44
OA	6.5	8.32	20.24	0.15	0.14	0.09	3.69	21.14	0.4	5.57	5.95

Table A.2.40. Results for 15 dB Music noise for low resolution speech signals.

DHO	GER20	GER10	GER05	FER20	FER10	FER05	V_UVE	UV_VE	THE	TLE	STD20
f1	0.74	2.6	4.09	-1	-0.57	-0.6	0	26.02	0	0	4.73
f2	2.06	4.13	12.01	-0.71	-0.3	0.29	0	34.91	0	0	7.29
f3	0	2.57	8.48	-0.87	-0.22	0.1	0	10.13	0	0	6.15
f4	10.22	12.88	19.02	-0.56	0.04	0.1	7.16	8.51	0	7.16	8.64
f5	4.09	5.56	15.79	-0.02	0.38	0.67	0.88	50.67	0	2.92	7.35
m1	13.08	24.7	40.68	0.2	0.1	-0.08	4.84	7.81	1.45	4.84	6.94
m2	3.11	8.7	27.95	1.78	2.58	1.5	0	32.82	0	0	7.2
m3	3.67	4.28	11.31	-0.81	-0.72	-0.36	3.67	16.07	0	3.67	3.69
m4	16.16	17.33	30.68	0.19	0.38	0.54	3.75	13.21	0.47	3.75	3.73
m5	11.08	16.89	27.7	-0.54	0.35	0.69	5.94	6.96	1.45	5.94	5.16
MA	9.42	14.38	27.67	0.16	0.54	0.46	3.64	15.37	0.67	3.64	5.34
FA	3.43	5.55	11.88	-0.63	-0.13	0.11	1.61	26.05	0	2.02	6.83
OA	6.42	9.96	19.77	-0.23	0.2	0.29	2.62	20.71	0.34	2.83	6.09

CORR	GER20	GER10	GER05	FER20	FER10	FER05	V_UVE	UV_VE	THE	TLE	STD20
f1	4.46	6.69	15.61	0.56	0.61	0.97	1.86	41.4	0.74	2.6	7.31
f2	5.63	9.76	24.39	0.45	-0.09	-0.07	1.31	26.51	0.94	3	9.83
f3	5.4	11.05	26.74	0.51	0.87	0.09	2.57	31.28	1.03	4.11	9.11
f4	19.02	20.86	34.97	-0.73	-0.69	-0.66	7.16	19.85	3.68	13.91	9.86
f5	0.88	2.05	20.47	-1.25	-1.49	-1.95	0.29	32.38	0.29	0.58	8.82
m1	33.66	34.38	38.26	0.34	0.37	0.39	14.04	32.23	18.4	14.04	3.11
m2	21.12	23.6	44.72	-2.37	-2.43	-1.14	6.21	44.93	10.56	6.21	6.44
m3	21.71	24.16	37.31	-0.87	-0.23	0.32	13.15	35.08	3.36	13.15	5.85
m4	39.81	39.81	47.54	0.47	0.47	0.25	24.82	30.98	12.18	24.82	2.93
m5	41.69	42.74	47.63	-0.09	-0.03	0.08	23.48	20.48	12.27	23.48	3.83
MA	31.6	32.94	43.09	-0.5	-0.37	-0.02	16.34	32.74	11.35	16.34	4.43
FA	7.08	10.08	24.44	-0.09	-0.16	-0.32	2.64	30.28	1.34	4.84	8.98
OA	19.34	21.51	33.76	-0.3	-0.26	-0.17	9.49	31.51	6.35	10.59	6.71

YIN	GER20	GER10	GER05	FER20	FER10	FER05	V_UVE	UV_VE	THE	TLE	STD20
f1	89.22	94.05	98.14	8.12	3.68	4.58	0	100	1.49	86.25	20.38
f2	85.37	88.74	92.68	2.91	-1.36	-0.26	0	100	0.19	82.36	18.13
f3	76.61	83.29	91	5.39	0.1	0.65	0	100	0.26	71.72	16.51
f4	76.28	82.62	90.39	5.8	-0.07	1.44	0	99.74	1.84	68.3	21.18
f5	87.72	91.23	92.11	7.92	-2.42	-0.91	0	100	0	85.09	17.91
m1	14.29	21.31	32.93	0.73	0.56	0.38	0	100	4.36	6.54	5.52
m2	34.78	39.75	51.55	2.03	1.15	0.65	0	100	1.86	30.43	7.36
m3	22.32	29.05	39.14	0.78	-0.02	-0.13	0	100	0.92	16.21	7.46
m4	4.92	7.03	15.46	0.56	0.61	0.45	0	100	2.81	0	3.11
m5	12.27	17.68	24.67	0.28	0.4	0.42	0	99.87	5.28	2.77	4.99
MA	17.72	22.96	32.75	0.87	0.54	0.35	0	99.97	3.05	11.19	5.69
FA	83.04	87.99	92.86	6.03	-0.01	1.1	0	99.95	0.75	78.74	18.82
OA	50.38	55.47	62.81	3.45	0.26	0.73	0	99.96	1.9	44.97	12.25

PRAAT	GER20	GER10	GER05	FER20	FER10	FER05	V_UVE	UV_VE	THE	TLE	STD20
f1	14.13	16.36	23.42	0.84	0.88	0.83	1.12	39.59	0.37	13.38	6.9
f2	12.01	15.95	30.21	0.7	0.03	0.05	0.19	24.35	0.19	10.13	10.05
f3	6.17	12.08	26.99	0.89	1.38	0.3	0.51	32.37	0.77	5.14	9.28
f4	16.97	18.2	31.08	-0.14	-0.21	-0.3	6.34	19.85	0.2	16.16	9.24
f5	3.8	4.97	22.81	-1.11	-1.36	-1.93	0	32.05	0	3.8	8.69
m1	11.86	14.04	22.76	0.71	0.75	0.56	9.69	32.89	1.69	9.69	3.85
m2	14.29	17.39	42.24	-2.17	-1.9	-0.73	4.97	41.85	4.97	4.97	6.68
m3	12.23	15.29	30.58	-0.61	0.09	0.19	7.34	35.08	0	7.34	6.03
m4	13.35	15.46	28.81	0.91	0.59	0.3	11.71	31.44	0.23	11.71	3.74
m5	22.3	23.88	33.25	0.36	0.3	0.25	11.61	18.61	5.8	11.61	4.05
MA	14.81	17.21	31.53	-0.16	-0.03	0.11	9.06	31.97	2.54	9.06	4.87
FA	10.62	13.51	26.9	0.24	0.14	-0.21	1.63	29.64	0.31	9.72	8.83
OA	12.71	15.36	29.21	0.04	0.06	-0.05	5.35	30.81	1.42	9.39	6.85

Table A.2.41. Results for 10 dB Music noise for low resolution speech signals.

DHO	GER20	GER10	GER05	FER20	FER10	FER05	V_UVE	UV_VE	THE	TLE	STD20
f1	0.74	1.49	3.72	-0.68	-0.48	-0.47	0	33.03	0	0	4.49
f2	3.19	5.25	13.51	-0.65	-0.19	0.3	0.38	55.6	0	0.38	7.48
f3	0.26	2.83	8.23	-0.84	-0.19	0.14	0	21.88	0	0	6.16
f4	14.72	17.79	23.93	-0.22	-0.02	0.15	9.61	17.78	0	9.61	9.53
f5	1.17	2.63	14.33	-0.23	0.15	0.7	0.58	52.85	0	0.58	7.48
m1	14.53	26.88	42.86	0.63	0.01	-0.17	3.87	33.55	4.36	3.87	7.27
m2	5.59	11.8	30.43	1.87	2.62	1.5	1.24	40.09	0	1.24	7.36
m3	8.87	11.01	18.04	-0.3	-0.49	-0.19	6.73	21.97	0	6.73	4.42
m4	14.05	16.16	29.51	-0.05	0.29	0.49	3.75	31.21	0.23	3.75	3.94
m5	17.15	24.54	34.83	-1.2	0.13	0.61	11.35	19.54	1.85	11.35	5.72
MA	12.04	18.08	31.13	0.19	0.51	0.45	5.39	29.27	1.29	5.39	5.74
FA	4.02	6	12.74	-0.52	-0.14	0.16	2.11	36.23	0	2.11	7.03
OA	8.03	12.04	21.94	-0.17	0.18	0.31	3.75	32.75	0.64	3.75	6.39

CORR	GER20	GER10	GER05	FER20	FER10	FER05	V_UVE	UV_VE	THE	TLE	STD20
f1	11.52	15.24	28.25	0.52	0.94	1.15	4.83	46.83	1.12	9.29	9
f2	13.51	19.51	35.46	0.53	0.03	0.17	4.13	33.41	1.5	8.63	11.17
f3	11.83	23.39	43.19	0.81	1.62	0.7	7.46	34.9	1.54	8.48	12.34
f4	36.2	38.24	51.53	0.04	-0.26	-0.26	14.31	27.06	5.73	26.38	11.1
f5	8.19	11.4	30.41	-1.09	-1.78	-1.67	3.51	35.4	2.05	5.85	10.06
m1	50.36	51.09	54.96	0.11	0.16	0.26	17.92	38.54	30.27	17.92	3.71
m2	27.95	31.06	51.55	-2.97	-2.27	-1.16	11.18	56.17	9.32	11.18	6.8
m3	33.94	37	47.71	-1.35	-0.44	0.38	17.43	43.61	7.03	17.43	6.31
m4	55.5	55.5	60.66	0.2	0.2	0.11	29.98	46.24	22.01	29.98	3.11
m5	56.73	57.39	60.69	-0.63	-0.47	-0.07	26.52	26.64	19.53	26.52	3.73
MA	44.9	46.41	55.11	-0.93	-0.56	-0.09	20.6	42.24	17.63	20.6	4.73
FA	16.25	21.56	37.77	0.16	0.11	0.02	6.85	35.52	2.39	11.73	10.73
OA	30.57	33.98	46.44	-0.38	-0.23	-0.04	13.73	38.88	10.01	16.17	7.73

YIN	GER20	GER10	GER05	FER20	FER10	FER05	V_UVE	UV_VE	THE	TLE	STD20
f1	82.16	88.1	92.94	3.54	-1.73	-0.15	0	100	2.23	75.09	18.51
f2	78.99	84.24	88.93	5.26	-1.05	-0.65	0	100	0	76.17	16.53
f3	63.75	71.98	81.49	3.57	-0.65	-0.35	0	100	0.26	57.84	14.64
f4	64.62	72.39	82	3.69	-0.95	-0.19	0	100	2.25	56.85	18.35
f5	84.8	88.01	92.11	5.77	-0.62	-0.97	0	100	0	81.58	17.74
m1	19.85	29.54	39.95	0.5	0.58	0.3	0	100	7.02	7.99	6.25
m2	36.65	45.96	58.39	2.03	0.15	0.7	0	100	1.86	29.19	9.04
m3	24.46	30.58	43.43	0.42	-0.01	-0.38	0	100	3.06	12.84	7.36
m4	9.37	12.88	20.61	0.67	0.45	0.31	0	100	5.85	0.23	3.65
m5	21.37	28.5	38.79	-0.01	0.39	0.38	0	100	11.21	2.9	5.55
MA	22.34	29.49	40.23	0.72	0.31	0.26	0	100	5.8	10.63	6.37
FA	74.86	80.95	87.49	4.37	-1	-0.47	0	100	0.95	69.51	17.15
OA	48.6	55.22	63.86	2.54	-0.34	-0.1	0	100	3.37	40.07	11.76

PRAAT	GER20	GER10	GER05	FER20	FER10	FER05	V_UVE	UV_VE	THE	TLE	STD20
f1	20.45	25.65	35.32	0.37	0.78	1.07	4.09	46.83	0.37	19.7	8.94
f2	19.7	25.7	42.96	1.86	0.45	0.05	1.88	33.41	0.56	16.51	11.26
f3	10.54	21.59	42.67	2.96	1.88	0.6	3.34	35.26	0.77	8.74	11.79
f4	37.42	38.85	50.31	0.08	-0.25	-0.26	13.5	25.52	0	31.9	10.01
f5	29.24	31.87	50.58	-0.32	-1.43	-1.05	2.05	36.07	0	27.49	10.77
m1	23.49	25.18	34.38	0.44	0.47	0.45	10.9	38.37	9.44	10.9	4.04
m2	22.98	28.57	47.83	-3.86	-2.64	-1.14	6.83	53.08	7.45	6.83	7.2
m3	20.18	23.85	38.84	-0.83	-0.02	0.22	8.56	40	0.31	10.09	6.64
m4	22.95	23.65	36.07	0.44	0.34	0.23	14.99	42.37	4.22	14.99	3.38
m5	42.08	43.54	50.53	-0.09	0	0.09	21.37	24.5	9.5	21.37	4.24
MA	26.34	28.96	41.53	-0.78	-0.37	-0.03	12.53	39.66	6.18	12.84	5.1
FA	23.47	28.73	44.37	0.99	0.28	0.08	4.97	35.42	0.34	20.87	10.56
OA	24.9	28.85	42.95	0.11	-0.04	0.03	8.75	37.54	3.26	16.85	7.83

Table A.2.42. Results for 5 dB Music noise for low resolution speech signals.

DHO	GER20	GER10	GER05	FER20	FER10	FER05	V_UVE	UV_VE	THE	TLE	STD20
f1	2.23	3.72	5.95	-0.79	-0.57	-0.6	2.23	55.66	0	2.23	4.59
f2	0.75	3.56	12.95	-0.83	-0.37	0.21	0	59.05	0	0	7.83
f3	1.8	5.14	10.8	-1.02	-0.46	-0.04	1.03	59.13	0	1.03	6.38
f4	13.91	21.27	30.06	1.35	0.38	0.46	9	50.26	0	9	12.58
f5	4.68	5.85	17.84	-0.29	0.03	0.77	4.09	69.63	0	4.09	7.36
m1	19.85	32.2	47.7	0.3	-0.16	-0.35	5.57	47.67	2.66	5.57	7.55
m2	5.59	11.8	33.54	2.81	2.55	1.55	1.24	47.58	0.62	1.24	7.25
m3	7.03	9.17	18.65	-0.66	-0.63	-0.38	0.92	39.02	0	0.92	4.84
m4	29.74	33.02	47.07	-0.06	0.41	0.65	6.79	60.14	5.85	6.79	4.76
m5	14.12	20.98	34.56	-0.8	0.04	0.51	5.28	44.44	4.62	5.28	5.36
MA	15.27	21.44	36.31	0.32	0.44	0.4	3.96	47.77	2.75	3.96	5.95
FA	4.67	7.91	15.52	-0.32	-0.2	0.16	3.27	58.75	0	3.27	7.75
OA	9.97	14.67	25.91	0	0.12	0.28	3.61	53.26	1.38	3.61	6.85

CORR	GER20	GER10	GER05	FER20	FER10	FER05	V_UVE	UV_VE	THE	TLE	STD20
f1	23.79	32.71	48.7	1.23	1.4	1.39	13.75	59.73	1.86	18.22	12.02
f2	26.08	35.08	51.59	1.26	-0.09	-0.69	11.26	45.91	2.44	17.26	12.97
f3	19.54	39.33	57.58	3.09	2.74	1.04	13.37	49.19	2.57	14.65	15.74
f4	55.83	60.33	70.55	0.17	0.16	-0.54	19.43	46.39	9.2	38.85	14.48
f5	23.39	28.95	48.54	-0.45	-1.81	-1.64	8.77	49.33	4.68	14.91	12.71
m1	72.88	72.88	75.79	-0.49	-0.49	-0.38	25.67	54.49	41.16	25.67	3.71
m2	47.2	51.55	68.32	-4.95	-3.25	-1.19	23.6	66.08	12.42	23.6	8.43
m3	55.35	60.24	67.89	-3.3	-1.17	0.27	22.02	59.34	14.37	22.02	7.77
m4	77.99	77.99	81.03	-0.24	-0.24	-0.12	37.47	63.55	33.26	37.47	3.55
m5	78.89	79.42	81.53	-1.36	-0.9	-0.3	30.21	47.93	26.52	30.21	4.62
MA	66.46	68.42	74.91	-2.07	-1.21	-0.34	27.79	58.28	25.55	27.79	5.61
FA	29.73	39.28	55.39	1.06	0.48	-0.09	13.32	50.11	4.15	20.78	13.58
OA	48.09	53.85	65.15	-0.5	-0.37	-0.22	20.55	54.19	14.85	24.29	9.6

YIN	GER20	GER10	GER05	FER20	FER10	FER05	V_UVE	UV_VE	THE	TLE	STD20
f1	58.36	68.4	78.07	1.73	-1.31	-0.93	0	100	1.86	50.93	15.95
f2	65.1	72.42	81.24	3.49	0.75	-0.08	0	100	0.38	58.72	16.29
f3	44.22	52.19	63.75	0.5	-0.83	-0.21	0	100	1.29	37.28	12.51
f4	60.33	70.35	81.6	1.99	-0.76	-0.14	0	100	3.89	46.01	20.63
f5	70.76	79.82	88.01	3.89	-0.16	-0.23	0	100	1.46	64.33	19.03
m1	26.63	36.08	49.15	0.59	0.53	0.36	0	100	17.19	4.12	7.14
m2	24.84	36.65	49.69	0.59	0.09	0.31	0	100	8.07	11.8	9.79
m3	39.45	51.07	59.94	-0.13	-0.21	-0.5	0	100	11.93	13.15	9.89
m4	19.44	26.93	40.52	0.68	0.59	0.36	0	100	13.11	0.23	5.15
m5	37.2	45.91	56.33	0.12	0.5	0.52	0	100	21.5	2.37	7.05
MA	29.51	39.33	51.13	0.37	0.3	0.21	0	100	14.36	6.34	7.8
FA	59.75	68.64	78.53	2.32	-0.46	-0.32	0	100	1.77	51.45	16.88
OA	44.63	53.98	64.83	1.35	-0.08	-0.05	0	100	8.07	28.89	12.34

PRAAT	GER20	GER10	GER05	FER20	FER10	FER05	V_UVE	UV_VE	THE	TLE	STD20
f1	30.11	38.29	52.79	0.28	0.39	1.09	12.27	56.79	0.37	26.39	11.49
f2	36.59	45.03	59.1	3.38	0.63	0.13	9.01	42.24	0.19	29.83	13.43
f3	16.45	37.02	57.33	6.26	2.58	1.09	8.74	46.29	0.77	14.14	15.06
f4	60.94	63.8	72.19	1.4	0.7	-0.21	18.4	39.69	0.2	48.26	12.92
f5	33.63	37.13	57.31	-0.76	-2.18	-0.88	6.73	45.13	0	30.41	12.09
m1	45.52	46.73	54.72	-0.12	0.1	0.28	15.25	50.17	21.55	16.71	4.84
m2	30.43	37.89	56.52	-5.16	-3.26	-1.32	9.94	64.1	4.97	13.04	8.08
m3	40.06	46.48	58.1	-2.95	-0.7	0.2	17.43	52.46	1.22	19.88	8.47
m4	37.7	40.05	51.76	0.65	0.34	0.09	24.59	60.36	9.37	24.59	4.16
m5	65.3	66.23	70.05	-0.92	-0.45	-0.1	24.14	41.23	15.44	24.67	4.67
MA	43.8	47.48	58.23	-1.7	-0.79	-0.17	18.27	53.66	10.51	19.78	6.05
FA	35.54	44.25	59.74	2.11	0.42	0.24	11.03	46.03	0.31	29.81	13
OA	39.67	45.87	58.99	0.21	-0.19	0.04	14.65	49.85	5.41	24.79	9.52

Table A.2.43. Results for 0 dB Music noise for low resolution speech signals.

DHO	GER20	GER10	GER05	FER20	FER10	FER05	V_UVE	UV_VE	THE	TLE	STD20
f1	4.09	10.04	12.27	-0.98	-1.03	-0.93	0.74	74.66	0	0.74	9.14
f2	4.13	9.01	18.95	0.38	0.14	0.46	2.06	63.79	0	2.06	9.86
f3	3.08	10.54	17.22	-2.29	-0.67	-0.32	0	73.24	0	0	8.59
f4	22.49	39.26	48.88	5.93	0.9	-0.1	11.04	67.27	0	11.04	17.72
f5	3.22	9.36	20.18	0.02	0.62	0.8	2.34	68.29	0	2.34	10.54
m1	22.52	33.9	49.88	-0.73	-0.19	-0.44	4.36	70.27	4.84	4.36	7.16
m2	11.8	19.88	41.61	3.39	2.75	0.99	0	61.01	0	0	8.05
m3	28.44	34.25	41.9	0.2	-0.84	-0.52	7.95	62.3	0	7.95	7.05
m4	23.42	31.15	43.33	-0.19	0.23	0.68	6.79	74.03	8.43	6.79	5.36
m5	26.12	37.34	49.87	-0.74	0.08	0.46	10.55	67.34	5.41	10.55	6.58
MA	22.46	31.3	45.32	0.39	0.41	0.23	5.93	66.99	3.74	5.93	6.84
FA	7.4	15.64	23.5	0.61	-0.01	-0.02	3.24	69.45	0	3.24	11.17
OA	14.93	23.47	34.41	0.5	0.2	0.11	4.58	68.22	1.87	4.58	9.01

CORR	GER20	GER10	GER05	FER20	FER10	FER05	V_UVE	UV_VE	THE	TLE	STD20
f1	38.29	59.11	72.49	1.76	0.68	1.37	19.33	71.04	4.83	23.79	17.88
f2	51.03	63.98	72.98	2.62	0.15	-0.86	22.89	63.36	4.32	34.52	16.85
f3	32.9	59.13	74.55	5.28	2.96	1.16	19.02	64.01	5.91	20.57	18.8
f4	70.14	77.71	85.48	3.87	1.91	0.53	24.34	63.66	9.2	48.67	20.66
f5	44.44	55.85	70.76	1.71	-1.42	-1	19.3	64.93	4.09	29.82	17.57
m1	89.83	90.56	92.01	-1.98	-0.52	-0.59	27.36	67.11	49.39	27.36	7.04
m2	63.35	70.81	85.71	-9.07	-5.43	-2.18	29.19	71.37	15.53	29.19	9.6
m3	68.81	76.15	82.87	-6.16	-1.83	0.26	22.32	70.49	20.49	22.32	9.59
m4	96.49	96.49	96.72	-1.07	-1.07	-0.64	37.47	72.21	47.54	37.47	2.87
m5	93.14	93.8	94.59	-3.54	-1.79	-0.54	28.1	63.59	33.38	28.1	6.57
MA	82.32	85.56	90.38	-4.36	-2.13	-0.74	28.89	68.95	33.27	28.89	7.14
FA	47.36	63.15	75.25	3.05	0.86	0.24	20.98	65.4	5.67	31.47	18.35
OA	64.84	74.36	82.82	-0.66	-0.64	-0.25	24.93	67.18	19.47	30.18	12.74

YIN	GER20	GER10	GER05	FER20	FER10	FER05	V_UVE	UV_VE	THE	TLE	STD20
f1	49.81	57.99	66.91	1.06	-0.2	0.27	0	100	6.32	34.2	13.37
f2	47.09	58.54	70.54	3.71	1.74	0.12	0	100	1.69	38.84	16.68
f3	37.28	51.93	70.44	1.5	-0.45	-0.53	0	100	4.88	22.37	15.42
f4	58.49	71.17	84.46	5.76	3.31	0.5	0	100	6.34	37.22	21.49
f5	57.31	67.25	78.65	2.23	-0.57	-0.12	0	100	4.97	43.27	17.55
m1	50.85	60.53	72.64	-0.02	0.73	0.45	0	100	33.17	3.87	8.09
m2	38.51	54.66	65.84	1.99	-0.28	-0.78	0	100	14.91	7.45	11.89
m3	54.13	70.95	80.12	-1.38	0.41	0.16	0	100	18.35	11.93	13.5
m4	46.14	59.25	71.9	0.06	0.65	0.35	0	100	33.02	0.7	7.45
m5	57.26	64.91	76.65	0.59	1.08	0.36	0	100	33.51	2.51	8.37
MA	49.38	62.06	73.43	0.25	0.52	0.11	0	100	26.59	5.29	9.86
FA	50	61.37	74.2	2.85	0.77	0.05	0	100	4.84	35.18	16.9
OA	49.69	61.72	73.82	1.55	0.64	0.08	0	100	15.72	20.24	13.38

PRAAT	GER20	GER10	GER05	FER20	FER10	FER05	V_UVE	UV_VE	THE	TLE	STD20
f1	40.89	61.34	75.84	7.32	2.48	1.04	16.73	69.68	0.37	33.83	16.73
f2	60.98	73.36	82.18	9.53	2.02	0.54	21.01	59.7	0.19	47.65	16.28
f3	28.79	53.98	74.55	8.73	3.49	0.84	13.88	61.48	0.51	24.68	17.14
f4	76.89	80.57	86.71	8.53	4.92	1.54	23.52	61.08	0.2	60.33	15.84
f5	54.39	61.7	76.61	0.94	-2.13	-1.1	19.01	62.58	0	47.37	16.11
m1	81.11	82.57	85.47	-1.49	-0.53	0.27	22.03	62.29	42.86	26.39	6.91
m2	54.66	63.98	81.37	-8.66	-5.39	-2.24	27.33	70.26	2.48	31.68	8.87
m3	55.96	66.67	77.37	-6.69	-2.14	-0.09	19.27	68.2	3.67	21.41	9.82
m4	72.83	75.41	81.97	1.18	0.23	0	33.49	71.3	25.76	33.72	5.29
m5	87.47	88.26	89.84	-2.32	-1.01	-0.24	26.78	60.51	17.55	28.23	6.44
MA	70.41	75.38	83.2	-3.6	-1.77	-0.46	25.78	66.51	18.46	28.29	7.47
FA	52.39	66.19	79.18	7.01	2.16	0.57	18.83	62.91	0.26	42.77	16.42
OA	61.4	70.78	81.19	1.71	0.19	0.06	22.3	64.71	9.36	35.53	11.94

Table A.2.44. Results for -5 dB Music noise for low resolution speech signals.

Appendix 3

THE COMPLETE RESULTS FOR MULTIPLE PITCH TRACK EXPERIMENTS

This appendix presents the complete results for the multiple pitch track estimation experiments. For better readability, the various systems are referred to by brief names, which were also used throughout the main text. For reference to these, please see chapter 2 and chapter 5.

The various error measures and their names, as used in the tables were described in detail in chapter 5.

The various error measures and their names, as used in the tables were described in detail in chapter 5. These are abbreviated in the tables as follows.

GEE20 is Gross Error Rate within 20% of the reference pitch estimate.

GEE10 is Gross Error Rate within 10% of the reference pitch estimate.

GEE05 is Gross Error Rate within 5% of the reference pitch estimate.

FEE20 is Fine Error Rate within 20% of the reference pitch estimate.

FEE10 is Fine Error Rate within 10% of the reference pitch estimate.

FEE05 is Fine Error Rate within 5% of the reference pitch estimate.

V_UVE is percentage voiced to unvoiced error measure.

UV_VE is the percentage unvoiced to voiced error measure.

THE is the percentage of “too-high” errors.

TLE is the percentage of “too-low” errors.

STD20 is the standard deviation of FEE20 errors.

The Foreground track average performance is represented by the name FAVG.

The Background track average performance is represented by the name BAVG.

For double pitch tracking systems (PMPT and DHO), the prefix **F_** before the utterance type means the foreground track performance, while the prefix **B_** means background track performance.

DHO	GEE20	GEE10	GEE05	FEE20	FEE10	FEE05	V_UVE	UV_VE	THE	TLE	STD20
F_v0	12.5	23.75	45	-0.77	-0.22	-0.57	11.88	0	0	11.88	7.15
B_v0	21.19	25.19	28.34	-0.23	-0.23	-0.88	20.19	0	0	19.19	5.75
F_v1	11.18	22.37	41.45	-0.61	-0.48	-0.71	11.18	0	0	11.18	6.82
B_v1	13.27	17.27	21.27	1.52	1.52	1.52	12.27	0	0	11.27	3.87
F_v2	4.88	12.2	46.34	-0.97	0.14	-0.53	0	6.25	0	0	7.31
B_v2	36.54	40.54	44.54	-3.43	-3.43	-3.43	35.54	0	0	34.54	2.21
F_v3	4.5	9.91	31.53	-0.39	0.23	-0.83	0.9	0	0	0.9	6.62
B_v3	6.33	10.33	14.33	-2.79	-2.79	-2.79	5.33	0	0	4.33	2.35
F_v4	5.84	12.34	28.57	-0.39	0.67	0.3	1.95	0	0.65	1.95	5.72
B_v4	15.45	19.45	20.2	1.97	1.97	1.71	14.45	0	0	13.45	8.27
F_v5	2.5	6.88	15.63	-0.92	-0.75	-0.77	0	0	0	0	5.18
B_v5	0	4	8	0	0	0	-1	0	0	-2	0
F_v6	0	0.6	21.69	-0.84	-0.69	-0.37	0	100	0	0	5.21
B_v6	0	4	8	0	0	0	-1	0	0	-2	0
F_v7	0	3.03	18.18	-1.29	-0.97	-0.54	0	0	0	0	4.49
B_v7	0	4	8	0	0	0	-1	0	0	-2	0
F_v8	0.79	1.59	19.05	-0.75	-0.6	-0.66	0.79	0	0	0.79	4.69
B_v8	0	4	8	0	0	0	-1	0	0	-2	0
F_v9	1.7	1.7	16.48	-0.55	-0.55	-0.44	0.57	0	0	0.57	3.85
B_v9	2.26	6.26	9.5	11.2	11.2	9.97	1.26	0	0	0.26	1.77
FAVG	4.39	9.44	28.39	-0.75	-0.32	-0.51	2.73	10.63	0.06	2.73	5.7
BAVG	9.5	13.5	17.02	0.82	0.82	0.61	8.5	0	0	7.5	2.42

PMPT	GEE20	GEE10	GEE05	FEE20	FEE10	FEE05	V_UVE	UV_VE	THE	TLE	STD20
F_v0	8.13	11.88	27.5	-0.56	-0.32	-0.41	3.13	100	5	3.13	4.66
B_v0	71.19	76.27	81.36	-0.99	1.41	1.57	55.93	23.53	0	59.32	18.57
F_v1	0	1.97	15.13	-0.05	0.02	-0.11	0	0	0	0	4.41
B_v1	73.45	79.65	83.19	-2.08	5.19	3.7	55.75	23.53	0	55.75	20.29
F_v2	0	3.66	24.39	-1.04	-0.43	-0.59	0	31.25	0	0	5.77
B_v2	59.62	63.46	73.08	10.56	8.14	4.28	59.62	23.53	0	59.62	10.65
F_v3	0.9	5.41	20.72	-0.59	0.09	-0.34	0.9	100	0	0.9	5.07
B_v3	64.56	72.15	77.22	-0.95	5	3.57	54.43	23.53	0	59.49	19.12
F_v4	4.55	9.09	22.73	-0.61	0	0.22	0.65	100	3.9	0.65	4.31
B_v4	76.42	85.37	88.62	-9.83	0.62	0.23	60.98	11.76	0	63.41	20.38
F_v5	5.63	8.75	17.5	-0.68	-0.19	-0.63	0	100	5	0	4.93
B_v5	74.14	83.62	86.21	-8.34	2.97	1.75	58.62	23.53	0	63.79	21.19
F_v6	0	0	12.05	-0.31	-0.31	-0.16	0	100	0	0	3.52
B_v6	71.67	80.83	84.17	-7.32	3.16	3.42	63.33	0	0	63.33	21.32
F_v7	1.01	1.01	8.08	-0.68	-0.68	-0.2	1.01	0	0	1.01	3.03
B_v7	60.78	62.75	66.67	8.32	6.92	4.94	60.78	5.88	0	60.78	9.54
F_v8	0	0.79	8.73	-0.48	-0.37	-0.42	0	33.33	0	0	3.45
B_v8	75.58	81.4	88.37	-0.5	6.42	6.22	59.3	18.75	0	59.3	21.52
F_v9	2.27	2.27	5.11	-0.27	-0.27	-0.27	0	0	2.27	0	2.79
B_v9	88.72	93.23	93.23	-13.45	1.3	1.3	69.92	23.53	0	70.68	18.73
FAVG	2.25	4.48	16.19	-0.53	-0.24	-0.29	0.57	56.46	1.62	0.57	4.2
BAVG	71.61	77.87	82.21	-2.46	4.11	3.1	59.87	17.76	0	61.55	18.13

Table A.3.1 (continued on the next page).

CORR	GEE20	GEE10	GEE05	FEE20	FEE10	FEE05	V_UVE	UV_VE	THE	TLE	STD20
v0	6.88	6.88	14.38	-0.18	-0.18	-0.31	6.88	100	0	6.88	2.55
v1	3.29	4.61	12.5	-1.35	-1.17	-0.5	3.29	0	0	3.29	3.13
v2	8.54	15.85	19.51	-1.83	-0.58	-0.23	2.44	37.5	6.1	2.44	4.86
v3	7.21	9.91	16.22	-1.4	-0.81	-0.16	1.8	100	5.41	1.8	4.43
v4	3.25	3.9	7.14	-0.12	-0.02	0.39	1.95	84.62	1.3	1.95	2.91
v5	3.13	6.88	11.88	-0.55	-0.04	-0.11	0	100	2.5	0	4.93
v6	0	0	3.61	-0.6	-0.6	-0.5	0	100	0	0	2.15
v7	0	0	13.13	-1.93	-1.93	-0.89	0	0	0	0	3.38
v8	3.17	3.17	7.94	-1.4	-1.4	-1	0	0	3.17	0	2.51
v9	1.14	2.84	4.55	-0.5	-0.4	-0.3	0	0	1.14	0	2.52
FAVG	3.66	5.4	11.08	-0.98	-0.71	-0.36	1.64	52.21	1.96	1.64	3.34

YIN	GEE20	GEE10	GEE05	FEE20	FEE10	FEE05	V_UVE	UV_VE	THE	TLE	STD20
v0	1.25	3.75	18.13	-0.16	-0.12	-0.3	0	100	0	0	4.48
v1	1.32	5.26	17.11	-0.26	-0.71	-0.5	0	0	0	0.66	5.74
v2	7.32	18.29	32.93	-1.78	-0.91	0.17	0	100	2.44	3.66	6.42
v3	0	6.31	9.01	-0.79	-0.42	-0.34	0	100	0	0	4.95
v4	3.25	6.49	15.58	-0.01	0.45	0.49	0	100	0.65	1.95	3.98
v5	3.13	8.75	15.63	-0.36	-0.2	-0.02	0	100	0	0	5.18
v6	4.82	6.02	10.84	-0.05	-0.36	-0.59	0	100	0	2.41	4.15
v7	8.08	9.09	17.17	-0.27	-0.51	-0.37	0	0	0	6.06	3.9
v8	2.38	7.94	9.52	0.1	-0.19	-0.33	0	100	0.79	1.59	4.84
v9	1.14	1.7	7.95	0.18	0.05	0.09	0	100	0.57	0	3.39
FAVG	3.27	7.36	15.39	-0.34	-0.29	-0.17	0	80	0.45	1.63	4.7

PRAAT	GEE20	GEE10	GEE05	FEE20	FEE10	FEE05	V_UVE	UV_VE	THE	TLE	STD20
v0	3.13	3.13	9.38	0.34	0.34	-0.03	3.13	100	0	3.13	2.29
v1	0	1.97	7.89	-0.97	-0.7	-0.21	0	0	0	0	3.12
v2	0	13.41	18.29	-2.1	-0.51	0.04	0	43.75	0	0	5.78
v3	0	0.9	8.11	-0.75	-0.6	-0.07	0	92.86	0	0	3.22
v4	0	1.95	4.55	0.04	0.32	0.61	0	92.31	0	0	3.02
v5	1.25	5.63	10	0.4	0.36	0.27	0	100	0	0	4.66
v6	0	0	4.22	-0.07	-0.07	-0.2	0	100	0	0	2.28
v7	0	0	12.12	-1.14	-1.14	-0.23	0	0	0	0	3.66
v8	0	0	5.56	-1.09	-1.09	-0.63	0	0	0	0	2.7
v9	0	0.57	3.41	-0.35	-0.27	-0.1	0	0	0	0	2.13
FAVG	0.44	2.76	8.35	-0.57	-0.34	-0.05	0.31	52.89	0	0.31	3.28

Table A.3.1. Performance with background utterance n7 at 5 dB SNR (high resolution signal).

DHO	GEE20	GEE10	GEE05	FEE20	FEE10	FEE05	V_UVE	UV_VE	THE	TLE	STD20
F_v0	1.25	12.5	36.25	-0.81	-0.35	-0.58	1.25	0	0	1.25	6.8
B_v0	2.54	6.54	10.54	-5.88	-5.88	-5.88	1.54	0	0	0.54	3.97
F_v1	12.5	21.71	41.45	-0.87	-0.42	-0.67	12.5	0	0	12.5	6.62
B_v1	15.04	19.04	23.04	-1.48	-1.48	-1.48	14.04	0	0	13.04	3.52
F_v2	4.88	12.2	47.56	-0.92	0.18	-0.38	0	6.25	0	0	7.39
B_v2	36.54	40.54	44.54	-3.42	-3.42	-3.42	35.54	0	0	34.54	2.27
F_v3	4.5	9.91	31.53	-0.42	0.19	-0.89	0.9	0	0	0.9	6.65
B_v3	5.06	9.06	13.06	-2.9	-2.9	-2.9	4.06	0	0	3.06	2.32
F_v4	5.19	12.34	25.97	-0.62	0.55	0.38	1.3	0	0.65	1.3	5.79
B_v4	13.82	17.82	21.01	-0.08	-0.08	0.61	12.82	0	0	11.82	5.86
F_v5	1.88	6.88	15.63	-1.07	-0.73	-0.76	0	0	0	0	5.57
B_v5	0	4	8	0	0	0	-1	0	0	-2	0
F_v6	0	0.6	22.89	-0.83	-0.68	-0.48	0	100	0	0	5.21
B_v6	0	4	8	0	0	0	-1	0	0	-2	0
F_v7	0	3.03	18.18	-1.3	-0.96	-0.53	0	0	0	0	4.54
B_v7	0	4	8	0	0	0	-1	0	0	-2	0
F_v8	0.79	1.59	19.05	-0.77	-0.6	-0.66	0.79	0	0	0.79	4.73
B_v8	0	4	8	0	0	0	-1	0	0	-2	0
F_v9	1.7	1.7	15.91	-0.56	-0.56	-0.48	0.57	0	0	0.57	3.85
B_v9	2.26	6.26	9.5	11.95	11.95	10.26	1.26	0	0	0.26	2.4
FAVG	3.27	8.25	27.44	-0.82	-0.34	-0.51	1.73	10.63	0.06	1.73	5.71
BAVG	7.53	11.53	15.37	-0.18	-0.18	-0.28	6.53	0	0	5.53	2.03

PMPT	GEE20	GEE10	GEE05	FEE20	FEE10	FEE05	V_UVE	UV_VE	THE	TLE	STD20
F_v0	0.63	7.5	23.75	-0.88	-0.33	-0.35	0.63	0	0	0.63	4.87
B_v0	83.33	83.33	88.1	-2.46	-2.46	-0.46	83.33	11.76	0	83.33	3.29
F_v1	0	1.97	15.79	-0.15	-0.08	-0.15	0	0	0	0	4.45
B_v1	81.4	81.4	81.4	-1.16	-1.16	-1.16	81.4	12.12	0	81.4	1.1
F_v2	0	4.88	21.95	-0.98	-0.25	-0.47	0	81.25	0	0	5.71
B_v2	61.54	65.38	73.08	-3.22	-2.54	-1.21	61.54	0	0	61.54	3.33
F_v3	0	4.5	20.72	-0.65	0.02	-0.32	0	64.29	0	0	5.17
B_v3	65.71	65.71	77.14	-3.59	-3.59	-1.63	65.71	15.38	0	65.71	3.08
F_v4	1.95	6.49	20.13	-0.54	0.05	0.34	1.95	30.77	0	1.95	4.31
B_v4	82.61	82.61	86.96	-2.67	-2.67	-1.34	82.61	16.18	0	82.61	2.61
F_v5	1.25	3.75	12.5	-0.75	-0.24	-0.57	0	0	0	0	4.64
B_v5	75	75	88.64	-4.65	-4.65	-1.34	75	16.67	0	75	3.26
F_v6	0	0	10.24	-0.41	-0.41	-0.3	0	100	0	0	3.48
B_v6	81.4	81.4	83.72	-1.97	-1.97	-1.26	81.4	10.29	0	81.4	2.27
F_v7	0	0	7.07	-0.33	-0.33	0.18	0	0	0	0	3.22
B_v7	69.23	69.23	76.92	-2.54	-2.54	-0.76	69.23	0	0	69.23	3.15
F_v8	0	0.79	6.35	-0.47	-0.36	-0.21	0	100	0	0	3.57
B_v8	80	80	80	-1.15	-1.15	-1.15	80	17.54	0	80	1.23
F_v9	1.7	1.7	4.55	-0.24	-0.24	-0.25	1.7	0	0	1.7	2.86
B_v9	86	86	88	-1.79	-1.79	-0.93	86	14.29	0	86	2.52
FAVG	0.55	3.16	14.3	-0.54	-0.22	-0.21	0.43	37.63	0	0.43	4.23
BAVG	76.62	77.01	82.39	-2.52	-2.45	-1.12	76.62	11.42	0	76.62	2.58

CORR	GEE20	GEE10	GEE05	FEE20	FEE10	FEE05	V_UVE	UV_VE	THE	TLE	STD20
v0	23.75	25.63	31.88	-0.71	-0.54	-0.26	7.5	100	15	7.5	3.58
v1	14.47	21.05	30.92	-2.36	-1.17	-0.42	3.95	0	10.53	3.95	5.23
v2	18.29	28.05	29.27	-2.63	-0.73	-0.55	2.44	37.5	15.85	2.44	5.65
v3	21.62	29.73	39.64	-3.74	-1.91	-0.56	2.7	100	18.92	2.7	6.5
v4	12.34	18.18	22.08	-1.12	0.15	0.52	4.55	100	7.79	4.55	5.49
v5	18.13	26.88	31.88	-1.92	-0.58	-0.11	2.5	100	13.13	2.5	5.95
v6	4.82	4.82	13.86	-0.51	-0.51	-0.31	0	100	3.01	0	3.09
v7	13.13	13.13	29.29	-2.19	-2.19	-0.72	5.05	0	6.06	5.05	4.15
v8	15.87	19.05	28.57	-2.52	-2	-0.99	0.79	0	15.08	0.79	4.28
v9	17.61	21.02	27.27	-1.29	-0.88	-0.37	1.7	33.33	15.91	1.7	4.08
FAVG	16	20.75	28.46	-1.9	-1.04	-0.38	3.12	57.08	12.13	3.12	4.8

Table A.3.2 (continued on the next page).

YIN	GEE20	GEE10	GEE05	FEE20	FEE10	FEE05	V_UVE	UV_VE	THE	TLE	STD20
v0	6.88	13.13	29.38	-0.22	-0.5	-0.56	0	100	2.5	1.25	5.79
v1	4.61	14.47	27.63	-1.65	-0.96	-0.67	0	0	0	3.29	6.63
v2	12.2	21.95	32.93	-1.1	-0.33	0.08	0	100	3.66	2.44	6.92
v3	7.21	22.52	36.94	0.08	-1.08	-0.7	0	100	0.9	0	8.34
v4	3.9	8.44	22.08	-0.95	-0.24	0.46	0	100	1.95	0.65	5.19
v5	3.13	15.63	30	-0.28	0.06	0.02	0	100	0	0	7.35
v6	8.43	12.05	25.9	0.93	0.25	-0.47	0	100	0	6.02	6.03
v7	4.04	9.09	23.23	-0.74	-1.16	-0.42	0	0	0	1.01	6.48
v8	11.9	15.87	26.19	-1.58	-0.89	-0.34	0	100	0.79	2.38	4.82
v9	7.95	14.2	23.3	0.3	-0.19	0.18	0	100	2.27	0	5.32
FAVG	7.02	14.74	27.76	-0.52	-0.5	-0.24	0	80	1.21	1.7	6.29

PRAAT	GEE20	GEE10	GEE05	FEE20	FEE10	FEE05	V_UVE	UV_VE	THE	TLE	STD20
v0	6.88	11.88	25.63	0.32	-0.06	-0.07	3.75	100	3.13	3.75	5.02
v1	0	8.55	19.74	-2.18	-1.04	-0.12	0	0	0	0	4.88
v2	3.66	21.95	26.83	-2.66	-0.7	-0.21	3.66	43.75	0	3.66	6.77
v3	0.9	19.82	30.63	-1.3	-1.3	-0.18	0.9	92.86	0	0.9	8.15
v4	7.79	12.99	18.18	-0.66	0.46	0.69	1.95	100	5.84	1.95	5.21
v5	1.25	23.75	31.25	0.98	0.46	0.16	0	100	0	0	8.57
v6	0	2.41	12.05	0.11	-0.22	-0.05	0	100	0	0	3.77
v7	0	5.05	24.24	-0.79	-1.83	-0.57	0	0	0	0	6.3
v8	0	3.97	15.08	-2.28	-1.73	-0.75	0	0	0	0	4.32
v9	0	5.68	15.34	-0.1	-0.64	-0.03	0	0	0	0	4.86
FAVG	2.05	11.6	21.9	-0.85	-0.66	-0.11	1.03	53.66	0.9	1.03	5.79

Table A.3.2. Performance with background utterance n8 at 5 dB SNR (high resolution signal).

DHO	GEE20	GEE10	GEE05	FEE20	FEE10	FEE05	V_UVE	UV_VE	THE	TLE	STD20
F_v0	11.88	23.13	45.63	-0.9	-0.38	-0.81	11.25	0	0	11.25	7.11
B_v0	17.8	21.8	25.8	-2.11	-2.11	-2.11	16.8	0	0	15.8	4.33
F_v1	25	34.21	53.95	-0.74	-0.39	-0.98	24.34	0	0	24.34	7.27
B_v1	9.73	13.73	17.73	-0.72	-0.72	-0.72	8.73	0	0	7.73	4.35
F_v2	4.88	12.2	47.56	-0.93	0.17	-0.36	0	0	0	0	7.34
B_v2	40.38	44.38	48.38	-3.12	-3.12	-3.12	39.38	0	0	38.38	2.42
F_v3	6.31	10.81	32.43	-0.23	0.23	-0.69	1.8	0	0	1.8	6.47
B_v3	11.39	15.39	19.39	-5.16	-5.16	-5.16	10.39	0	0	9.39	2.51
F_v4	5.84	12.34	27.27	-0.4	0.67	0.39	1.95	0	0.65	1.95	5.72
B_v4	9.76	13.76	16.94	-1.35	-1.35	-0.51	8.76	0	0	7.76	5.99
F_v5	1.88	6.88	15	-0.99	-0.69	-0.75	0	0	0	0	5.4
B_v5	0	4	8	0	0	0	-1	0	0	-2	0
F_v6	0.6	0.6	22.29	-0.68	-0.68	-0.41	0	100	0	0	4.84
B_v6	0	4	8	0	0	0	-1	0	0	-2	0
F_v7	0	3.03	21.21	-1.29	-0.97	-0.67	0	0	0	0	4.51
B_v7	0	4	8	0	0	0	-1	0	0	-2	0
F_v8	0.79	1.59	19.05	-0.75	-0.61	-0.67	0.79	0	0	0.79	4.65
B_v8	0	4	8	0	0	0	-1	0	0	-2	0
F_v9	1.7	1.7	15.91	-0.56	-0.56	-0.47	0.57	0	0	0.57	3.84
B_v9	1.5	5.5	9.5	9.67	9.67	9.67	0.5	0	0	-0.5	1.08
FAVG	5.89	10.65	30.03	-0.75	-0.32	-0.54	4.07	10	0.06	4.07	5.72
BAVG	9.06	13.06	16.98	-0.28	-0.28	-0.2	8.06	0	0	7.06	2.07

PMPT	GEE20	GEE10	GEE05	FEE20	FEE10	FEE05	V_UVE	UV_VE	THE	TLE	STD20
F_v0	1.25	5	22.5	-0.62	-0.39	-0.54	1.25	100	0	1.25	4.57
B_v0	82.73	83.64	89.09	3.86	5.3	5.58	73.64	40.91	0	73.64	11.72
F_v1	0	1.32	15.13	-0.02	-0.02	-0.16	0	0	0	0	4.4
B_v1	74.77	79.44	85.05	-3.12	1.78	2.37	74.77	25	0	74.77	14.39
F_v2	0	4.88	25.61	-1.01	-0.29	-0.47	0	56.25	0	0	5.75
B_v2	67.69	67.69	72.31	6.12	6.12	4.56	55.38	46.15	0	55.38	5.6
F_v3	0.9	5.41	19.82	-0.69	-0.02	-0.41	0.9	100	0	0.9	5.1
B_v3	71.23	71.23	78.08	2.85	2.85	4.35	71.23	9.52	0	71.23	9.9
F_v4	1.95	6.49	20.78	-0.52	0.08	0.28	1.95	100	0	1.95	4.33
B_v4	85.84	87.61	92.92	2.03	6.01	5.72	85.84	42.86	0	85.84	15.13
F_v5	1.25	3.75	11.25	-0.74	-0.46	-0.62	0	100	0	0	4.44
B_v5	79.09	83.64	88.18	-1.35	4.81	6.11	75.45	10	0	75.45	14.72
F_v6	0	0	11.45	-0.31	-0.31	-0.18	0	100	0	0	3.54
B_v6	90.35	93.86	97.37	-17.14	-12.32	-6.14	85.09	15	0	85.09	7.8
F_v7	0	0	8.08	-0.57	-0.57	-0.02	0	0	0	0	3.16
B_v7	78.46	78.46	83.08	5.38	5.38	2.55	72.31	58.33	0	72.31	7.56
F_v8	0	0.79	9.52	-0.34	-0.23	-0.36	0	100	0	0	3.67
B_v8	97.5	97.5	100	-16.26	-16.26	0	95	52.63	0	95	1.77
F_v9	1.14	1.14	4.55	-0.25	-0.25	-0.29	1.14	0	0	1.14	2.84
B_v9	84.07	86.73	91.15	-4.26	-0.32	1.4	84.07	50	0	84.07	13.86
FAVG	0.65	2.88	14.87	-0.51	-0.25	-0.28	0.52	65.63	0	0.52	4.18
BAVG	81.17	82.98	87.72	-2.19	0.33	2.65	77.28	35.04	0	77.28	10.25

CORR	GEE20	GEE10	GEE05	FEE20	FEE10	FEE05	V_UVE	UV_VE	THE	TLE	STD20
v0	55.63	55.63	56.88	-0.24	-0.24	0.06	10.63	100	45	10.63	2.2
v1	43.42	46.71	52.63	-2.52	-1.66	-0.82	13.82	0	29.61	13.82	4.53
v2	29.27	37.8	39.02	-2.81	-0.78	-0.57	4.88	37.5	24.39	4.88	6.02
v3	43.24	47.75	52.25	-2.43	-1.07	-0.25	2.7	100	40.54	2.7	5.43
v4	40.26	46.75	50	-2.36	-0.52	-0.08	5.84	100	34.42	5.84	5.93
v5	50.63	51.25	55	-0.9	-0.72	-0.69	8.75	100	39.38	8.75	3.22
v6	30.72	31.33	38.55	-0.71	-0.55	0.03	8.43	100	21.69	8.43	3.84
v7	30.3	34.34	47.47	-4	-3.21	-1.46	11.11	0	19.19	11.11	5.16
v8	34.13	34.92	43.65	-2.41	-2.27	-1.05	2.38	0	31.75	2.38	4
v9	38.64	43.18	49.43	-1.86	-0.75	-0.27	3.98	33.33	34.09	3.98	4.9
FAVG	39.62	42.97	48.49	-2.02	-1.18	-0.51	7.25	57.08	32	7.25	4.52

Table A.3.3 (continued on the next page).

YIN	GEE20	GEE10	GEE05	FEE20	FEE10	FEE05	V_UVE	UV_VE	THE	TLE	STD20
v0	20.63	29.38	46.25	-0.8	-0.41	-0.41	0	100	8.75	0.63	6.6
v1	18.42	36.84	48.03	-1.73	-1.13	-0.85	0	0	6.58	3.29	9.34
v2	24.39	36.59	45.12	-0.48	-0.76	-0.13	0	100	6.1	1.22	8.45
v3	26.13	39.64	54.95	-3.49	-1.54	-1.11	0	100	1.8	0.9	8
v4	17.53	25.32	37.01	-0.57	0.28	0.83	0	100	5.19	1.3	6.29
v5	18.75	36.25	49.38	-0.34	0.35	0.05	0	100	5.63	0	8.95
v6	9.04	19.28	37.95	1.47	0.53	-0.35	0	100	2.41	2.41	7.85
v7	15.15	21.21	35.35	-1.8	-1.42	-0.66	0	0	1.01	2.02	7.22
v8	29.37	39.68	52.38	-1.69	-1.13	-0.4	0	100	10.32	5.56	7.64
v9	27.84	35.8	49.43	-0.49	-0.12	0.26	0	100	8.52	1.14	7.62
FAVG	20.72	32	45.59	-0.99	-0.53	-0.28	0	80	5.63	1.85	7.8

PRAAT	GEE20	GEE10	GEE05	FEE20	FEE10	FEE05	V_UVE	UV_VE	THE	TLE	STD20
v0	21.25	30.63	41.88	1.67	0.85	0.23	6.25	100	6.25	11.25	6.25
v1	4.61	25	37.5	-4.98	-1.59	-0.34	4.61	0	0	4.61	7.45
v2	20.73	32.93	37.8	-3.39	-0.86	-0.27	3.66	25	17.07	3.66	6.5
v3	9.91	33.33	48.65	-1.75	-1.89	-0.08	0.9	92.86	9.01	0.9	9.52
v4	9.09	21.43	29.22	-1.59	0.74	0.8	0.65	100	5.84	0.65	7.08
v5	10	36.25	42.5	0.63	-0.03	0	0	100	0	0	9.94
v6	6.02	13.86	26.51	1.21	-0.04	0.35	0	100	0	0	6.34
v7	15.15	22.22	44.44	-4.06	-2.69	-0.87	9.09	0	0	9.09	6.59
v8	24.6	32.54	43.65	-3.74	-2.35	-0.9	2.38	0	19.84	2.38	5.57
v9	13.64	26.7	39.2	-1.16	-1.02	-0.16	1.7	0	6.82	1.7	7.63
FAVG	13.5	27.49	39.14	-1.72	-0.89	-0.13	2.92	51.79	6.48	3.42	7.29

Table A.3.3. Performance with background utterance n9 at 5 dB SNR (high resolution signal).

DHO	GEE20	GEE10	GEE05	FEE20	FEE10	FEE05	V_UVE	UV_VE	THE	TLE	STD20
F_v0	1.25	12.5	37.5	-0.83	-0.37	-0.52	1.25	0	0	1.25	6.84
B_v0	14.29	18.29	22.29	0.05	0.05	0.05	13.29	0	0	12.29	0.62
F_v1	19.08	27.63	48.03	-0.69	-0.26	-0.8	18.42	0	0	18.42	7
B_v1	9.3	10.98	14.98	-3.83	-0.65	-0.65	5.98	0	0	4.98	5.52
F_v2	4.88	12.2	47.56	-0.96	0.15	-0.38	0	0	0	0	7.38
B_v2	7.69	11.69	15.69	-0.7	-0.7	-0.7	6.69	0	0	5.69	0.35
F_v3	4.5	9.91	29.73	-0.41	0.21	-0.64	0.9	0	0	0.9	6.61
B_v3	17.14	21.14	22.29	-1.61	-1.61	-0.73	16.14	0	0	15.14	2.46
F_v4	5.19	12.34	27.27	-0.72	0.44	0.22	1.3	0	0.65	1.3	5.73
B_v4	28.26	23.57	27.57	-3.21	-1.67	-1.67	18.57	0	0	17.57	6.74
F_v5	1.88	6.88	15.63	-1.06	-0.74	-0.76	0	50	0	0	5.46
B_v5	2.27	6.27	10.27	-0.74	-0.74	-0.74	1.27	0	0	0.27	0
F_v6	0.6	0.6	21.69	-0.69	-0.69	-0.37	0	100	0	0	4.85
B_v6	20.93	24.93	28.93	-0.32	-0.32	-0.32	19.93	0	0	18.93	0.65
F_v7	0	2.02	19.19	-1.29	-1.07	-0.48	0	0	0	0	4.5
B_v7	46.15	50.15	54.15	-0.5	-0.5	-0.5	45.15	0	0	44.15	0.99
F_v8	0.79	1.59	19.05	-0.74	-0.59	-0.65	0.79	0	0	0.79	4.7
B_v8	0	4	8	0	0	0	-1	0	0	0	-2
F_v9	1.7	1.7	15.34	-0.55	-0.55	-0.58	0.57	0	0	0.57	3.85
B_v9	4	8	12	-0.61	-0.61	-0.61	3	0	4	2	0.08
FAVG	3.99	8.74	28.1	-0.79	-0.35	-0.5	2.32	15	0.06	2.32	5.69
BAVG	15	17.9	21.62	-1.15	-0.68	-0.59	12.9	0	0.4	11.9	1.74

PMPT	GEE20	GEE10	GEE05	FEE20	FEE10	FEE05	V_UVE	UV_VE	THE	TLE	STD20
F_v0	6.25	10	23.75	-0.53	-0.28	-0.39	1.25	100	5	1.25	4.64
B_v0	37.29	37.29	38.14	-1	-1	-1.23	37.29	0	0	37.29	4.22
F_v1	0	0.66	15.13	-0.06	0.01	-0.18	0	0	0	0	4.27
B_v1	28.32	28.32	30.97	-0.91	-0.91	-0.83	28.32	0	0	28.32	5.06
F_v2	0	4.88	21.95	-1.03	-0.33	-0.59	0	100	0	0	5.62
B_v2	59.62	59.62	67.31	4.2	4.2	0.8	59.62	0	0	59.62	8.65
F_v3	0.9	5.41	19.82	-0.61	0.08	-0.43	0.9	100	0	0.9	5.08
B_v3	39.24	39.24	40.51	-0.29	-0.29	-0.72	39.24	0	0	39.24	5.83
F_v4	3.9	8.44	22.08	-0.6	0.01	0.14	0.65	100	3.25	0.65	4.3
B_v4	36.59	36.59	37.4	-1.75	-1.75	-1.62	34.96	0	0	36.59	3.17
F_v5	5.63	8.75	15.63	-0.68	-0.2	-0.66	0	100	5	0	4.89
B_v5	38.79	38.79	39.66	-1.92	-1.92	-1.78	33.62	0	0	38.79	4.25
F_v6	0	0	11.45	-0.38	-0.38	-0.26	0	100	0	0	3.52
B_v6	29.17	29.17	30	-0.52	-0.52	-0.68	29.17	0	0	29.17	5.01
F_v7	1.01	1.01	9.09	-0.63	-0.63	-0.1	1.01	0	0	1.01	2.96
B_v7	31.37	31.37	33.33	0.07	0.07	-0.39	31.37	0	0	31.37	5.57
F_v8	0	0.79	7.94	-0.41	-0.3	-0.25	0	100	0	0	3.52
B_v8	38.37	38.37	40.7	0.13	0.13	-0.47	38.37	0	0	38.37	5.65
F_v9	3.41	3.41	6.25	-0.27	-0.27	-0.28	0	0	3.41	0	2.76
B_v9	69.17	69.17	69.17	-0.7	-0.7	-0.7	68.42	0	0	69.17	2.59
FAVG	2.11	4.33	15.31	-0.52	-0.23	-0.3	0.38	70	1.67	0.38	4.15
BAVG	40.79	40.79	42.72	-0.27	-0.27	-0.76	40.04	0	0	40.79	5

CORR	GEE20	GEE10	GEE05	FEE20	FEE10	FEE05	V_UVE	UV_VE	THE	TLE	STD20
v0	2.5	3.75	6.88	-0.73	-0.6	-0.37	2.5	25	0	2.5	1.93
v1	0.66	0.66	8.55	-0.8	-0.8	-0.44	0.66	0	0	0.66	2.48
v2	3.66	3.66	3.66	-0.05	-0.05	-0.05	3.66	43.75	0	3.66	1.44
v3	1.8	1.8	1.8	-0.39	-0.39	-0.39	1.8	85.71	0	1.8	0.69
v4	0	0	0.65	-0.26	-0.26	-0.22	0	46.15	0	0	1.19
v5	1.25	3.75	6.25	-0.31	-0.47	-0.61	1.25	0	0	1.25	2.61
v6	0	0.6	1.81	-0.23	-0.3	-0.4	0	100	0	0	1.74
v7	0	0	3.03	-0.07	-0.07	-0.24	0	0	0	0	2.08
v8	0	0.79	1.59	-0.32	-0.19	-0.23	0	100	0	0	1.91
v9	0	0	0	-0.09	-0.09	-0.09	0	0	0	0	0.92
FAVG	0.99	1.5	3.42	-0.33	-0.32	-0.3	0.99	40.06	0	0.99	1.7

Table A.3.4 (continued on the next page).

YIN	GEE20	GEE10	GEE05	FEE20	FEE10	FEE05	V_UVE	UV_VE	THE	TLE	STD20
v0	1.25	4.38	17.5	-0.52	-0.14	-0.11	0	100	0	0	3.95
v1	3.95	7.89	19.08	0.2	-0.43	-0.43	0	0	0	2.63	5.71
v2	2.44	2.44	7.32	-0.12	-0.12	-0.15	0	100	0	2.44	3.58
v3	0	0	3.6	0.04	0.04	0.2	0	100	0	0	2.91
v4	1.95	3.9	10.39	-0.28	0.04	0.33	0	100	0	0.65	3.51
v5	1.88	4.38	11.25	-0.22	0.03	-0.3	0	100	0	0	4.25
v6	6.02	7.23	12.05	-0.05	-0.28	-0.37	0	100	0	4.22	3.65
v7	1.01	1.01	3.03	0.58	0.58	0.56	0	0	0	0	2.73
v8	5.56	7.14	12.7	0.34	0.25	-0.13	0	100	0	3.17	4.01
v9	2.27	2.84	6.25	0.21	0.33	0.15	0	100	0	0.57	2.94
FAVG	2.63	4.12	10.32	0.02	0.03	-0.03	0	80	0	1.37	3.72

PRAAT	GEE20	GEE10	GEE05	FEE20	FEE10	FEE05	V_UVE	UV_VE	THE	TLE	STD20
v0	0.63	2.5	5.63	-0.41	-0.21	-0.06	0.63	25	0	0.63	2.01
v1	0	0.66	5.92	-0.42	-0.5	-0.34	0	0	0	0	2.4
v2	0	2.44	2.44	0.21	0.48	0.48	0	56.25	0	0	2.16
v3	0	0	0	0.03	0.03	0.03	0	92.86	0	0	0.73
v4	0	0	0	-0.03	-0.03	-0.03	0	53.85	0	0	0.93
v5	0	3.13	6.25	0.47	0.08	-0.13	0	0	0	0	2.88
v6	0	0.6	1.81	0.04	-0.03	-0.13	0	100	0	0	1.61
v7	0	0	3.03	0.24	0.24	0.07	0	0	0	0	1.97
v8	0	0	1.59	0.14	0.14	0.01	0	100	0	0	1.29
v9	0	0	0	0.12	0.12	0.12	0	0	0	0	0.74
FAVG	0.06	0.93	2.67	0.04	0.03	0	0.06	42.8	0	0.06	1.67

Table A.3.4. Performance with background utterance n7 at 0 dB SNR (high resolution signal).

DHO	GEE20	GEE10	GEE05	FEE20	FEE10	FEE05	V_UVE	UV_VE	THE	TLE	STD20
F_v0	1.88	12.5	37.5	-0.82	-0.44	-0.67	1.88	0	0	1.88	6.86
B_v0	0	4	8	0	0	0	-1	0	0	-2	0
F_v1	15.13	24.34	45.39	-0.83	-0.55	-0.76	15.13	0	0	15.13	6.88
B_v1	0	4	8	0	0	0	-1	0	0	-2	0
F_v2	4.88	12.2	45.12	-0.95	0.16	-0.31	0	6.25	0	0	7.35
B_v2	0	4	8	0	0	0	-1	0	0	-2	0
F_v3	4.5	9.91	32.43	-0.4	0.22	-0.95	0.9	0	0	0.9	6.61
B_v3	2.86	6.86	10.86	-1.9	-1.9	-1.9	1.86	0	0	0.86	0
F_v4	5.19	11.69	26.62	-0.46	0.6	0.34	1.95	0	0	1.95	5.71
B_v4	0	4	8	0	0	0	-1	0	0	-2	0
F_v5	1.88	6.88	15	-1.01	-0.68	-0.74	0	0	0	0	5.49
B_v5	15.91	19.91	23.91	-0.05	-0.05	-0.05	14.91	0	0	13.91	0.49
F_v6	0	0.6	21.69	-0.84	-0.69	-0.36	0	0	0	0	5.2
B_v6	0	4	8	0	0	0	-1	0	0	-2	0
F_v7	0	3.03	19.19	-1.31	-0.97	-0.48	0	0	0	0	4.53
B_v7	23.08	27.08	31.08	0.05	0.05	0.05	22.08	0	0	21.08	0.46
F_v8	0.79	1.59	19.05	-0.76	-0.6	-0.66	0.79	0	0	0.79	4.68
B_v8	11.43	15.43	16.57	1.23	1.23	-1.04	10.43	0	2.86	9.43	4
F_v9	1.7	1.7	15.91	-0.56	-0.56	-0.47	0.57	0	0	0.57	3.84
B_v9	4	8	12	-1.05	-1.05	-1.05	3	0	0	2	0.01
FAVG	3.6	8.44	27.79	-0.79	-0.35	-0.51	2.12	0.63	0	2.12	5.71
BAVG	5.73	9.73	13.44	-0.17	-0.17	-0.4	4.73	0	0.29	3.73	0.5

PMPT	GEE20	GEE10	GEE05	FEE20	FEE10	FEE05	V_UVE	UV_VE	THE	TLE	STD20
F_v0	0.63	4.38	22.5	-0.64	-0.42	-0.39	0.63	0	0	0.63	4.55
B_v0	47.62	47.62	57.14	-1.79	-1.79	-0.76	47.62	0	0	47.62	2.42
F_v1	0	1.32	14.47	-0.13	0.02	-0.11	0	0	0	0	4.41
B_v1	46.51	46.51	55.81	-1.9	-1.9	-1.14	46.51	0	0	46.51	2.26
F_v2	0	4.88	29.27	-1.11	-0.38	-0.3	0	31.25	0	0	5.8
B_v2	65.38	65.38	65.38	-0.29	-0.29	-0.29	65.38	0	0	65.38	0.45
F_v3	0.9	5.41	19.82	-0.59	0.1	-0.26	0.9	100	0	0.9	5.11
B_v3	60	60	62.86	-0.85	-0.85	-0.45	60	1.92	0	60	1.96
F_v4	1.3	5.84	19.48	-0.48	0.12	0.36	1.3	23.08	0	1.3	4.31
B_v4	56.52	56.52	60.87	-1.21	-1.21	-0.64	56.52	2.94	0	56.52	2.03
F_v5	1.25	3.75	11.25	-0.68	-0.18	-0.6	0	0	0	0	4.53
B_v5	59.09	61.36	72.73	-2.73	-2.32	-0.18	59.09	6.06	0	59.09	3.73
F_v6	0	0	11.45	-0.42	-0.42	-0.31	0	100	0	0	3.53
B_v6	65.12	65.12	69.77	-0.96	-0.96	-0.26	65.12	0	0	65.12	1.91
F_v7	0	0	7.07	-0.67	-0.67	-0.19	0	0	0	0	3.05
B_v7	50	50	50	-0.02	-0.02	-0.02	50	0	0	50	1
F_v8	0	0.79	7.94	-0.42	-0.3	-0.27	0	0	0	0	3.61
B_v8	60	60	60	-0.4	-0.4	-0.4	60	0	0	60	0.86
F_v9	1.7	1.7	4.55	-0.28	-0.28	-0.29	1.7	0	0	1.7	2.83
B_v9	58	58	58	-0.71	-0.71	-0.71	58	0	0	58	1.4
FAVG	0.58	2.81	14.78	-0.54	-0.24	-0.24	0.45	25.43	0	0.45	4.17
BAVG	56.82	57.05	61.26	-1.09	-1.04	-0.48	56.82	1.09	0	56.82	1.8

CORR	GEE20	GEE10	GEE05	FEE20	FEE10	FEE05	V_UVE	UV_VE	THE	TLE	STD20
v0	3.13	5	8.75	-0.85	-0.65	-0.47	3.13	25	0	3.13	2.28
v1	7.24	9.21	15.13	-0.23	-0.56	-0.34	5.26	0	0	5.26	3.11
v2	1.22	3.66	4.88	-0.1	0.18	0.25	1.22	37.5	0	1.22	2.92
v3	8.11	8.11	15.32	-0.09	-0.09	-0.21	8.11	85.71	0	8.11	3.13
v4	3.25	4.55	6.49	0.03	-0.15	-0.09	3.25	46.15	0	3.25	2.26
v5	3.13	8.13	10.63	0.12	-0.5	-0.63	3.13	0	0	3.13	3.54
v6	7.83	10.24	12.05	0.09	-0.26	-0.32	5.42	100	0	5.42	2.85
v7	9.09	9.09	14.14	0.28	0.28	0.1	9.09	0	0	9.09	2.25
v8	1.59	3.17	4.76	-0.8	-0.49	-0.39	1.59	100	0	1.59	2.9
v9	1.14	1.14	1.7	-0.07	-0.07	-0.1	1.14	0	0	1.14	1.27
FAVG	4.57	6.23	9.38	-0.16	-0.23	-0.22	4.13	39.44	0	4.13	2.65

Table A.3.5 (continued on the next page).

YIN	GEE20	GEE10	GEE05	FEE20	FEE10	FEE05	V_UVE	UV_VE	THE	TLE	STD20
v0	1.88	5.63	18.75	-0.27	-0.29	-0.22	0	100	0	0.63	4.64
v1	5.92	11.18	25.66	1.15	-0.01	-0.13	0	0	0	3.95	6.42
v2	3.66	8.54	19.51	-0.13	0.15	-0.07	0	100	0	3.66	5.47
v3	2.7	2.7	12.61	0.41	0.41	0.19	0	100	0	2.7	3.81
v4	3.25	4.55	10.39	0.33	0.33	0.3	0	100	0	1.95	3.57
v5	3.75	11.25	18.75	0.75	0.24	-0.21	0	100	0	1.25	5.55
v6	9.04	13.86	20.48	0.93	-0.13	-0.31	0	100	0	3.61	5.72
v7	3.03	7.07	12.12	1.27	0.85	0.66	0	0	0	1.01	4.59
v8	7.14	9.52	14.29	0.15	0.27	0.13	0	100	0	4.76	3.66
v9	1.14	1.7	4.55	0.39	0.52	0.3	0	100	0	0	3.07
FAVG	4.15	7.6	15.71	0.5	0.23	0.06	0	80	0	2.35	4.65

PRAAT	GEE20	GEE10	GEE05	FEE20	FEE10	FEE05	V_UVE	UV_VE	THE	TLE	STD20
v0	1.25	3.75	8.13	-0.58	-0.32	-0.08	1.25	25	0	1.25	2.45
v1	4.61	6.58	13.16	-0.09	-0.4	-0.13	0	0	0	0	3.06
v2	0	3.66	6.1	0.3	0.74	0.73	0	62.5	0	0	3.19
v3	1.8	1.8	9.01	0.5	0.5	-0.03	1.8	92.86	0	1.8	2.98
v4	0	1.95	5.84	0.35	0.06	0.05	0	53.85	0	0	2.91
v5	0	5.63	9.38	0.91	0.21	-0.18	0	0	0	0	3.8
v6	3.61	7.83	10.24	0.65	0.09	-0.12	0	100	0	0	3.22
v7	0	0	11.11	0.94	0.94	0.47	0	0	0	0	3.39
v8	0	2.38	3.17	0.26	0	-0.08	0	100	0	0	2.33
v9	0	0	1.14	0.18	0.18	0.12	0	0	0	0	1.33
FAVG	1.13	3.36	7.73	0.34	0.2	0.08	0.31	43.42	0	0.31	2.87

Table A.3.5. Performance with background utterance n8 at 0 dB SNR (high resolution signal).

DHO	GEE20	GEE10	GEE05	FEE20	FEE10	FEE05	V_UVE	UV_VE	THE	TLE	STD20
F_v0	1.88	13.13	38.13	-0.82	-0.35	-0.74	1.88	0	0	1.88	6.84
B_v0	2.38	6.38	10.38	-1.1	-1.1	-1.1	1.38	0	0	0.38	0
F_v1	17.11	25	46.05	-0.59	-0.23	-0.74	17.11	0	0	17.11	6.92
B_v1	0	4	8	0	0	0	-1	0	0	-2	0
F_v2	4.88	12.2	46.34	-0.96	0.15	-0.24	0	12.5	0	0	7.35
B_v2	15.38	7.85	11.85	-9.94	-1.55	-1.55	2.85	0	0	1.85	4.84
F_v3	4.5	9.91	33.33	-0.41	0.2	-0.95	0.9	0	0	0.9	6.62
B_v3	20	24	28	-0.4	-0.4	-0.4	19	0	0	18	0.78
F_v4	5.19	12.34	27.27	-0.65	0.53	0.25	1.3	0	0	1.3	5.84
B_v4	6.52	4	8	-4.41	0	0	-1	0	0	-2	12.32
F_v5	1.88	6.88	15	-1	-0.68	-0.75	0	0	0	0	5.42
B_v5	13.64	6.27	10.27	-11.4	-0.91	-0.91	1.27	0	0	0.27	4.78
F_v6	0.6	0.6	22.89	-0.69	-0.69	-0.49	0	100	0	0	4.85
B_v6	0	4	8	0	0	0	-1	0	0	-2	0
F_v7	0	3.03	18.18	-1.29	-0.95	-0.37	0	0	0	0	4.54
B_v7	0	4	8	0	0	0	-1	0	0	-2	0
F_v8	0.79	1.59	19.05	-0.75	-0.6	-0.66	0.79	0	0	0.79	4.7
B_v8	20	24	28	0.02	0.02	0.02	19	0	0	18	1.05
F_v9	1.7	1.7	16.48	-0.56	-0.56	-0.43	0.57	0	0	0.57	3.86
B_v9	10	14	18	-0.39	-0.39	-0.39	9	0	0	8	1.17
FAVG	3.85	8.64	28.27	-0.77	-0.32	-0.51	2.25	11.25	0	2.25	5.69
BAVG	8.79	9.85	13.85	-2.76	-0.43	-0.43	4.85	0	0	3.85	2.49

PMPT	GEE20	GEE10	GEE05	FEE20	FEE10	FEE05	V_UVE	UV_VE	THE	TLE	STD20
F_v0	14.38	18.13	34.38	-0.57	-0.3	-0.34	0.63	100	13.75	0.63	4.88
B_v0	51.82	53.64	56.36	0.44	1.34	1.06	51.82	0	0	51.82	7.82
F_v1	1.97	2.63	15.13	-0.28	-0.21	-0.01	0	0	0.66	0	4.2
B_v1	41.12	41.12	43.93	1.38	1.38	1.15	32.71	0	0	41.12	6.28
F_v2	0	6.1	26.83	-1.3	-0.42	-0.37	0	43.75	0	0	5.93
B_v2	10.77	13.85	20	1.92	2.91	1.98	10.77	0	0	10.77	8.16
F_v3	0	4.5	19.82	-0.78	-0.08	-0.51	0	35.71	0	0	5.14
B_v3	38.36	38.36	45.21	1.06	1.06	1.46	38.36	0	0	38.36	7.14
F_v4	5.84	11.69	24.03	-0.61	0.2	0.32	1.95	100	3.9	1.95	4.53
B_v4	61.95	61.95	65.49	3.94	3.94	2.65	61.95	0	0	61.95	6.9
F_v5	1.25	3.75	11.88	-0.71	-0.45	-0.66	0	100	0	0	4.35
B_v5	42.73	42.73	45.45	2.29	2.29	1.61	38.18	0	0	42.73	5.59
F_v6	0	0	12.05	-0.33	-0.33	-0.17	0	100	0	0	3.58
B_v6	42.11	42.11	47.37	2.87	2.87	2.09	42.11	0	0	42.11	6.7
F_v7	0	0	8.08	-0.71	-0.71	-0.2	0	0	0	0	3.04
B_v7	47.69	47.69	50.77	2.21	2.21	1.39	47.69	0	0	47.69	7
F_v8	0	0.79	7.94	-0.33	-0.22	-0.3	0	100	0	0	3.65
B_v8	57.5	57.5	61.25	2.62	2.62	2.21	57.5	0	0	57.5	8.25
F_v9	0	0	3.41	-0.34	-0.34	-0.38	0	0	0	0	2.81
B_v9	74.34	74.34	78.76	3.77	3.77	3.62	74.34	0	0	74.34	6.97
FAVG	2.34	4.76	16.4	-0.6	-0.29	-0.26	0.26	58	1.83	0.26	4.21
BAVG	46.8	47.3	51.5	2.25	2.44	1.92	45.5	0	0	46.8	7.08

CORR	GEE20	GEE10	GEE05	FEE20	FEE10	FEE05	V_UVE	UV_VE	THE	TLE	STD20
v0	14.38	20.63	28.13	0.33	-0.42	-0.45	13.75	25	0	14.38	5.15
v1	15.13	17.76	24.34	0.01	-0.55	-0.39	9.87	0	0	9.87	4.06
v2	13.41	17.07	24.39	-0.69	-0.2	0.1	3.66	37.5	1.22	3.66	4.33
v3	24.32	24.32	27.93	0.08	0.08	-0.2	16.22	85.71	0	19.82	2.77
v4	12.99	22.73	26.62	1.6	0.21	-0.08	9.74	46.15	0	9.74	5.73
v5	16.25	26.25	28.75	1.59	-0.25	-0.6	11.25	0	0.63	15.63	5.66
v6	24.1	28.31	30.12	0.59	-0.15	-0.2	13.86	100	0	18.67	3.52
v7	14.14	14.14	19.19	0.5	0.5	0.13	7.07	0	0	7.07	2.32
v8	23.02	24.6	26.19	-0.12	-0.53	-0.39	11.11	100	1.59	15.08	3.27
v9	13.07	17.61	20.45	1.26	0.13	-0.1	7.39	0	0.57	9.09	5.06
FAVG	17.08	21.34	25.61	0.51	-0.12	-0.22	10.39	39.44	0.4	12.3	4.19

Table A.3.6 (continued on the next page).

YIN	GEE20	GEE10	GEE05	FEE20	FEE10	FEE05	V_UVE	UV_VE	THE	TLE	STD20
v0	3.75	12.5	31.88	0.73	0.29	-0.14	0	100	0	1.88	6.03
v1	9.21	15.13	30.26	1.05	-0.35	-0.05	0	0	0	3.95	6.94
v2	7.32	21.95	35.37	0.81	-0.01	0.03	0	100	0	3.66	7.68
v3	5.41	16.22	27.03	1.96	0.33	0.39	0	100	0	0.9	6.98
v4	4.55	15.58	27.27	2.6	1.07	0.45	0	100	0	1.95	6.51
v5	7.5	21.25	31.88	1.8	0.22	-0.36	0	100	0	3.13	7.29
v6	17.47	25.3	31.93	1.28	-0.34	-0.29	0	100	0	9.64	6.23
v7	11.11	15.15	24.24	1.28	0.94	0.7	0	0	0	4.04	5.05
v8	15.87	21.43	29.37	1.3	0.07	0.06	0	100	0	7.14	5.78
v9	6.25	13.07	22.73	1.64	0.6	-0.17	0	100	0	1.14	5.84
FAVG	8.84	17.76	29.19	1.44	0.28	0.06	0	80	0	3.74	6.43

PRAAT	GEE20	GEE10	GEE05	FEE20	FEE10	FEE05	V_UVE	UV_VE	THE	TLE	STD20
v0	7.5	16.88	26.25	0.95	0.1	-0.09	3.13	25	0	6.25	5.82
v1	9.21	13.82	22.37	0.27	-0.63	-0.17	1.97	0	0	3.29	4.59
v2	10.98	14.63	20.73	-0.29	0.19	0.63	0	56.25	0	0	4.12
v3	18.02	22.52	25.23	1.35	0.46	0.14	3.6	92.86	0	9.91	4.46
v4	5.84	19.48	25.32	2.89	0.56	0.28	0	53.85	0	0.65	6.83
v5	10.63	25.63	29.38	3.15	0.48	0.06	5	0	0	9.38	6.54
v6	19.88	24.1	26.51	0.86	0.18	0.08	8.43	100	0	13.86	3.36
v7	12.12	13.13	18.18	1.06	0.91	0.54	3.03	0	0	3.03	2.62
v8	18.25	24.6	24.6	1.14	-0.06	-0.06	4.76	100	0	11.9	4.49
v9	9.09	15.34	20.45	1.84	0.36	0.06	0	0	0	3.41	5.79
FAVG	12.15	19.01	23.9	1.32	0.26	0.15	2.99	42.8	0	6.17	4.86

Table A.3.6. Performance with background utterance n9 at 0 dB SNR (high resolution signal).

DHO	GEE20	GEE10	GEE05	FEE20	FEE10	FEE05	V_UVE	UV_VE	THE	TLE	STD20
F_v0	2.5	13.75	38.75	-0.84	-0.37	-0.64	2.5	0	0	2.5	6.85
B_v0	3.64	7.64	8	18.72	18.72	0	2.64	0	0	1.64	3.07
F_v1	25	34.87	53.29	-0.52	-0.29	-0.72	24.34	0	0	24.34	7.4
B_v1	28.04	32.04	32.3	-1.35	-1.35	0.7	27.04	0	0	26.04	6.84
F_v2	4.88	12.2	43.9	-0.88	0.23	-0.31	0	6.25	0	0	7.39
B_v2	7.69	11.69	14.15	-7.34	-7.34	-4.33	6.69	0	0	5.69	6.87
F_v3	4.5	9.91	31.53	-0.5	0.12	-0.94	0.9	0	0	0.9	6.59
B_v3	0	4	8	0	0	0	-1	0	0	-2	0
F_v4	5.84	12.34	26.62	-0.56	0.5	0.34	1.95	0	0.65	1.95	5.66
B_v4	16.81	20.81	20.39	8.18	8.18	4.47	15.81	0	0	14.81	7.09
F_v5	1.88	6.88	15.63	-1.05	-0.74	-0.76	0	50	0	0	5.44
B_v5	0.91	4.91	8	22.46	22.46	0	-0.09	0	0	-1.09	0
F_v6	0.6	0.6	21.69	-0.69	-0.69	-0.37	0	0	0	0	4.84
B_v6	0.88	4.88	8.88	5.25	5.25	5.25	-0.12	0	0	-1.12	0
F_v7	0	3.03	20.2	-1.26	-0.94	-0.7	0	0	0	0	4.51
B_v7	0	4	8	0	0	0	-1	0	0	-2	0
F_v8	0.79	1.59	19.05	-0.76	-0.61	-0.67	0.79	0	0	0.79	4.68
B_v8	3.75	7.75	10.5	-10.49	-10.49	-8.97	2.75	0	0	1.75	2.43
F_v9	1.7	1.7	15.91	-0.54	-0.54	-0.45	0.57	0	0	0.57	3.83
B_v9	0	4	8	0	0	0	-1	0	0	-2	0
FAVG	4.77	9.69	28.66	-0.76	-0.33	-0.52	3.11	5.63	0.06	3.11	5.72
BAVG	6.17	10.17	12.62	3.54	3.54	-0.29	5.17	0	0	4.17	2.63

PMPT	GEE20	GEE10	GEE05	FEE20	FEE10	FEE05	V_UVE	UV_VE	THE	TLE	STD20
F_v0	7.5	11.88	26.88	-0.56	-0.42	-0.45	3.13	100	4.38	3.13	4.66
B_v0	70.34	76.27	78.81	0.94	4.68	3.99	53.39	23.53	0	53.39	18.19
F_v1	1.97	3.29	17.11	-0.27	-0.11	-0.28	0	0	1.97	0	4.23
B_v1	72.57	77.88	82.3	-0.1	5.41	3.79	52.21	23.53	0	60.18	18.56
F_v2	0	3.66	29.27	-1.03	-0.47	-0.42	0	31.25	0	0	5.82
B_v2	55.77	57.69	63.46	8.17	6.94	4.68	55.77	23.53	0	55.77	9.35
F_v3	0.9	5.41	19.82	-0.59	0.09	-0.4	0.9	100	0	0.9	5.11
B_v3	54.43	64.56	68.35	-2.74	4.6	3.94	48.1	23.53	0	53.16	19.79
F_v4	7.14	11.69	25.32	-0.76	-0.13	0.08	0	100	7.14	0	4.34
B_v4	76.42	86.18	88.62	-10.32	1.91	0.23	53.66	23.53	0	59.35	21
F_v5	11.25	11.88	17.5	-0.09	-0.19	-0.68	1.25	100	10	1.25	3.42
B_v5	76.72	84.48	87.07	-7.43	1.26	-0.38	53.45	23.53	0	64.66	19.6
F_v6	0	0	13.25	-0.39	-0.39	-0.33	0	100	0	0	3.48
B_v6	69.17	79.17	81.67	-6.21	4.09	3.34	49.17	23.53	0	49.17	20.34
F_v7	0	0	7.07	-0.8	-0.8	-0.36	0	0	0	0	2.93
B_v7	50.98	52.94	56.86	8.06	6.94	5.4	50.98	23.53	0	50.98	8.88
F_v8	8.73	8.73	16.67	-0.46	-0.46	-0.51	0	0	8.73	0	3.22
B_v8	74.42	80.23	83.72	-2.35	2.4	0.97	41.86	18.75	0	51.16	18.04
F_v9	4.55	4.55	8.52	-0.22	-0.22	-0.24	0	0	4.55	0	2.83
B_v9	78.2	87.22	89.47	-10.93	1.97	0.25	50.38	23.53	0	54.14	21.8
FAVG	4.2	6.11	18.14	-0.52	-0.31	-0.36	0.53	53.13	3.68	0.53	4
BAVG	67.9	74.66	78.03	-2.29	4.02	2.62	50.9	23.05	0	55.2	17.55

CORR	GEE20	GEE10	GEE05	FEE20	FEE10	FEE05	V_UVE	UV_VE	THE	TLE	STD20
v0	5.63	6.88	14.38	-1.01	-0.86	-0.33	5	100	0.63	5	2.77
v1	0.66	0.66	1.32	-0.12	-0.12	-0.15	0.66	0	0	0.66	1.7
v2	4.88	7.32	10.98	-0.85	-0.58	-0.53	3.66	43.75	1.22	3.66	2.72
v3	6.31	7.21	10.81	-0.98	-0.78	-0.6	2.7	100	3.6	2.7	2.75
v4	3.25	3.25	3.25	-0.35	-0.35	-0.35	3.25	100	0	3.25	1.25
v5	0.63	6.25	11.88	-2.12	-1.17	-0.79	0	100	0	0	4.51
v6	3.01	3.01	3.61	-0.05	-0.05	-0.08	0	100	0	0	1.27
v7	0	0	2.02	-0.06	-0.06	0.1	0	0	0	0	1.99
v8	2.38	7.14	14.29	-0.23	0.37	-0.09	0	100	0	0	4.25
v9	2.27	3.41	7.95	-0.53	-0.37	-0.15	2.27	0	0	2.27	2.75
FAVG	2.9	4.51	8.05	-0.63	-0.4	-0.3	1.75	64.38	0.54	1.75	2.6

Table A.3.7 (continued on the next page).

YIN	GEE20	GEE10	GEE05	FEE20	FEE10	FEE05	V_UVE	UV_VE	THE	TLE	STD20
v0	1.25	3.13	20	-0.64	-0.56	-0.52	0	100	0	0	4.05
v1	1.97	3.29	11.84	0.28	0.04	-0.24	0	0	0	1.32	4.2
v2	9.76	12.2	28.05	-0.41	-0.48	-0.49	0	100	1.22	3.66	4.8
v3	4.5	7.21	17.12	-0.5	-0.62	-0.44	0	100	1.8	0	4.18
v4	0.65	4.55	9.09	0.04	0.15	0.12	0	100	0	0	4
v5	1.88	9.38	14.38	-1.94	-0.8	-0.4	0	100	0	0	4.86
v6	10.84	11.45	16.27	-0.27	-0.16	-0.36	0	100	0	6.63	2.94
v7	0	2.02	3.03	0.24	0.56	0.64	0	0	0	0	3.34
v8	0.79	7.94	17.46	0.01	0.88	0.22	0	100	0	0	5.72
v9	1.14	3.41	10.23	0.03	-0.04	-0.05	0	66.67	0	0	3.77
FAVG	3.28	6.45	14.75	-0.32	-0.1	-0.15	0	76.67	0.3	1.16	4.19

PRAAT	GEE20	GEE10	GEE05	FEE20	FEE10	FEE05	V_UVE	UV_VE	THE	TLE	STD20
v0	0	2.5	8.13	-0.76	-0.48	-0.12	0	100	0	0	2.83
v1	0	0	1.97	0.32	0.32	0.23	0	0	0	0	1.41
v2	0	6.1	9.76	-0.65	-0.3	-0.25	0	50	0	0	3.7
v3	4.5	4.5	7.21	-0.41	-0.41	-0.32	0.9	100	3.6	0.9	1.93
v4	0	0	0	-0.08	-0.08	-0.08	0	100	0	0	1.25
v5	0.63	6.88	9.38	-1.7	-0.61	-0.43	0	100	0	0	4.68
v6	3.01	3.01	3.61	0.23	0.23	0.19	0	100	0	0	1.06
v7	0	0	1.01	0.55	0.55	0.63	0	0	0	0	1.71
v8	0.79	6.35	14.29	0.17	0.64	0.16	0	100	0	0	4.51
v9	0	1.7	6.82	-0.07	-0.01	-0.03	0	0	0	0	2.78
FAVG	0.89	3.1	6.22	-0.24	-0.01	0	0.09	65	0.36	0.09	2.59

Table A.3.7. Performance with background utterance n7 at -5 dB SNR (high resolution signal).

DHO	GEE20	GEE10	GEE05	FEE20	FEE10	FEE05	V_UVE	UV_VE	THE	TLE	STD20
F_v0	13.13	23.75	45.63	-0.77	-0.36	-0.44	11.88	0	0	11.88	7.04
B_v0	23.64	27.64	27.09	6	6	3.22	22.64	0	0	21.64	6.45
F_v1	17.11	26.97	46.71	-0.54	-0.35	-0.88	17.11	0	0	17.11	7.06
B_v1	20.56	23.63	25.76	-3.42	-2.56	-1.35	18.63	0	0	17.63	6.9
F_v2	4.88	12.2	47.56	-0.91	0.19	-0.34	0	6.25	0	0	7.37
B_v2	7.69	11.69	14.15	-6.42	-6.42	-4.06	6.69	0	0	5.69	5.68
F_v3	4.5	9.91	31.53	-0.38	0.23	-0.85	0.9	0	0	0.9	6.64
B_v3	0	4	8	0	0	0	-1	0	0	-2	0
F_v4	5.19	11.69	29.22	-0.43	0.62	0.25	1.95	0	0	1.95	5.72
B_v4	15.04	19.04	18.62	8.81	8.81	4.78	14.04	0	0	13.04	7.27
F_v5	1.88	6.88	16.25	-1.05	-0.72	-0.73	0	0	0	0	5.52
B_v5	1.82	5.82	8	22.76	22.76	0	0.82	0	0	-0.18	0.68
F_v6	0	0.6	21.69	-0.77	-0.69	-0.37	0	100	0	0	4.95
B_v6	0	4	8	0	0	0	-1	0	0	-2	0
F_v7	0	3.03	20.2	-1.27	-0.95	-0.72	0	0	0	0	4.53
B_v7	0	4	8	0	0	0	-1	0	0	-2	0
F_v8	0.79	1.59	19.05	-0.75	-0.6	-0.65	0.79	0	0	0.79	4.68
B_v8	3.75	7.75	9.25	-10.25	-10.25	-7.66	2.75	0	0	1.75	2.05
F_v9	1.7	1.7	14.77	-0.54	-0.54	-0.45	0.57	0	0	0.57	3.83
B_v9	0	4	8	0	0	0	-1	0	0	-2	0
FAVG	4.92	9.83	29.26	-0.74	-0.32	-0.52	3.32	10.63	0	3.32	5.74
BAVG	7.25	11.16	13.49	1.75	1.83	-0.51	6.16	0	0	5.16	2.9

PMPT	GEE20	GEE10	GEE05	FEE20	FEE10	FEE05	V_UVE	UV_VE	THE	TLE	STD20
F_v0	5	14.38	30	-0.91	-0.37	-0.2	0	0	0	3.75	5.14
B_v0	73.81	73.81	85.71	-3.12	-3.12	-0.37	73.81	14.71	0	73.81	3.5
F_v1	0	1.32	15.13	-0.19	-0.02	-0.29	0	0	0	0	4.56
B_v1	67.44	69.77	79.07	-3.58	-3.1	-1.3	67.44	12.12	0	67.44	3.25
F_v2	0	4.88	23.17	-1.06	-0.36	-0.55	0	81.25	0	0	5.65
B_v2	50	50	69.23	-3.45	-3.45	-1.36	50	0	0	50	2.83
F_v3	0	4.5	21.62	-0.81	-0.15	-0.62	0	71.43	0	0	5.26
B_v3	48.57	51.43	65.71	-3.04	-2.67	-1.22	48.57	15.38	0	48.57	3.06
F_v4	1.95	7.14	19.48	-0.45	0.26	0.43	1.95	53.85	0	1.95	4.51
B_v4	65.22	65.22	78.26	-3.91	-3.91	-2.16	65.22	17.65	0	65.22	2.61
F_v5	8.13	12.5	17.5	-0.67	-0.54	-0.58	0	0	0	4.38	5.01
B_v5	70.45	70.45	79.55	-3.64	-3.64	-1.92	63.64	18.18	0	63.64	2.96
F_v6	6.63	12.05	24.1	0.44	-0.28	-0.12	0	100	0	6.63	4.68
B_v6	93.02	95.35	95.35	-7.65	-2.92	-2.92	90.7	13.24	0	90.7	6.7
F_v7	0	0	8.08	-0.32	-0.32	0.3	0	0	0	0	3.43
B_v7	53.85	57.69	65.38	-3.09	-2.55	-1.36	53.85	0	0	53.85	3.23
F_v8	0	2.38	8.73	-0.13	-0.31	-0.25	0	100	0	0	4.4
B_v8	57.14	57.14	71.43	-3.62	-3.62	-1.74	57.14	21.05	0	57.14	3.01
F_v9	1.7	1.7	6.25	-0.2	-0.2	-0.32	1.7	0	0	1.7	3.03
B_v9	68	76	86	1.51	-3.8	-1.24	68	17.14	0	68	9.64
FAVG	2.34	6.08	17.41	-0.43	-0.23	-0.22	0.37	40.65	0	1.84	4.57
BAVG	64.75	66.69	77.57	-3.36	-3.28	-1.56	63.84	12.95	0	63.84	4.08

CORR	GEE20	GEE10	GEE05	FEE20	FEE10	FEE05	V_UVE	UV_VE	THE	TLE	STD20
v0	22.5	30	38.75	-1.73	-0.42	-0.43	5.63	100	15.63	5.63	4.94
v1	7.89	13.82	24.34	1.28	-0.23	0.13	3.29	0	3.95	3.29	6.65
v2	12.2	29.27	41.46	2.09	-0.43	-0.58	1.22	50	7.32	1.22	9.2
v3	11.71	21.62	27.93	-1	-1.13	-0.92	2.7	100	5.41	2.7	6.8
v4	20.78	22.73	26.62	-0.55	-0.14	-0.33	14.94	100	4.55	14.94	3.64
v5	7.5	15	27.5	-2.37	-1.26	-0.87	2.5	100	5	2.5	4.97
v6	13.25	15.66	22.29	0.36	0.53	0.08	3.61	100	2.41	3.61	4.65
v7	4.04	11.11	20.2	1.6	0.82	0.28	4.04	0	0	4.04	5.54
v8	11.9	15.08	25.4	-0.26	0.35	0.02	3.97	100	7.14	3.97	4.95
v9	9.66	14.77	25	-1.64	-0.77	-0.11	4.55	0	4.55	4.55	4.81
FAVG	12.14	18.91	27.95	-0.22	-0.27	-0.27	4.64	65	5.59	4.64	5.61

Table A.3.8 (continued on the next page).

YIN	GEE20	GEE10	GEE05	FEE20	FEE10	FEE05	V_UVE	UV_VE	THE	TLE	STD20
v0	8.75	21.88	43.75	-2.16	-1.35	-0.64	0	100	2.5	0	6.55
v1	5.26	6.58	26.97	0.42	0.22	0.2	0	0	0	4.61	4.74
v2	9.76	21.95	46.34	1.35	-0.1	-0.31	0	100	0	4.88	8.55
v3	9.91	23.42	36.04	-0.32	-1.27	-1.04	0	100	3.6	0.9	7.72
v4	5.84	18.18	30.52	-0.45	0.48	0.3	0	100	1.3	0	6.15
v5	3.13	17.5	30	-3.13	-0.92	-0.54	0	100	0	0	6.47
v6	10.24	14.46	24.7	-0.05	0.42	-0.22	0	100	0	3.01	5.52
v7	2.02	11.11	23.23	1.59	0.7	1.18	0	0	0	0	6.47
v8	2.38	11.11	26.19	1.71	0.64	0.14	0	100	0	0	6.05
v9	2.27	11.36	25	-1.47	-0.41	-0.16	0	100	0	0.57	6.14
FAVG	5.96	15.76	31.27	-0.25	-0.16	-0.11	0	80	0.74	1.4	6.44

PRAAT	GEE20	GEE10	GEE05	FEE20	FEE10	FEE05	V_UVE	UV_VE	THE	TLE	STD20
v0	8.13	23.13	36.25	-2.55	-0.43	-0.08	1.88	100	3.13	1.88	6.13
v1	0	7.89	19.08	1.84	0.11	0.2	0	0	0	0	6.74
v2	0	24.39	39.02	1.62	-0.35	-0.49	0	50	0	0	9.57
v3	7.21	18.92	27.03	-0.83	-0.49	-0.5	0.9	100	3.6	0.9	7.2
v4	8.44	18.18	24.03	-0.34	0.32	-0.07	5.19	100	0	5.19	5.65
v5	3.75	14.38	26.88	-2.53	-0.73	-0.42	1.88	100	0	1.88	6.18
v6	7.23	10.84	18.07	1.13	0.86	0.25	0	100	0	0	4.84
v7	0	9.09	19.19	1.5	1.27	0.65	0	0	0	0	6.27
v8	4.76	9.52	23.81	0.32	0.97	0.26	3.17	100	0	3.17	5.39
v9	3.41	10.8	22.73	-1.44	-0.39	0.01	2.27	0	0	2.27	5.37
FAVG	4.29	14.71	25.61	-0.13	0.11	-0.02	1.53	65	0.67	1.53	6.33

Table A.3.8. Performance with background utterance n8 at -5 dB SNR (high resolution signal).

DHO	GEE20	GEE10	GEE05	FEE20	FEE10	FEE05	V_UVE	UV_VE	THE	TLE	STD20
F_v0	11.25	22.5	43.75	-0.85	-0.33	-0.56	10.63	0	0	10.63	7.13
B_v0	23.64	27.64	27.09	6.13	6.13	3.41	22.64	0	0	21.64	6.38
F_v1	13.16	21.71	44.74	-0.62	-0.19	-0.87	13.16	0	0	13.16	6.81
B_v1	27.1	30.17	31.36	-2.28	-1.53	0	25.17	0	0	24.17	7.12
F_v2	4.88	12.2	47.56	-0.96	0.14	-0.36	0	0	0	0	7.35
B_v2	7.69	11.69	14.15	-7.18	-7.18	-4.43	6.69	0	0	5.69	6.45
F_v3	4.5	9.91	34.23	-0.5	0.11	-1.13	0.9	0	0	0.9	6.58
B_v3	0	4	8	0	0	0	-1	0	0	-2	0
F_v4	5.19	12.34	28.57	-0.41	0.73	0.23	1.95	0	0	1.95	5.7
B_v4	15.04	19.04	18.62	8.7	8.7	4.56	14.04	0	0	13.04	7.38
F_v5	1.88	6.88	15	-1.01	-0.68	-0.74	0	0	0	0	5.52
B_v5	1.82	5.82	8	22.22	22.22	0	0.82	0	0	-0.18	0.17
F_v6	0.6	0.6	22.89	-0.69	-0.69	-0.48	0	100	0	0	4.85
B_v6	0	4	8	0	0	0	-1	0	0	-2	0
F_v7	0	3.03	19.19	-1.29	-0.97	-0.63	0	0	0	0	4.54
B_v7	0	4	8	0	0	0	-1	0	0	-2	0
F_v8	0.79	1.59	19.05	-0.74	-0.6	-0.65	0.79	0	0	0.79	4.66
B_v8	3.75	7.75	10.5	-10.33	-10.33	-8.85	2.75	0	0	1.75	2.36
F_v9	1.7	1.7	15.91	-0.57	-0.57	-0.48	0.57	0	0	0.57	3.85
B_v9	0	4	8	0	0	0	-1	0	0	-2	0
FAVG	4.4	9.25	29.09	-0.76	-0.3	-0.57	2.8	10	0	2.8	5.7
BAVG	7.9	11.81	14.17	1.73	1.8	-0.53	6.81	0	0	5.81	2.99

PMPT	GEE20	GEE10	GEE05	FEE20	FEE10	FEE05	V_UVE	UV_VE	THE	TLE	STD20
F_v0	5.63	9.38	25	-0.65	-0.41	-0.52	3.75	100	1.88	3.75	4.57
B_v0	69.09	69.09	72.73	1.82	1.82	2.15	48.18	68.18	0	48.18	7.49
F_v1	0	1.32	15.79	-0.04	-0.05	-0.12	0	0	0	0	4.42
B_v1	65.42	70.09	75.7	-3.39	0.37	2.03	53.27	80	0	53.27	13.43
F_v2	0	4.88	25.61	-0.97	-0.26	-0.42	0	62.5	0	0	5.76
B_v2	46.15	46.15	52.31	3.84	3.84	2.4	26.15	53.85	0	26.15	6.5
F_v3	0	4.5	20.72	-0.76	-0.1	-0.45	0	100	0	0	5.06
B_v3	42.47	42.47	52.05	1.45	1.45	1.56	24.66	28.57	0	24.66	8.53
F_v4	0.65	5.19	18.18	-0.58	0	0.26	0.65	100	0	0.65	4.24
B_v4	71.68	74.34	78.76	-1.13	2.72	2.7	61.95	71.43	0	61.95	14.74
F_v5	1.25	3.75	10.63	-0.71	-0.46	-0.68	0	100	0	0	4.32
B_v5	66.36	71.82	77.27	-1.9	2.45	3.47	50.91	70	0	50.91	13.13
F_v6	0	0	9.64	-0.21	-0.21	-0.2	0	100	0	0	3.45
B_v6	67.54	73.68	79.82	-5.42	-0.4	1.8	56.14	70	0	56.14	14.35
F_v7	0	0	8.08	-0.66	-0.66	-0.12	0	0	0	0	3.22
B_v7	60	60	63.08	2.8	2.8	1.75	38.46	66.67	0	38.46	7.01
F_v8	0	0.79	8.73	-0.38	-0.26	-0.32	0	100	0	0	3.62
B_v8	67.5	68.75	73.75	-0.56	1.15	1.19	55	73.68	0	55	12.77
F_v9	0	0	3.41	-0.29	-0.29	-0.32	0	0	0	0	2.78
B_v9	74.34	76.99	84.07	-1.93	0.61	3.07	69.91	60	0	69.91	12.62
FAVG	0.75	2.98	14.58	-0.53	-0.27	-0.29	0.44	66.25	0.19	0.44	4.14
BAVG	63.06	65.34	70.95	-0.44	1.68	2.21	48.46	64.24	0	48.46	11.06

CORR	GEE20	GEE10	GEE05	FEE20	FEE10	FEE05	V_UVE	UV_VE	THE	TLE	STD20
v0	43.75	46.25	53.75	-1.39	-0.71	-0.66	8.13	100	35	8.13	4.49
v1	27.63	32.89	47.37	-0.03	-0.93	0.34	3.95	0	19.08	3.95	6.71
v2	18.29	47.56	59.76	2.43	-0.63	-0.9	1.22	68.75	17.07	1.22	12.2
v3	18.92	28.83	45.05	-1.71	-2.07	-0.93	0	100	16.22	0	7.56
v4	34.42	38.96	47.4	-0.56	-0.05	-0.62	8.44	100	25.97	8.44	5.29
v5	28.75	35.63	45	-1.56	-0.74	-1.21	8.13	100	20	8.13	5.58
v6	31.93	38.55	45.18	-0.57	0.19	0.12	2.41	100	21.69	2.41	5.75
v7	17.17	20.2	39.39	0.3	0.5	0.37	4.04	0	12.12	4.04	5.99
v8	26.98	28.57	34.92	-0.43	-0.16	-0.2	2.38	100	24.6	2.38	4.47
v9	38.07	41.48	46.59	-1.42	-0.65	-0.52	7.95	0	28.41	7.95	4.54
FAVG	28.59	35.89	46.44	-0.49	-0.52	-0.42	4.66	66.88	22.02	4.66	6.26

Table A.3.9 (continued on the next page).

YIN	GEE20	GEE10	GEE05	FEE20	FEE10	FEE05	V_UVE	UV_VE	THE	TLE	STD20
v0	21.25	36.88	56.25	-1.59	-1.57	-1.4	0	100	6.25	1.88	7.69
v1	9.21	17.11	46.05	0.09	0.06	0.18	0	0	0.66	5.26	7.45
v2	23.17	51.22	65.85	0.03	-0.28	-0.44	0	100	1.22	9.76	12.23
v3	14.41	29.73	49.55	-1.64	-2.37	-1.6	0	100	5.41	3.6	9.08
v4	11.04	29.87	52.6	-0.35	0.87	-0.48	0	100	1.95	0.65	8.94
v5	13.13	30.63	45.63	-2.58	-1.11	-0.94	0	100	4.38	1.25	8.05
v6	18.67	32.53	47.59	0.48	1.26	0.17	0	100	2.41	5.42	8.12
v7	7.07	19.19	42.42	1.34	0.72	0.55	0	0	0	1.01	7.79
v8	19.05	27.78	44.44	-0.67	-0.27	-0.17	0	100	0	7.14	7.96
v9	16.48	30.11	44.89	-2.63	-0.91	0	0	66.67	0	2.84	7.91
FAVG	15.35	30.5	49.53	-0.75	-0.36	-0.41	0	76.67	2.23	3.88	8.52

PRAAT	GEE20	GEE10	GEE05	FEE20	FEE10	FEE05	V_UVE	UV_VE	THE	TLE	STD20
v0	14.38	31.25	48.75	-1.6	-0.38	-0.43	0.63	100	8.75	0.63	7.75
v1	0.66	13.16	38.16	1.34	-0.63	0.38	0	0	0	0	8.25
v2	0	36.59	51.22	1.58	-0.36	-0.48	0	56.25	0	0	11.79
v3	11.71	22.52	42.34	-1.83	-1.91	-0.59	0.9	100	5.41	0.9	7.82
v4	12.34	31.17	46.1	0.04	0.75	-0.19	5.19	100	0.65	5.19	8.39
v5	15	30.63	44.38	-3.13	-0.98	-0.45	9.38	100	0	9.38	7.55
v6	10.84	24.7	39.76	0.1	1	0.48	0	100	0	2.41	8.61
v7	0	14.14	36.36	1.91	1.21	0.77	0	0	0	0	8.05
v8	9.52	13.49	33.33	0.13	0.92	0.58	3.17	100	3.97	3.17	6.14
v9	15.34	26.7	40.91	-2.67	-0.66	-0.12	10.8	0	0	10.8	7.01
FAVG	8.98	24.43	42.13	-0.41	-0.1	0	3.01	65.63	1.88	3.25	8.14

Table A.3.9. Performance with background utterance n9 at -5 dB SNR (high resolution signal).

DHO	GEE20	GEE10	GEE05	FEE20	FEE10	FEE05	V_UVE	UV_VE	THE	TLE	STD20
F_v0	8.92	20.38	43.31	-0.71	-0.5	-0.5	8.28	0	0	8.28	6.63
B_v0	0	4	8	0	0	0	-1	0	0	-2	0
F_v1	5.88	16.34	29.41	-0.65	-0.38	-0.51	5.88	0	0	5.88	6.53
B_v1	5.31	9.31	13.31	-4.75	-4.75	-4.75	4.31	0	0	3.31	1.67
F_v2	7.41	19.75	48.15	-1.25	0.34	-0.3	2.47	0	0	2.47	8.17
B_v2	38.46	42.46	46.46	-2.83	-2.83	-2.83	37.46	0	0	36.46	1.95
F_v3	8.18	15.45	38.18	-0.77	0.5	-1.02	5.45	0	0	5.45	7.1
B_v3	18.75	22.75	26.75	-4.55	-4.55	-4.55	17.75	0	0	16.75	3.28
F_v4	5.88	12.42	26.14	-0.19	0.93	0.3	2.61	0	0	2.61	5.71
B_v4	4.88	8.88	12.07	7.12	7.12	5.65	3.88	0	0	2.88	3.95
F_v5	2.5	8.13	16.25	-0.89	-0.94	-0.82	0.63	50	0	0.63	5.7
B_v5	0	4	8	0	0	0	-1	0	0	-2	0
F_v6	0.6	3.01	25.3	-1.23	-0.91	-0.69	0.6	0	0	0.6	5.5
B_v6	0	4	8	0	0	0	-1	0	0	-2	0
F_v7	1.02	7.14	23.47	-1.73	-0.95	-0.91	1.02	0	0	1.02	5.25
B_v7	0	4	8	0	0	0	-1	0	0	-2	0
F_v8	0.81	1.61	25	-0.94	-0.85	-0.65	0.81	40	0	0.81	4.97
B_v8	0	4	8	0	0	0	-1	0	0	-2	0
F_v9	2.29	3.43	20	-0.56	-0.44	-0.56	2.29	25	0	2.29	4.35
B_v9	4.48	8.48	11.73	10.43	10.43	9.55	3.48	0	0	2.48	2.54
FAVG	4.35	10.77	29.52	-0.89	-0.32	-0.56	3	11.5	0	3	5.99
BAVG	7.19	11.19	15.03	0.54	0.54	0.31	6.19	0	0	5.19	1.34

CORR	GEE20	GEE10	GEE05	FEE20	FEE10	FEE05	V_UVE	UV_VE	THE	TLE	STD20
v0	11.46	11.46	12.1	-0.29	-0.29	-0.25	2.55	80	8.92	2.55	1.31
v1	4.58	6.54	11.76	-1.22	-0.95	-0.51	3.92	0	0.65	3.92	3.03
v2	9.88	16.05	19.75	-1.68	-0.53	-0.11	1.23	25	8.64	1.23	4.98
v3	12.73	14.55	18.18	-0.83	-0.41	0	0.91	100	8.18	0.91	3.8
v4	7.84	7.84	9.15	0.07	0.07	0.24	3.92	78.57	3.92	3.92	1.93
v5	9.38	11.25	14.38	-0.45	-0.33	-0.07	1.88	100	6.88	1.88	3.3
v6	2.41	2.41	3.61	-0.52	-0.52	-0.44	2.41	100	0	2.41	1.76
v7	0	0	8.16	-1.27	-1.27	-0.6	0	0	0	0	2.87
v8	8.87	8.87	10.48	-0.98	-0.98	-0.85	0.81	0	8.06	0.81	2.28
v9	7.43	8	8	-0.24	-0.15	-0.15	1.71	0	5.71	1.71	1.77
FAVG	7.46	8.7	11.56	-0.74	-0.54	-0.27	1.93	48.36	5.1	1.93	2.7

Table A.3.10 (continued on the next page).

YIN	GEE20	GEE10	GEE05	FEE20	FEE10	FEE05	V_UVE	UV_VE	THE	TLE	STD20
v0	5.1	5.73	20.38	0.13	0.04	0.06	0	100	2.55	1.27	3.96
v1	12.42	16.34	24.18	0.46	-0.2	-0.26	0	0	0	11.76	5.51
v2	24.69	37.04	50.62	0.45	-0.51	-0.68	0	100	2.47	13.58	8.55
v3	11.82	13.64	20.91	-0.54	-0.17	-0.51	0	100	0	7.27	4.37
v4	7.84	11.76	17.65	0.23	0.1	0.21	0	100	1.31	5.23	4.77
v5	12.5	16.88	26.25	-1.04	-0.36	-0.23	0	100	0	5.63	5.32
v6	16.27	19.88	27.11	0.39	0.28	-0.19	0	100	0	14.46	5.32
v7	13.27	17.35	18.37	0.12	-0.44	-0.31	0	0	0	10.2	4.88
v8	24.19	28.23	34.68	0.72	0.59	0	0	100	0.81	16.94	4.59
v9	9.71	14.86	20	0.44	0.13	0.26	0	100	0	6.86	5.26
FAVG	13.78	18.17	26.01	0.14	-0.06	-0.17	0	80	0.71	9.32	5.25

PRAAT	GEE20	GEE10	GEE05	FEE20	FEE10	FEE05	V_UVE	UV_VE	THE	TLE	STD20
v0	1.91	1.91	5.1	0.1	0.1	-0.13	1.91	80	0	1.91	1.78
v1	0	1.96	7.19	-0.86	-0.58	-0.16	0	0	0	0	2.99
v2	7.41	14.81	19.75	-1.06	-0.16	0.1	1.23	18.75	3.7	1.23	5.14
v3	0	0.91	6.36	-0.5	-0.35	-0.02	0	100	0	0	2.86
v4	2.61	2.61	3.92	0.3	0.3	0.46	1.31	78.57	1.31	1.31	1.84
v5	0.63	6.25	10	0.8	0.24	0.23	0	100	0	0	4.94
v6	0	0	4.22	-0.08	-0.08	-0.22	0	100	0	0	2.14
v7	0	0	9.18	-0.92	-0.92	-0.28	0	0	0	0	3.1
v8	0	0	3.23	-0.84	-0.84	-0.6	0	0	0	0	2.21
v9	0	0.57	4	-0.26	-0.18	-0.05	0	0	0	0	2.12
FAVG	1.26	2.9	7.29	-0.33	-0.25	-0.07	0.45	47.73	0.5	0.45	2.91

Table A.3.10. Performance with background utterance n7 at 5 dB SNR (low resolution signal).

DHO	GEE20	GEE10	GEE05	FEE20	FEE10	FEE05	V_UVE	UV_VE	THE	TLE	STD20
F_v0	0.64	12.1	36.31	-0.85	-0.67	-0.74	0.64	0	0	0.64	6.54
B_v0	0	4	8	0	0	0	-1	0	0	-2	0
F_v1	5.23	15.69	28.1	-0.59	-0.36	-0.5	5.23	0	0	5.23	6.41
B_v1	22.12	26.12	28.35	-3.2	-3.2	-2.46	21.12	0	0	20.12	4.21
F_v2	4.94	14.81	44.44	-1.17	0.04	-0.44	0	0	0	0	7.93
B_v2	40.38	44.38	48.38	-3.1	-3.1	-3.1	39.38	0	0	38.38	2.13
F_v3	8.18	15.45	38.18	-0.68	0.59	-0.94	5.45	0	0	5.45	7.13
B_v3	17.5	21.5	25.5	-4.9	-4.9	-4.9	16.5	0	0	15.5	3.12
F_v4	6.54	13.73	25.49	-0.28	0.94	0.37	2.61	0	0	2.61	5.81
B_v4	4.88	8.88	12.07	6.64	6.64	5.08	3.88	0	0	2.88	4.36
F_v5	2.5	8.13	16.25	-0.95	-0.99	-0.88	0.63	0	0	0.63	5.85
B_v5	0	4	8	0	0	0	-1	0	0	-2	0
F_v6	0.6	3.61	24.1	-1.31	-0.93	-0.87	0.6	0	0	0.6	5.42
B_v6	0	4	8	0	0	0	-1	0	0	-2	0
F_v7	1.02	7.14	24.49	-1.76	-0.97	-0.88	1.02	0	0	1.02	5.31
B_v7	0	4	8	0	0	0	-1	0	0	-2	0
F_v8	1.61	4.03	23.39	-1.03	-0.76	-0.77	1.61	40	0	1.61	5
B_v8	1.16	5.16	9.16	6.95	6.95	6.95	0.16	0	0	-0.84	0
F_v9	1.71	3.43	20	-0.74	-0.47	-0.58	1.14	0	0	1.14	4.72
B_v9	6.72	10.72	14.72	3.41	3.41	3.41	5.72	0	0	4.72	5.26
FAVG	3.3	9.81	28.08	-0.94	-0.36	-0.62	1.89	4	0	1.89	6.01
BAVG	9.28	13.28	17.02	0.58	0.58	0.5	8.28	0	0	7.28	1.91

CORR	GEE20	GEE10	GEE05	FEE20	FEE10	FEE05	V_UVE	UV_VE	THE	TLE	STD20
v0	31.85	31.85	33.76	-0.2	-0.2	-0.2	3.18	80	24.84	3.18	1.66
v1	20.26	22.88	29.41	-1.56	-1.12	-0.39	2.61	0	17.65	2.61	3.76
v2	19.75	28.4	30.86	-2.55	-0.83	-0.38	2.47	25	17.28	2.47	5.68
v3	30.91	35.45	40.91	-2.21	-1	-0.23	2.73	100	27.27	2.73	5.37
v4	18.95	21.57	24.18	-0.68	-0.08	0.31	4.58	100	13.73	4.58	4.24
v5	32.5	33.13	36.25	-0.4	-0.51	-0.25	0.63	100	28.75	0.63	2.67
v6	12.65	12.65	15.06	-0.61	-0.61	-0.5	0	100	11.45	0	2.35
v7	11.22	11.22	24.49	-1.97	-1.97	-0.61	2.04	0	9.18	2.04	4
v8	26.61	26.61	33.06	-1.47	-1.47	-0.75	2.42	0	24.19	2.42	2.96
v9	28	29.14	30.29	-0.64	-0.41	-0.28	1.71	0	26.29	1.71	2.49
FAVG	23.27	25.29	29.83	-1.23	-0.82	-0.33	2.24	50.5	20.06	2.24	3.52

Table A.3.11 (continued on the next page).

YIN	GEE20	GEE10	GEE05	FEE20	FEE10	FEE05	V_UVE	UV_VE	THE	TLE	STD20
v0	15.29	19.75	32.48	-0.44	0.09	-0.13	0	100	3.18	6.37	5.37
v1	16.34	23.53	32.03	-0.05	-0.8	-0.23	0	0	1.31	11.11	6.63
v2	30.86	41.98	48.15	0.32	0.42	-0.23	0	100	4.94	17.28	7.49
v3	21.82	29.09	40.91	-1.29	-0.92	-1.08	0	100	1.82	10.91	7.25
v4	9.8	13.73	27.45	-0.81	-0.38	-0.01	0	100	1.96	3.92	5.22
v5	14.38	21.25	30	-0.37	-0.08	-0.04	0	100	1.25	8.75	6.26
v6	16.87	20.48	30.12	0.91	0.32	0.01	0	100	0	15.06	5.55
v7	4.08	9.18	19.39	-0.9	-0.78	-0.36	0	0	0	2.04	5.56
v8	21.77	27.42	33.06	-0.7	-0.17	-0.02	0	100	0.81	12.1	5.57
v9	12	18.29	22.29	0.57	0.14	0.06	0	100	0	7.43	5.18
FAVG	16.32	22.47	31.59	-0.28	-0.22	-0.2	0	80	1.53	9.5	6.01

PRAAT	GEE20	GEE10	GEE05	FEE20	FEE10	FEE05	V_UVE	UV_VE	THE	TLE	STD20
v0	3.18	7.64	19.75	0.27	0	-0.17	0.64	80	2.55	0.64	4.61
v1	0	6.54	16.99	-0.79	-0.92	-0.07	0	0	0	0	4.6
v2	7.41	16.05	22.22	-1.86	-0.82	-0.12	1.23	18.75	3.7	1.23	5.52
v3	6.36	22.73	33.64	-0.83	-1.18	0.08	0	100	0	0	8.16
v4	7.19	11.11	15.69	-0.56	0.29	0.61	1.96	100	5.23	1.96	4.79
v5	10.63	21.25	25.63	-0.41	-0.01	0.25	0	100	0	5	6.59
v6	0	0.6	8.43	-0.05	-0.12	-0.11	0	100	0	0	3.01
v7	0	0	17.35	-1.29	-1.29	-0.27	0	0	0	0	4.3
v8	9.68	12.9	21.77	-2.11	-1.6	-0.75	0.81	0	8.87	0.81	4.17
v9	6.29	9.14	14.86	-0.43	-0.5	-0.26	1.14	0	5.14	1.14	3.55
FAVG	5.07	10.8	19.63	-0.81	-0.62	-0.08	0.58	49.88	2.55	1.08	4.93

Table A.3.11. Performance with background utterance n8 at 5 dB SNR (low resolution signal).

DHO	GEE20	GEE10	GEE05	FEE20	FEE10	FEE05	V_UVE	UV_VE	THE	TLE	STD20
F_v0	7.01	19.11	41.4	-0.85	-0.56	-0.63	7.01	0	0	7.01	6.73
B_v0	0	4	8	0	0	0	-1	0	0	-2	0
F_v1	5.88	15.69	29.41	-0.66	-0.51	-0.52	5.88	0	0	5.88	6.4
B_v1	16.81	20.81	22.16	-2.66	-2.66	-2.44	15.81	0	0	14.81	6.52
F_v2	4.94	14.81	46.91	-1	0.2	-0.7	0	0	0	0	7.89
B_v2	38.46	42.46	46.46	-3.04	-3.04	-3.04	37.46	0	0	36.46	2.28
F_v3	4.55	10.91	32.73	-0.59	0.18	-0.97	0.91	0	0	0.91	6.77
B_v3	7.5	11.5	15.5	-2.63	-2.63	-2.63	6.5	0	0	5.5	1.97
F_v4	7.84	15.03	26.8	-0.43	0.85	0.41	4.58	0	0	4.58	5.96
B_v4	5.69	9.69	13.69	5.96	5.96	5.96	4.69	0	0	3.69	2.51
F_v5	3.13	8.13	16.25	-0.81	-0.97	-0.86	0.63	50	0	0.63	5.63
B_v5	0.86	4.86	8.86	9.58	9.58	9.58	-0.14	0	0	-1.14	0
F_v6	0.6	3.61	24.1	-1.21	-0.83	-0.86	0.6	0	0	0.6	5.43
B_v6	0.83	4.83	8.83	6.39	6.39	6.39	-0.17	0	0	-1.17	0
F_v7	1.02	7.14	22.45	-1.78	-0.98	-0.85	1.02	0	0	1.02	5.29
B_v7	0	4	8	0	0	0	-1	0	0	-2	0
F_v8	0.81	2.42	24.19	-0.94	-0.77	-0.96	0.81	60	0	0.81	4.93
B_v8	1.16	5.16	9.16	9.92	9.92	9.92	0.16	0	0	-0.84	0
F_v9	1.71	4.57	22.29	-0.88	-0.47	-0.53	1.14	0	0	1.14	4.84
B_v9	6.72	10.72	14.72	3.46	3.46	3.46	5.72	0	0	4.72	5.26
FAVG	3.75	10.14	28.65	-0.91	-0.39	-0.65	2.26	11	0	2.26	5.99
BAVG	7.8	11.8	15.54	2.7	2.7	2.72	6.8	0	0	5.8	1.85

CORR	GEE20	GEE10	GEE05	FEE20	FEE10	FEE05	V_UVE	UV_VE	THE	TLE	STD20
v0	50.32	50.32	50.32	0.02	0.02	0.02	3.82	100	45.22	3.82	1.07
v1	39.87	41.18	44.44	-1.3	-1	-0.54	1.96	0	37.91	1.96	3.04
v2	29.63	35.8	38.27	-2.36	-0.82	-0.41	1.23	25	28.4	1.23	5.64
v3	45.45	49.09	52.73	-2.08	-0.91	-0.1	2.73	100	42.73	2.73	5.31
v4	43.14	45.1	47.06	-1.3	-0.73	-0.28	1.31	100	41.18	1.31	3.94
v5	48.75	48.75	48.75	-0.51	-0.51	-0.51	1.25	100	45.63	1.25	1.36
v6	31.93	31.93	33.73	-0.28	-0.28	-0.01	2.41	100	28.31	2.41	2.71
v7	28.57	29.59	40.82	-2.68	-2.51	-0.88	2.04	0	26.53	2.04	4.42
v8	40.32	40.32	45.16	-1.52	-1.52	-0.75	3.23	0	37.1	3.23	3.37
v9	41.71	42.29	45.14	-0.6	-0.47	-0.2	0	0	40.57	0	2.55
FAVG	39.97	41.44	44.64	-1.26	-0.87	-0.37	2	52.5	37.36	2	3.34

Table A.3.12 (continued on the next page).

YIN	GEE20	GEE10	GEE05	FEE20	FEE10	FEE05	V_UVE	UV_VE	THE	TLE	STD20
v0	26.75	33.76	52.23	-1.18	-0.55	-0.04	0	100	7.01	10.19	6.21
v1	31.37	43.14	50.98	-0.41	-0.74	-0.52	0	0	9.15	16.99	8.6
v2	33.33	41.98	49.38	1.56	-0.04	-0.08	0	100	1.23	19.75	7.58
v3	38.18	45.45	60.91	-0.03	-0.8	-1.55	0	100	6.36	20.91	8.51
v4	26.8	33.99	49.67	-1.13	-0.71	-0.1	0	100	7.84	8.5	6.48
v5	30	41.25	51.88	-0.43	-0.05	-0.67	0	100	5	20.63	8.63
v6	33.73	42.77	53.01	0.16	0.28	-0.05	0	100	1.2	20.48	8.74
v7	16.33	20.41	33.67	-1.46	-1.04	-0.79	0	0	0	11.22	5.18
v8	40.32	53.23	65.32	0.66	0.26	-0.22	0	100	6.45	20.97	9.75
v9	32	43.43	53.14	-0.03	-0.07	0.08	0	100	3.43	18.86	8.38
FAVG	30.88	39.94	52.02	-0.23	-0.35	-0.4	0	80	4.77	16.85	7.81

PRAAT	GEE20	GEE10	GEE05	FEE20	FEE10	FEE05	V_UVE	UV_VE	THE	TLE	STD20
v0	32.48	34.39	35.67	0.63	0.25	0.21	1.91	80	17.83	8.28	2.74
v1	22.22	28.76	35.29	-2.49	-1.13	-0.34	1.96	0	13.73	1.96	5.4
v2	18.52	29.63	34.57	-3.06	-0.88	-0.3	2.47	25	16.05	2.47	6.06
v3	18.18	41.82	50	-1.06	-0.94	0.18	2.73	100	15.45	2.73	9.63
v4	18.95	26.14	28.1	-1.31	0.42	0.61	3.27	100	12.42	3.27	5.98
v5	33.75	40.63	45	-1.13	0.03	0.17	3.13	100	19.38	9.38	5.55
v6	15.06	16.27	25.9	0.15	-0.2	0.35	0	100	0	0	4.46
v7	14.29	17.35	36.73	-3.4	-2.92	-0.89	2.04	0	12.24	2.04	4.82
v8	31.45	34.68	44.35	-2.71	-2.04	-0.93	3.23	0	25	3.23	4.73
v9	34.86	37.71	44.57	-1.36	-0.74	0.12	4.57	0	30.29	4.57	4.34
FAVG	23.98	30.74	38.02	-1.57	-0.82	-0.08	2.53	50.5	16.24	3.79	5.37

Table A.3.12. Performance with background utterance n9 at 5 dB SNR (low resolution signal).

DHO	GEE20	GEE10	GEE05	FEE20	FEE10	FEE05	V_UVE	UV_VE	THE	TLE	STD20
F_v0	5.73	17.2	39.49	-0.85	-0.67	-0.61	5.73	0	0	5.73	6.67
B_v0	24.39	28.39	32.39	-0.14	-0.14	-0.14	23.39	0	0	22.39	0.65
F_v1	5.23	15.69	28.1	-0.64	-0.38	-0.39	5.23	0	0	5.23	6.45
B_v1	14.29	18.29	22.29	0.14	0.14	0.14	13.29	0	0	12.29	0.55
F_v2	4.94	17.28	46.91	-1.26	0.27	-0.54	0	0	0	0	8.06
B_v2	34.62	38.62	42.62	0.03	0.03	0.03	33.62	0	0	32.62	0.73
F_v3	3.64	9.09	30	-0.41	0.21	-0.98	0.91	0	0	0.91	6.61
B_v3	29.41	33.41	37.41	-0.08	-0.08	-0.08	28.41	0	0	27.41	0.67
F_v4	7.19	14.38	27.45	-0.44	0.84	0.42	3.92	0	0	3.92	5.99
B_v4	24.44	12.89	16.89	-7.93	-0.53	-0.53	7.89	0	0	6.89	5.63
F_v5	3.75	8.13	16.25	-0.63	-0.96	-0.84	3.13	50	0	3.13	5.24
B_v5	11.9	15.9	19.9	0.66	0.66	0.66	10.9	0	0	9.9	0.47
F_v6	1.2	3.01	25.9	-1.14	-0.93	-0.88	1.2	100	0	1.2	5.3
B_v6	28.57	32.57	36.57	-0.21	-0.21	-0.21	27.57	1.45	0	26.57	0.69
F_v7	1.02	7.14	24.49	-1.84	-1.04	-0.97	1.02	0	0	1.02	5.32
B_v7	23.08	27.08	31.08	-0.2	-0.2	-0.2	22.08	0	0	21.08	0.73
F_v8	1.61	2.42	25.81	-1.15	-1.05	-0.85	1.61	40	0	1.61	5.02
B_v8	17.14	21.14	25.14	0.15	0.15	0.15	16.14	0	0	15.14	0.54
F_v9	1.71	4	20.57	-0.87	-0.54	-0.64	1.14	25	0	1.14	4.84
B_v9	8.33	12.33	16.33	-0.47	-0.47	-0.47	7.33	0	0	6.33	0.59
FAVG	3.6	9.83	28.5	-0.92	-0.42	-0.63	2.39	21.5	0	2.39	5.95
BAVG	21.62	24.06	28.06	-0.81	-0.06	-0.06	19.06	0.14	0	18.06	1.13

CORR	GEE20	GEE10	GEE05	FEE20	FEE10	FEE05	V_UVE	UV_VE	THE	TLE	STD20
v0	7.01	8.92	10.83	-0.62	-0.42	-0.27	3.82	0	3.18	3.82	1.87
v1	7.19	7.19	9.8	-0.4	-0.4	-0.18	7.19	0	0	7.19	1.8
v2	3.7	3.7	3.7	-0.02	-0.02	-0.02	3.7	18.75	0	3.7	1.46
v3	0.91	0.91	0.91	-0.23	-0.23	-0.23	0	71.43	0.91	0	0.83
v4	1.31	1.31	1.31	-0.29	-0.29	-0.29	1.31	42.86	0	1.31	0.83
v5	4.38	5.63	7.5	-0.22	-0.2	-0.39	4.38	0	0	4.38	2.2
v6	1.2	1.81	3.01	-0.11	-0.18	-0.28	1.2	100	0	1.2	1.73
v7	1.02	1.02	5.1	-0.06	-0.06	-0.18	1.02	0	0	1.02	2.07
v8	1.61	2.42	2.42	-0.11	-0.2	-0.2	1.61	80	0	1.61	1.55
v9	0	0	0	-0.02	-0.02	-0.02	0	25	0	0	0.97
FAVG	2.83	3.29	4.46	-0.21	-0.2	-0.21	2.42	33.8	0.41	2.42	1.53

Table A.3.13 (continued on the next page).

YIN	GEE20	GEE10	GEE05	FEE20	FEE10	FEE05	V_UVE	UV_VE	THE	TLE	STD20
v0	4.46	7.01	20.38	-0.31	-0.02	-0.26	0	100	0	3.82	4.08
v1	10.46	13.73	22.22	0.36	0.09	-0.12	0	0	0	9.15	5.17
v2	18.52	25.93	34.57	1.56	0.5	-0.49	0	100	0	13.58	7.02
v3	13.64	16.36	26.36	0.84	0.17	-0.18	0	100	0	10.91	5.41
v4	7.19	11.11	18.95	0.19	0.12	0.17	0	100	0	5.88	5.13
v5	6.88	10.63	18.75	-0.26	0.01	-0.35	0	100	0	5	4.98
v6	23.49	27.71	32.53	0.48	0.07	-0.32	0	100	0	22.29	5.94
v7	14.29	16.33	17.35	0.46	0.46	0.39	0	0	0	13.27	3.39
v8	19.35	23.39	30.65	0.36	0.15	-0.51	0	100	0	16.94	5.52
v9	9.71	13.14	16.57	1.18	0.36	0.29	0	100	0	7.43	5.07
FAVG	12.8	16.53	23.83	0.48	0.19	-0.14	0	80	0	10.83	5.17

PRAAT	GEE20	GEE10	GEE05	FEE20	FEE10	FEE05	V_UVE	UV_VE	THE	TLE	STD20
v0	0.64	3.18	5.73	-0.42	-0.16	-0.06	0.64	20	0	0.64	2.01
v1	0	0.65	3.92	-0.16	-0.24	-0.19	0	0	0	0	2.17
v2	0	0	0	0.4	0.4	0.4	0	18.75	0	0	1.24
v3	0	0	0	-0.02	-0.02	-0.02	0	85.71	0	0	0.36
v4	0	0	0.65	0	0	-0.05	0	42.86	0	0	0.83
v5	0	3.13	5	0.43	0.05	-0.13	0	0	0	0	2.64
v6	0	0.6	1.81	0.02	-0.05	-0.14	0	100	0	0	1.57
v7	0	0	3.06	0.27	0.27	0.1	0	0	0	0	1.96
v8	0	1.61	2.42	0.23	0.06	0	0	80	0	0	1.7
v9	0	0	0	0.11	0.11	0.11	0	25	0	0	0.65
FAVG	0.06	0.92	2.26	0.09	0.04	0	0.06	37.23	0	0.06	1.51

Table A.3.13. Performance with background utterance n7 at 0 dB SNR (low resolution signal).

DHO	GEE20	GEE10	GEE05	FEE20	FEE10	FEE05	V_UVE	UV_VE	THE	TLE	STD20
F_v0	0.64	12.1	35.03	-0.82	-0.65	-0.52	0.64	0	0	0.64	6.53
B_v0	24.39	28.39	32.39	-0.05	-0.05	-0.05	23.39	0	0	22.39	0.68
F_v1	5.23	15.69	29.41	-0.73	-0.44	-0.47	5.23	0	0	5.23	6.58
B_v1	16.67	20.67	24.67	0.06	0.06	0.06	15.67	0	0	14.67	0.42
F_v2	4.94	16.05	48.15	-1.23	0.1	-0.73	0	0	0	0	8.01
B_v2	26.92	7.85	11.85	8.81	-1.08	-1.08	2.85	0	0	1.85	4.06
F_v3	3.64	10	31.82	-0.36	0.11	-0.82	0.91	0	0	0.91	6.69
B_v3	32.35	36.35	40.35	-0.27	-0.27	-0.27	31.35	0	0	30.35	0.7
F_v4	5.23	11.11	23.53	-0.22	0.85	0.38	1.96	0	0	1.96	5.72
B_v4	13.33	17.33	21.33	0.21	0.21	0.21	12.33	0	0	11.33	0.3
F_v5	3.13	8.13	16.25	-0.83	-0.98	-0.87	0.63	50	0	0.63	5.67
B_v5	23.81	27.81	31.81	-0.04	-0.04	-0.04	22.81	0	0	21.81	0.59
F_v6	1.2	3.61	25.9	-1.12	-0.85	-0.71	1.2	100	0	1.2	5.26
B_v6	26.19	30.19	34.19	-0.16	-0.16	-0.16	25.19	0	0	24.19	0.6
F_v7	1.02	7.14	23.47	-1.83	-1.05	-0.88	1.02	0	0	1.02	5.25
B_v7	23.08	27.08	31.08	0.02	0.02	0.02	22.08	0	0	21.08	0.77
F_v8	1.61	3.23	25.81	-1.09	-0.91	-0.9	1.61	40	0	1.61	5.01
B_v8	28.57	32.57	36.57	-0.25	-0.25	-0.25	27.57	1.72	0	26.57	0.57
F_v9	1.71	4.57	21.14	-0.81	-0.42	-0.56	1.14	0	0	1.14	4.77
B_v9	14.58	18.58	22.58	0.14	0.14	0.14	13.58	0	0	12.58	0.41
FAVG	2.83	9.16	28.05	-0.9	-0.42	-0.61	1.43	19	0	1.43	5.95
BAVG	22.99	24.68	28.68	0.85	-0.14	-0.14	19.68	0.17	0	18.68	0.91

CORR	GEE20	GEE10	GEE05	FEE20	FEE10	FEE05	V_UVE	UV_VE	THE	TLE	STD20
v0	9.55	12.1	14.01	-0.71	-0.43	-0.28	6.37	20	3.18	6.37	2.12
v1	10.46	12.42	18.3	0.2	-0.15	-0.16	5.88	0	0	5.88	3.14
v2	2.47	4.94	4.94	0	0.28	0.28	2.47	18.75	0	2.47	2.65
v3	8.18	8.18	11.82	-0.18	-0.18	-0.25	7.27	85.71	0.91	7.27	2.31
v4	6.54	7.19	7.84	-0.16	-0.26	-0.22	6.54	50	0	6.54	1.67
v5	6.25	8.75	12.5	0.11	-0.03	-0.46	5.63	0	0	6.25	3.17
v6	9.04	12.05	13.86	0.44	0	-0.07	4.22	100	0	6.63	2.99
v7	11.22	11.22	17.35	0.29	0.29	0	11.22	0	0	11.22	2.4
v8	4.84	4.84	5.65	-0.35	-0.35	-0.29	0.81	80	2.42	0.81	1.61
v9	2.29	2.29	2.86	-0.06	-0.06	-0.11	1.14	25	1.14	1.14	1.27
FAVG	7.08	8.4	10.91	-0.04	-0.09	-0.15	5.15	37.95	0.77	5.46	2.33

Table A.3.14 (continued on the next page).

YIN	GEE20	GEE10	GEE05	FEE20	FEE10	FEE05	V_UVE	UV_VE	THE	TLE	STD20
v0	1.27	5.73	19.11	-0.04	0.06	-0.22	0	100	0	0.64	5.2
v1	15.69	18.3	28.1	0.82	0.09	0.02	0	0	0	11.76	5.41
v2	18.52	27.16	38.27	0.9	0.45	-0.18	0	100	0	14.81	7.4
v3	11.82	14.55	24.55	0.89	0.55	0.09	0	100	0	8.18	5.3
v4	7.84	10.46	17.65	-0.13	0.13	0.1	0	100	0	3.92	4.3
v5	5.63	11.25	21.25	0.35	0.14	-0.15	0	100	0	3.13	5.74
v6	24.7	30.72	37.95	0.64	-0.13	-0.45	0	100	0	19.88	6.65
v7	14.29	16.33	22.45	0.69	0.68	0.63	0	0	0	13.27	3.97
v8	15.32	20.16	28.23	0.92	-0.02	-0.48	0	100	0	12.9	5.24
v9	6.86	8.57	13.14	1.01	0.59	0.43	0	100	0	4.57	4.41
FAVG	12.19	16.32	25.07	0.6	0.25	-0.02	0	80	0	9.31	5.36

PRAAT	GEE20	GEE10	GEE05	FEE20	FEE10	FEE05	V_UVE	UV_VE	THE	TLE	STD20
v0	0.64	3.18	5.73	-0.47	-0.21	-0.12	0.64	20	0	0.64	2.16
v1	5.88	9.15	15.69	0.45	-0.09	-0.04	0	0	0	0.65	3.45
v2	0	3.7	4.94	0.15	0.59	0.66	0	18.75	0	0	2.91
v3	0	0	6.36	0.57	0.57	-0.01	0	92.86	0	0	2.78
v4	3.27	5.23	7.84	0.26	-0.05	0.06	3.27	42.86	0	3.27	2.6
v5	3.75	10	12.5	0.95	0.15	-0.03	0	0	0	3.13	3.59
v6	9.04	13.25	15.06	0.83	0.22	0.06	0	100	0	4.82	3.1
v7	9.18	9.18	13.27	0.68	0.68	0.41	9.18	0	0	9.18	2.04
v8	3.23	4.84	6.45	0.14	-0.04	-0.08	0	80	0	0	2.11
v9	0	0	1.14	0.21	0.21	0.14	0	25	0	0	1.2
FAVG	3.5	5.85	8.9	0.38	0.2	0.11	1.31	37.95	0	2.17	2.59

Table A.3.14. Performance with background utterance n8 at 0 dB SNR (low resolution signal).

DHO	GEE20	GEE10	GEE05	FEE20	FEE10	FEE05	V_UVE	UV_VE	THE	TLE	STD20
F_v0	13.38	26.11	46.5	-0.94	-0.47	-0.3	12.74	0	0	12.74	7.09
B_v0	24.39	28.39	32.39	-0.22	-0.22	-0.22	23.39	0	0	22.39	0.7
F_v1	6.54	16.34	29.41	-0.72	-0.35	-0.47	6.54	0	0	6.54	6.42
B_v1	23.81	27.81	31.81	0.03	0.03	0.03	22.81	0	0	21.81	0.61
F_v2	4.94	14.81	44.44	-1.02	0.17	-0.68	0	0	0	0	7.87
B_v2	30.77	34.77	38.77	-0.45	-0.45	-0.45	29.77	0	0	28.77	1.04
F_v3	3.64	8.18	29.09	-0.43	0.34	-0.83	0.91	0	0	0.91	6.5
B_v3	11.76	15.76	19.76	0.14	0.14	0.14	10.76	0	0	9.76	0.6
F_v4	8.5	14.38	27.45	-0.22	0.87	0.54	4.58	0	0	4.58	5.79
B_v4	28.89	32.89	36.89	-0.26	-0.26	-0.26	27.89	0	0	26.89	0.86
F_v5	4.38	8.13	16.88	-0.45	-0.92	-0.86	3.13	50	0	3.13	4.94
B_v5	19.05	23.05	27.05	0.11	0.11	0.11	18.05	0	0	17.05	0.7
F_v6	1.2	3.61	23.49	-1.13	-0.86	-0.83	1.2	100	0	1.2	5.28
B_v6	21.43	25.43	29.43	-0.16	-0.16	-0.16	20.43	0	0	19.43	0.65
F_v7	1.02	7.14	23.47	-1.75	-0.97	-0.93	1.02	0	0	1.02	5.29
B_v7	42.31	46.31	50.31	-0.15	-0.15	-0.15	41.31	0	0	40.31	0.65
F_v8	1.61	3.23	25	-1.05	-0.87	-0.83	1.61	20	0	1.61	4.96
B_v8	31.43	35.43	39.43	-0.3	-0.3	-0.3	30.43	0	0	29.43	0.65
F_v9	2.86	3.43	21.71	-0.76	-0.69	-0.66	2.86	0	0	2.86	4.5
B_v9	25	29	33	-0.47	-0.47	-0.47	24	0	0	23	0.73
FAVG	4.81	10.54	28.74	-0.85	-0.37	-0.58	3.46	17	0	3.46	5.86
BAVG	25.88	29.88	33.88	-0.17	-0.17	-0.17	24.88	0	0	23.88	0.72

CORR	GEE20	GEE10	GEE05	FEE20	FEE10	FEE05	V_UVE	UV_VE	THE	TLE	STD20
v0	20.38	25.48	29.94	0.17	-0.13	-0.19	13.38	40	5.1	15.29	4.34
v1	17.65	20.26	26.14	0.29	-0.3	-0.25	11.11	0	0.65	12.42	4.11
v2	16.05	20.99	23.46	-0.37	0.42	0	7.41	18.75	1.23	7.41	4.26
v3	26.36	27.27	29.09	0.28	0.12	-0.13	17.27	92.86	0.91	21.82	2.5
v4	16.99	24.18	26.8	1.03	-0.22	-0.16	15.03	57.14	0	15.03	5.46
v5	17.5	24.38	28.13	1.21	0.14	-0.37	11.25	0	1.88	15	5.1
v6	19.88	24.1	25.3	0.82	0.13	0	7.83	100	0.6	13.25	3.34
v7	21.43	21.43	26.53	0.71	0.71	0.31	13.27	0	0	13.27	2.31
v8	26.61	26.61	26.61	-0.27	-0.27	-0.27	10.48	80	2.42	16.94	1.2
v9	18.86	23.43	24.57	1.16	-0.06	-0.17	8.57	25	4	9.71	5.13
FAVG	20.17	23.81	26.66	0.5	0.06	-0.12	11.56	41.38	1.68	14.01	3.77

Table A.3.15 (continued on the next page).

YIN	GEE20	GEE10	GEE05	FEE20	FEE10	FEE05	V_UVE	UV_VE	THE	TLE	STD20
v0	10.19	19.75	36.31	0.25	-0.56	-0.61	0	100	2.55	2.55	6.36
v1	18.3	20.26	32.03	0.8	0.34	0.14	0	0	1.31	7.84	5.07
v2	20.99	30.86	44.44	0.53	0.92	0.06	0	100	0	14.81	7.48
v3	11.82	23.64	39.09	2.17	0.83	-0.35	0	100	0.91	7.27	7.47
v4	7.84	13.73	24.18	1.17	0.74	0.2	0	100	0	3.92	5.41
v5	10	20.63	31.88	0.7	0.27	-0.48	0	100	0	4.38	7.16
v6	31.93	40.36	50	1.73	0.68	-0.09	0	100	0	23.49	7.45
v7	16.33	26.53	35.71	2.85	0.96	0.93	0	0	2.04	9.18	7.14
v8	22.58	30.65	40.32	1.06	0.4	-0.52	0	100	0.81	14.52	6.47
v9	13.14	17.71	25.71	1.2	0.21	-0.27	0	100	0.57	5.71	5.55
FAVG	16.31	24.41	35.97	1.25	0.48	-0.1	0	80	0.82	9.37	6.55

PRAAT	GEE20	GEE10	GEE05	FEE20	FEE10	FEE05	V_UVE	UV_VE	THE	TLE	STD20
v0	8.92	17.83	25.48	1.08	0.32	-0.06	3.82	20	0	7.64	5.4
v1	10.46	15.03	22.22	0.41	-0.5	-0.16	3.27	0	0	4.58	4.54
v2	9.88	14.81	20.99	-0.29	-0.01	0.53	0	18.75	0	0	4.14
v3	17.27	20.91	22.73	1.09	0.33	0.13	0	85.71	0	6.36	3.99
v4	9.8	22.22	25.49	2.32	0.43	0.13	3.27	50	0	4.58	6.53
v5	8.13	19.38	21.25	1.99	0.12	-0.1	1.25	0	0	6.25	5.55
v6	16.87	21.08	23.49	0.96	0.31	0.1	4.82	100	0	10.84	3.2
v7	19.39	19.39	24.49	0.95	0.95	0.55	9.18	0	0	9.18	2.21
v8	17.74	22.58	23.39	1.05	0.08	-0.03	2.42	80	0	11.29	4.31
v9	9.71	16	19.43	1.88	0.4	0.12	0.57	25	0	4	5.7
FAVG	12.82	18.92	22.9	1.15	0.24	0.12	2.86	37.95	0	6.47	4.56

Table A.3.15. Performance with background utterance n9 at 0 dB SNR (low resolution signal).

DHO	GEE20	GEE10	GEE05	FEE20	FEE10	FEE05	V_UVE	UV_VE	THE	TLE	STD20
F_v0	1.27	12.1	36.31	-0.84	-0.67	-0.58	0.64	0	0	0.64	6.64
B_v0	4.46	8.46	8	18.12	18.12	0	3.46	0	0	2.46	4.04
F_v1	5.88	15.69	30.07	-0.75	-0.62	-0.48	5.88	0	0	5.88	6.4
B_v1	21.3	25.3	29.3	3.79	3.79	3.79	20.3	0	0	19.3	3.72
F_v2	3.7	14.81	48.15	-1.36	0.08	-0.6	0	0	0	0	8.2
B_v2	6.15	10.15	12.62	-7.34	-7.34	-5.59	5.15	0	0	4.15	3.95
F_v3	3.64	8.18	30.91	-0.45	0.32	-0.95	0.91	0	0	0.91	6.55
B_v3	0	4	8	0	0	0	-1	0	0	-2	0
F_v4	6.54	12.42	26.14	-0.11	0.94	0.25	2.61	0	0	2.61	5.72
B_v4	14.78	18.78	18.43	8.68	8.68	4.3	13.78	0	0	12.78	7.78
F_v5	3.13	8.13	15.63	-0.78	-0.96	-0.78	0.63	0	0	0.63	5.55
B_v5	1.8	5.8	8	22.81	22.81	0	0.8	0	0	-0.2	0.48
F_v6	1.2	3.61	24.7	-1.09	-0.83	-0.8	1.2	100	0	1.2	5.22
B_v6	0	4	8	0	0	0	-1	0	0	-2	0
F_v7	1.02	7.14	22.45	-1.83	-1.08	-0.95	1.02	0	0	1.02	5.19
B_v7	0	4	8	0	0	0	-1	0	0	-2	0
F_v8	1.61	3.23	25	-1.03	-0.85	-0.94	1.61	20	0	1.61	4.94
B_v8	3.75	7.75	10.5	-10.16	-10.16	-8.81	2.75	0	0	1.75	2.22
F_v9	2.29	4.57	20.57	-0.71	-0.43	-0.61	2.29	0	0	2.29	4.58
B_v9	1.74	5.74	9.74	-0.05	-0.05	-0.05	0.74	0	0	-0.26	6.13
FAVG	3.03	8.99	27.99	-0.9	-0.41	-0.64	1.68	12	0	1.68	5.9
BAVG	5.4	9.4	12.06	3.59	3.59	-0.64	4.4	0	0	3.4	2.83

CORR	GEE20	GEE10	GEE05	FEE20	FEE10	FEE05	V_UVE	UV_VE	THE	TLE	STD20
v0	22.29	22.29	23.57	-0.3	-0.3	-0.17	7.01	100	15.29	7.01	1.53
v1	8.5	8.5	8.5	-0.15	-0.15	-0.15	3.27	0	5.23	3.27	1.48
v2	9.88	9.88	9.88	-0.37	-0.37	-0.37	3.7	18.75	3.7	3.7	1.82
v3	10.91	10.91	13.64	-0.64	-0.64	-0.41	2.73	100	8.18	2.73	1.66
v4	4.58	4.58	4.58	-0.31	-0.31	-0.31	4.58	100	0	4.58	1.05
v5	5	8.13	13.13	-1.34	-0.93	-0.57	0.63	100	4.38	0.63	2.93
v6	4.22	4.22	5.42	-0.02	-0.02	-0.02	1.81	100	1.2	1.81	1.49
v7	2.04	2.04	3.06	0.07	0.07	0.16	2.04	0	0	2.04	1.74
v8	12.1	12.1	13.71	0.03	0.03	0.01	0	80	12.1	0	1.77
v9	5.71	6.29	8.57	-0.41	-0.34	-0.23	3.43	25	2.29	3.43	2.2
FAVG	8.52	8.89	10.4	-0.34	-0.3	-0.21	2.92	62.38	5.24	2.92	1.77

Table A.3.16 (continued on the next page).

YIN	GEE20	GEE10	GEE05	FEE20	FEE10	FEE05	V_UVE	UV_VE	THE	TLE	STD20
v0	10.83	13.38	29.3	-0.11	-0.22	-0.11	0	100	3.18	1.27	4.49
v1	7.19	12.42	20.92	0.63	-0.41	-0.48	0	0	0	5.88	5.61
v2	20.99	27.16	37.04	-0.07	-0.54	-0.71	0	100	1.23	14.81	6.89
v3	12.73	16.36	29.09	0.08	-0.35	-0.83	0	100	4.55	4.55	5.56
v4	7.84	11.11	16.34	0	0.35	0.24	0	100	0	6.54	4.5
v5	8.75	13.13	20.63	-1.85	-1.24	-0.49	0	100	0.63	5.63	5.18
v6	19.88	22.89	33.73	0.3	0.1	-0.21	0	100	0	14.46	5.18
v7	10.2	10.2	13.27	0.47	0.47	0.38	0	0	0	8.16	3.3
v8	18.55	23.39	35.48	1.35	1	-0.17	0	100	2.42	12.1	5.52
v9	12	15.43	24	0.58	0.04	-0.15	0	100	0.57	8	5.05
FAVG	12.9	16.55	25.98	0.14	-0.08	-0.25	0	80	1.26	8.14	5.13

PRAAT	GEE20	GEE10	GEE05	FEE20	FEE10	FEE05	V_UVE	UV_VE	THE	TLE	STD20
v0	6.37	7.64	9.55	-0.29	-0.15	0	1.27	100	5.1	1.27	1.9
v1	0	0	0	0.15	0.15	0.15	0	0	0	0	1.32
v2	3.7	4.94	4.94	-0.15	-0.35	-0.35	0	12.5	1.23	0	2.46
v3	5.45	5.45	7.27	-0.42	-0.42	-0.26	0.91	100	4.55	0.91	1.5
v4	0	0	0	-0.06	-0.06	-0.06	0	100	0	0	0.99
v5	2.5	6.25	12.5	-1.35	-0.71	-0.23	0	100	0	0	3.86
v6	3.01	3.01	3.61	0.2	0.2	0.16	0	100	0	0	1.03
v7	0	0	0	0.51	0.51	0.51	0	0	0	0	1.35
v8	4.03	8.06	13.71	-0.28	0.44	0.16	0	80	0	0	4.01
v9	2.29	3.43	5.71	-0.2	-0.06	0.05	2.29	25	0	2.29	2.32
FAVG	2.74	3.88	5.73	-0.19	-0.04	0.01	0.45	61.75	1.09	0.45	2.08

Table A.3.16. Performance with background utterance n7 at -5 dB SNR (low resolution signal).

DHO	GEE20	GEE10	GEE05	FEE20	FEE10	FEE05	V_UVE	UV_VE	THE	TLE	STD20
F_v0	14.01	25.48	47.13	-0.85	-0.54	-0.46	12.74	0	0	12.74	7.05
B_v0	26.79	30.79	30.32	5.31	5.31	2.75	25.79	0	0	24.79	6.67
F_v1	5.23	15.69	29.41	-0.7	-0.47	-0.67	5.23	0	0	5.23	6.43
B_v1	14.81	18.81	22.81	4.76	4.76	4.76	13.81	0	0	12.81	3.16
F_v2	4.94	16.05	48.15	-1.32	0.02	-0.79	0	0	0	0	8.06
B_v2	6.15	10.15	12.62	-6.94	-6.94	-5.43	5.15	0	0	4.15	4.07
F_v3	3.64	10	28.18	-0.36	0.11	-0.76	0.91	0	0	0.91	6.65
B_v3	2.67	6.67	10.67	-1.2	-1.2	-1.2	1.67	0	0	0.67	0.52
F_v4	8.5	15.03	27.45	-0.25	0.91	0.44	4.58	0	0	4.58	5.79
B_v4	22.61	26.61	26.26	7.08	7.08	4.22	21.61	0	0	20.61	6.68
F_v5	2.5	8.13	16.25	-0.93	-0.96	-0.85	0.63	50	0	0.63	5.78
B_v5	0.9	4	8	24.66	0	0	-1	0	0	-2	0
F_v6	0.6	3.61	23.49	-1.21	-0.83	-0.68	0.6	0	0	0.6	5.39
B_v6	0.87	4.87	8.87	5.21	5.21	5.21	-0.13	0	0	-1.13	0
F_v7	1.02	7.14	23.47	-1.83	-1.05	-0.86	1.02	0	0	1.02	5.24
B_v7	0	4	8	0	0	0	-1	0	0	-2	0
F_v8	1.61	2.42	24.19	-1.02	-0.93	-0.84	1.61	40	0	1.61	4.95
B_v8	2.5	6.5	10.5	-8.81	-8.81	-8.81	1.5	0	0	0.5	1.39
F_v9	1.71	2.29	21.71	-0.73	-0.67	-0.57	1.14	0	0	1.14	4.46
B_v9	0	4	8	0	0	0	-1	0	0	-2	0
FAVG	4.38	10.58	28.94	-0.92	-0.44	-0.6	2.85	9	0	2.85	5.98
BAVG	7.73	11.64	14.6	3.01	0.54	0.15	6.64	0	0	5.64	2.25

CORR	GEE20	GEE10	GEE05	FEE20	FEE10	FEE05	V_UVE	UV_VE	THE	TLE	STD20
v0	40.13	40.13	43.31	0.04	0.04	-0.19	3.82	100	35.03	3.82	1.78
v1	20.92	25.49	30.07	1.44	-0.06	0.27	5.23	0	14.38	5.23	6.63
v2	18.52	28.4	38.27	2.59	-0.03	-0.4	2.47	18.75	12.35	2.47	8.09
v3	17.27	23.64	30	-0.02	-0.51	-0.54	4.55	100	11.82	4.55	6.08
v4	27.45	27.45	28.76	-0.38	-0.38	-0.39	12.42	100	15.03	12.42	1.81
v5	20	23.13	28.13	-1.21	-0.73	-0.62	3.13	100	16.88	3.13	3.39
v6	22.29	22.29	25.3	0.32	0.32	0.26	7.83	100	12.05	7.83	2.19
v7	14.29	15.31	23.47	0.75	0.56	0.34	4.08	0	10.2	4.08	4.08
v8	21.77	21.77	25	-0.04	-0.04	0.05	1.61	80	20.16	1.61	2.72
v9	22.29	22.29	28.57	-0.59	-0.59	-0.15	3.43	25	18.86	3.43	2.92
FAVG	22.49	24.99	30.09	0.29	-0.14	-0.14	4.86	62.38	16.68	4.86	3.97

Table A.3.17 (continued on the next page).

YIN	GEE20	GEE10	GEE05	FEE20	FEE10	FEE05	V_UVE	UV_VE	THE	TLE	STD20
v0	28.66	35.67	50.96	0.36	-0.11	-0.28	0	100	19.75	1.91	6.83
v1	15.03	24.84	32.68	1.14	-0.2	-0.5	0	0	2.61	5.23	7.67
v2	19.75	33.33	51.85	3.94	0.42	0.12	0	100	1.23	12.35	9.67
v3	18.18	29.09	44.55	0.54	-1.31	-1.34	0	100	5.45	4.55	7.64
v4	25.49	31.37	39.22	0.11	0.72	0.5	0	100	9.15	7.84	5.26
v5	16.25	22.5	36.88	-1.55	-1.4	-0.56	0	100	6.88	5.63	6.27
v6	25.3	31.33	42.77	1.15	0.65	0.19	0	100	3.01	12.65	7.26
v7	12.24	22.45	25.51	1.47	0.74	0.65	0	0	2.04	6.12	7.05
v8	25.81	29.84	42.74	0.37	-0.2	-0.47	0	100	0.81	12.9	6.24
v9	21.14	26.29	34.86	-0.17	-0.49	0.09	0	100	4.57	8	6.58
FAVG	20.79	28.67	40.2	0.74	-0.12	-0.16	0	80	5.55	7.72	7.05

PRAAT	GEE20	GEE10	GEE05	FEE20	FEE10	FEE05	V_UVE	UV_VE	THE	TLE	STD20
v0	28.03	29.3	37.58	-0.03	0.17	-0.08	4.46	100	23.57	4.46	2.96
v1	1.31	6.54	15.69	1.56	0.31	0.28	1.31	0	0	1.31	6.03
v2	4.94	19.75	32.1	2.53	-0.13	-0.45	0	12.5	1.23	0	8.34
v3	11.82	20	27.27	-0.24	-0.7	-0.36	3.64	100	5.45	3.64	6.25
v4	19.61	19.61	22.22	0.12	0.12	-0.03	9.8	100	9.8	9.8	1.94
v5	9.38	15	23.75	-1.65	-0.69	-0.36	3.75	100	3.13	3.75	4.75
v6	9.64	12.65	18.07	0.85	0.8	0.29	3.61	100	0	3.61	4.84
v7	4.08	9.18	18.37	1.92	1.15	0.59	4.08	0	0	4.08	4.97
v8	10.48	14.52	22.58	-0.3	0.49	0.3	4.03	80	0	4.03	4.88
v9	6.29	12	21.71	-1.34	-0.42	-0.03	4.57	25	0	4.57	4.71
FAVG	10.56	15.85	23.93	0.34	0.11	0.02	3.93	61.75	4.32	3.93	4.97

Table A.3.17. Performance with background utterance n8 at -5 dB SNR (low resolution signal).

DHO	GEE20	GEE10	GEE05	FEE20	FEE10	FEE05	V_UVE	UV_VE	THE	TLE	STD20
F_v0	0.64	12.1	36.94	-0.68	-0.5	-0.55	0.64	0	0	0.64	6.62
B_v0	6.25	10.25	9.79	15.21	15.21	7.63	5.25	0	0	4.25	5.85
F_v1	5.23	15.03	29.41	-0.72	-0.56	-0.54	5.23	0	0	5.23	6.37
B_v1	15.74	19.74	23.74	3.99	3.99	3.99	14.74	0	0	13.74	4.21
F_v2	4.94	14.81	44.44	-1.04	0.18	-0.35	0	6.25	0	0	7.87
B_v2	6.15	10.15	14.15	-5.71	-5.71	-5.71	5.15	0	0	4.15	3.16
F_v3	3.64	8.18	30.91	-0.48	0.29	-0.97	0.91	0	0	0.91	6.54
B_v3	0	4	8	0	0	0	-1	0	0	-2	0
F_v4	8.5	15.03	27.45	-0.26	0.88	0.36	4.58	0	0	4.58	5.75
B_v4	22.61	26.61	26.26	7.1	7.1	4.16	21.61	0	0	20.61	6.78
F_v5	3.75	8.13	16.25	-0.62	-0.95	-0.74	2.5	0	0	2.5	5.2
B_v5	1.8	5.8	8	22.97	22.97	0	0.8	0	0	-0.2	0.44
F_v6	1.2	3.61	24.7	-1.13	-0.85	-0.74	1.2	0	0	1.2	5.33
B_v6	1.74	5.74	9.74	7.33	7.33	7.33	0.74	0	0	-0.26	2.12
F_v7	1.02	7.14	23.47	-1.8	-1.03	-1.02	1.02	0	0	1.02	5.2
B_v7	0	4	8	0	0	0	-1	0	0	-2	0
F_v8	1.61	3.23	25	-1.04	-0.86	-0.81	1.61	20	0	1.61	4.98
B_v8	3.75	7.75	9.25	-10.82	-10.82	-7.68	2.75	0	0	1.75	2.59
F_v9	1.71	3.43	21.14	-0.72	-0.53	-0.61	1.71	0	0	1.71	4.45
B_v9	0	4	8	0	0	0	-1	0	0	-2	0
FAVG	3.22	9.07	27.97	-0.85	-0.39	-0.6	1.94	2.63	0	1.94	5.83
BAVG	5.8	9.8	12.49	4.01	4.01	0.97	4.8	0	0	3.8	2.52

CORR	GEE20	GEE10	GEE05	FEE20	FEE10	FEE05	V_UVE	UV_VE	THE	TLE	STD20
v0	57.32	57.32	59.24	-0.18	-0.18	-0.47	4.46	100	52.87	4.46	1.98
v1	39.87	41.18	52.29	-0.12	-0.62	0.15	1.31	0	31.37	1.31	5.32
v2	35.8	48.15	54.32	4.11	0	-0.26	1.23	25	32.1	1.23	9.28
v3	30.91	33.64	46.36	-1.63	-1.85	-0.79	0.91	100	30	0.91	5.48
v4	47.71	48.37	52.29	-0.88	-0.73	-0.54	8.5	100	39.22	8.5	2.98
v5	43.13	45.63	48.75	-1.92	-1.38	-0.92	2.5	100	40.63	2.5	3.43
v6	48.8	50.6	51.2	-0.65	-0.12	-0.03	3.61	100	39.16	3.61	3.31
v7	35.71	37.76	48.98	-0.94	-0.53	-0.29	5.1	0	29.59	5.1	4.85
v8	36.29	36.29	37.1	-0.23	-0.23	-0.14	0.81	80	35.48	0.81	2.65
v9	45.71	45.71	49.71	-0.78	-0.78	-0.28	5.14	25	40.57	5.14	3.05
FAVG	42.13	44.46	50.02	-0.32	-0.64	-0.36	3.36	63	37.1	3.36	4.23

Table A.3.18 (continued on the next page).

YIN	GEE20	GEE10	GEE05	FEE20	FEE10	FEE05	V_UVE	UV_VE	THE	TLE	STD20
v0	38.85	49.68	68.79	-1.17	-1.24	-0.63	0	100	19.75	5.1	8.64
v1	29.41	39.22	53.59	0.77	0.39	0.46	0	0	13.07	3.92	8.09
v2	33.33	51.85	67.9	3.78	1.62	1.12	0	100	8.64	11.11	10.47
v3	23.64	37.27	61.82	0.16	-1.08	-0.87	0	100	5.45	9.09	9.66
v4	39.22	48.37	60.13	1.08	0.66	0.04	0	100	18.95	11.11	7.19
v5	23.13	39.38	48.13	-2.49	-1.55	-0.83	0	100	10	7.5	8.02
v6	39.76	54.22	68.07	1.57	1.17	-0.33	0	100	7.23	13.86	10.3
v7	40.82	51.02	65.31	2.93	1.05	0.76	0	0	8.16	12.24	9.61
v8	35.48	48.39	58.87	0.29	0.18	0.05	0	100	0.81	16.13	9.33
v9	35.43	43.43	59.43	-0.54	-1.05	0	0	100	9.71	12	7.92
FAVG	33.91	46.28	61.2	0.64	0.01	-0.02	0	80	10.18	10.21	8.92

PRAAT	GEE20	GEE10	GEE05	FEE20	FEE10	FEE05	V_UVE	UV_VE	THE	TLE	STD20
v0	40.76	42.68	51.59	-0.22	0.13	-0.46	5.1	100	32.48	8.28	3.72
v1	19.61	29.41	41.18	2.03	-0.3	0.52	1.31	0	17.65	1.31	8.32
v2	7.41	37.04	46.91	2.06	-0.29	-0.31	0	25	1.23	0	11.43
v3	11.82	22.73	41.82	-1.6	-1.64	-0.53	0.91	100	5.45	0.91	7.7
v4	35.29	35.95	45.1	0.81	0.7	-0.07	13.07	100	22.22	13.07	3.72
v5	22.5	33.13	45	-1.93	-1.03	-0.54	3.75	100	16.25	3.75	6.28
v6	21.08	29.52	39.16	1.61	1.34	0.52	3.61	100	10.84	3.61	7.66
v7	4.08	16.33	33.67	1.24	0.84	0.73	4.08	0	0	4.08	7.45
v8	23.39	23.39	33.87	0.49	0.49	0.48	1.61	80	21.77	1.61	4.31
v9	31.43	33.14	44.57	-1.25	-0.95	-0.14	9.71	25	21.71	9.71	4.38
FAVG	21.74	30.33	42.29	0.33	-0.07	0.02	4.32	63	14.96	4.63	6.5

Table A.3.18. Performance with background utterance n9 at -5 dB SNR (low resolution signal).

REFERENCES

- [Allen, 1996] Allen, J. B., "Harvey Fletcher's role in the creation of communication acoustics," *Journal of the Acoustical Society of America*, 99(4), 1825-1839, 1996.
- [Assmann, Summerfield, 1990] Assmann, P. F., and Summerfield, Q., "Modeling the perception of concurrent vowels: Vowels with different fundamental frequencies," *Journal of the Acoustical Society of America*, 88, 680-697, 1990.
- [Atal et al, 1971] Atal, B. S., Hanauer, S. L., "Speech analysis and synthesis by linear prediction of speech wave," *Journal of the Acoustical Society of America*, 50(2), 637-655, 1971.
- [Barlow, 1959] Barlow, H.B., "Sensory mechanisms, the reduction of redundancy, and intelligence," in *The Mechanisation of Thought Processes* London: Her Majesty's Stationery Office, 535-539, 1959.
- [Barlow, 1961] Barlow, H.B., "The coding of sensory messages," chapter XIII, in *Current Problems in Animal Behaviour*, Thorpe and Zangwill, editors, Cambridge University Press, 330-360, 1961.
- [von Békésy, 1960] Békésy, G. von, *Experiments in Hearing*, McGraw-Hill, New York, 1960.
- [von Békésy, 1963] Békésy, G. von, "Three experiments concerned with pitch perception," *Journal of the Acoustical Society of America*, 35, 602-606, 1963.
- [Bergem, 1993] Bergem, D., "Acoustic vowel reduction as a function of sentence accent, word stress and word class on the quality of vowels," *Speech Communications*, 12, 1-23, 1993.
- [Bilsen, 1966] Bilsen, F. A., "Repetition pitch: monaural interaction of a sound with the repetition of the same, but phase shifted sound," *Acoustica*, 17, 265-300, 1966.
- [Boer, 1956] Boer, E. de., "On the residue in hearing," Unpublished doctoral dissertation, University of Amsterdam, 1956.

- [Boer, 1977] Boer, E. de., "Pitch theories unified," in *Psychophysics and Physiology of Hearing*. Editors: Evans, E., F., Wilson, J. P., Academic Press, London, 1977.
- [Boersma, 1993] Boersma, P., "Accurate short-term analysis of the fundamental frequency and the harmonics to noise ratio of a sampled sound," *Proceedings of the Institute of Phonetic Sciences*, 17, 97-110, 1993.
- [Boersma, 2002] Boersma, P., "Doing phonetics with computer," Web page: www.praat.org. The Official PRAAT software home-page on the internet.
- [Bregman, 1990] Bregman, A. S., *Auditory Scene Analysis*, Cambridge, MA: MIT Press, 1990.
- [Brown, Cooke, 1994] Brown, G. J., Cooke, M., "Computational auditory scene analysis," *Computer Speech and Language*, 8, 297-336, 1994.
- [Brown, Cooke, 1995] Brown, G. J., Cooke, M., "A neural oscillator model of primitive auditory grouping," *Proceedings of IEEE Signal Processing Workshop on Applications of Signal Processing to Audio and Acoustics*, New York, 53-56, 1995.
- [Brown, Wang, 1997] Brown, G. J., Wang, D. L., "Modelling the perceptual segregation of double vowels with a network of neural oscillators," *Neural Networks*, 10 (9), 1547-1558, 1997.
- [Cariani, 1997] Cariani P., "Temporal coding of sensory information," in: Bower, JM, editors, *Computational Neuroscience: Trends in Research*, 591-598, New York: Plenum 1997.
- [Cariani, 1999] Cariani P., "Temporal coding of periodicity pitch in the auditory system: An overview," *Neural Plasticity* 6(4), 147-172, 1999.
- [Cariani, Delgutte, 1996] Cariani, P. A., Delgutte, B., "Neural Correlates of the pitch of complex tones. I. Pitch and pitch salience," *Journal of Neurophysiology*, 76, 1698-1716, 1996.
- [Carylon, 1996] Carlyon, R. P. "Encoding the fundamental frequency of a complex tone in the presence of a spectrally overlapping masker," *Journal of the Acoustical Society of America*, 99, 517-524, 1996.

- [Carylon, Shackleton, 1994] Carlyon, R. P., Shackleton, T. M., "Comparing the fundamental frequencies of resolved and unresolved harmonics: Evidence for two pitch mechanisms" *Journal of the Acoustical Society of America*, 95, 3541–3554, 1994.
- [Cooke, 1993] Cooke, M. P., *Modeling auditory processing and organization*, Cambridge University Press, UK, 1993.
- [Cooke, 2002] Cooke, M. P., <http://www.dcs.shef.ac.uk/~martin/>, Sheffield, UK, 2002.
- [Culling et al, 1994] Culling, J. F. and Darwin, C. J. "Perceptual and computational separation of simultaneous vowels: cues arising from low frequency beating," *Journal of the Acoustical Society of America*, 95, 1559-1569, 1994.
- [Darwin, 1984] Darwin, C. J., "Perceiving Vowels in the presence of another sounds: Constraints on formant perception," *Journal of the Acoustical Society of America*, 76 (6), 1984.
- [de Cheveigné, 1993] de Cheveigné, A., "Separation of concurrent harmonic sounds: Fundamental frequency estimation and time domain cancellation model of auditory processing," *Journal of the Acoustical Society of America*, 93, 3271-3290, 1993.
- [de Cheveigné, 1998] de Cheveigné, A., "Cancellation Model of Pitch Perception," *Journal of the Acoustical Society of America*, 103, 1261-1271, 1998.
- [de Cheveigné, 2002], de Cheveigné, A., *Home page*. On the world wide web, <http://www.ircam.fr/equipes/pcm/cheveign/>
- [de Cheveigné, Kawahara, 2002] de Cheveigné, A., Kawahara, H., "YIN, a fundamental frequency estimator for speech and music," *Journal of the Acoustical Society of America*, 111(4), April, 2002.
- [Delgutte, 1995] Delgutte B., "Physiological models for basic auditory percepts," in Hawkins H, McMullin T, Popper AN, Fay R. R., editors, *Auditory Computation*. New York: Springer Verlag, 157-220, 1995.

- [Denham, 2001] Denham, S. L., "Cortical synaptic depression and auditory perception," in *Computational Models of Auditory Function*, Editors Greenberg, S., Slaney, M. IOS Press, 2001.
- [Dau et al, 1997] Dau, T., Kollmeier, B., Kohlrausch, A., "Modeling auditory processing of amplitude modulation. I-II," *Journal of the Acoustical Society of America*, 102, 2892-2919, 1997.
- [Dubnowski et al, 1976] Dubnowski, J., J., Schafer, R., W., Rabiner, L., R., "Real time digital hardware pitch detector," *IEEE Transactions Acoustics, Speech and Signal Processing*, vol. ASSP-24, 2-8, 1976.
- [Dusterhoff, Black, 1997] Dusterhoff, K. and Black, A., "Generating F0 contours for speech synthesis using the Tilt intonation theory," proceedings, ESCA Workshop of Intonation, Athens, Greece, September, 1997.
- [Duifhuis et al, 1982] Duifhuis, H., Willems, L., F., Sluyter, R., J., "Measurement of pitch in speech: an implementation of Goldstein's theory of pitch perception," *Journal of the Acoustical Society of America*, 71, 1568-1580, 1982.
- [Fletcher, 1934] Fletcher, H., "Loudness, pitch and timbre of musical tones and their relation to the intensity, the frequency and the overtone structure," *Journal of the Acoustical Society of America*, 6, 59-69, 1934.
- [Fujimura, 1979] Fujimura, O., "An analysis of english syllables as cores and affixes", *Zeitschrift fur Phonetik*, 4/5, 471-476, 1979.
- [Furui, 1996] Furui, S., "An overview of speaker recognition technology", in *Automatic Speech and Speaker Recognition*, Editors Lee, C., H., Soong, F., K., Paliwal, K., K. Kluwer, Boston, 1996.
- [Glasberg, Moore, 1990] Glasberg, B. R., Moore, B. C. J., "Derivation of auditory filter shapes from notched-noise data," *Hearing Res.*, 47, 103-138, 1990.
- [Goldstein, 1973] Goldstein, J.L., "An optimum processor theory for the central formation of pitch of complex tones," *The Journal of the Acoustical Society of America*, 54, 1496-1516, 1973.

- [Goodwin, 1992] Goodwin, M., M., "Adaptive signal models: theory, algorithms and audio applications," PhD. Thesis, MIT, 1992.
- [Gray, 1990] Gray, A. A., "On a modification of the Helmholtz theory of hearing." *Journal of Anat. Physiology*, London, 34, 324-350, 1990.
- [Greenberg et al., 1996] Greenberg, S., Hollenback, J., Ellis, D., "Insights into Spoken Language Gleaned from Phonetic Transcription of the Switchboard Corpus," *Proc. ICSLP*, 32-35, 1996.
- [Guenther, 1995] Guenther, F.H., "Speech sound acquisition, coarticulation, and rate effects in a neural network model of speech production". *Psychological Review*, 102, , pp. 594-621, 1995.
- [Helmholtz, 1870] Helmholtz, H. v., "Die lehre von den tonempfindungen als physiologische grundlage für theorie der musik," translated into English by Ellis, A. J., *On the Sensations of Tone*, Dover, New York, 1954.
- [Hermes, 1988] Hermes, D. J., "Measurement of Pitch by subharmonic summation," *Journal of the Acoustical Society of America*, 83, 257-264, 1988.
- [Hermes, 1993] Hermes, D. J., "Pitch Analysis" in *Visual Representation of Speech Signals*, Edited by Cooke, M., Beet, S., and Crawford, M., John Wiley and Sons Ltd., 1993.
- [Hess, 1983] Hess, W., *Pitch Determination of Speech Signals*, Springer-Verlag, 1983.
- [Hirsch, Ehrlicher, 1995] Hirsch, H. G., Ehrlicher, C., "Noise estimation techniques for robust speech recognition," in *Proceedings of ICASSP*, 1, 143-156, 1995.
- [Irino, Unoki, 2001] Irino, T., Unoki, M., "An analysis synthesis auditory filterbank based on an IIR gammachirp filter," in *Computational Models of Auditory Function*, Editors: Greenberg, S., Slayney, M., IOS Press, 2001.
- [Itakura, 1975] Itakura, F., "Line spectral representation of linear predictor coefficients of speech signals," *Journal of the Acoustical Society of America*, 57, Supplement 1, S35, 1975.

- [Kammler, 2000] Kammler, D. W., *A First Course in Fourier Analysis*, Prentice Hall, New Jersey, 2000.
- [Kawahara, 1996] Kawahara, H., "Auditory analysis and speech communication," available on internet at: "<http://www.hip.atr.co.jp/departments/Dept1/progress95/progress95.html>", May, 1996.
- [Kiang et al, 1965] Kiang, N. Y., Watanabe, T., Thomas, E. C., Clark, L. F., "Discharge of auditory fibers in cats auditory nerve," *Res. Mon.*, 35, MIT, Cambridge, MA.
- [Klatt, 1980] Klatt, D. H., "Software for cascade/parallel formant synthesizer," *Journal of the Acoustical Society of America*, 67, 838-844, 1980.
- [Hermansky, 1990] Hermansky, H., "Perceptual linear predictive (plp) analysis for speech," *Journal of The Acoustical Society of America*, 87, 1738-1752, 1990.
- [Kaernbach, Demany, 1998] Kaernbach, C., Demany, L., "Psychophysical evidence against the autocorrelation theory of auditory temporal processing," *Journal of the Acoustical Society of America*, 104 (4), 2298-2306, 1998.
- [Konishi, 1993] Konishi, M., "Listening with two ears," *Scientific American*, 34-41, April, 1993
- [Licklider, 1951] Licklider, J. C. R., "A duplex theory of pitch perception," *Experientia*, 7, 128-134, 1951.
- [Licklider, 1956] Licklider, J. C. R., "Auditory Frequency Analysis," in *Information Theory*, editor, Cherry, C., Butterworths and Academic Press, London, New York, 1956.
- [Lyon, 1983] Lyon, R. F., "A computational model of binaural localization and separation," *Proceedings of IEEE ICASSP*, Boston, MA, 1983.
- [McCabe, Denham, 1997] McCabe, S. L., Denham, M. J., "A model of auditory streaming," *Journal of the Acoustical Society of America*, 101, 1611-1621, 1997.

- [Meddis, Hewitt, 1991] Meddis, R., Hewitt, M. J., "Virtual pitch and phase sensitivity of a computer model of the auditory periphery. I. Pitch Identification," *Journal of the Acoustical Society of America*, 89, 2866-2882, 1991.
- [Meddis, Hewitt, 1992] Meddis, R., Hewitt, M. J., "Modelling and identification of concurrent vowels with different fundamental frequencies," *Journal of the Acoustical Society of America*, 91, 233-24, 1992.
- [Meddis, O'Mard, 1997] Meddis, R., O'Mard, L., "A unitary model of pitch perception," *Journal of the Acoustical Society of America*, 102(3), 1811-1819, 1997.
- [Moore, 1997] Moore, B. C. J., *An Introduction to the Physiology of Hearing*, Academic Press, New York, 1997.
- [Moore et al, 1985] Moore, B. C. J., Glasberg, B. R., Peters, R. W., "Relative dominance of individual partials in determining the pitch of complex tones," *Journal of the Acoustical Society of America*, 77, 1853-1860, 1985.
- [Nishihara, Crossley, 1988] Nishihara, H. K., Crossley, P. A., "Measuring photolithographic overlay accuracy and critical dimensions by correlating binarized laplacian of gaussian convolutions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 10 (3), 17-30, January 1988.
- [Noll, 1967] Noll, A. M., "Cepstrum pitch determination," *Journal of the Acoustical Society of America*, 41, 293-309, 1967.
- [Ohgushi, 1978] Ohgushi, K., "On the role of spatial and temporal cues in the perception of the pitch of complex tones," *Journal of the Acoustical Society of America*, 64, 764-771, 1978.
- [Oppenheim, Schafer, 1975] Oppenheim, A.V., Schafer, A., W., *Digital Signal Processing*, Prentice-Hall, Englewood Cliffs, New Jersey, Chapter 10, pp. 490 – 500, 1975,
- [Pain, 1976] Pain, H. J., *The Physics of Vibrations and Waves*, 2nd edition, Wiley, London, 1976.
- [Park, 1964] Park, D. D. A., *Introduction to the quantum theory*, McGraw Hill, New York, 1964.

- [Patterson et al, 1988] Patterson, R. D., Holdsworth, J., Nimmo-Smith, I., Rice, P., "Implementing a gammatone Filterbank". APU Report 2341, Cambridge, Applied Psychology Unit.
- [Patterson, Holdsworth, 1991] Patterson, R. D., Holdsworth, J., "A functional model of neural activity patterns and auditory images," in *Advances in Speech Hearing and Language Processing*, vol. 3, JAI Press, London, 1991.
- [[Patterson et al., 1996] Patterson, R. P., Handel, S., Yost, W. A., and Datta, A. J., "The relative strength of tone and noise components of iterated rippled noise," *Journal of the Acoustical Society of America*, 100, 3286-3294, 1996.
- [Pick et al, 1989] Pick, H.L., Siegel, G.M., Fox, P.W., Garber, S.R., Kearney, J.K., "Inhibiting the Lombard effect," *Journal of the Acoustical Society of America*, 85 (2), 894-900, 1989.
- [Pickles, 1988] Pickles, J. O., *An Introduction to the Physiology of Hearing*. London, Academic Press, 1988.
- [Plante et al, 1995] Plante, F., Meyer, G. F., Ainsworth, W. A., "A Pitch Extraction Reference Database," in ESCA Eurospeech' 95; 4th European Conference on Speech Communication and Technology, Madrid, September, 1995.
- [Plomp, 1965] Plomp, R., "Detectability threshold for combination tones," *Journal of the Acoustical Society of America*, 37, 1110-1123, 1965.
- [Plomp, 1976] Plomp, R., *Aspects of Tone Sensations*, Academic Press, London, 1976.
- [Rabiner et al, 1976] Rabiner, L. R., Cheng, M., J., Rosenberg, A. E., McGoneal, C. A., "A comparative performance study of several pitch detection algorithms," *IEEE Transactions on Acoustics, Speech and Signal Processing*, Vol. ASSP-24, No. 5, October, 1976.
- [Rabiner, Juang, 1993] Rabiner, L. R., Juang, B. H., *Fundamentals of Speech Recognition*, Prentice Hall, Englewood Cliffs, New Jersey, 1993.

- [Richmond, Gawne, 1998] Richmond, B. J., Gawne, T. J., "The relationship between neuronal codes and cortical organization," in: *Neuronal Ensembles: Strategies for Recording and Decoding*, editors: Eichenbaum, H. B, Davis, J. L., New York: Wiley, 1998.
- [Rorabaugh, 1997] Rorabaugh, C. B., *Digital Filter Designer's Handbook*. McGraw-Hill, 1997.
- [Rose et al, 1967] Rose, J. E, Brugge J. R, Anderson, D. J., Hind, J. E., "Phase-locked response to low-frequency tones in single auditory nerve fibers of the squirrel monkey," *Journal of Neurophysiology*, 30, 769-793, 1967.
- [Russel, 1987] Russel, I. J., "The physiology of the organ of corti," *British Medical Bulletin*, 43, 4, 802-820, 1987.
- [Ryugo, 1992] Ryugo, D.K., "The auditory nerve: peripheral innervation, cell body morphology, and central projections," in *The Mammalian Auditory Pathway: Neuroanatomy*, editors: Webster, D. B., Popper, A. N., Fay, R. R, New York, Springer-Verlag, 1992.
- [Scheirer, 2000] Scheirer, E. D., *Music-Listening Systems*, PhD thesis, MIT, 2000.
- [Scheffers, 1983] Scheffers, M. T., "Sifting vowels: Auditory pitch analysis and sound segregation," Ph.D. Thesis, Groningen University, The Netherlands, 1983.
- [Sejnowski, Rosenberg, 1986] Sejnowski, T., and Rosenberg, C., "NetTalk: A Parallel Network That Learns to Read Aloud," *NeuroComputing*, 1986.
- [Shamma, Klein, 2000] Shamma, S., Klein, D., "The case of the missing pitch templates: How harmonic templates emerge in the early auditory system," *Journal of the Acoustical Society of America*, 107 (5), 2631-2645, 2000.
- [Shouten, 1940] Schouten, J. F., "The residue and the mechanism of hearing". *Proceedings Koninkl. Ned. Akad. Wetenschap.* 43, 991-999, 1940.
- [Shouten et al, 1962] Shouten, J. F., Ritsma, R. J., Cardozo, B. J., "Pitch of the residue," *Journal of the Acoustical Society of America*, 34, 1418-1424, 1962.

- [Shroeder, 1999] Shroeder, M. R., *Computer Speech*, Springer Verlag, 1999.
- [Singh, 1997] Singh, S., *Fermats Last Theorem*, Fourth Estate, 1997.
- [Silipo, Greenberg, 1999] Silipo, R. and Greenberg, S., "Automatic transcription of prosodic stress for spontaneous english discourse. The phonetics of spontaneous speech," ICPhS-99, San Francisco, CA, August 1999.
- [Slaney, 1988] Slayney, M., "Lyon's Cochlear Model," Apple Computer Technical Report Number 13, 1988. Currently available at :
<http://web.interval.com/~malcolm/pubs.html#LyonCochlear>
- [Slaney, 1998] Slaney, M., *Auditory Toolbox Version 2*, Technical Report Number 1998-010, Interval Research Corp., 1998.
- [Slaney, Lyon, 1990] Slaney, M., Lyon, R. F., "A perceptual pitch detector," In Proceeding of International Conference on Acoustics Speech and Signal Processing (p. 357-360 vol.1), 1990.
- [Slaney, Lyon, 1993] Slaney, M., Lyon, R. F., "On the importance of time—A temporal representation of sound," in *Visual Representations of Speech Signals*, Cooke, M., Beete, S., and Crawford, M., editors, John Wiley and Sons, Sussex, England, 1993.
- [Small, 1970] Small, A., W., "Periodicity Pitch," in *Foundations of Modern Auditory Theory*, editors: Tobias, J. V., Academic Press, New York, 1970.
- [Sondhi, 1968] Sondhi, M., M., "New methods of pitch extraction," IEEE Transactions on Audio Electroacoustics, (*Special issue in speech communication and processing – part II*) vol. AU-16, 262-266, 1968.
- [Strope at al, 2001] Strope, B. P., Alwan, A.. A, "Modeling the perception of pitch-rate amplitude modulation in noise", in *Computational Models of Auditory Function*, Editors Greenberg, S., Slaney, M. IOS Press, 2001.

- [Terhardt, 1974] Terhardt, E., "Pitch consonance and harmony," *Journal of the Acoustical Society of America*, 55, 1061-1069, 1974.
- [Terhardt, 1980] Terhardt, E., "Towards understanding pitch perception: problems, concepts and solutions," in *Psychophysical, Physiological and Behavioral Studies of Hearing*, editors: Brink, G. van der, Bilsen, F. A., Delft University, 1980.
- [Terhardt et al, 1982] Terhardt, E., Stoll, G. & Seewann, M., "Algorithm for extraction of pitch and pitch salience from complex tonal signals". *Journal of the Acoustical Society of America*, 71, 679-688, 1982.
- [Terhardt, 1991] Terhardt, E., "Music perception and sensory information acquisition: Relationships and low-level analogies," *Music Perception*, 8, 217-240, 1991.
- [Terhardt, 2002] Terhardt, E., *Pitch perception*, on the internet, available at <http://www.mmk.ei.tum.de/persons/ter/top/pitch.html>, world-wide-web, 2002.
- [Tolonen, et al, 2000] Tolonen, T., Karjalainen, M., "A computationally efficient multipitch analysis model," *IEEE Transactions on speech and audio processing*, vol. 8, no. 6, 708-716, 2000.
- [van den Brink, 1974] van den Brink, G., "Monotic and dichotic pitch matchings with complex sounds," in: *Facts and Models in Hearing*, Zwicker, E., Terhardt, E., editors, Springer, Berlin/Heidelberg, 178-188, 1974.
- [Wang, 1993] Wang, D., "Modelling global synchrony in the visual cortex by locally coupled neural oscillators", *Proceedings of the 15th Annual Conference of the Cognitive Science Society*, Boulder, CO, pp 1058-1063, 1993.
- [Weintraub, 1986] Weintraub, M., "A computational model for separating two simultaneous talkers," in *Proceedings of IEEE ICASSP*, 81-84, 1986.
- [Whitefield, 1970] Whitefield, I. C., "Neural integration and pitch perception," in *Excitatory Synaptic Mechanisms*, edited by Andersen, P., Jansen, J. K. S., Oslo, 1970.

- [Wu et al, 2002] Wu M., Wang D.L., Brown G. J., "A multi-pitch tracking algorithm for noisy speech," Proceedings of *ICASSP-02*, 1.369-1.372, 2002.
- [Yates et al., 1985] Yates, G. K., Robertson, D., Johnstone, B. M., "Very rapid adaptation in the guinea pig auditory nerve," *Hearing Research*, 17, 1-12, 1985.
- [Yost et al, 1978] Yost, W. A., Hill, R., Perez-Falcon, T., "Pitch and pitch discrimination of broadband signals with rippled power spectra," *Journal of the Acoustical Society of America*, 63, 1166-1173, 1978.
- [Yost et al, 1996] Yost, W. A., Patterson, R. D., Sheft, S., "A time domain description for the pitch strength of iterated rippled noise," *Journal of the Acoustical Society of America*, 99, 1066-1078, 1996.
- [Zwicker, Fastl, 1990] Zwicker, E., Fastl, H., *Psychoacoustics, Facts and Models*, Springer-Verlag, Munich, 1990.