Analysis of Very Low Quality Speech for Mask-Based Enhancement

Sira Gonzalez

Communication and Signal Processing Group
Deparment of Electrical and Electronic Engineering

Imperial College London

A thesis submitted for the degree of

Doctor of Philosophy

September 2013

Copyright Declaration

Statement of Originality

I hereby certify that this thesis is the outcome of the research conducted by myself under supervision from Mike Brookes in the Department of Electrical and Electronic Engineering at Imperial College London. Any work that has been previously published and included in this thesis has been fully acknowledged in accordance with the standard referencing practices of this discipline. I declare that this thesis has not been submitted for any degree at any other University or Institution.

Abstract

The complexity of the speech enhancement problem has motivated many different solutions. However, most techniques address situations in which the target speech is fully intelligible and the background noise energy is low in comparison with that of the speech. Thus while current enhancement algorithms can improve the perceived quality, the intelligibility of the speech is not increased significantly and may even be reduced.

Recent research shows that intelligibility of very noisy speech can be improved by the use of a binary mask, in which a binary weight is applied to each time-frequency bin of the input spectrogram. There are several alternative goals for the binary mask estimator, based either on the Signal-to-Noise Ratio (SNR) of each time-frequency bin or on the speech signal characteristics alone. Our approach to the binary mask estimation problem aims to preserve the important speech cues independently of the noise present by identifying time-frequency regions that contain significant speech energy.

The speech power spectrum varies greatly for different types of speech sound. The energy of voiced speech sounds is concentrated in the harmonics of the fundamental frequency while that of unvoiced sounds is, in contrast, distributed across a broad range of frequencies. To identify the presence of speech energy in a noisy speech signal we have therefore developed two detection algorithms. The first is a robust algorithm that identifies voiced speech segments and estimates their fundamental frequency. The second detects the presence of sibilants and estimates their energy distribution. In addition, we have developed a robust algorithm to estimate the active level of the speech. The outputs of these algorithms are combined with other features estimated from the noisy speech to form the input to a classifier which estimates a mask that accurately reflects the time-frequency distribution of speech energy even at low SNR levels. We evaluate a mask-based speech enhancer on a range of speech and noise signals and demonstrate a consistent increase in an objective intelligibility measure with respect to noisy speech.

Acknowledgements

First of all, this thesis would not exist without the help and encouragement of my supervisor, Mr Mike Brookes, whose advice has played a significant role in the research presented in this thesis. The regular meetings have always been a source of great support during the productive, and especially during the less productive, periods of the PhD. His patience and optimism have always been greatly appreciated. Typically I am not good at making compliments, but I hope that after reading these ones he will write me a nice letter of recommendation.

I have to thank people from the lab who have made the long hours at Imperial enjoyable, especially Daniel and James. Also, I am very grateful to the Speech and Audio group, whose meetings have always been a source of new ideas. I would like to add a special thanks to my boyfriend, David, who, apart from putting up with me, has provided me with most of the coffee needed to survive the writing up period.

Last, but not least, a huge thanks to my family, particularly to my parents, who have always supported me no matter what.

Contents

List of Figures

List of Tables

List of Acronyms

ACF Autocorrelation Function 42

AI Articulation Index 17

AMS Amplitude Modulation Spectogram 34

AR Autoregressive 8

ASR Automatic Speech Recognition 31

CART Classification and Regression Tree 22

CASA Computational Auditory Scene Analysis 10

CODEC Coder-Decoder 2

CRLB Cramér-Rao Lower Bound 41

DET Detection Error Trade-off 62

DNN Deep Neural Network 35

DP Dynamic Programming 42

EM Estimation-Maximization 41

ERB Equivalent Rectangular Bandwidth v

FT Filter and Threshold 35

GFCC Gammatone Frequency Cepstral Coefficient 34

GMM Gaussian Mixture Model 14

HMM Hidden Markov Model 41

IBM Ideal Binary Mask 13

KL	Karhunen-Loéve	8
KLT	Karhunen-Loéve Transform	7
LC	Local Criterion	13
LTASS	Long Term Average Speech Spectrum	5
MAP	Maximum A-Posteriori	40
MFCC	Mel-frequency Cepstral Coefficient	35
ML	Maximum Likelihood	40
MMSE	Minimum Mean Squared Error	6
MOS	Mean Opinion Score	15
MP	Matching Pursuit	35
PCA	Principal Components Analysis	7
PESQ	Perceptual Evaluation of Speech Quality	16
POLQA	Perceptual Objective Listening Quality Analysis	17
PSD	Power Spectral Density	76
RASTA-PLP	Relative Spectral Transform and Perceptual Linear Prediction	34
RIR	Room Impulse Response	71
SII	Speech Intelligibility Index	17
SNR	Signal-to-Noise Ratio	4
SPP	Speech Presence Probability	6
STFT	Short Time Fourier Transform	8
STI	Speech Transmission Index	18
STOI	Short-Time Objective Intelligibility	18
SVM	Support Vector Machine	14
TBM	Target Binary Mask	25
UTBM	Universal Target Binary Mask	23

VAD Voice Activity Detector 6

List of publications

The following publications were produced during the course of this work:

Journal papers

1. S. Gonzalez and M. Brookes "PEFAC – A Pitch Estimation Algorithm Robust to High Levels of Noise" submitted to IEEE Trans. Audio, Speech, Lang. Process.

Conference papers

1. S. Gonzalez and M. Brookes "A Pitch Estimation Filter Robust to High Levels of Noise (PEFAC)" in Proc. European Signal Processing Conf. , Page(s): 451-455, Barcelona, Spain, Aug 2011

2. S. Gonzalez and M. Brookes "Sibilant Speech Detection in Noise" in Proc. Interspeech Conf. , Portland, USA, Sep 2012

3. S. Gonzalez and M. Brookes "Speech Active Level Estimation in Noisy Conditions" in Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP), Vancouver, Canada, May 2013

4. S. Gonzalez and M. Brookes "Mask-based Enhancement for Very Low Quality Speech" submitted to IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP), Florence, Italy, 2014

Chapter 1

Introduction

The use of systems involving speech-based communication technology is now ubiquitous; such systems include mobile phones, hearing aids and video-conferencing technology. The perceived quality, and in more severe cases the intelligibility, of the speech signal in these systems is reduced when they are used under the adverse noise conditions encountered in real environments such as offices, crowded public spaces, or railway stations.

To illustrate the passage of a speech signal from talker to listener, a typical single-channel speech recording chain is shown in Fig. 1.1. The desired speech signal passes through a convolutive acoustic channel before reaching the microphone, where it is combined with sound from other acoustic sources in the environment and it is transduced into the electronic domain. The speech signal can become degraded by further additive noise as well as by possible non-linear distortion within the electronic domain.

It is convenient to classify speech signal degradations into the following three classes which differ in their causes and potential remedies:

(i) additive background noise that can arise in either the electronic or acoustic domains, although serious signal degradation is normally caused only by acoustic noise from unwanted sources in the environment;

(ii) convolutive effects including echo and reverberation; and

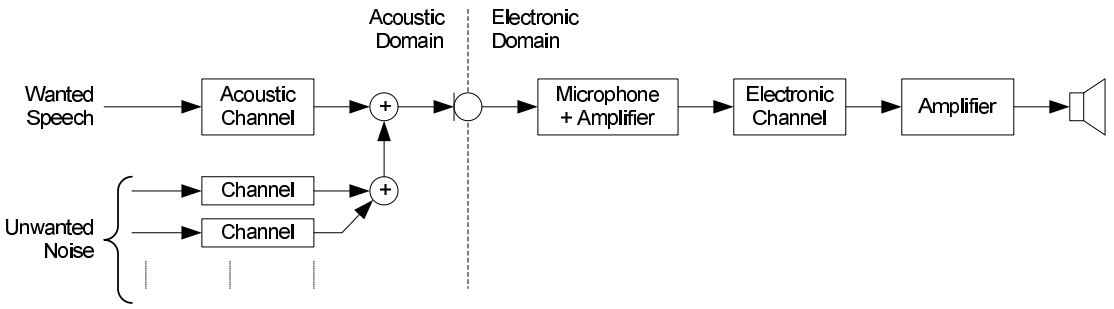


Figure 1.1: Typical speech recording chain.

(iii) non-linear speech distortion which may, for example, be introduced by amplitude limiting or clipping in the microphone, amplifier or Coder-Decoder (CODEC).

In recent decades a diverse range of solutions has been proposed to address these degradation effects. Speech enhancement techniques aim to restore corrupted speech signals by removing or compensating for degradation without damaging the speech signal itself. The work in this thesis is concerned with the enhancement of single-channel speech signals that have been corrupted by levels of additive noise that are high enough to affect the intelligibility of the speech.

In this chapter, we highlight some properties of speech and noise signals, outline the basis of common speech enhancement algorithms and provide an overview of evaluation methods. Finally we state the research motivation and aims, we summarise the layout of the thesis and highlight the thesis contributions.

1.1 Characteristics of speech signals

Speech sounds can be broadly divided into two categories: voiced and unvoiced. Voiced sounds are produced when the vocal folds are vibrating, producing a quasi-periodic signal, while unvoiced sounds are articulated without vibration of the vocal folds. Speech consists of a sequence of vowels and consonants together with brief

silences between phonemes and words [113]. Vowels are created by a voiced sound without any constriction in the vocal tract. Consonants, however, can be originated by a voiced or an unvoiced sound and are classified [93] as:

Stops: which occur when the air flow is blocked and suddenly released.

Nasals: produced when the air is stopped in the oral cavity but not through the nasal cavity.

Approximants: produced when there is a constriction but not narrow enough to result in a turbulence.

Fricatives: a narrow constriction in the vocal tract resulting in a turbulent air flow.

Each of the phonemes included in the different classes share common spectral characteristics. In Fig. 1.2 we illustrate a speech spectrogram labelled with the five different classes: vowels (V), stops (S), nasals (N), approximants (A) and fricatives (F). As we can observe, there are noticeable differences between the spectral shape of some of the classes. Vowels, together with the nasal and approximant voiced consonants, have clear horizontal striations corresponding to the fundamental frequency and its harmonics. Fricatives, however, have an aperiodic noise pattern, especially in higher frequency regions whereas stops are characterised by a silent interval followed by a burst of noise. Stops and fricatives can either be voiced or unvoiced, but in both cases the spectral distribution is similar. The time and energy distribution of the different phoneme classes calculated over the training set of the TIMIT database [37] is shown in Fig. 1.3. In the time distribution, Fig. 1.3(a), we can observe how vowels occupy 52% of the time, followed by fricatives, approximants, stops and nasals. Vowels are also the predominant phoneme class in the energy distribution, Fig. 1.3(b), where they account for 83% of the total energy, followed by approximants (11%) and fricatives (4%). Stops and nasals only account for approximately 1% of the energy each.

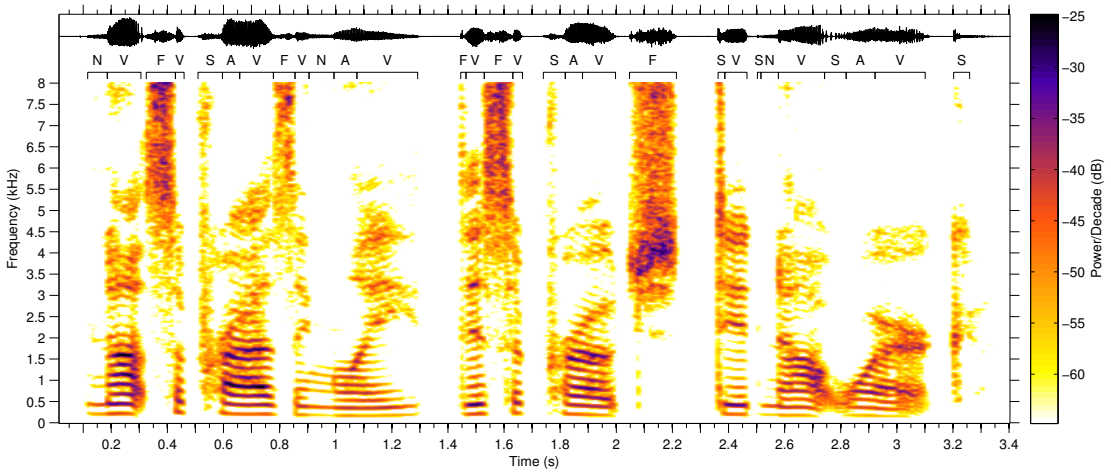
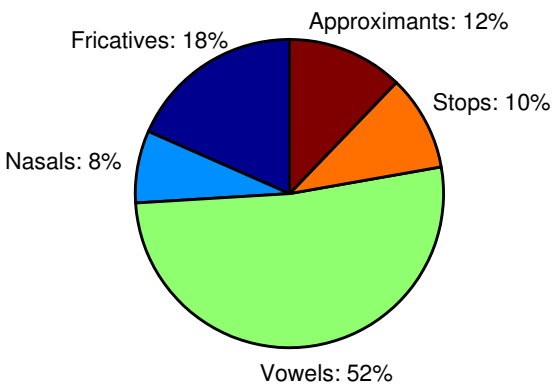


Figure 1.2: Speech spectrogram of the sentence: 'Not surprisingly this approach did not work' divided into five different classes of phonemes: vowels (V), stops (S), nasals (N), approximants (A) and fricatives (F).

(a) Time distribution



(a) Energy distribution

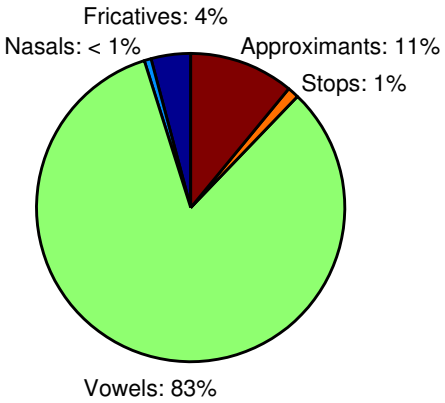


Figure 1.3: (a) Time and (b) energy distribution of the different phoneme classes calculated over the training set of the TIMIT database [37].

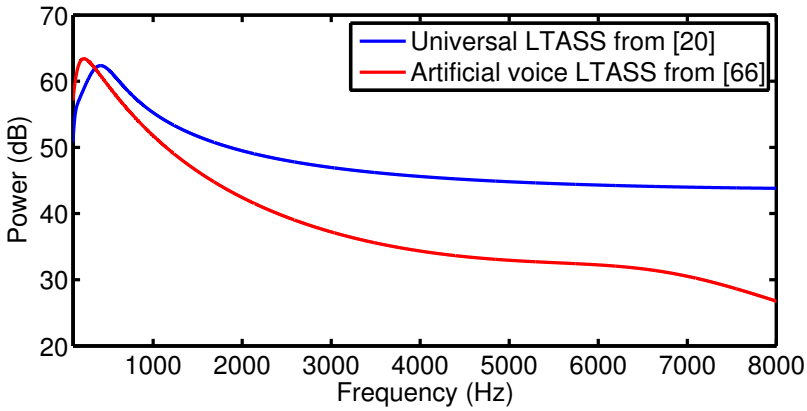


Figure 1.4: Comparison of the universal LTASS recommended in Table II of [19] and the LTASS of the artificial voice [71] defined in (1.1).

1.1.1 Long term average speech spectrum

The frequency distribution of the time-averaged power spectrum of speech is known as the Long Term Average Speech Spectrum (LTASS). It was found in [19] that the LTASS of speech signals was largely independent of language and could be represented therefore by a universal LTASS. Relative to this universal LTASS, the frequency-averaged standard deviation of an individual speaker's LTASS was found to be about 3 dB. An LTASS of the artificial voice, which is aimed at reproducing the characteristics of real speech over the bandwidth 100 Hz − 8 kHz was proposed in [71]. The LTASS of the artificial voice is defined as

$$L_{dB}(f) = -376.44 + 465.439(\log_{10} f) - 157.745(\log_{10} f)^2 + 16.7124(\log_{10} f)^3 \quad (1.1)$$

where $L_{dB}(f)$ is the normalised power spectra in dB relative to $1\,\mathrm{pW/m^2}$ sound intensity per Hertz at the frequency f.

A comparison of the proposed universal LTASS in Table II of [19] and the LTASS of the artificial voice in (1.1) [71] is shown in Fig. 1.4. Although there are some differences in their spectral power density distributions, most of the power is, in both cases, concentrated in frequencies below 1000 Hz.

1.2 Characteristics and estimation of noise signals

Noise, in contrast to speech, can originate from any kind of source and have any spectral and temporal characteristics. There are, however, some common assumptions made about the noise when approaching the speech enhancement problem:

(i) the power spectrum of noise is more stationary than that of speech, and

(ii) speech and noise are statistically independent.

Many speech enhancement techniques require an estimation of the noise power spectrum, or, equivalently, the SNR at each time-frequency bin. The accuracy of the noise estimation technique has a major impact on both the quality and intelligibility performance of the processed speech.

The first noise estimation approaches used Voice Activity Detector (VAD) estimators to identify noise-only intervals. The noise could be then calculated by a temporal average during the speech absences using an averaging time-constant that depends on the assumed stationarity of the noise. A detailed review of several VAD estimators can be found in [16].

A minimum statistics approach was introduced to estimate the noise in [107, 108]. The basis of this approach is that over a given time-interval there will be pauses in the speech in every frequency band and consequently the minimum value of the noisy speech spectrum within a frequency band will correspond to the noise power.

The noise power spectrum can also be calculated by using a Minimum Mean Squared Error (MMSE) estimator. In [48], an MMSE estimator was used to minimise the power of the difference function between the estimated and the true noise power spectrum. This algorithm was found to perform best in a comparative evaluation of several noise estimation algorithms in [136]. The work in [48] has been further extended in [39], where a soft decision Speech Presence Probability (SPP) was used to update the noise adequately. While decreasing the computational complexity of the original algorithm, the estimation accuracy was maintained.

1.3 Single channel speech enhancement

In this thesis, we are concerned with single-channel speech enhancement in which only a single microphone is used. If, in contrast, an array of microphones is available, the speech enhancement problem can be approached differently and the SNR of the desired signal can be improved by coherent averaging or beamforming [12].

Numerous approaches for single-channel speech enhancement, mainly driven by the requirements of telecommunications companies and hearing aid manufacturers, have been developed over many years. A number of speech enhancement algorithms operate in the time domain and typically use adaptive filters [124, 125] or Kalman filters [41, 154, 130]. The majority of algorithms, however, perform the enhancement in a transform domain in which both speech and noise signals are sparse and are therefore more easily separated; these algorithms are described in more detail below.

In this thesis, we represent the noisy speech signal in the time-domain as $y(\tau)$ and we assume that it can be decomposed as

$$y(\tau) = s(\tau) + n(\tau) \tag{1.2}$$

where τ is a sample index and $s(\tau)$ and $n(\tau)$ are the time-domain speech and noise signals respectively.

1.3.1 Enhancement in the Karhunen-Loève domain

The Karhunen-Loéve Transform (KLT), also known as Principal Components Analysis (PCA) [121], is a statistical procedure which allows the orthogonal transformation of a set of observations of a number of correlated variables into a set of values of linearly uncorrelated variables called principal components. The number of principal components is less than or equal to the number of original variables.

In speech enhancement, subspace methods use the KLT to decompose the noisy signal vector within a frame into mutually orthogonal subspaces that are dominated by speech and noise energy respectively. Under the assumption that speech inter-

vals of around 20 ms can be treated as time-invariant and modelled by a low order Autoregressive (AR) process, the vector of speech samples within a frame of this length lies within a low-order subspace. If this subspace is identified, the speech samples can be constrained to lie within it by applying an orthogonal projection onto this subspace.

A speech enhancement algorithm introduced in [28] dealt with white noise by retaining only a specific number of singular values after applying singular value decomposition. Subspace enhancement became popular following [33], in which the noise components in the speech subspace were also removed. The method assumes that the noise is white, and uses an eigendecomposition of the autocovariance matrix of the noisy speech, which consists of the sum of a low-rank matrix arising from the speech and a multiple of the identity matrix arising from the noise. The approach has been further developed in [64, 114] to deal with coloured noise.

Although the KLT provides good speech and noise separability, the transformation into the Karhunen-Loéve (KL) domain is computationally expensive because a different transformation must be determined for each frame.

1.3.2 Enhancement in the time-frequency domain

The dominant domain in which speech enhancement algorithms operate is the time-frequency domain. The reason for this is that transforming the signal into the time-frequency domain is much less computationally expensive than the KL transform, but still provides a separation between speech and noise. Several approaches follow the steps shown in Fig. 1.5 and enhance the signal by applying a time-frequency gain modification.

The most common signal transformation is the Short Time Fourier Transform (STFT), where the first step consists of splitting the discrete input signal, $y(\tau)$, up into frames such that for frame t,

$$y(t,u) = y(\tau)w(\tau - tL) \tag{1.3}$$

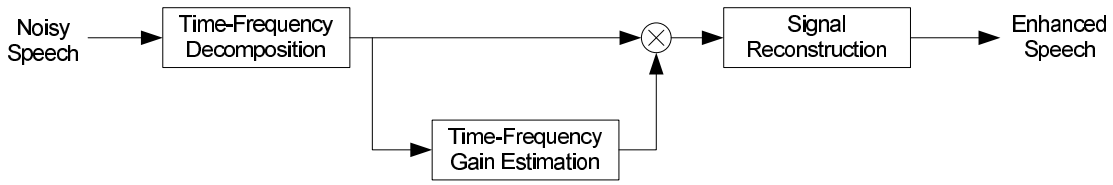


Figure 1.5: Block diagram of time-frequency gain modification techniques.

where τ is a sample index, $w(\tau)$ is a windowing function with finite support N, $u = \tau - tL$ and L is the interframe increment. The window function, $w(\tau)$, is used to avoid spectral artefacts due to discontinuities at the frame boundaries. After performing the decomposition of the signal into overlapping frames, the Fourier transform is calculated on each frame to obtain the STFT

$$Y^{\circ}(t,f) = \sum_{u=0}^{N-1} y(t,u)e^{-j2\pi fu} \tag{1.4}$$

If no further processing is done, the original signal can be perfectly reconstructed from $Y^{\circ}(t,f)$ by applying the inverse Fourier transform and joining the frames up using overlap-add processing [1, 2]

$$y(\tau) = \sum_{t} y(t, \tau - tL)\nu(\tau - tL) \tag{1.5}$$

where $\nu(\tau)$ represents the synthesis window, often chosen to be the same as the analysis window $w(\tau)$. The condition for perfect reconstruction is that the product of the analysis and synthesis windows sums to 1, such that

$$\sum_{t} w(\tau - tL)\nu(\tau - tL) = 1 \tag{1.6}$$

Motivated by the measured characteristics of the inner ear, Patterson et al. [120] proposed a gammatone filterbank as an alternative way of performing the time-frequency decomposition of a speech signal. The impulse response of the filter centred

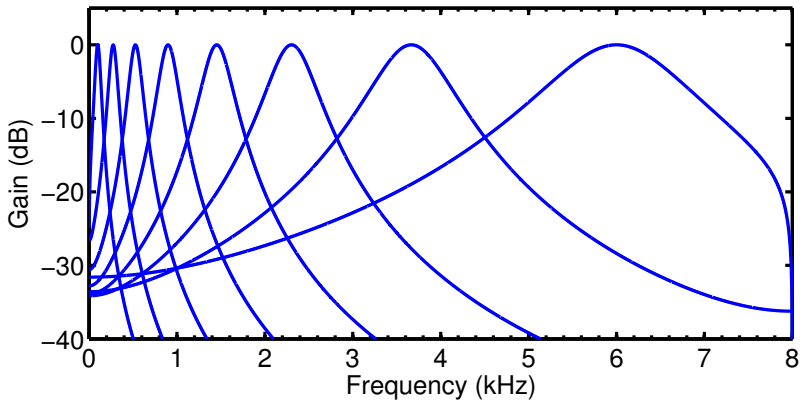


Figure 1.6: Response of a gammatone filterbank composed of 8 gammatone filters equally spaced on the ERB rate scale.

at frequency f_c can be expressed as

$$g(\tau) = \tau^{n-1} \exp(-2\pi b\tau) \cos(2\pi f_c \tau + \phi) \tag{1.7}$$

where n is the order, b is a bandwidth parameter and ϕ is the phase. Figure 1.6 illustrates the frequency response of a gammatone filterbank containing 8 filters equally spaced on the Equivalent Rectangular Bandwidth (ERB) rate scale [115]. The output signal of each gammatone filter is later divided into overlapping time frames, as seen in equation (1.3). Gammatone filterbanks are often used in Computational Auditory Scene Analysis (CASA) approaches, which are inspired by the processing performed by the human auditory system. To follow the same steps as our auditory system, in CASA approaches the output of each gammatone filter is usually further processed to model the inner hair cells [112]. The disadvantage of using a gammatone filterbank to perform a time-frequency decomposition is that perfect reconstruction of the signal is not possible [51].

1.3.2.1 Spectral subtraction

The spectral subtraction approach was introduced in [11] and it is based on the assumption that the complex spectrum of the input signal, $Y^{\circ}(t, f)$, can be expressed as the sum of the speech signal complex spectra, $S^{\circ}(t, f)$, and that of the background

noise, $N^{\circ}(t,f)$, such that $Y^{\circ}(t,f) = S^{\circ}(t,f) + N^{\circ}(t,f)$. If we are able to estimate the noise, we can then recover the speech signal by a simple subtraction. The phase of the noise, however, is usually unknown but it is shown in [31] that, under certain modelling assumptions, the optimal MMSE estimate of the phase of $S^{\circ}(t,f)$ is the phase of the noisy speech component, $Y^{\circ}(t,f)$. Accordingly most enhancers modify the magnitude of the noisy speech spectral components while leaving the phase unaltered. This process can be expressed as

$$\widehat{S}^{\circ}(t,f) = G(t,f)Y^{\circ}(t,f) \tag{1.8}$$

where $G(t,f)$ represents a real-valued gain function. In the simplest form of spectral subtraction this gain is defined by

$$G_{SS}(t,f) = \max\left\{\frac{|Y^{\circ}(t,f)| - |\widehat{N}^{\circ}(t,f)|}{|Y^{\circ}(t,f)|}, 0\right\} \tag{1.9}$$

where $|\widehat{N}^{\circ}(t,f)|$ represents the noise amplitude estimate. Because of errors in the noise estimate, the enhanced speech will have residual noise, either broad-band or narrow-band. Narrow-band residual noise is commonly known as musical noise due to the tonal components it generates. Many modifications to the gain function have been proposed since then in the literature to attenuate residual noise [8, 141], leading to a more general gain function

$$G_{SS}(t,f) = \max\left\{\frac{\left(|Y^{\circ}(t,f)|^{\lambda} - \eta|\widehat{N}^{\circ}(t,f)|^{\lambda}\right)^{1/\lambda}}{|Y^{\circ}(t,f)|}, \beta|\widehat{N}^{\circ}(t,f)|\right\} \tag{1.10}$$

where λ controls the domain in which the gain is calculated, β sets the noise floor and η the noise oversubtraction. Recently, a theoretical analysis of the amount of musical noise generated by spectral subtraction was performed in [67], where it was found that a small λ leads to a musical noise reduction. A subjective evaluation confirmed this finding, where the lowest tested λ was equal to 0.05.

1.3.2.2 Minimum mean square error estimators

Many speech enhancement methods, following the same structure as spectral subtraction, apply a gain function to the noisy time-frequency spectrogram. This gain function is often calculated using MMSE estimators to minimise a specific cost function given assumed models for the speech and noise processes. Systems based on Wiener filtering [99, 110] minimise the power of the difference between the estimated and clean speech power spectra. The gain function of the Wiener filter is given by

$$G_{WF}(t,f) = \frac{\xi(t,f)}{\xi(t,f)+1} \tag{1.11}$$

where $\xi(t,f)$ is the a-priori SNR, defined by

$$\xi(t,f) = \frac{S(t,f)}{N(t,f)} \tag{1.12}$$

where $S(t,f)$ and $N(t,f)$ represent the power spectrogram of speech and noise respectively.

In [31], the aim of the enhancer is to optimise the estimate of the real spectral amplitudes under the assumption that speech and noise spectral components are statistically independent Gaussian random variables. The same authors further extended their algorithm in [32] where they minimised the mean-square error of the log-spectral amplitude. The authors reported that this results in lower residual background noise and improved perceived quality. The gain function can be expressed as

$$G_{MMSE}(t,f) = \frac{\xi(t,f)}{\xi(t,f)+1} \exp\left(\frac{1}{2}\int_{\upsilon(t,f)}^{\infty} \frac{e^{-z}}{z} dz\right) \tag{1.13}$$

where

$$\upsilon(t,f) = \frac{\xi(t,f)}{1+\xi(t,f)}\gamma(t,f)$$

$\gamma(t,f)$ is the a-posteriori SNR

$$\gamma(t,f) = \frac{Y(t,f)}{N(t,f)}$$

and $Y(t, f)$ represents the power spectrogram of the noisy input signal.

To improve the performance of the MMSE algorithm, perceptual masking models have also been introduced [47]. Their motivation is to incorporate the concept of frequency or temporal auditory masking within the human auditory system to remove only the audible noise in the speech signal.

The assumption in [31, 32] that speech and noise spectral components can be modelled as independent Gaussian random variables does not hold when the correlation length of the speech is larger than the analysis window. To overcome this problem, several researchers [101, 109] have extended the MMSE approach under the assumption of a super-Gaussian distribution for speech and/or noise. They have found that this leads to a reduction in residual noise but sometimes at the expense of poorer noise quality.

1.3.2.3 Binary masks

The time-frequency gain modification approaches described above apply a gain function, $G(t, f)$, to each time-frequency cell whose value normally varies continuously over the range 0 to 1. A binary mask enhancer, in contrast, uses a gain function that takes one of two values, 1 and ϵ, where ϵ is a small value typically in the range 0 to 0.1. The most widely used goal is to estimate the so-called Ideal Binary Mask (IBM) [122], which is defined as

$$\mathrm{IBM}(t, f) = \begin{cases} 1 & \text{if } S_{dB}(t, f) > N_{dB}(t, f) + \mathrm{LC}, \\ 0 & \text{otherwise.} \end{cases} \tag{1.14}$$

where $S_{dB}(t, f)$ and $N_{dB}(t, f)$ are the power of the speech and noise signals in decibels respectively and the Local Criterion (LC) is the threshold above which the time-frequency bin is believed to be dominated by the target signal, often set to 0 dB.

There are several motives for the use of binary masks. First, the enhancement problem has been changed from one of estimation to one of classification which allows the use of classification techniques from detection theory and machine learning. Second, it is known from psychoacoustics that the ear perceives only the domin-

ant signal within each frequency band and that weaker signals are masked by the strongest one. Thus it makes sense to attenuate time-frequency cells in which the SNR is so poor that they do not contribute to intelligibility. Third, it has been shown experimentally that arbitrarily noisy speech can be made fully intelligible by using an appropriate binary mask derived from the true speech and noise spectrograms [87]. Fourth, within time-frequency regions where $G = 1$, the enhancer will avoid introducing modulation artefacts and will preserve low level signal components that may contribute to intelligibility even though they cannot be detected explicitly by the algorithm.

A speech enhancer using binary masks was introduced in [83, 82]. The classification of each time-frequency cell was on the basis of the likelihood ratio of two Gaussian Mixture Models (GMMs) trained respectively on training data cells whose local SNR was above and below a threshold. For each frequency channel, the 45-element input feature vector comprised a 15-element modulation spectrum for that channel together with its time and frequency derivatives. The enhancer was evaluated on noisy speech at -5 and 0 dB and consistently improved the subjective intelligibility. Binary masks for enhancement have also been estimated using Support Vector Machines (SVMs) [45], deep belief networks [149] and sparse coding techniques [91]. A more detailed discussion of binary masks and the methods used to estimate them is given in Chapter 2.

1.3.2.4 Gain curves comparison

A comparison of the gain curves of the algorithms outlined in this section is shown in Fig. 1.7 where they are plotted against the a-priori SNR, ξ. As we can observe, for high values of ξ, all algorithms tend to a maximum gain of 1. For values of ξ higher than 5 dB, the Wiener filter and the log-spectral amplitude MMSE estimator behave in a similar way, while for lower ξ values, the Wiener filter attenuates the signal more aggressively. Spectral subtraction has the most gradually changing gain while the binary mask gain changes abruptly from 1 to 0 when ξ becomes negative (assuming LC = 0 dB). Besides the differences in the gain curves, the performance

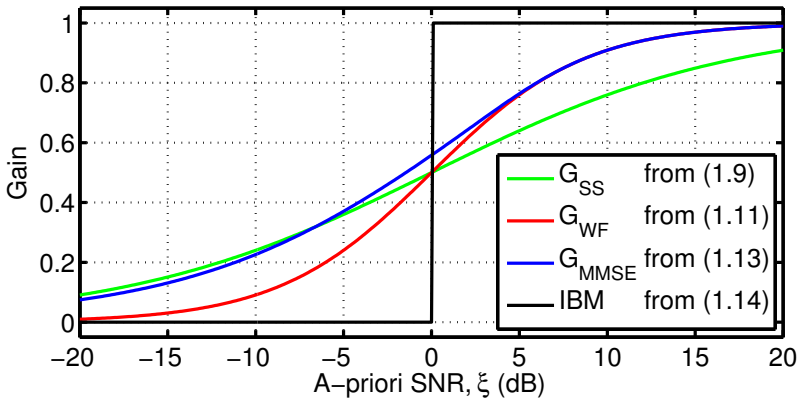


Figure 1.7: Gain curves of different time-frequency domain speech enhancers.

of all algorithms significantly depends on the reliability of the estimated noise power spectrum and/or a-priori SNR.

1.4 Evaluation of speech enhancement systems

The performance of a speech enhancement procedure can be evaluated according to two different perceptual criteria: speech quality and speech intelligibility. Speech quality assesses how comfortable the listener is when listening to the signal. Various characteristics affect the speech quality, such as the level of the residual noise and the final distortion of the signal. In contrast, speech intelligibility is characterised by the percentage of an utterance that a listener is able to identify correctly.

The methods used to evaluate either speech quality or intelligibility can be divided into two groups: subjective methods and objective methods. Subjective methods require the participation of human listeners, and can use absolute scoring if a single stimulus is evaluated at each time or preference scoring if a comparison is made between two or more signals. A popular absolute scoring quality measure is provided by the Mean Opinion Score (MOS) [70]. The MOS value is calculated as the average score provided by a number of trained listeners who rate the quality of the speech on a scale from 1 (bad) to 5 (excellent). Subjective intelligibility evaluation typically requires listeners to identify words that are placed in an unpredictable context (e.g. "The birch canoe slid on the smooth planks") with the intelligibility score taken as

the percentage of content words correctly identified.

Subjective methods, although the only way to obtain true measurements of speech quality and intelligibility, are expensive in both time and resources. Objective measures, in contrast, do not require any external evaluation and estimate the intelligibility or quality using some analysis of the signal, providing an efficient approach for evaluation. Objective measures can be subdivided into (i) non-intrusive methods, which only use the degraded signal for the analysis, and (ii) intrusive methods, which also require the original clean speech signal.

In the next subsections, we focus exclusively on objective measures to evaluate both speech quality and intelligibility and their correlation with subjective ratings. In the scope of this research, the original signal is available and therefore intrusive methods are our main interest for evaluation purposes.

1.4.1 Objective methods for speech quality evaluation

Speech quality objective measures are widely used to evaluate the performance of speech enhancement techniques. The simplest and most widespread intrusive quality measures are based on the SNR. There are many different variations, some of which can be found in [138], but the most popular is the segmental-SNR. The segmental-SNR is calculated by splitting the signals into frames and later averaging the calculated SNR in dB in all the frames that contain speech. Although of low computational complexity, a study published in [62] found that, when used after a speech enhancer, it does not correlate well with subjective quality scores at moderate SNRs of 5 dB and 10 dB.

An intrusive objective method for assessing the quality of noisy speech is defined in [72], named Perceptual Evaluation of Speech Quality (PESQ) and standardised as ITU-T P.862. The PESQ algorithm provides a quality score on a scale from -0.5 to 4.5 by imitating the process the sound undergoes in our auditory system. This score can be converted to the MOS scale [70] by a mapping function defined in [73], such

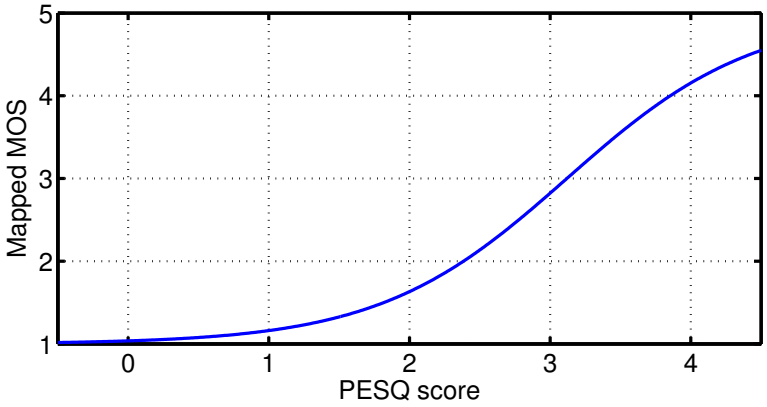


Figure 1.8: Mapping function from the PESQ score to the MOS scale.

that

$$\text{MOS}_{\text{est}} = 0.999 + \frac{4}{1 + e^{-1.4945c + 4.6607}} \tag{1.15}$$

where c is the PESQ quality score. This mapping function is illustrated in Fig. 1.8, where we observe that, above a PESQ score of 1, the PESQ score is approximately linearly related to MOS.

The performance of PESQ on processed speech using speech enhancement algorithms was evaluated in [85], finding high correlations, between 0.83 and 0.96, with the subjective measures. More recently, an extension of the PESQ algorithm, the Perceptual Objective Listening Quality Analysis (POLQA), was developed to add new capabilities and handle higher bandwidths and was standardised as ITU-T P.863 [74].

1.4.2 Objective methods for speech intelligibility evaluation

Over the years, several intrusive methods for speech intelligibility evaluation have been developed to identify how understandable the speech signal is to the listener. One of the earliest approaches was proposed in [34], which led to a standard method for calculating the Articulation Index (AI), ANSI 3.5–1969, [92, 3] and later the Speech Intelligibility Index (SII), ANSI 3.5–1997, [4]. The idea behind these methods is to estimate the speech information that is audible across different frequency bands and weight the output according to the contribution of that particular frequency

band to speech intelligibility. The Speech Transmission Index (STI) [132, 53, 133] adapted the idea behind AI and SII to measure the effect on intelligibility of a transmission channel, dealing with the effects of reverberation and non-linear degradations by measuring the reduction in signal modulation. Various evaluations of STI-based algorithms performed in [53, 102, 42] show that although a good intelligibility correlation is achieved for speech corrupted with additive noise or reverberation, they are unable to predict the effects of speech enhancement algorithms on intelligibility.

With the aim of predicting intelligibility after non-linear processing, a number of intrusive approaches based on correlation measures between the clean and processed signal have been developed. In [42] the authors proposed a normalised correlation measure which gave reasonable results for predicting non-linear processed speech intelligibility. The coherence between the signals was also proposed in [80] to estimate noise and distortion effects, achieving a better prediction performance than that of the SII. Many other algorithms based on correlation methods have subsequently been proposed since [52, 103, 77, 134]. An assessment of various intelligibility evaluation methods is performed in [50], where the results indicate that the intelligibility after speech enhancement algorithms is best predicted by the Short-Time Objective Intelligibility (STOI) measure [134]. STOI first applies the STFT to the input signals and interpolates it into a log-frequency scale. The linear correlation coefficient between the clean and modified time-frequency bins is then calculated over approximately 400 ms windows and averaged over all bands and frames. This intrusive algorithm provides a value between 0 and 1 which is expected to have a monotonic relationship with the speech intelligibility and can be mapped to an absolute intelligibility prediction score with the logistic function

$$\mathrm{INT}_{\mathrm{est}} = \frac{100}{1 + \exp(ad + b)} \tag{1.16}$$

where a and b are free constants (set to -14.5435 and 7.0792 for the Dantale corpus [143]) and d is the STOI value. The mapping function from (1.16) is plotted in Fig. 1.9. We can observe that for STOI values above 0.7, almost perfect intelligibility

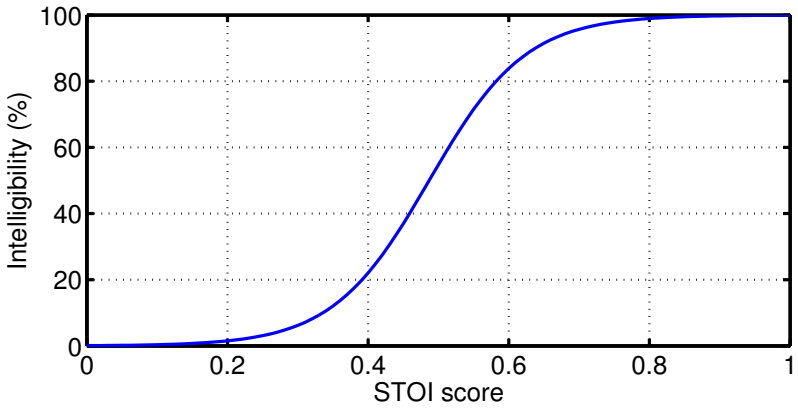


Figure 1.9: Mapping function from the STOI score to the predicted intelligibility for the Dantale corpus [143].

is predicted, while below 0.3 the speech is too corrupted to be understood.

1.5 Performance of speech enhancement algorithms

An example of how intelligibility and quality degrade with additive white noise is shown in Fig. 1.10. STOI [134] is used to estimate the speech intelligibility and PESQ [72] to estimate the quality. Figure 1.10 shows how, in presence of white noise, the estimated quality degrades steadily below 50 dB SNR. The predicted intelligibility, however, remains high for positive SNRs but decreases rapidly below 0 dB SNR. Many speech enhancement techniques operate in the SNR range where the speech intelligibility is still high while the speech quality has decreased substantially. The aim of the speech enhancer in this SNR range is to improve the speech quality while maintaining its intelligibility.

A detailed analysis of the performance of several speech enhancement algorithms both in terms of speech quality and intelligibility can be found in [66]. It was found that, even though the characteristics of each method differ, no enhancement system was capable of improving both quality and intelligibility. The algorithms which performed best in terms of quality were not the same ones that performed best in terms of speech intelligibility. Furthermore, no algorithm provided significant improvements in intelligibility. Previous experiments by [5] reached the same conclusion showing

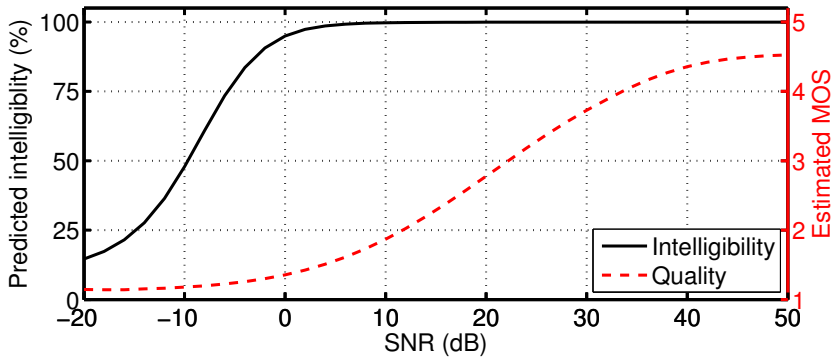


Figure 1.10: Average estimated MOS values using PESQ and predicted intelligibility using STOI values over 100 files of the test set of the TIMIT database [37], where the utterances have been corrupted with white noise at different SNR levels.

that mostly intelligibility gets worse although perceived quality may improve. While the previous studies were performed only for the English language, an intelligibility evaluation for Chinese, Japanese and English is performed in [96]. The results vary significantly between languages, but the general conclusion, again, was the inability of the algorithms to improve substantially the intelligibility. We can conclude that the current speech enhancement techniques are appropriate for positive SNRs where the main task is to improve the speech quality, but they are inappropriate for negative SNRs, where making speech intelligible is more important than improving its quality.

1.6 Research motivations and aims

The goal of this research project is to improve the intelligibility of very low-quality speech. Each kind of speech degradation has different characteristics and it is complicated to develop an approach that can cater at the same time with all of them. The work in this thesis is concerned with additive background noise due to its major contribution to speech intelligibility degradation.

We have seen in Section 1.5 that no current speech enhancement approach has been able to improve speech intelligibility. However, several studies [18, 86] have shown the potential of time-frequency binary masks to enhance speech intelligibility. A binary gain is a special case of a continuous gain and so the performance of an ideal binary

gain system cannot exceed that of an ideal continuous gain system, and, as pointed out in [75, 104], if the a priori SNR is known, the performance of binary masks is lower than that of algorithms applying a continuous gain. However, the advantage of binary masks is that they permit entirely new approaches to the speech intelligibility enhancement problem, which may now be seen as a classification problem rather than as an estimation problem. In this thesis, we aim to study and develop this potential to estimate a binary mask which is able to enhance the intelligibility of the corrupted speech.

Based on the idea that a binary mask based only on the speech can provide good intelligibility performance, as shown by the target binary mask performance in [86], our approach to the binary mask estimation problem aims to preserve the important speech cues independently of any noise that is present. As we have seen, the time-frequency regions that contain significant speech energy depend heavily on the kind of speech sound produced, and, to locate the speech energy in the time-frequency domain we need to identify voiced speech and its fundamental frequency and to estimate the location and energy distribution of the unvoiced sounds. All the extracted information can be combined for the binary mask estimation. In order to make our algorithm independent of the input speech level, we also need to estimate the speech active level and normalise the input appropriately.

1.7 Thesis overview

A detailed explanation of the different binary mask targets is provided in Chapter 2. We define a new time-frequency binary mask target that is both noise and speaker independent and we explore the ways in which the binary mask estimation problem has been approached in the literature.

In Chapter 3 we present PEFAC, a fundamental frequency estimation algorithm that is able to identify voiced frames and estimate pitch reliably even at negative SNRs. The algorithm combines a normalisation stage (to remove channel dependency and to attenuate narrow-band noise components) with a harmonic summing filter

applied in the log-frequency power spectral domain. A voiced speech probability is computed from the likelihood ratio of two classifiers, one for voiced speech and one for unvoiced speech/silence. We compare the performance of our algorithm with that of other widely used algorithms and demonstrate that it performs exceptionally well in both high and low levels of additive noise.

A new method for speech active level estimation which combines a novel algorithm based on voiced speech energy extraction with the standardised ITU-T Recommendation P.56 is described in Chapter 4. At poor SNRs, the algorithm estimates the active level by identifying intervals of voiced speech and summing the energy of the pitch harmonics in the time-frequency domain while rejecting that of the noise. We compare the performance of our method with that of ITU-T P.56 on the TIMIT database and demonstrate that it performs well in both high and low levels of additive noise

We focus on unvoiced speech in Chapter 5, where we introduce an algorithm for identifying the location of sibilant phones in noisy speech. Our algorithm does not attempt to identify sibilant onsets and offsets directly but instead detects a sustained increase in power over the entire duration of a sibilant phone. The normalised estimate of the sibilant power forms the input to two Gaussian mixture models that are trained on sibilant and non-sibilant frames respectively. The likelihood ratio of the two models is then used to classify each frame. We evaluate the performance of our algorithm on the TIMIT database and demonstrate that the classification accuracy is over 80% at 0 dB signal to noise ratio for additive white noise.

All the information extracted by the algorithms explained in Chapters 3, 4 and 5 are combined in Chapter 6 with a noise estimate to form the feature vector to the mask estimator. We use the Classification and Regression Tree (CART) approach to estimate the mask and we show that, for noise types included in the training, the proposed method is able to achieve substantial improvements in the predicted intelligibility using the STOI algorithm for SNRs as low as -5 dB.

Finally, Chapter 7 concludes the thesis and proposes future work.

1.8 Thesis contributions

To the best of the author's knowledge, the original contributions of this thesis are:

1. The proposal and evalulation of the Universal Target Binary Mask (UTBM).
2. The PEFAC algorithm, a pitch estimation algorithm robust to high levels of noise.
3. A speech activel level estimation algorithm in noisy conditions.
4. A method for detecting sibilant speech in noise.
5. A mask-based speech enhancer able to improve the predicted intelligibility of low quality speech.

Chapter 2

Time-frequency binary masks

Time-frequency binary masks aim to identify regions of the time-frequency plane that contain information from the target sound. They were first introduced in the field of speech recognition to identify noise-dominated regions of the time-frequency domain. Either these regions can then be ignored completely in subsequent processing or else the "missing data" that they should contain can be estimated prior to performing recognition [23]. They have also been used in the field of Computational Auditory Scene Analysis (CASA) as a way of segregating a single source from a complex auditory scene by selecting only those time-frequency cells in which the wanted source is dominant [144]. More recently, they have been used as a time-frequency gain function in speech enhancement [83].

The most popular binary mask is the Ideal Binary Mask (IBM), where the decision whether to retain a time-frequency bin depends on its SNR. The IBM was proposed as the goal of CASA in [144], supported by its consistency with the auditory masking effect, in which if two sounds are within the same critical frequency band [155] the weaker signal is masked and eliminated from our perception. Within the field of speech enhancement, binary masks have generated a lot of interest in the last few years as they provide the possibility of approaching the problem as a classification rather than as an estimation problem. As a classification problem, it can benefit from the modelling power of machine learning techniques.

In this chapter, we describe the IBM together with alternative goals for the binary

mask estimation problem. A literature review of the algorithms for binary mask estimation is also presented.

2.1 Goals of mask estimation

The parameters which determine the rejection/acceptance of a time-frequency bin vary according to different binary mask definitions. The original goal of the binary mask estimation was to identify the regions where the SNR was higher than 0 dB [144, 98]. Later research [146, 86] has shown that the optimum SNR threshold to maximise intelligibility depends on the global SNR of the noisy input speech.

In recent years, an alternative goal has been proposed [86], which aims at retaining time-frequency regions with significant speech energy for speech intelligibility. The definition of "significant speech energy" for speech intelligibility is complex, as it should depend on both time and frequency information, and the problem is simplified by identifying the time-frequency bins in which power is above a specific threshold.

In this section, we define two existing time-frequency masks: the IBM which is a function of the local SNR and the Target Binary Mask (TBM) which depends on the LTASS of the speaker. We propose a variation of the TBM, the UTBM, and we show it has a similar performance to that of the TBM while removing dependency on the speaker.

2.1.1 Ideal binary mask (IBM)

The IBM is defined in terms of the SNR at each time-frequency bin. If $S_{dB}(t, f)$ is the power of the desired stream measured in decibels at frame t and frequency f and $N_{dB}(t, f)$ is the corresponding power of the interference, the mask is defined by

$$\mathrm{IBM}(t, f) = \begin{cases} 1 & \text{if } S_{dB}(t, f) > N_{dB}(t, f) + \mathrm{LC}, \\ 0 & \text{otherwise.} \end{cases} \tag{2.1}$$

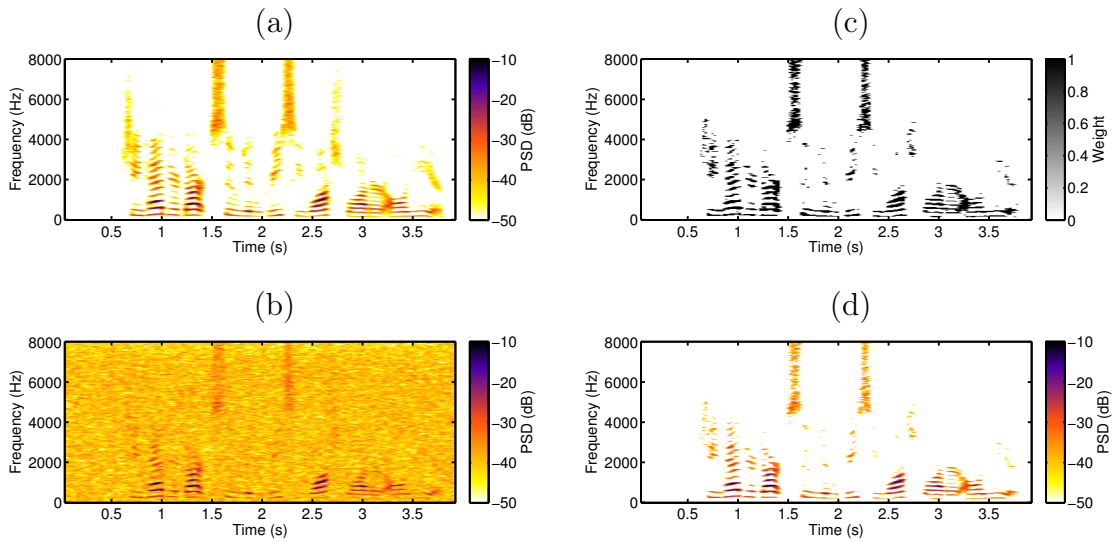


Figure 2.1: (a) Spectrogram of the clean speech, (b) spectrogram of the speech corrupted with white noise at 0 dB SNR, (c) ideal binary mask with 0 dB LC, and (d) segregated speech spectrogram.

where the Local Criterion (LC) is the SNR threshold above which the time-frequency bin is assumed to be dominated by the target signal. In [144] this definition was justified based on its flexibility, unambiguity and its consistency with the auditory masking effect. It has been found in many studies [18, 97] that applying an IBM to noisy speech can provide perfect intelligibility for a range of LC values. An example to illustrate the IBM is shown in Fig. 2.1. Figure 2.1(a) shows the spectrogram of a female speaker saying "She had your dark suit in greasy wash water all year" and Fig. 2.1(b) illustrates the speech spectrogram corrupted with additive white noise at 0 dB SNR. Using 0 dB LC, the obtained ideal binary mask is shown in Fig. 2.1(c) and the segregated speech spectrogram in Fig. 2.1(d).

The optimal performance of the IBM with 0 dB LC is shown in [98] in terms of SNR gain at three different levels: time-frequency unit level, time frame level and global level. According to [98], the task for a sound separation system is to estimate the IBM with 0 dB LC from the noisy speech signal. In terms of intelligibility, however, several studies [146, 86] have shown that the best results are achieved when the LC is chosen to be similar to the input SNR. A model for intelligibility as a function of SNR and LC is presented in [87] based on measurements described in [86]. The

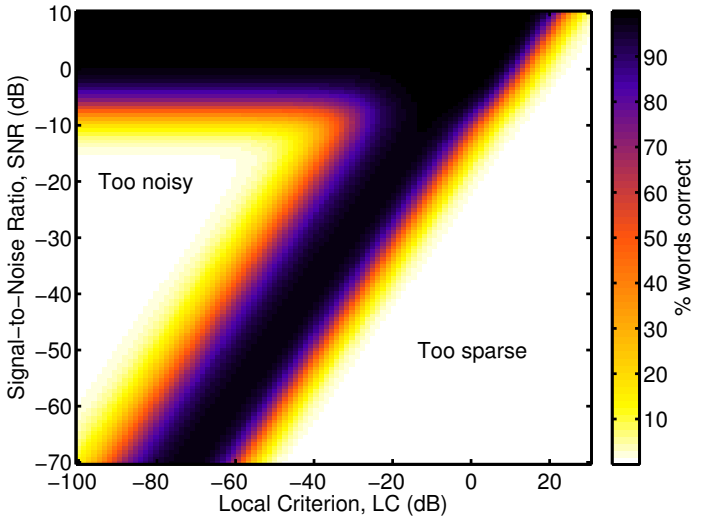


Figure 2.2: Model of intelligibility versus SNR and LC from [87]. The dark areas correspond to high intelligibility.

hypothesis underlying the model is that the auditory system combines two sources of information. The first source is the noisy speech whose intelligibility depends on the SNR while the second source, the noise vocoded signal, consists of noise that has been modulated by the speech information in the mask pattern. The intelligibility of this second source depends on the difference between the mask threshold and the level of the speech. If the LC is set too high relative to the speech level, the mask pattern will become very sparse and the intelligibility of both information sources will degrade. The model is plotted in Fig. 2.2 using the model parameters determined for speech-shaped noise; the dark areas show regions of intelligibility. As we can observe in the figure, above an SNR threshold (approximately $-5\,\mathrm{dB}$), the noisy speech is already understandable and will remain intelligible unless the LC value is so high that the processed signal is too sparse. For any SNR below this threshold there is a range of LC values for which perfect intelligibility is possible, centred approximately on the value of the input SNR.

Research performed by [147] evaluated the intelligibility performance of IBM vocoded noise for different numbers of frequency channels. It was shown that a relatively coarse time-frequency resolution, with as few as 16 frequency bands equally spaced

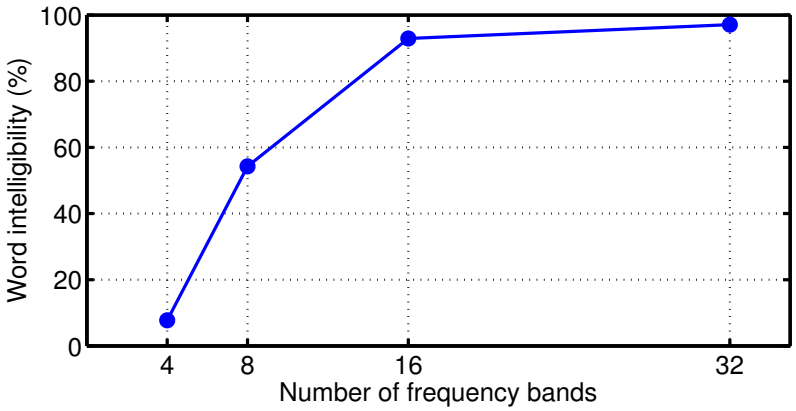


Figure 2.3: Word intelligibility scores versus number of frequency bands from [147].

on the Equivalent Rectangular Bandwidth (ERB) rate scale [115], was sufficient for a high recognition rate, as shown in Fig. 2.3.

A study in [14] evaluated the effects of a tempered version of the IBM where the attenuation function is gradual instead of binary and limited to 0.1. Although the tempered version improves the processed speech naturalness and provides less noise annoyance, its performance is not as good as the performance of the IBM in terms of intelligibility.

2.1.2 Target binary mask (TBM)

The finding in [147] that vocoded noise using the IBM is understandable implies that the binary mask carries by itself all the information needed for intelligibility. This indicates that the binary mask should not depend on the noise, but rather focus on preserving the speech information necessary for intelligibility.

A binary mask based only on the speech was first proposed in [5]. With the aim of preserving 99 % of the speech energy, a threshold was set for all frequency bands above which the binary mask was equal to 1. In [86], inspired by the results obtained in [147], the authors proposed the TBM. The TBM, whose threshold varies across frequencies, is defined as

$$\text{TBM}(t,f) = \begin{cases} 1 & \text{if } S_{dB}(t,f) > r_{dB}(f) + \text{LC}, \\ 0 & \text{otherwise.} \end{cases} \tag{2.2}$$

where $r_{dB}(f)$ is the LTASS of the speaker. In this way, the TBM is calculated by comparing the energy of the target speech with the long-term average energy of speech from the same speaker. The intelligibility performance of the TBM was evaluated in [86], where the authors show that this mask is capable of providing the same or better intelligibility results as the IBM.

2.1.2.1 Universal target binary mask (UTBM)

Although the TBM removes dependency from the noise, its definition still relies on the LTASS of the speaker. In this section, we propose an alternative to the TBM, the UTBM, which removes dependency on both the noise and the speaker by using a universal LTASS. As mentioned in Section 1.1.1, the LTASS of speech signals is largely independent of language and can be represented by a universal LTASS [19]. Consequently, instead of the LTASS of the speaker, the LTASS of the artificial voice [71] defined in (1.1) is used to define the mask such that:

$$\text{UTBM}(t,f) = \begin{cases} 1 & \text{if } S_{dB}(t,f) > L_{dB}(f) + \alpha + \text{LC}, \\ 0 & \text{otherwise.} \end{cases} \tag{2.3}$$

where $\alpha = l - P_L$ is a variable to adjust the power in dB of the threshold function, P_L, to that of the speech active level, l, [68].

We evaluate the intelligibility of the UTBM by using STOI measure [134] which can accurately predict the intelligibility achieved using the TBM and the IBM. Figure 2.4 shows the average predicted intelligibility for the noisy speech and for the enhanced speech using the TBM and the UTBM (the LC value is equal to 0 in both cases) when the speech is corrupted with all noise types from the RSG-10 database [131] at different SNRs. As we can observe, the predicted intelligibility is close to 100% for all the evaluated SNRs for both masks. A comparison between the predicted intelligibility versus LC for TBM and UTBM is provided in Fig. 2.5. The TBM

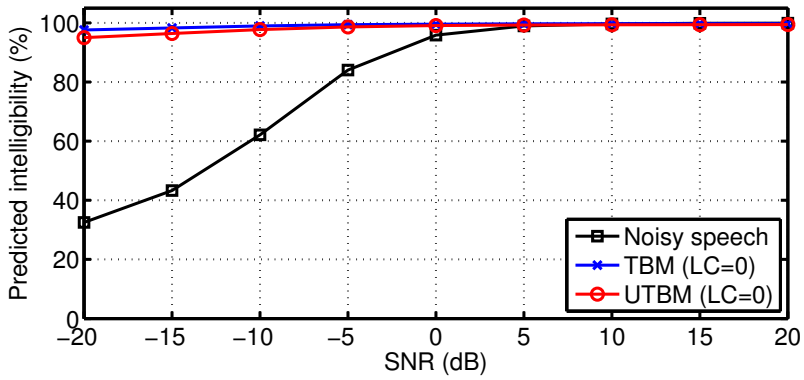


Figure 2.4: Average predicted intelligibility using STOI over 98 speech segments of 5 s duration from 4 speakers from the SAM database, where the utterances have been corrupted with different noise types from the RSG-10 database [131] at different SNR levels.

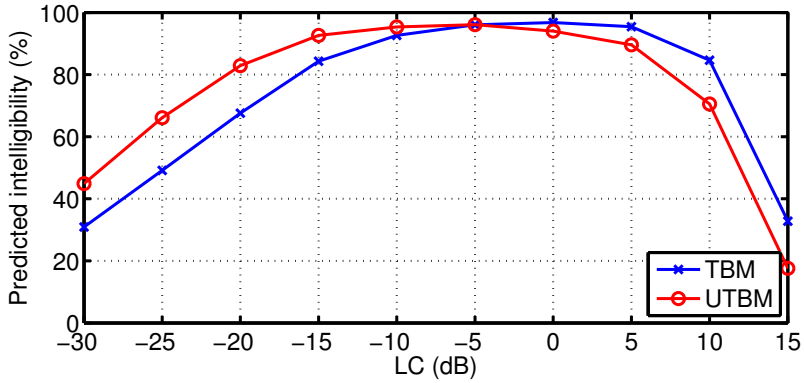


Figure 2.5: Average predicted intelligibility using STOI over 98 speech segments of 5 s duration from 4 speaker from the SAM database [20]. The calculated TBM and UTBM for different LC values have been gated through speech shape noise.

and UTBM were calculated for different LC values and gated through speech shape noise. We can observe that both masks follow a similar intelligibility pattern with a horizontal shift of 5 dB; best predicted intelligibility is achieved at -5 dB LC for TBM and at 0 dB LC for UTBM. We can summarise the crucial property of a mask by noting that speech will be intelligible as long as its most important features are preserved, independently of the background noise.

2.2 Evaluation of mask estimation

A critical part of the binary mask estimation problem is the evaluation: there is no universally accepted measure to evaluate the performance of a binary mask estimator. The different objectives of the system in which the binary mask is to be used makes it difficult to define a measure that is universally appropriate. While sometimes the aim is to improve the performance of Automatic Speech Recognition (ASR), other systems aim to improve the intelligibility of the speech.

A subjective evaluation of the speech intelligibility is both expensive and time consuming, and objective measures are often used for the binary mask evaluation. Up to now, many binary mask estimation techniques are evaluated in terms of SNR improvement. Unfortunately, this measurement does not correlate directly with the speech intelligibility or quality at poor SNRs.

Another measure which has been more recently introduced is the Hit minus False Alarms (HIT-FA) rates. A study in [97] showed that not all the errors introduced in the binary mask have the same effects on intelligibility: false alarms (originally 0 labelled as 1) degrade intelligibility more than misses (originally 1 labelled as 0). Later, a correlation between HIT-FA and speech intelligibility was discovered in [83]. Based on these results, HIT-FA error percentages in unit labelling are usually provided to evaluate the estimated binary mask. The disadvantage of HIT-FA rates is that they provide no comparison between the intelligibility of the noisy speech and the masked speech, therefore they give no information about whether the technique has improved the original intelligibility. Another disadvantage is that direct comparison between algorithms is not possible if they pursue different binary mask definitions. The use of recent objective measures for speech intelligibility such as STOI [134], which has shown good intelligibility correlation for binary masks, solves both problems as they operate on the segregated speech and they provide a good comparison between the original and the processed speech.

2.3 Mask estimation techniques

Most binary mask estimation approaches have aimed to pursue the IBM with 0 dB LC, which means that the target dominated regions need to be estimated. There are different perspectives from which the estimation can be approached. While some techniques try to estimate the noise and from it the SNR, others have sought to identify the speech energy.

2.3.1 SNR-based masking

An estimation of the binary mask from the ratio of the noisy speech spectrum to the estimated noise was presented in Cooke et al. [23]. Working in the time-frequency domain, the estimated noise was computed with a simple averaging over the initial frames of the files when only noise was present. With a threshold of 0 dB, they called this the "negative energy criterion". Other thresholds were studied too, based on the local SNR criterion. However, the results were similar to spectral subtraction and musical noise, individual narrow-band spectral spikes which generate tonal noise, was present in the final segregation.

Several binary mask estimation techniques, which were adapted from other speech enhancement techniques centred around SNR estimation, were evaluated in [63]. The a-priori SNR was calculated using the gain functions of spectral subtraction and several MMSE-based techniques. The hit and false alarm rates were calculated for positive SNRs and the best performance was achieved for the statistical-based algorithms [31, 32, 100].

2.3.2 Identification of speech energy

Speech characteristics were exploited in Seltzer et al. [127] for binary mask estimation. The algorithm first classified voiced and unvoiced speech frames using a pitch estimator based on RAPT [137]. Seven different features were used to identify reliable voiced speech:

(a) the ratio of energy at the harmonics of the voiced speech to the energy outside the harmonics, which decreases with the presence of noise;

(b) the ratio of the second largest to the largest peak in the autocorrelation, also decreasing with poor SNR;

(c) the subband to fullband energy ratio, measuring the effect of noise on a particular subband and on the overall contour;

(d) the kurtosis of the subband signal that will decrease at poor SNRs as the signal becomes more Gaussian;

(e) the spectral flatness in the region of the subband, this being given by the variance of the subband energy in a neighbourhood;

(f) the subband energy to subband noise floor ratio; and

(g) the estimated SNR based on spectral subtraction.

For unvoiced frames a reduced set of features was used that excluded those dependent on the pitch, i.e. (a) and (b). The mask was calculated using a two-class (reliable and corrupt) Bayesian classifier for each type of speech: unvoiced and voiced. Each subband of the spectrogram was processed individually and a classifier was trained for each one. A binary and/or continuous mask was estimated using this technique with the purpose of increasing the speech recognition scheme performance. In this application, the speech quality and/or intelligibility measurements were not relevant and the lowest tested SNR was 0 dB.

Following the work in [127], cepstral coefficients were used in [84] inside each subband along with its derivatives as features for the classifier. Additional features included a spectral flatness measure and the ratio of the energy in the harmonics voiced speech to the energy outside the harmonics. The training was done using coloured noise and a restoration method for voiced and unvoiced frames misclassification is introduced. The results were evaluated in terms of ASR recognition accuracy for positive SNRs and showed an improvement equivalent to about 5 dB in SNR.

An algorithm based on the modulation domain was proposed in Kim et al. [83], where no speech/noise detection or noise statistics were required. The algorithm consisted of a two-class Bayesian classifier that divided time-frequency units into target-dominated and masker-dominated groups. Amplitude Modulation Spectograms (AMSs) [90], which are neurophysiologically and psychoacustically motivated and capture information about amplitude and frequency modulations, were used along with their time and frequency derivatives to train two GMMs that represented the distribution of the feature vector of each class. SNR levels of -5 and 0 dB were used for the experiments in which the same noise signals were used for training and testing. Using subjective tests, an improvement in intelligibility was claimed equivalent to up to 5 dB of SNR. This work was further extended in [82], where a fast adaptation to new noise environments was introduced by implementing an incremental training approach.

A classification approach using SVMs was used in Han and Wang [45]. Pitch-based features and AMSs were used to train a radial basis function SVM. A re-thresholding was done to the output of the SVM to maximize the HIT-FA rates in each channel. Finally, an auditory segmentation stage taked advantage of information in neighbouring time-frequency bins to estimate the final mask. The results showed that this method achieved higher HIT-FA rates on seen noise types than the previous approach proposed by Kim et al. in [83], which used GMMs for classification. The authors further improved this algorithm in [46], where the binary mask estimation was accommodated to unseen conditions. The feature set was extended by including Relative Spectral Transform and Perceptual Linear Prediction (RASTA-PLP) features. In order to adapt to different SNRs or noises, the SVM decision was mapped to a number between 0 and 1 and the threshold above which the time-frequency bin is considered to belong to the speech was accordingly changed. Contextual information was also used and the results, shown in terms of SNR improvement and HIT-FA percentages, showed an improvement for most tested SNRs with respect to previous approaches.

A detailed analysis of different features for the ideal binary mask estimation was performed in Wang et al. [148]. Pitch-based features, AMSs, Gammatone Fre-

quency Cepstral Coefficients (GFCCs), Mel-frequency Cepstral Coefficients (MFCCs), RASTA-PLP were explored as inputs to a Gaussian kernel SVM for classification. The results were evaluated in terms of HIT-FA for matched and unmatched noise conditions. Individually, GFCCs and RASTA-PLP obtained the best results for matched-noise conditions and for unmatched conditions respectively.

With the aim of modelling temporal dynamics, Wang and Wang [149] employed linear-chain structured perceptrons. Their algorithm first extracted a set of features containing AMSs, RASTA-PLP, MFCCs, pitch-based features and delta functions. As the performance of structured perceptrons is largely dependent on the linearly separability of the features, Deep Neural Networks (DNNs) were used to learn linearly separable features functions from the input set of features. The HIT-FA rates for both seen and unseen noises outperformed previous approaches which used either GMMs or SVMs as classifiers. The same authors, aiming to improve the binary mask estimation on unseen conditions, proposed an algorithm in [150] which also used DNNs. The technique extracted the same set of features studied in [148] and DNNs were again employed to learn linearly separable features. In this case, a linear SVM was used for classification and the experiments showed, in terms of HIT-FA results, a better generalisation than that achieved with a Gaussian-kernel SVM. No comparison was made between the approach presented in [149] and in [148].

A new method for binary mask estimation was introduced [91], where sparse coding techniques were used. The authors chose a dictionary which consisted of gammatone functions [129] and used the Matching Pursuit (MP) greedy algorithm to minimize the number of non-zero coefficients. The Filter and Threshold (FT) algorithm, less computationally expensive, was also evaluated. The performance of both algorithms was shown in terms of predicted intelligibility using the STOI algorithm. However, neither of the algorithms was able to increase the predicted intelligibility.

2.3.2.1 Voiced speech segregation

One of the earliest approaches for voiced speech segregation was proposed by [119], based on the fundamental frequency estimation algorithm described by [126]. The

idea underpinning [126] was to create a subharmonic histogram knowing that the fundamental frequency can be calculated with high precision by dividing the frequency of a harmonic by its harmonic number. In [119], using this subharmonic histogram and a pitch tracker, the authors developed a method for separating voiced speech from interfering voiced speech. No background noise other than an interfering speaker was considered.

A system to separate harmonic sounds based on association cues in human auditory organisation was presented in [142]. The system first extracted sinusoidal spectral components and then calculated perceptual distances between sinusoidal trajectories focusing on the synchronous changes of the components and their harmonic concordance. Finally, these trajectories were classified into different sound sources minimising the distances between trajectories inside a class. As in [119], the authors only considered harmonic interfering sounds.

Following the steps of CASA-based models to segregate voiced speech [24, 145], the algorithm in [56, 57] used temporal continuity and measures of the correlation between adjacent frequency channels to identify regions dominated by a periodic signal. The pitch of the target speech was estimated and then used to label each region dominated by a periodic signal either as target or interference. The algorithm was tested for different noise types with SNRs equal or higher than 0 dB, but no direct measure of speech intelligibility or quality was conducted.

A tandem algorithm for pitch estimation and voiced speech segregation was proposed in [60]. After an approximate estimate of pitch contours, voiced speech time-frequency bins were identified and then used to improve the pitch estimate. This process is iterated until it converges or a maximum number of iterations is reached. The results show an SNR improvement over previous work by the same authors [56].

2.3.2.2 Unvoiced speech segregation

Although there are many approaches for the segregation of voiced speech, methods for unvoiced speech segregation are less well developed. An algorithm for stop consonants separation was proposed in [55, 54], where the authors focused on identifying stop

bursts by detecting their onsets. Information on their auditory spectrum, relative intensity and intensity decay time is also integrated to differentiate stops from other signals. The final classification was performed using a Bayesian decision rule.

A method for unvoiced speech segregation which targeted both stops and fricatives was proposed in [59]. In this approach, the noisy speech was divided into different segments based on the onsets and offsets of auditory events [58]. After removing the segments dominated by voiced speech or by periodic or quasiperiodic signals, two multilayer perceptrons were used as classifiers to identify the segments dominated by unvoiced speech.

Another way of segregating unvoiced speech was proposed in [61]. After the voiced speech and periodic noise were segregated, the non-periodic noise was estimated during the neighbouring voiced intervals. Spectral subtraction was then used to estimate unvoiced segments, which were then classified between unvoiced speech segments and interference segments based on the lower and upper frequency bound of the segment using thresholding or Bayesian classification.

2.4 Summary

In this chapter, we have discussed alternative binary mask targets and methods of estimating them. The time-frequency bin selection of binary mask targets can be based on the SNR of the time-frequency bin (IBM), or based only on the speech power (TBM and UTBM). We have seen that binary masks based only on the speech power have similar intelligibility performance to the IBM, showing a new way of understanding the action of a binary mask. The binary mask estimation problem can thus be approached as a speech power identification problem, independently of the noise present.

Many approaches to estimate a binary mask have been attempted, mainly aiming to estimate the IBM with 0 dB LC. While some methods focus on the noise estimation, others concentrate efforts on extracting information from the speech. Several approaches have recently investigated the potential of different machine learning tech-

niques to perform the time-frequency bin classification. However, the different evaluation techniques and conditions used for performance evaluation make the comparison of the algorithms very difficult. In general, binary mask estimation is a problem that has attracted considerable interest in the last years, and new techniques aimed at lower SNRs, adaptation to unseen conditions and intelligibility improvement are in constant development.

Chapter 3

Pitch estimation algorithm robust to high levels of noise (PEFAC)

The estimation of fundamental frequency, or pitch[1], is a key component of voiced speech segregation and therefore an essential element for a binary mask estimator based on identifying time-frequency regions containing speech energy. The estimation of the fundamental frequency also plays an important role in many other speech processing applications and numerous approaches have been described in the literature.

In situations where there is a high level of acoustic noise or where the distance between the microphone and speaker is large, the SNR of an acquired speech signal can be very poor. In such circumstances the performance of pitch estimation algorithms degrades [128], and may become unusable below 0 dB SNR. In recent years a number of noise-robust algorithms have been proposed but reliable fundamental frequency estimation at low SNRs remains a challenging problem.

This chapter presents a fundamental frequency estimation algorithm, PEFAC, that is able to identify voiced frames and estimate pitch reliably even at negative SNRs. The algorithm combines a normalization stage, to remove channel dependency and to attenuate narrow-band noise components, with a harmonic summing filter applied in the log-frequency power spectral domain, the impulse response of which is

[1]In this thesis we treat "pitch" and "fundamental frequency" as synonyms.

chosen to sum the energy of the fundamental frequency harmonics while attenuating smoothly-varying noise components. Temporal continuity constraints are applied to the selected pitch candidates and a voiced speech probability is computed from the likelihood ratio of two classifiers, one for voiced speech and one for unvoiced speech/silence. We compare the performance of our algorithm with that of other widely used algorithms and demonstrate that it performs well in both high and low levels of additive noise.

3.1 Introduction

Many pitch estimation algorithms have been proposed in the literature; these may be divided into parametric and non-parametric algorithms. While parametric algorithms assume an explicit model for the noisy speech, non-parametric methods do not make such assumptions. In this section, we first review the literature of existing pitch estimation algorithms, we explain the intrinsic difficulties in estimating pitch and finally we provide an overview of the proposed PEFAC algorithm. In this chapter, we are concerned with the specific problem of tracking the pitch of voiced speech from a single speaker.

3.1.1 Parametric pitch estimators

The parametric algorithms define a parametric stochastic model for a noisy speech signal with the pitch, or its equivalent, as one of the parameters. The pitch is then estimated by calculating the MMSE or Maximum Likelihood (ML) estimate of the model parameters from the observed signal. By incorporating prior distributions for the parameters, Bayes' theorem can be used to obtain a Maximum A-Posteriori (MAP) estimate. A good description of several parametric methods is contained in [21]. A widely used time-domain parametric model for voiced speech consists of a harmonic series comprising sinusoidal components at integer multiples of the pitch; in the HMUSIC algorithm [22] this is combined with a white noise model and the algorithm simultaneously estimates both the pitch and the number of harmonics

present in the signal. Instead of operating in the time domain, several authors define a parametric model of the power spectrum obtained by applying either the STFT (see Sec. 1.3.2), or an alternative time-frequency transform, to the noisy speech signal. In [152], the pitch is quantized into discrete values (including an unvoiced state) and a separate GMM is trained to represent the log power spectrum for each pitch possibility. This is then used in a factorial Hidden Markov Model (HMM) to track the pitch of one or more sources. It was found that the use of speaker-dependent or gender-dependent models improved the tracking performance of multiple speakers significantly. In [44], the instantaneous frequency of each STFT bin is extracted and a statistical model for each harmonic of a source is defined. The Estimation-Maximization (EM) algorithm [27] is used to find the ML estimate of the pitches present in each frame and a multiple agent approach is then used to track the pitch of multiple sources. In [95] the power spectrum of each harmonic is modelled as a Gaussian distribution while the noise spectrum is similarly modelled as a sum of overlapping Gaussians on a uniform grid. The time-evolution of each harmonic amplitude is represented as a sum of overlapping Gaussians while that of the pitch as a cubic spline. The EM algorithm is again used to determine the ML model parameters and the method yields parametric models not only of the voiced speech but also of the smoothed noise spectrum. The advantages of the parametric approach to pitch estimation are that the assumptions about the signal are explicit, the limitations of an algorithm are often predictable, the performance can be optimal in a well defined sense and in some cases a Cramér-Rao Lower Bound (CRLB) can be calculated or estimated [21]. The disadvantage of the approach is that the performance may degrade when the, often quite strong, modelling assumptions are not satisfied.

3.1.2 Non-parametric pitch estimators

Non-parametric algorithms avoid using explicit signal models and identify the pitch of a signal either from its harmonic structure in the frequency domain, its periodicity in the time domain or from the periodicity of individual frequency bins in the time-frequency domain. Two widely used pitch estimation algorithms that operate in

the time domain are RAPT [137] and YIN [26]. RAPT calculates the normalized Autocorrelation Function (ACF) and selects its peaks as pitch period candidates. Dynamic Programming (DP) is then used to identify the voiced frames and to select the best sequence of pitch candidates. The YIN algorithm uses the squared difference function, closely related to the ACF, to identify pitch candidates. Neighbouring candidates within a short time interval are taken into account to select the best local estimate. YIN does not perform voiced/unvoiced classification and provides a pitch estimate for each frame using quadratic interpolation to obtain subsample resolution. Instead of the ACF, the cross correlation of two adjacent single-period waveform segments is used by [111] and [30]; this gives better time resolution at high pitch frequencies. Autocorrelation-based pitch detectors perform well in moderate noise levels since the ACF of an aperiodic noise source typically falls off rapidly with lag. At negative SNRs, however, a voiced speech signal whose energy is dominated by low-order harmonics will not generate a distinct peak in the ACF and, as will be seen in Sec. 3.4.2, reliable pitch estimation becomes impossible.

Instead of taking the ACF of the full-band signal, [123] uses an auditory filterbank to divide the signal into subbands. In each low frequency band the ACF is calculated directly while in the high frequency bands, which normally include multiple harmonics, the ACF is taken of the signal envelope. The advantage of this multiband approach is that subbands that are dominated by noise or that lack a reliable ACF peak can be deleted before the subband ACFs are combined to give an overall pitch estimate. This idea has been extended in [153] and later in [76] where multiple pitch candidates are obtained from each frame and a tracking algorithm based on an HMM is used to find the optimal sequence of zero, one or two sources thereby implicitly performing voiced/voiceless discrimination.

Non-parametric algorithms operating in the frequency domain typically identify harmonic peaks in the short-time amplitude, log-amplitude or power spectrum. The width of each peak depends on the window used in the spectral analysis, the harmonic number and the rate of change of pitch. The idea of creating a subharmonic histogram by assuming each peak in the spectrum to be a potential pitch harmonic

was introduced in [126] and later extended in [119]. Because the harmonic number is unknown, multiple possibilities are considered for each peak. The "harmonic sum" and "harmonic product" spectra generalize this idea by summing versions of the power or log power spectrum that have been compressed in frequency by a sequence of integer factors [126, 118]. In both cases, the peak of the resulting sum defines the pitch estimate. It was found in [118] that the harmonic sum and product spectra had similar performance for pitch detection and that both outperformed a parametric ML method, which was prone to octave errors in the presence of noise. In [106], instead of identifying isolated peaks, comb-filters corresponding to different fundamental frequencies are applied to the power spectrum of the speech to calculate a weighted sum of the harmonic powers. The highest peak at the output is achieved when the fundamental frequency of the comb-filter matches the pitch. If the spectrum is transformed into the log-frequency domain, the spacing of the comb filter tines becomes non-uniform but does not now depend on the pitch; this allows a more efficient implementation. A harmonic-summation method in the log-frequency domain is proposed in [49], in which the spectrum is shifted along the log-frequency axis, weighted and summed. Following the pitch estimation, frames are classified as voiced or unvoiced based on the correlation coefficient between adjacent pitch periods. In a similar approach, [17] convolves the spectrum in the log-frequency domain with a train of harmonically spaced delta functions and selects the highest peak. Three harmonic summing algorithms for multipitch estimation were described in [88] for music signals; these were later extended in [89] to use an auditory front end which gave a small improvement in some cases. An advantage of harmonic summation methods is that since most of the energy of a voiced speech signal is normally concentrated into a small number of harmonic peaks, these remain detectable even at poor SNRs.

3.1.3 Temporal continuity constraints

It is worth noting that the task of estimating pitch is inherently ill-conditioned; the pitch, f_0, of a periodic signal will, for example, be halved by the addition of an arbitrarily small component at $1.5f_0$. Because of this, all pitch estimation algorithms

are inevitably prone to errors in which the true pitch is multiplied or divided by two (octave errors) or, more generally, by any simple rational number. Therefore, a large number of pitch tracking algorithms apply temporal continuity constraints to the pitch estimate which can be effective at suppressing octave errors. Many algorithms divide the input signal into frames and identify multiple pitch candidates in each frame, often associating a measure of confidence or likelihood with each candidate. By defining the probabilities of inter-frame pitch transitions and of voicing onsets and offsets, it is possible to use DP to determine one or more maximum likelihood pitch tracks within the framework of an HMM. The use of DP for pitch tracking was introduced in [7] and extended in [116] and [117], which incorporated a pitch transition cost (equivalent to negative log likelihood) proportional to the absolute time derivative of pitch. Instead, [137] used a cost proportional to the derivative of log pitch and also applied a reduced cost to octave jumps. A complication is that, particularly at the end of voiced segments, the true pitch of speech may become irregular, make abrupt octave jumps or show bicyclic behaviour in which odd and even larynx cycles have different periods [29]. Although DP can compensate for pitch estimation errors at the frame level, the use of a strong continuity constraint may itself introduce errors and is no substitute for high accuracy in the raw pitch estimation.

3.1.4 Overview of PEFAC

In this chapter, we present PEFAC (Pitch Estimation Filter with Amplitude Compression[2]), a non-parametric frequency domain algorithm for single pitch estimation that is robust to high levels of noise. Our algorithm estimates the fundamental frequency of each frame by convolving its power spectrum in the log-frequency domain with a filter that sums the energy of the pitch harmonics. Unlike previous harmonic summing algorithms, the filter impulse response is designed to integrate the broadened harmonic peaks while rejecting additive noise that has a smoothly varying power spectrum. This improves the SNR of the filter output and contributes significantly to the noise-robustness of the algorithm. Prior to this filtering operation, a

[2]The MATLAB code of the proposed algorithm is available in [15].

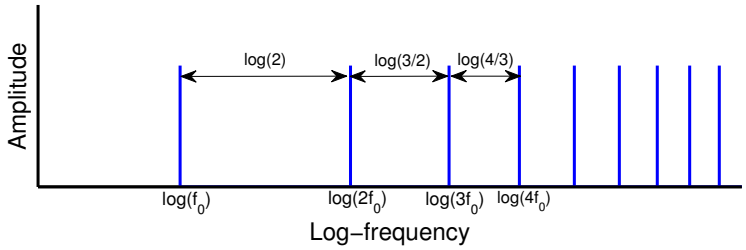


Figure 3.1: Power spectral density of a periodic source with pitch f_0 in the log-frequency domain.

novel spectral normalization is applied to reduce channel dependency and attenuate narrow-band. This normalization removes dependency on the input signal power and improves noise-robustness. The PEFAC algorithm provides a pitch estimate for each frame and, in addition, provides an estimate of voicing probability.

3.2 The PEFAC algorithm

For a periodic source with pitch f_0 in stationary noise, the power spectral density in the log-frequency domain is given by

$$Y(q) = \sum_{k=1}^{K} b_k \delta(q - \log k - \log f_0) + N(q) \tag{3.1}$$

where $q = \log f$. In (3.1), b_k represents the power of the k^{th} harmonic, $N(q)$ the power spectral density of the unwanted noise, δ the Dirac delta function and K the number of harmonics. As shown in Fig. 3.1, the spacing of the harmonics in the log-frequency domain does not depend on f_0 and their energy can therefore be summed by convolving $Y(q)$ with a matched filter [139] whose reversed impulse response is

$$h_i(q) = \sum_{k=1}^{K} \delta(q - \log k). \tag{3.2}$$

The convolution $Y(q) * h_i(-q)$ will result in a peak at $q_0 = \log f_0$ together with additional peaks corresponding to simple rational multiples and sub-multiples of f_0. In principle therefore, the pitch, f_0, can be found by taking the highest peak in the output of the filter.

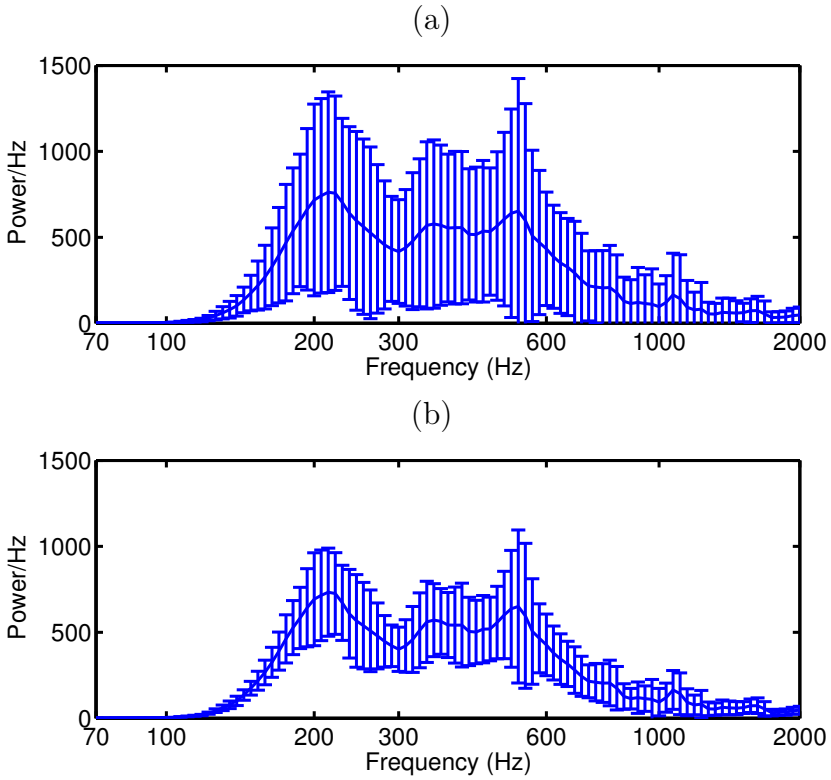


Figure 3.2: Alternative methods of computing the average power spectrum of a 120 s speech file. (a) Mean and standard deviation of the averaged speech spectrum over intervals of 3 s, (b) mean and standard deviation of the averaged speech spectrum over intervals of 3 s smoothed over 0.15 octaves in the log-frequency domain.

In practice, both speech and noise are non-stationary and we process the noisy signal in overlapping frames. The idealized filter defined by (3.2) is now unsuitable for pitch estimation because the spectral peaks are broadened and the filter output is adversely affected by additive noise and the channel response. In the PEFAC algorithm, described below, the approach outlined above is developed into a robust pitch estimation algorithm.

3.2.1 Normalization

The first stage of the algorithm performs spectral normalization. The motivation for this is that if the shape of the average power spectrum of clean speech is known a priori, deviations from this shape indicate either a non-uniform channel response or the presence of noise.

As discussed in Section 1.1.1, it was found in [19] that the LTASS of speech

signals is largely independent of both language and talker and can be represented by a universal LTASS, relative to which the frequency-averaged standard deviation of an individual speaker's LTASS was found to be about 3 dB. Accordingly, we use the universal gender-independent LTASS recommended in Table II of [19] as the expected spectral shape of the clean speech power spectrum, and we denote it by $L(q)$. The LTASS of an individual speaker was determined in [19] by averaging over 64 s of speech. However, for pitch estimation applications, it is desirable to average the noisy periodogram over a shorter interval to adapt, for instance, to different speaker levels contained within the same recording. In Sec. 3.3, our experiments use an interval of about 3 s. To compensate for using a short time interval, the smoothed periodogram, $\check{Y}_t(q)$, is calculated by averaging in both the time and the log-frequency domains. Figure 3.2 illustrates how smoothing both in time and in log-frequency can compensate for using a shorter speech interval. Figure 3.2(a) shows the mean and standard deviation of the average speech spectrum of a speaker calculated over intervals of 3 s and Fig. 3.2(b) shows the speech spectrum of the same speaker averaged both over intervals of 3 s and over 0.15 octaves in the log-frequency domain. We observe that the standard deviation is much lower in Fig. 3.2(b) than that in Fig. 3.2(a), showing that the individual estimates are closer to the average speech spectrum.

In the spectral normalization stage, the periodogram of the observed signal at time frame t, $Y_t(q)$, is first smoothed in both time and frequency to give

$$\check{Y}_t(q) = g(t,q) * Y_t(q) \tag{3.3}$$

where $g(t,q)$ is the two-dimensional impulse response of the moving average filter. The normalized periodogram, $Y'_t(q)$, is then obtained as

$$Y'_t(q) = Y_t(q)\frac{L(q)}{\check{Y}_t(q)} \tag{3.4}$$

where $L(q)$ represents the universal LTASS spectrum from Table II of [19]. We can

write

$$\begin{aligned} g(t,q) * Y_t'(q) &= g(t,q) * \left(Y_t(q) \frac{L(q)}{\check{Y}_t(q)} \right) \\ &\approx (g(t,q) * Y_t(q)) \frac{L(q)}{\check{Y}_t(q)} = L(q) \end{aligned} \tag{3.5}$$

where the approximation assumes that both $L(q)$ and $\check{Y}_t(q)$ are sufficiently smooth that they do not change significantly within the support of $g(t,q)$.

From (3.5) we see that, following normalization, the smoothed periodogram of the observed signal will match the universal LTASS; this provides three benefits. First, in the case of a flat channel with no added noise, the procedure will normalize the power of the input signal to that of the $L(q)$ target but will otherwise have little effect since the spectral shape of the $L(q)$ target matches the average spectral shape of clean speech. Second, any time-invariant channel response applied to the noisy speech will affect $Y_t(q)$ and $\check{Y}_t(q)$ equally providing it is sufficiently smooth that it

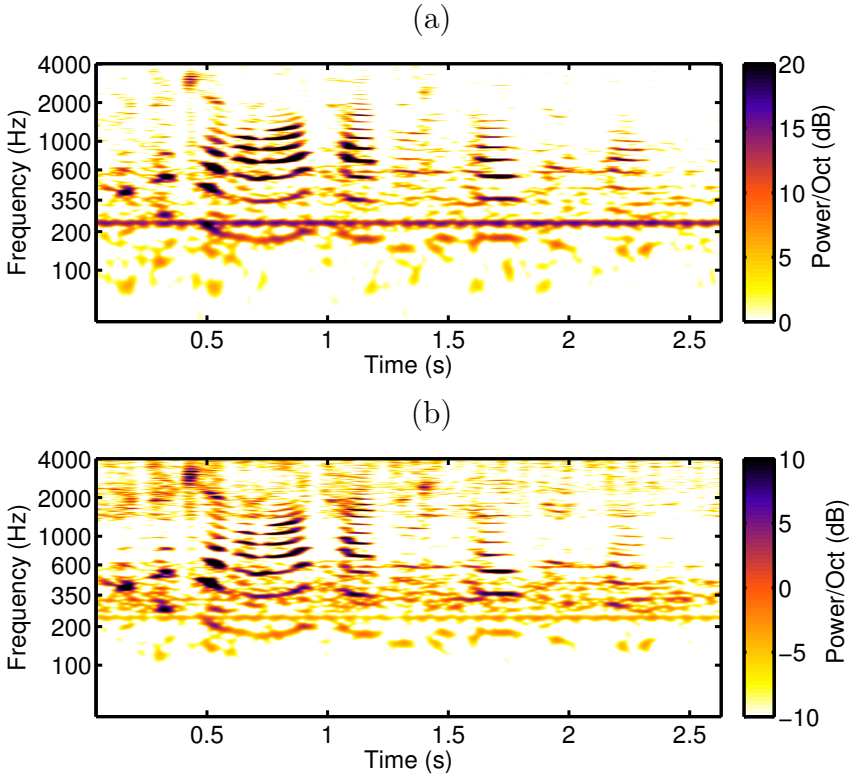


Figure 3.3: Periodogram of speech corrupted by narrow-band noise at 5 dB SNR before (a) and after (b) normalization.

does not change significantly within the support of $g(t,q)$. The normalization will therefore cancel out the effects of such a channel; this action is similar to that of the widely used technique of cepstral mean subtraction [35]. Third, the normalization will attenuate any additive noise components that are strong enough to distort the average spectrum. To show this, suppose that the noisy signal speech periodogram is $Y_t(q) = S_t(q) + N_t(q)$ where the speech and noise periodograms, $S_t(q)$ and $N_t(q)$ respectively, are assumed to add in the power domain. Following the smoothing operation, we have

$$\begin{aligned} \check{Y}_t(q) &= g(t,q) * S_t(q) + g(t,q) * N_t(q) \\ &\approx \alpha L(q) + \check{N}_t(q) \end{aligned} \tag{3.6}$$

where we have approximated the smoothed periodogram of the speech by $\alpha L(q)$ with α representing the power of the input speech relative to LTASS. The normalization gain factor in (3.4) can now be determined as

$$\frac{L(q)}{\check{Y}_t(q)} = \frac{\alpha^{-1}}{1 + \frac{\check{N}_t(q)}{\alpha L(q)}} \tag{3.7}$$

From (3.7) we see that the gain factor is a function of the SNR, $\alpha L(q)/\check{N}_t(q)$, of the smoothed noisy speech periodogram, $\check{Y}_t(q)$. At frequencies for which this SNR $\gg 1$, the gain factor approximates to α^{-1} thereby normalizing the input speech power. At frequencies having a low SNR, however, the gain will be less than α^{-1} and any regions of the periodogram for which the SNR $\ll 1$ will be heavily attenuated. To illustrate this effect, Fig. 3.3(a) shows the periodogram of a speech signal corrupted with noise that has a strong tonal component at 250 Hz while Fig. 3.3(b) shows the periodogram of the same speech segment after normalization. We see that the narrow-band noise is highly attenuated while the speech spectrum is slightly amplified at other frequencies.

To assess the effect of the normalization on the overall SNR of the signal, we assume that the average noisy speech spectrum is $\check{Y}(q) = L(q) + N(q)$ where speech is assumed to follow the universal LTASS spectrum, and $N(q)$ is the noise spectrum,

which is supposed not to change significantly within the support of $g(t,q)$. Writing L_q for $L(q)$ and N_q for $N(q)$, the original SNR is therefore $\int_q L_q dq / \int_q$. We now scale the noisy speech spectrum by $L_q/(L_q+N_q)$ which forces noisy spectrum to be standard LTASS so that the new SNR is

$$\frac{\int_q \frac{L_q^2}{L_q+N_q} dq}{\int_q \frac{L_q N_q}{L_q+N_q} dq}$$

The SNR has been improved by the normalization if

$$\begin{aligned}
&\frac{\int_q \frac{L_q^2}{L_q+N_q} dq}{\int_q \frac{L_q N_q}{L_q+N_q} dq} \geq \frac{\int_p L_p dp}{\int_p N_p dp} \\
\Leftrightarrow \quad &\int_q \frac{L_q^2}{L_q+N_q} dq \int_p N_p dp \geq \int_q \frac{L_q N_q}{L_q+N_q} dq \int_p L_p dp \\
\Leftrightarrow \quad &\iint_{q,p} \frac{L_q \left(L_q N_p - L_p N_q\right)}{L_q+N_q} dpdq \geq 0
\end{aligned} \tag{3.8}$$

where, for clarity, the frequency has been represented as a subscript. We can decompose the left side of the inequality

$$\begin{aligned}
&\iint_{q,p} \frac{L_q \left(L_q N_p - L_p N_q\right)}{L_q+N_q} dpdq = \\
&= \frac{1}{2} \iint_{q,p} \frac{L_q \left(L_q N_p - L_p N_q\right)}{L_q+N_q} dpdq + \frac{1}{2} \iint_{p,q} \frac{L_p \left(L_p N_q - L_q N_p\right)}{L_p+N_p} dqdp \\
&= \frac{1}{2} \iint_{q,p} \frac{\left(L_q N_p - L_p N_q\right)^2}{\left(L_q+N_q\right)\left(L_p+N_p\right)} dpdq \geq 0
\end{aligned}$$

since the integrand is always non-negative. For an LTASS speech signal, the normalization, therefore, will always improve the overall SNR unless $L(q)N(p) - L(q)N(p) = 0 \forall p,q$ that is if $L(q)/N(q)$ has the same value for all q.

3.2.2 Filter definition

Although the idealized matched filter defined in (3.2) comprises a sequence of delta functions, the width of each harmonic peak will, in practice, be broadened due to the analysis window and to the rate of change of f_0. Accordingly we use a filter with broadened peaks defined by

$$h_p(q) = \frac{1}{\gamma - \cos\left(2\pi e^q\right)} - \beta \tag{3.9}$$

for $\log(0.5) < q < \log(K + 0.5)$ and $h_p(q) = 0$ otherwise. The algorithm parameter γ controls the peak width while β is chosen so that $\int h_p(q)dq = 0$. The number of peaks, K, is discussed in Section 3.3; it needs to be large enough to include all harmonics with significant energy while avoiding a high response of $Y_t(q) * h_p(-q)$ at values of q corresponding to subharmonics of f_0. Figure 3.4(a) shows $h_p(q)$ for $\gamma = 1.8$ and $K = 10$. The Fourier transform of the filter is shown in Fig. 3.4(b), where we observe that, since $h_p(q)$ is chosen to have zero mean, the filter has a zero gain at DC. Moreover, the normalized gain equals -6 dB at 0.39 cycles per octave, meaning that

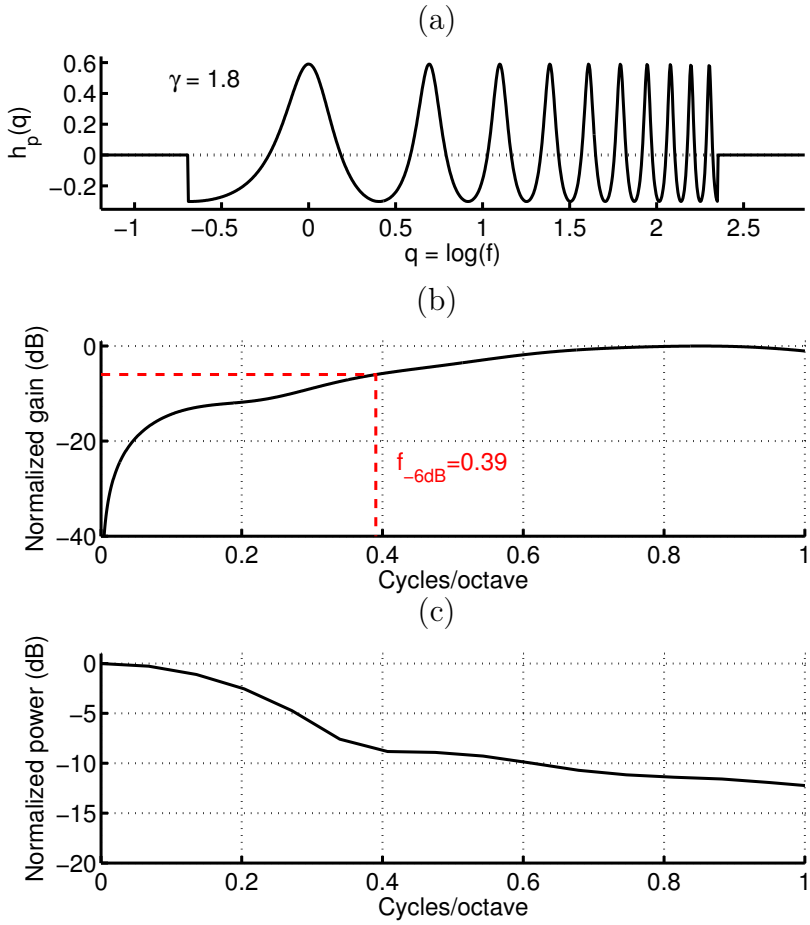
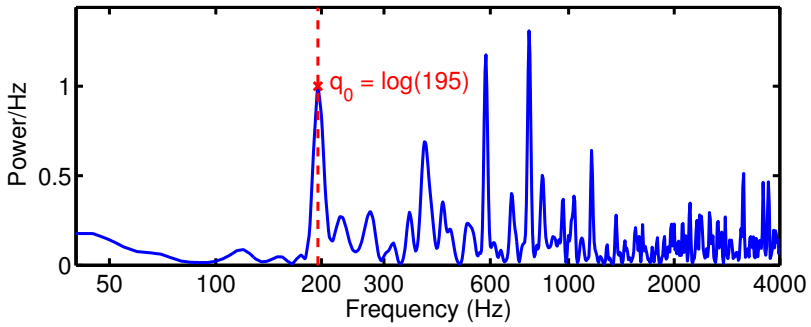


Figure 3.4: (a) The function $h_p(q)$ defined in (3.9), (b) its Fourier transform for $\gamma = 1.8$ and $K = 10$, and (c) the Fourier transform of the noise periodogram, $N(q)$ averaged over all noises in the RSG-10 database [131].

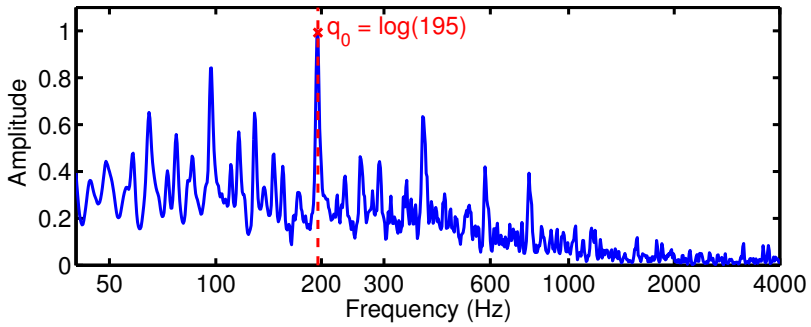
any noise whose power spectrum varies with log-frequency at a slower rate than this will be highly attenuated. Figure. 3.4(c) illustrates the distribution of energy versus cycles per octave averaged over all noises in the RSG-10 database [131], where it can be seen that the noise power is concentrated in the region below the 0.39 cycles per octave cutoff of the filter.

Fig. 3.5 shows an example that illustrates the properties of the filters defined by (3.2) and (3.9) when used for pitch estimation. Figure 3.5(a) shows the periodogram of a noisy voiced speech frame with a pitch of $f_0 = 195\,\text{Hz}$. The output of the idealized filter, defined by (3.2), is shown in Fig. 3.5(b) and it can be seen that although the highest peak is at the correct pitch, there are many additional peaks at both harmonically related and unrelated frequencies. The output from the proposed filter, defined by (3.9), is shown in Fig. 3.5(c), where we see that it almost entirely suppresses both the peaks due to noise and the peaks at integer multiples of f_0. The suppression of noise peaks is due to the high attenuation by the filter of noise components whose power spectrum varies more slowly than 0.39 cycles per octave, as seen in Fig. 3.4(b). The suppression of peaks at integer multiples of f_0 occurs because, at these frequencies, some of the pitch harmonics will be aligned with the negative regions of the filter impulse response, $h_p(-q)$, and will therefore contribute negatively to the output. As an example, when generating the output of the filter at $2f_0$, the odd harmonics of f_0 will be aligned with negative regions of the impulse response and so will partially cancel the contribution of the even harmonics which will be aligned with positive regions of the impulse response. Peaks at sub-harmonics of f_0 remain in Fig. 3.5(c) but have been attenuated; thus the relative amplitude of the peak at 98 Hz, the first subharmonic, has been reduced from 0.84 in Fig. 3.5(b) to 0.63 in Fig. 3.5(c). At any given subharmonic, f_0/n, both filters will include the energy from only the first K/n harmonics of f_0. The resultant peak will be lower than that at f_0 partly because fewer harmonics are included but also, in the case of the filter from (3.9), because the positive area associated with each harmonic in Fig. 3.4(a) is inversely proportional to the harmonic number. The comparative performance of the two filters on a large number of speech utterances is discussed in Section 3.4.2.

(a) Noisy voiced frame spectrum, $Y_t(q)$



(b) Output of filter (3.2)



(c) Output of filter (3.9)

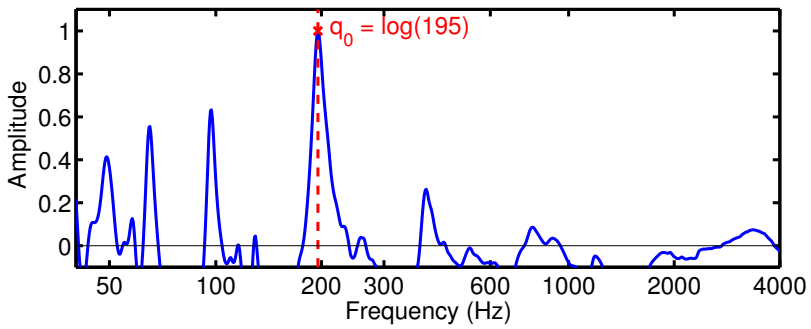


Figure 3.5: (a) The periodogram of a voiced frame corrupted with white noise at −8 dB SNR. The voiced frame, taken from the TIMIT database, contains the first vowel of 'unstuck' and has a fundamental frequency of 195 Hz. The output of the idealized filter (3.2) and the proposed filter (3.9) are shown in (b) and (c) respectively.

The PEFAC algorithm convolves the normalized periodogram, $Y_t'(q)$, with $h_p(-q)$ to give

$$Z_t(q) = Y_t'(q) * h_p(-q) \tag{3.10}$$

As noted above, $Z_t(q)$ will contain peaks corresponding to f_0 and to simple rational multiples and submultiples of f_0. We define $f_{t,n}$ and $a_{t,n}$ respectively as the frequency

and amplitude of the n^{th} highest peak of $Z_t(q)$. The frequencies, $f_{t,n}$, and peak amplitudes, $a_{t,n}$, are used below in Section 3.2.3 to estimate the voicing probability and in Section 3.2.4 to estimate pitch. If no temporal constraints are applied, the pitch estimation at each time frame t is taken as $f_{t,1}$.

3.2.3 Voiced speech probability

Estimation of the pitch is only meaningful in voiced speech segments, but identifying these reliably in the presence of high levels of noise is a challenging problem. Therefore, we have chosen to give separately an estimate of the fundamental frequency at every time-frame together with an estimated probability that the time-frame contains voiced speech.

This voicing probability is based on a 2-element feature vector calculated at each frame and comprising:

(a) the log-mean power of the normalized time-frame spectrum, $L_t = \log E_t$ such that $E_t = \left(\frac{1}{Q}\sum_{i=1}^{Q} Y_t'(q_i)\right)$, where Q represents the number of frequency bins in the log-frequency domain. Because voiced speech contains most speech energy, the mean power of a voiced frame is typically higher than the power of an unvoiced frame;

(b) the ratio of the sum of the highest three peaks in $Z_t(q)$ to E_t

$$r_t = \frac{\sum_{n=1}^{3} a_{t,n}}{E_t + \epsilon} \tag{3.11}$$

where ϵ is a small regularization constant. This ratio depends on the fraction of the frame's total power that is harmonically related. The highest three peaks, rather than only the highest one, are used in the numerator of (3.11) to give greater robustness to noise; a voiced frame will include several high peaks at f_0 and its sub-harmonics (see Fig. 3.5(c)).

Fig. 3.6 shows the histograms of the joint distribution of L_t and r_t for both unvoiced and voiced frames. We can observe that unvoiced speech frames typically have

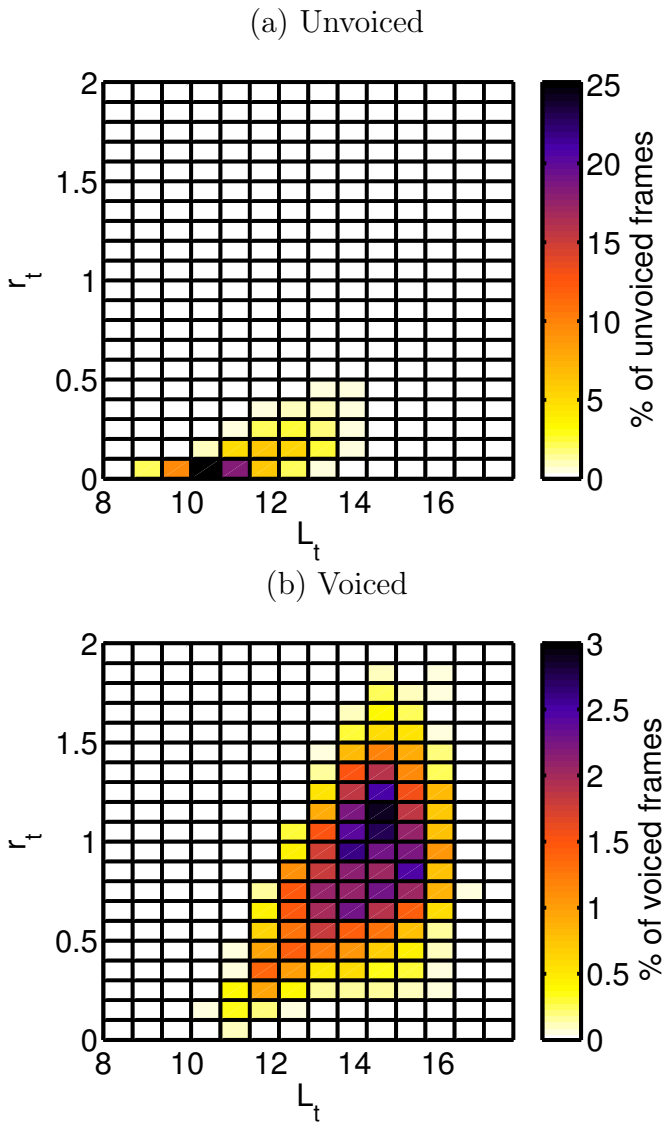


Figure 3.6: Histogram of the joint distribution of L_t and r_t for (a) unvoiced and (b) voiced frames. The frames, a total of 16832, are extracted from a subset of utterances of the TIMIT database training set mixed with white noise at +20 dB SNR.

lower L_t values than voiced speech frames and that r_t values are consistently low for unvoiced frames and higher and more variable for voiced frames.

Two GMMs are trained; one for voiced frames and the other for unvoiced frames. The input to both GMMs is the 2-element feature vector, $[L_t, r_t]$ and the voiced speech probability for each frame is calculated from their likelihood ratio, $P_t(voiced) = 1/(1 + p_{t,u}/p_{t,v})$, where $p_{t,u}$ and $p_{t,v}$ are the output probabilities at time-frame t from the unvoiced and voiced GMM respectively.

3.2.4 Temporal continuity constraints

In PEFAC up to three pitch candidates are identified in each frame and dynamic programming is used to select the sequence of pitch candidates that minimizes a cost function expressed as a weighted sum of three parameters:

(a) the relative amplitude of the peaks. The amplitude of the peaks, $a_{t,n}$, indicates the amount of harmonically related energy associated with the frequency of the spectral peak at $f_{t,n}$. We penalize the selection of lower amplitude peaks by including in the dynamic programming a cost term equal to $c_{t,n}^{(a)} = -\frac{a_{t,n}}{a_{t,1}}$.

(b) the rate of change of the fundamental frequency. To penalize rapid changes of fundamental frequency, we calculate the normalized rate of pitch change as

$$\Delta f_{t,nm} = \frac{2(f_{t,n} - f_{t-1,m})}{\Delta t(f_{t,n} + f_{t-1,m})}$$

where $f_{t,n}$ and $f_{t-1,m}$ are pitch candidates in frames t and $t-1$ respectively and Δt is the frame time increment. We introduce a cost term proportional to the squared deviation of $\Delta f_{t,nm}$ from its mean value determined from training data: $c_{t,nm}^{(f)} = (\Delta f_{t,nm} - \mu_{\Delta f})^2$.

(c) the deviation from the median pitch. Although this value is unknown, the median pitch at time t, $\tilde{f}_{t,0}$, can be estimated as the median frequency of the highest peak, $f_{t,1}$, in nearby frames that have a high voiced speech probability. The cost related to this measure, which provides robustness to outlier errors, is $c_{t,n}^{(m)} = \frac{|f_{t,n} - \tilde{f}_{t,0}|}{\tilde{f}_{t,0}}$.

The overall cost from candidate m in time-frame $t-1$ to candidate n in time-frame t can therefore be expressed as

$$c_{t,nm} = w_1 \cdot c_{t,n}^{(a)} + w_2 \cdot \min(c_{t,nm}^{(f)}, w_3) + w_4 \cdot c_{t,n}^{(m)} \tag{3.12}$$

where w_i are the weights associated with each parameter, with the exception of w_3, which acts as an upper limit for $c_{t,nm}^{(f)}$ to permit pitch changes between voicing spurts.

3.2.5 Fundamental frequency estimation

The complete PEFAC algorithm therefore comprises the following steps:

(i) transform the input signal to the time-frequency power spectrum domain, $Y_t(f)$, using the short-time Fourier transform (STFT);

(ii) interpolate the periodogram of each frame onto a log-spaced frequency grid, $Y_t(q)$;

(iii) calculate the normalized periodogram, $Y'_t(q)$ so that the smoothed spectrum $\check{Y}_t(q)$ equals $L(q)$,

$$Y'_t(q) = Y_t(q)\frac{L(q)}{\check{Y}_t(q)};$$

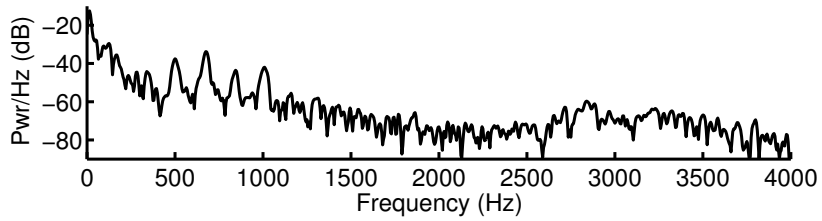
(iv) calculate $Z_t(q) = Y'_t(q) * h(-q)$ and select as pitch candidates the three highest peaks in the feasible range;

(v) estimate the voiced probability for each frame;

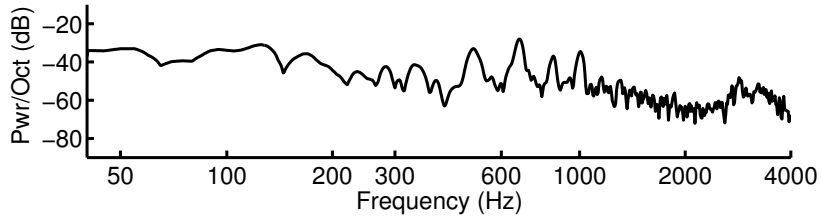
(vi) use dynamic programming to select the sequence of candidates with lowest cost.

Fig. 3.7 shows the output of the various algorithm steps for a frame of voiced speech with a pitch of 168 Hz corrupted by car noise. In Fig. 3.7(a) we see that the noise masks the first two pitch harmonics although harmonics 3 to 7 are visible as peaks. Figure 3.7(b) shows the same periodogram interpolated onto a logarithmic scale and restricted to the range 40 Hz to 4 kHz. The low frequency noise that masks the pitch in Fig. 3.7(a,b) has been greatly attenuated by the normalization stage in Fig. 3.7(c), which shows the normalized periodogram $Y'_t(q)$, while the peaks at harmonics 3 to 7 have been preserved. The dashed line shows the LTASS normalization target spectrum. Figure 3.7(d), which illustrates the output of the filter, shows a clear peak at 168 Hz despite its absence in Fig. 3.7(a), the original spectrum. Figure 3.7(d) highlights the three highest peaks of $Z_t(q)$, which in this case are harmonically related. They correspond to the pitch, the first subharmonic and the second harmonic respectively.

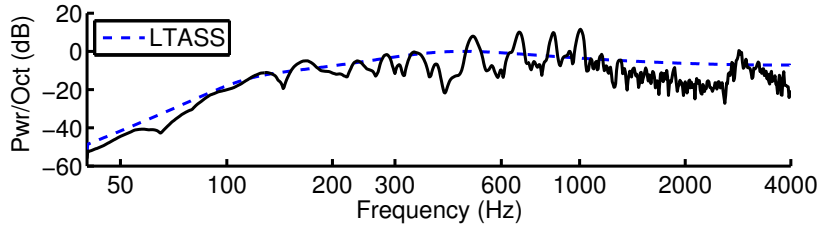
(a) $Y_t(f)$, output of step (i)



(b) $Y_t(q)$, output of step (ii)



(c) $Y_t'(q)$, output of step (iii)



(d) $Z_t(q)$, output of step (iv)

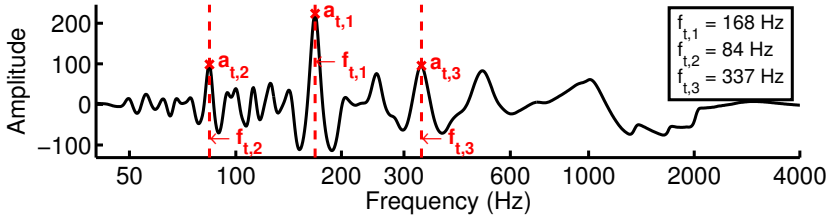


Figure 3.7: PEFAC processing steps for a single voiced frame of speech corrupted with car noise at −19 dB SNR. (a) Periodogram in dB, (b) periodogram in dB on a log-frequency grid, (c) normalized periodogram in dB on a log-frequency grid, and (d) output of the pitch extraction filter. The voiced frame, taken from the TIMIT database, contains the second vowel of 'himself' and has a fundamental frequency of 168 Hz.

3.3 Experiments

A subset of the training set from the TIMIT database [38] was used for training PEFAC. The training subset contains 16 male and 8 female speakers each reading 3 distinct sentences. The sampling frequency of the speech material is 16 kHz.

To determine the ground truth for the fundamental frequency and voicing, the autocorrelation method, the cross-correlation method and the sub-harmonic summation from Praat [10] together with the YIN [26] and RAPT [137, 15] algorithms were applied to the clean speech signals. A frame was identified as voiced if the majority of the algorithms gave the same pitch estimate and, in this case, the ground truth was taken as the mean of the estimates. The ground truth pitch track was superimposed on a spectrogram and for the small number of frames where there was visual disagreement (less than 4% of the total), the pitch was manually resolved.

For training, car, babble and white noise from the RSG-10 database [131] was added to the speech files to generate the noisy signals. The calculation of SNR used ITU-T P.56 [68, 15] for the speech level and unweighted power for the noise.

PEFAC includes a number of algorithm parameters whose values were determined empirically from the training data. The STFT uses a Hamming analysis window of 90 ms duration; this is long enough to resolve the pitch harmonics even for low values of f_0 but short enough to limit the pitch variation within a frame. The inter-frame time increment is 10 ms and each windowed input frame is zero-padded to 360 ms to aid the interpolation stage at low frequencies.

The spectrum of each frame is interpolated onto a dense logarithmic grid ranging from 10 Hz to 4 kHz with a frequency resolution of 0.58% corresponding to 120 samples per octave. Conceptually the sampled spectrum is first converted to a continuous spectrum using linear interpolation and this is then resampled using a variable width triangular sampling kernel as is used when forming mel-frequency cepstrum coefficients [25]. In practice the two stages are combined and the continuous spectrum is never calculated explicitly [15].

The smoothing filter used in (3.3) in the normalization step has a uniform impulse

response within its support, i.e. $g(t,q) = 1$ for $|t| < T_0$, $|q| < Q_0$. The support in the log-frequency, $2Q_0$, was chosen empirically as 0.15 octaves by maximizing the training set performance on a range of noise types. The support of the smoothing filter in frequency is a compromise between reducing the standard deviation of the smoothed spectrum and resolving narrow-band noise sources. Due to the short duration of the TIMIT and CSLU-VOICES utterances (typically of 3-5s duration), the averaging in the time axis is done over the entire utterance duration. The LTASS response used for $L(q)$ is derived from the tabulated values in Table II of reference [19] and are the average over 12 languages of many speakers. To obtain a continuous response, a 7th order IIR filter was fitted to the tabulated values [15].

Following normalization, a discrete convolution is performed between the sampled spectrum of each frame and the filter impulse response defined by (3.9). The filter impulse response is sampled onto the same dense grid as the spectrum and β is chosen to make its samples sum to zero. The optimum value of the parameter γ depends on the nature of the noise and the value 1.8 was chosen as the best compromise to maximize performance on the training set. The number of harmonics captured by the filter, K, was set to 10. Figure 3.8 shows the results of our algorithm on the training set for different values of K at different SNRs for white noise, where we observe that the number of harmonics captured by the filter is not critical above a threshold and that the algorithm performance has reached convergence at $K = 10$. When K is fixed

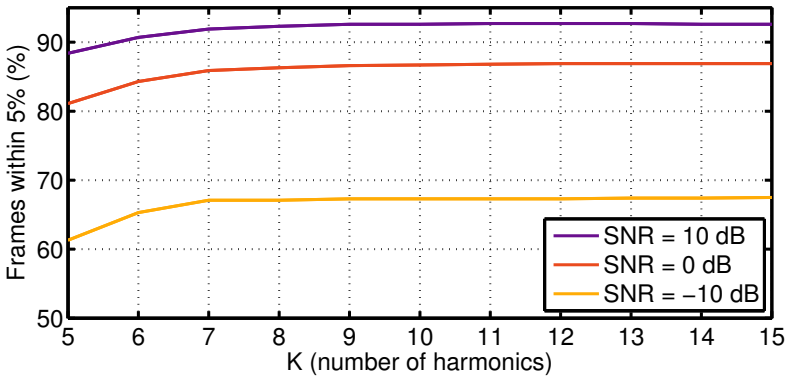


Figure 3.8: Variation of pitch estimation accuracy with the number of harmonics, K, for white noise at -10, 0 and $+10$ dB SNR on a subset of the training set of the TIMIT database.

at 10, the pitch is estimated within ±5% for 67.0% of voiced frames at −10 dB SNR. If, instead, K is chosen optimally for each frame, this figure would rise only to 69.7% indicating that even if K were chosen adaptively for each frame, the potential benefit is very small. At −10 dB SNR, for 94.7% of all voiced frames, the choice of K (in the range 10 to 15) has no effect on whether or not the estimated pitch of the frame is correct (< 5% error).

Two multivariate GMM models are trained on voiced and unvoiced frames respectively. Both GMMs use 6 mixtures with full-covariance matrices. The GMMs were trained with the subset of the TIMIT training set using the union of training data at various noisy conditions: adding white, car and babble noise at SNRs from −5 dB to +20 dB.

Table 3.1: Dynamic programming weights for equation (3.12)

w_1	w_2	w_3	w_4
1	0.019	0.007	0.825

Dynamic programming parameters are weighted to obtain the final cost, (3.12). These weights are calculated using discriminative training [78, 6]. For the training, three types of noises were used at an SNR range from −20 to +20 dB: white noise, car noise and babble noise. Table 3.1 shows the final weights used for the dynamic programming, where the dominant terms are the relative amplitude, w_1, and the deviation from the median pitch, w_4. The estimated median pitch at time t, $\tilde{f}_{t,0}$ in Section 3.2.4(c), was calculated as the median of $f_{t-\Delta t,1}$ of the frames for which $P_{t-\Delta t}(voiced) > 0.7$ where $0 < \Delta t < 2$ s.

3.4 Results

In this section, the performance of the proposed fundamental frequency estimator is evaluated on the core test set from the TIMIT database [38] and on the CSLU-VOICES corpus [79]. The TIMIT core test set contains 16 male and 8 female speakers each reading 8 sentences for a total of 192 sentences all with distinct texts. The CSLU-

VOICES corpus contains 7 male and 5 female speakers each reading 50 phonetically rich sentences, of which the 223 with manually verified and adjusted pitch marks were used for evaluation. Noise from the RSG-10 database [131] and from the ITU-T P.501 standard [69] was added to the speech utterances. Spectrograms of all the noise types used in training and testing are included in Appendix A.

3.4.1 Voiced speech activity detector

The performance of the voiced speech activity detector is illustrated in Fig. 3.9, where the Detection Error Trade-off (DET) curve [105] shows the miss probability versus the false alarm probability, from which a threshold for the classifier can be chosen according to different requirements. Figure 3.9 covers a likelihood ratio threshold ranging from 0.11 to 9. The circles in Fig. 3.9 indicate the results for a likelihood ratio threshold of unity. We note that at +20 dB SNR the performance of the algorithm is similar for all three noises. However, its performance degrades in a different way for each noise as the SNR is reduced. In the white noise case, shown in Fig. 3.9(a), for a likelihood ratio threshold of unity, the false alarm probability remains low even for negative SNRs, while the miss probability increases. Similar behaviour is obtained for car noise, Fig. 3.9(b), although the performance degradation when decreasing the SNR is less severe than for white noise. This is because the power spectrum of car noise is concentrated at low frequencies and speech harmonics at higher frequencies remain unmasked. However, the opposite behaviour is observed for babble noise, Fig. 3.9(c), where false alarm probability degrades rapidly with SNR. This behaviour is due to the background speech present in babble noise, which the algorithm identifies as voiced at low SNRs.

The voiced speech activity detector of PEFAC was compared to RAPT [137, 15] and Jin & Wang (J&W) [76]. Table 3.2 shows a performance comparison of PEFAC (using a likelihood ratio threshold of unity) with RAPT and J&W for white, car and babble noise. For each algorithm and noise type, the table shows the miss probability (P_{miss}) and the false alarm probability (P_{fa}) for SNRs in the range −20 dB to +20 dB. In each case, the algorithm with the lowest total error rate ($P_{miss} + P_{fa}$) has been

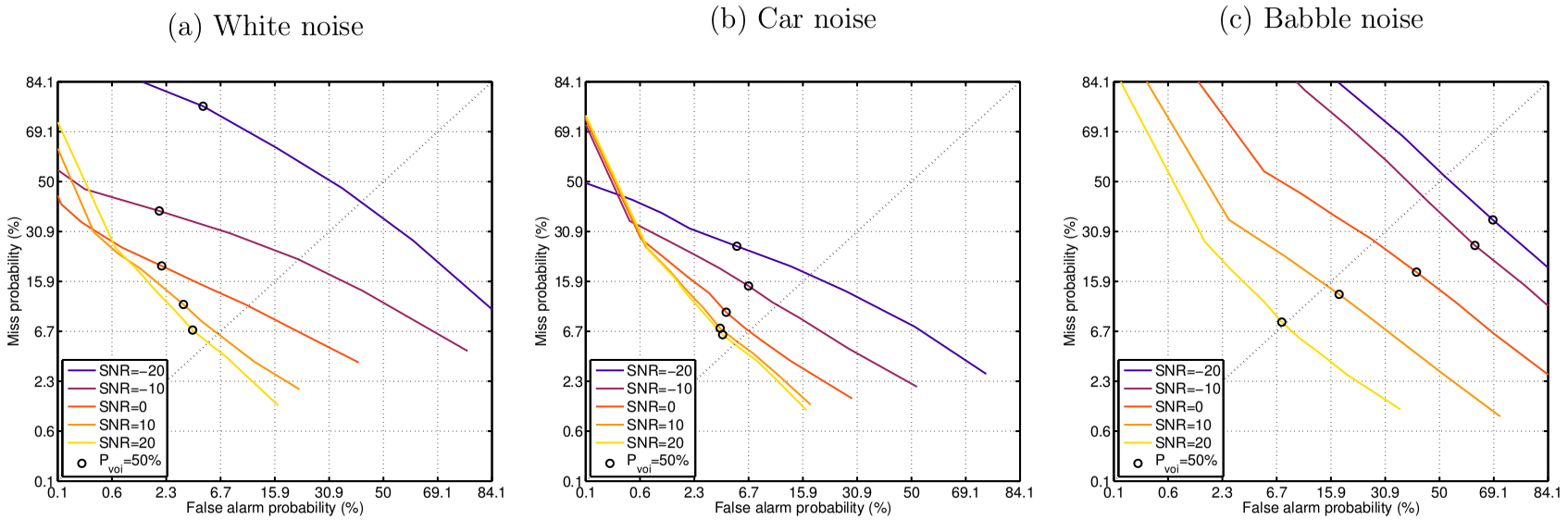


Figure 3.9: DET curve of the voiced activity detection algorithm obtained for −20 dB, −10 dB, 0 dB, +10 dB and +20 dB SNR for (a) white noise, (b) car noise, and (c) babble noise on the core test set of the TIMIT database. The circles indicate the results for a likelihood ratio threshold of unity.

Table 3.2: Voiced speech activity detection comparison

SNR		+20 dB		+10 dB		0 dB		-10 dB		-20 dB	
		$P_{miss}(\%)$	$P_{fa}(\%)$	$P_{miss}(\%)$	$P_{fa}(\%)$	$P_{miss}(\%)$	$P_{fa}(\%)$	$P_{miss}(\%)$	$P_{fa}(\%)$	$P_{miss}(\%)$	$P_{fa}(\%)$
White noise	PEFAC	6.85	4.45	10.92	3.68	19.94	2.47	**38.45**	**1.84**	**77.50**	**3.87**
	RAPT	**1.81**	**4.86**	14.96	1.81	59.60	0.06	99.88	0.00	100.00	0.00
	J&W	0.55	6.32	**1.56**	**4.72**	**12.08**	**5.50**	66.94	2.24	98.04	0.93
Car noise	PEFAC	6.27	4.59	7.10	4.07	9.53	5.06	14.79	9.03	25.88	9.32
	RAPT	**0.58**	**8.73**	4.04	6.54	23.55	5.23	70.19	4.82	96.74	4.69
	J&W	0.57	16.64	**0.89**	**7.40**	**2.53**	**3.35**	**8.81**	**2.41**	**29.05**	**2.19**
Babble noise	PEFAC	**7.96**	**8.80**	**12.93**	**18.93**	**18.24**	**41.19**	26.14	63.39	35.07	71.58
	RAPT	0.22	48.30	0.47	63.88	1.55	84.51	1.03	95.92	0.07	97.54
	J&W	0.94	38.81	2.96	56.03	8.47	66.47	**17.58**	**71.27**	**24.81**	**72.21**
Overall	PEFAC	**7.03**	**5.95**	**10.31**	**8.89**	**15.90**	**16.24**	**26.46**	**24.75**	**46.15**	**28.26**
	RAPT	0.87	20.63	6.49	24.08	28.23	29.93	57.03	33.58	65.60	34.08
	J&W	0.69	20.59	1.80	22.72	7.69	25.11	31.11	25.31	50.63	25.11

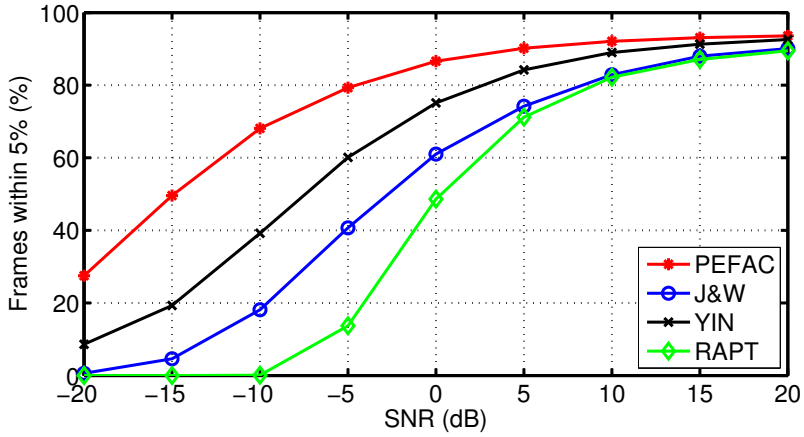
highlighted. The final row of the table shows the overall performance on all three noise types. The voiced speech activity detector of RAPT is very accurate at high SNRs, having the lowest total error rate at +20 dB SNR for white and car noise. However, its performance degrades rapidly and becomes poor at low SNRs. The accuracy of J&W at high SNRs depends heavily on the type of noise. J&W performs particularly well with car noise achieving the best performance at most SNRs. The PEFAC voiced speech activity detector is less dependent on noise type than the other algorithms and its overall error rate is consistently lower.

3.4.2 Pitch estimation

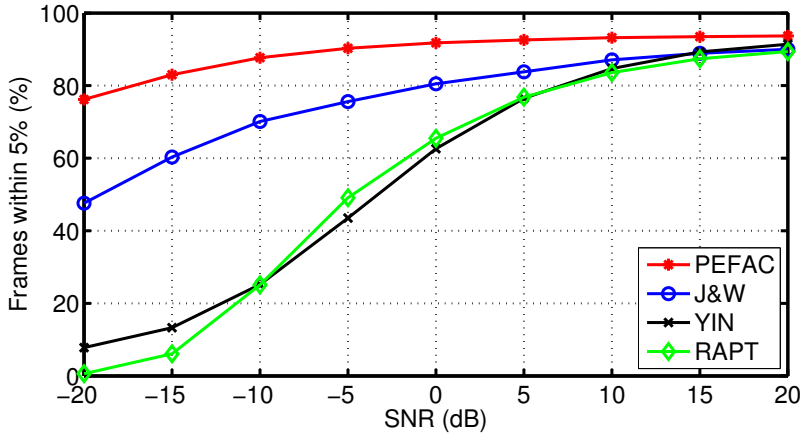
In this section, the performance of the proposed pitch estimator is evaluated. For performance comparison, RAPT [137, 15], YIN [26] and Jin & Wang (J&W) [76] were used. The first two of these are non-parametric time-domain algorithms while the third is a non-parametric algorithm operating in the time-frequency domain.

Evaluation of pitch estimation was restricted to voiced frames and a pitch estimate was classified as correct if it was within ±5% of the true value. The graphs in Fig. 3.10 show the performance of the algorithms for white (a), car (b) and babble (c) noise respectively on the core test set of the TIMIT database. The noises were taken from the RSG-10 database. It can be seen that at +20 dB SNR, all of the algorithms reach a performance plateau which varies slightly between algorithms. Although the two time-domain algorithms, YIN and RAPT, were not specifically designed for noise robustness, YIN in particular maintains its high performance in white noise down to 0 dB SNR. Below this level however, the performance of both algorithms degrades rapidly for all noise types. The J&W algorithm also degrades rapidly for white and babble noise, while having a robust performance to car noise. The proposed algorithm, PEFAC, has excellent performance at +20 dB SNR and retains this high performance at significantly lower SNR levels than the other algorithms. In addition to the TIMIT database, the algorithm was also evaluated on the CSLU-VOICES corpus [79]. Noises from the ITU-T P.501 standard [69] were added. The obtained performances are shown in Fig. 3.11 for cafeteria (a), metro (b) and street (c) noises. We observe that

(a) White noise



(b) Car noise



(c) Babble noise

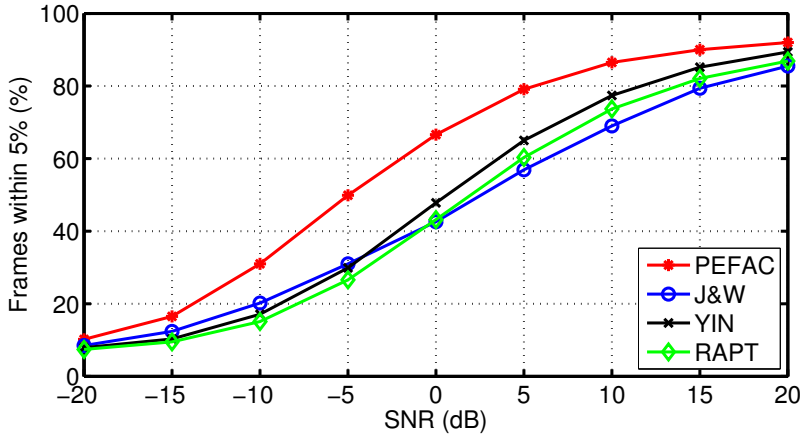
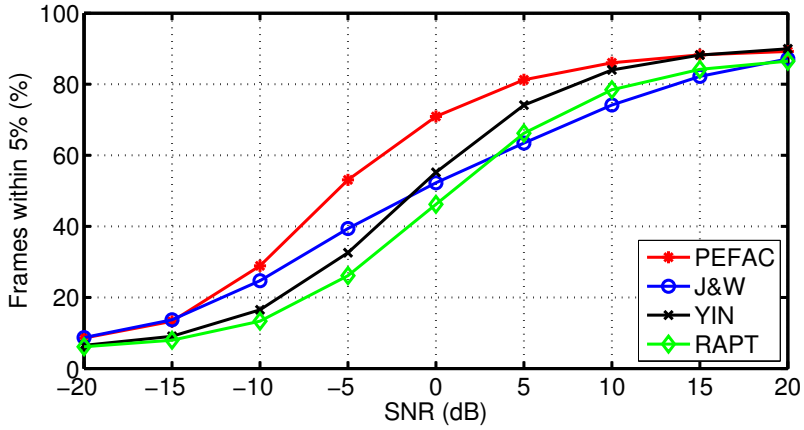
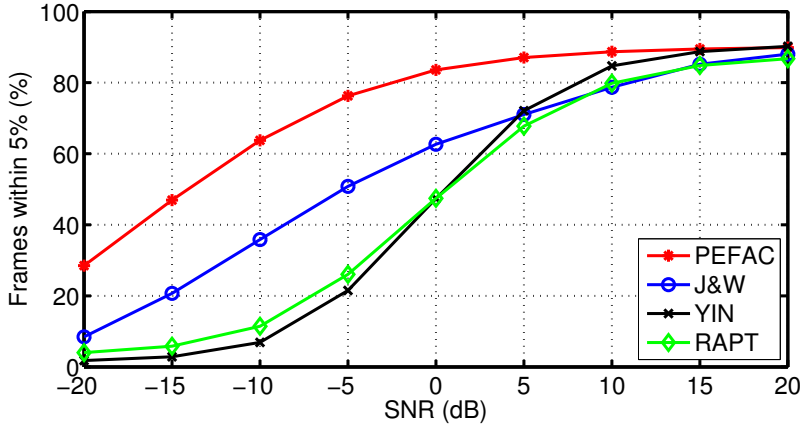


Figure 3.10: Variation of pitch estimation accuracy on the core test set of the TIMIT database with SNR for (a) white noise, (b) car noise, and (c) babble noise from the RSG-10 database [131]. The graphs show the percentage of correct frames (error below 5%) for each of the algorithms: PEFAC, J&W [76], YIN [26] and RAPT [137].

(a) Cafeteria noise



(b) Metro noise



(c) Street noise

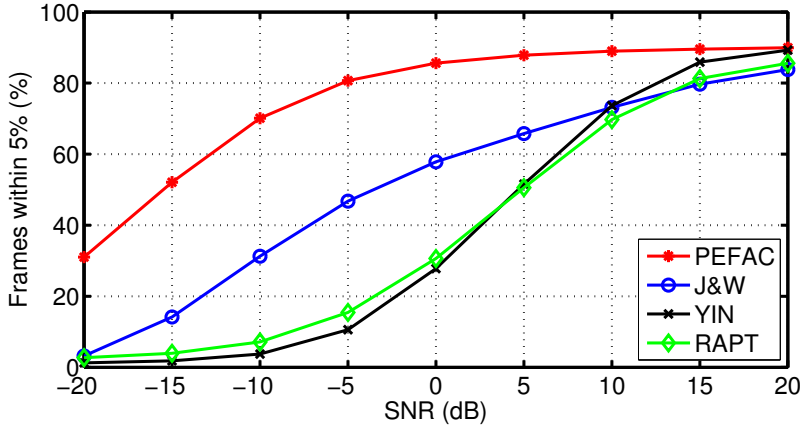


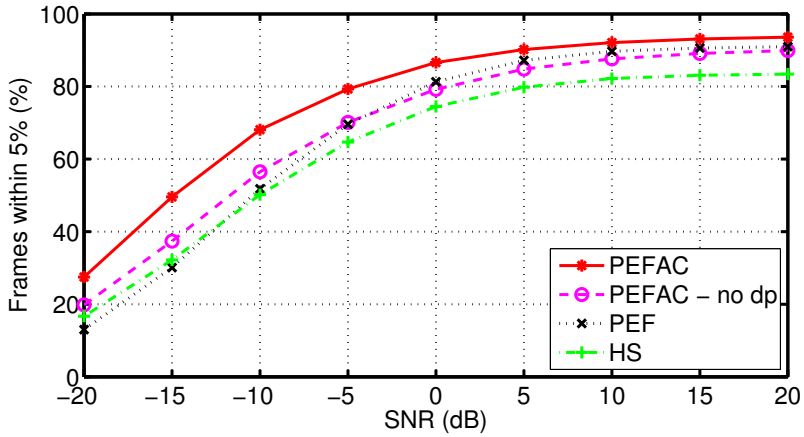
Figure 3.11: Variation of pitch estimation accuracy on the CSLU-VOICES corpus with SNR for (a) cafeteria noise, (b) metro noise, and (c) street noise from the ITU-T P.501 standard [69]. The graphs show the percentage of correct frames (error below 5%) for each of the algorithms: PEFAC, J&W [76], YIN [26] and RAPT [137].

PEFAC outperforms the other algorithms also on speech and noise databases not used in the training. As Fig. 3.10 and Fig. 3.11 illustrate, the robustness of PEFAC varies for different noise types. The PEFAC algorithm is robust to relatively narrow-band noises, such as car noise, as the normalization stage is able to attenuate them. However it is less robust to noises such as babble or cafeteria noise, whose power spectrum matches that of speech as the harmonic power is masked by the noise at low SNRs. Overall, the performance of PEFAC consistently exceeds that of the other algorithms.

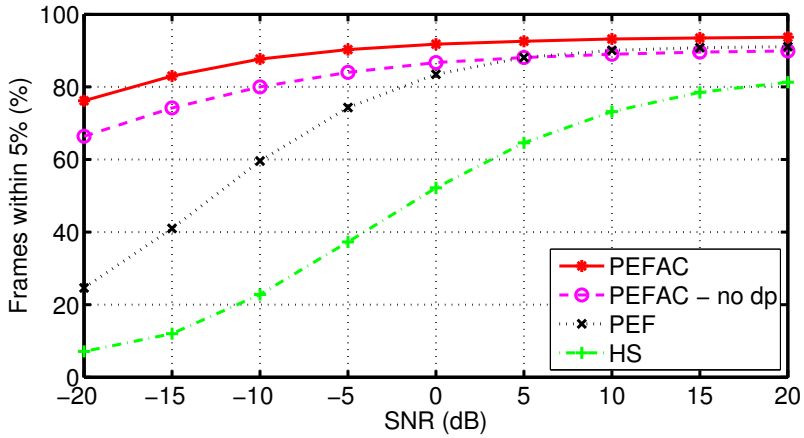
In Fig. 3.12 we show a breakdown of the algorithm performance. In each plot the performance of PEFAC is represented by the solid line, the dashed line shows the performance of PEFAC without the dynamic programming stage ("PEFAC - no dp") similar to the earlier version of the algorithm presented in [43], the dotted line shows the performance without the normalization stage using only the filter defined in (3.9) ("PEF") and the dash-dot line the performance using the filter defined in (3.2) ("HS"). It is shown in [135, 21] that, for a sufficiently long analysis window, HS is a close approximation to the maximum likelihood pitch estimate for a periodic signal in white Gaussian noise. It can be seen from Fig. 3.12(a) that HS and PEF have similar performance when the noise is indeed white Gaussian. For babble and car noise, however, the PEF algorithm is substantially better than HS. The normalization stage gives no benefit for babble noise, which already follows an LTASS spectrum, and gives only a small improvement at low SNRs for white noise. However for car noise, which includes a strong low frequency component, the benefit is very substantial. Finally the dynamic programming stage results in a small but worthwhile gain in all cases.

The distribution of the ratio of the estimated to the ground truth pitch, $\hat{f}_0/f_0$, is shown on a log-probability scale in Fig. 3.13 for white noise at -20, 0 and $+20$ dB SNR. As expected, peaks at half and double the fundamental frequency are present in the distribution although they are much lower than the main peak. We can also observe how the dash-dot vertical lines at $\pm 5\%$ encompass the main peak of the distribution. For low SNRs, the error distribution is relatively uniform apart from the main peak. The mean and standard deviation of the fine pitch errors (errors

(a) White noise



(b) Car noise



(c) Babble noise

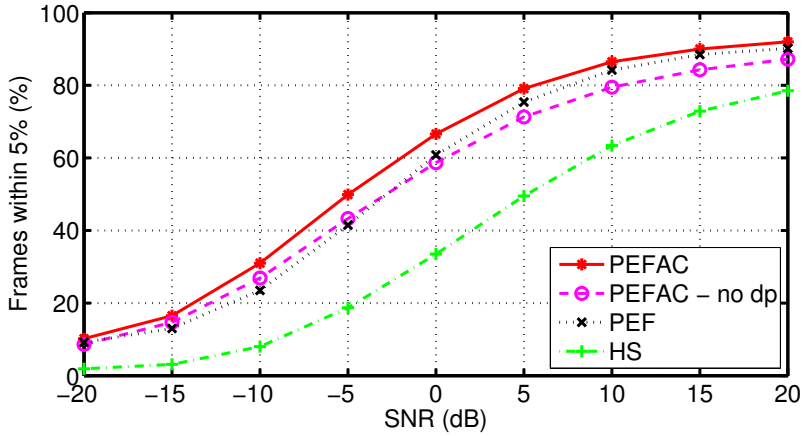


Figure 3.12: Variation of pitch estimation accuracy (error below 5%) with SNR for (a) white noise, (b) car noise, and (c) babble noise. The solid line shows the percentage of correct frames for PEFAC. The dashed line shows the performance of the algorithm without dynamic programming (PEFAC - no dp), the dotted line shows the performance of the algorithm without dynamic programming or normalization (PEF) and the dash-dot line the performance using only the filter defined in (3.2) (HS).

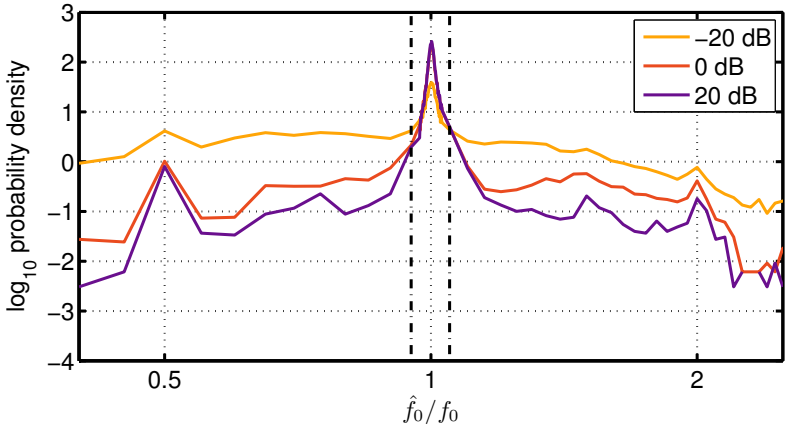


Figure 3.13: Log probability density distribution of the ratio of the estimated to the ground truth pitch, $\hat{f}_0/f_0$, at different SNRs for white noise on the core test set of the TIMIT database. The dash-dot vertical lines are at ±5%.

below ±5%) of our algorithm are shown in Table 3.3 for white noise. We see that on the TIMIT database there is a small bias of +0.24% at high SNRs which is not present on the CSLU-VOICES database. We believe that this bias may arise from small errors in the ground truth. The standard deviation of the fine pitch error is very similar on both databases and increases at lower SNRs as the error distribution becomes more uniform. We show the mean results averaged over white, babble and car noise for male and female speakers separately in Fig. 3.14 and we observe that, although the results are similar in both cases, the performance is consistently lower for male speakers at negative SNRs. The use of the universal LTASS as the target of the normalization stage attenuates low frequency components, which, in the case of male speakers, may include the fundamental frequency.

Table 3.3: Mean and standard deviation of the fine pitch error for white noise

	SNR (dB)	−20	−10	0	10	+20
TIMIT	Mean (%)	0.08	0.19	0.23	0.24	0.24
	Std (%)	1.93	1.37	1.21	1.17	1.18
CSLU	Mean (%)	-0.09	-0.06	-0.06	-0.03	-0.02
	Std (%)	2.05	1.54	1.37	1.29	1.28

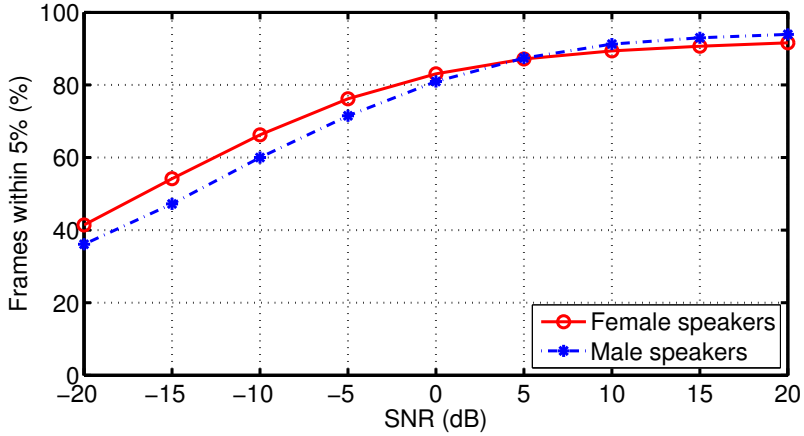


Figure 3.14: Variation of the mean pitch estimation accuracy on the core test set of the TIMIT database over white, babble and car noise with SNR for male and female speakers.

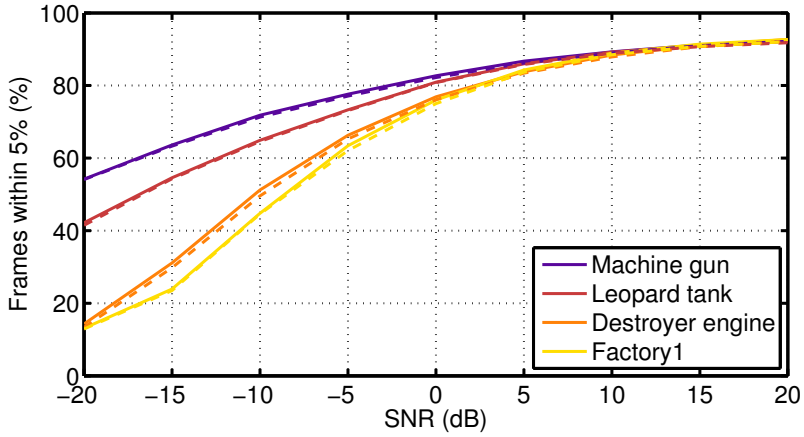


Figure 3.15: Variation of pitch estimation accuracy with SNR for different noise types from the RSG-10 database without (solid line) and with reverberation (dashed line) on the core test set of the TIMIT database.

We have evaluated the performance of PEFAC on all 15 noise types included in the RSG-10 database. The best and worst performances are respectively the car and babble noises already shown in Fig. 3.10. In Fig. 3.15 we show the performance for some other noise types, not used for the training, both with and without reverberation. The MARDY database [151] contains measured Room Impulse Responses (RIRs) for a number of source-to-microphone configurations. To create the reverberant speech we use a RIR corresponding to a direct line of sight configuration with a source-to-microphone distance of 3 m and a 60 dB decay time of 660 ms. Figure 3.15 shows

Table 3.4: Processing time (in seconds) per second of speech

	PEFAC	RAPT	YIN	J&W
MATLAB	0.200	0.462	0.194	
JAVA				22.304

that the performance accuracy of PEFAC is over 75% at positive SNRs for the four different noises. The dashed lines in Fig. 3.15 represent the pitch estimation accuracy in reverberant conditions, where the ground truth was the same as for anechoic speech. The results are very similar to non-reverberant conditions and, for machine gun and leopard tank noise, are almost indistinguishable.

Finally, as an indication of the comparative computational complexity of each algorithm, we have calculated the average processing time (in seconds) per second of speech on a PC having an Intel Xeon CPU with 2.27 GHz clock speed. As we can observe in Table 3.4, RAPT, YIN and PEFAC were all implemented in MATLAB and the processing time was less than half a second in each case, with YIN and PEFAC having a processing time close to 0.2 s for a second of speech. J&W, which is a multipitch algorithm, was implemented in JAVA and has the highest processing time, taking an average of 22.3 s to process a second of speech.

3.5 Summary

In this chapter we have presented the PEFAC pitch estimation algorithm and shown that it is able to give both a reliable pitch estimation and accurate voiced speech detection even at poor SNRs. The algorithm comprises a normalization stage that attenuates narrow-band noise components with a pitch estimation filter that rejects broadband noise that has a smooth power spectrum. Dynamic programming is used to impose soft temporal continuity constraints by selecting between pitch candidates in each frame. For voiced speech detection, two GMMs are trained on voiced and unvoiced frames respectively and the likelihood ratio of the two models is used to classify each frame.

The proposed pitch estimation algorithm has been evaluated on the TIMIT core test set and on the CSLU-VOICES corpus with a variety of noise types and consistently outperformed other widely used algorithms. It has also been evaluated on reverberant speech without a degradation in performance. The voiced activity detector has been shown to discriminate between voiced and unvoiced with a lower overall error rate than the detectors implemented by other competing algorithms.

Chapter 4

Speech active level estimation in noisy conditions

The active level of a speech signal is defined to be its average power during intervals when speech is present. The measurement of a signal's active level is an essential component in any application where the input speech power needs to be normalized, such as in non-intrusive metrics for quality assessment [81]. It is also important whenever a pre-trained speech model is combined with an estimated noise model as in the parallel model combination technique [140, 36] or to determine the SNR of an input signal. For binary mask estimation, the speech active level can be used to make the process independent of the initial speech level, as we shall explain in Chapter 6.

In this chapter, we present a new method for speech active level estimation which combines a novel algorithm based on voiced speech energy extraction with the standardized ITU-T Recommendation P.56 [68]. At poor signal-to-noise ratios, the algorithm estimates the active level by identifying intervals of voiced speech and summing the energy of the pitch harmonics in the time-frequency domain while rejecting that of the noise. We compare the performance of our method with that of ITU-T P.56 on the TIMIT database and demonstrate that it performs exceptionally well in both high and low levels of additive noise.

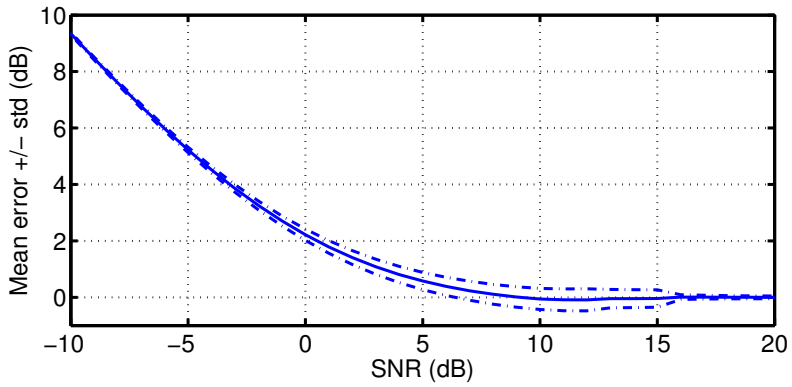


Figure 4.1: Variation of P.56 mean error (solid line) plus and minus the standard deviation (dash-dot line) with SNR for white noise on 1000 utterances from the training set of the TIMIT sentence database [37].

4.1 Standardized ITU-T recommendation

The ITU-T Recommendation P.56 [68] defines a standardized method for objectively measuring the speech active level. The procedure first low-pass filters the rectified signal to obtain its envelope. The speech is then defined to be active whenever the envelope has exceeded a specified threshold within the past 200 ms [9]. This threshold is circularly defined to be 15.9 dB below the active level (which equals the mean power during times when the speech is active). This algorithm performs extremely well at high SNRs since the speech pauses are easily detectable in the signal envelope from their low amplitude. However, at low SNRs, the speech pauses are difficult to identify and the algorithm falsely takes some or all of the noise energy to be speech. Figure 4.1 shows the mean error of the ITU-T P.56 algorithm as a function of SNR for white noise. We can observe how the performance increasingly deteriorates below 5 dB SNR, showing the need to develop a new speech level estimation approach based on speech characteristics that are robust to noise.

4.2 Harmonic summation algorithm

The majority of the energy in a speech signal is concentrated in the voiced intervals (see Fig. 1.3). In the time-frequency domain, most of the voiced speech energy is

located in a small number of harmonic peaks that remain detectable even at poor SNRs. In this section, we propose a method to estimate the speech active level at low SNRs from the energy of the harmonic peaks during voiced intervals.

We have shown in the previous chapter that we are able both to identify voiced speech intervals and to estimate the pitch, f_0, reliably even at negative SNRs. We assume that during voiced intervals, the speech can be represented as a periodic source at frequency f_0 so that our signal model in the Power Spectral Density (PSD) domain is given by (3.1), reproduced here for convenience,

$$Y(f) = \sum_{k=1}^{K} a_k \delta(f - kf_0) + N(f) \tag{4.1}$$

where $N(f)$ represents the power spectral density of the unwanted noise, a_k the power of the k^{th} harmonic and K is the number of harmonics. From equation (4.1) we note that, for this idealized signal model, all the speech energy is located at the harmonics of the fundamental frequency f_0. In practice, we process the noisy signal in overlapping frames and the energy of the harmonics is spread over a range of frequencies by the effects of the analysis window and the rate of change of f_0. To extract the energy of these harmonics, we need to identify the voiced speech intervals and, within these, estimate the value of f_0. In this chapter, we use PEFAC, the pitch estimation algorithm robust to high levels of noise which was presented in Chapter 3. We note that our proposed speech level estimation algorithm can equally be implemented using any other pitch estimator and that its robustness to noise depends heavily on the pitch estimator performance.

Once the voiced speech segments are identified and the fundamental frequency estimated, we need to measure the energy of the harmonics. For the energy of the k^{th} harmonic, we calculate a weighted integral of the frame power spectrum as $\int h_a(f - kf_0)Y(f)df$. The weighting function, $h_a(f)$, should be chosen such that:

(i) it gathers most of the harmonic energy while avoiding any interaction with adjacent harmonics,

(ii) it avoids including the energy of the noise in the harmonic energy estimate.

A weighting function that accomplish these requirements is the weighted Mexican hat wavelet, the negative normalized second derivative of a Gaussian function, which can be expressed as

$$h_a(f) = \left(1 - \frac{f^2}{\sigma^2}\right) e^{\frac{-f^2}{2\sigma^2}} \tag{4.2}$$

To accomplish the first property, the positive part of the weighting function needs to cover the width of the harmonic and its total length needs to be restricted not to interact with adjacent harmonics. To ensure this, the support of the weighting function should lie within $\pm \min f_0$. The width of the harmonic is mainly dependent on the window used to calculate the periodogram of the frame, as the signal frequency components, $Y_t(f)$, are convolved with the PSD of the window function, $W(f)$, to give $R_t(f) = Y_t(f) * W(f)$. Figure 4.2 compares the PSD of a Hamming window having the parameters defined in Section 4.4 (dash-dot line) with the weighting function defined in (4.2) (solid line) with $\sigma = 15$. We can observe the fulfilment of the two requirements, as the total length is only about 100 Hz and the positive part covers the width of the harmonic.

The second requirement, the minimization of the noise contribution to the estimated harmonic energy, is accomplished since the weighting function has the property that $\int h_a(f)df = 0$. This means that any smoothly varying noise spectrum will be

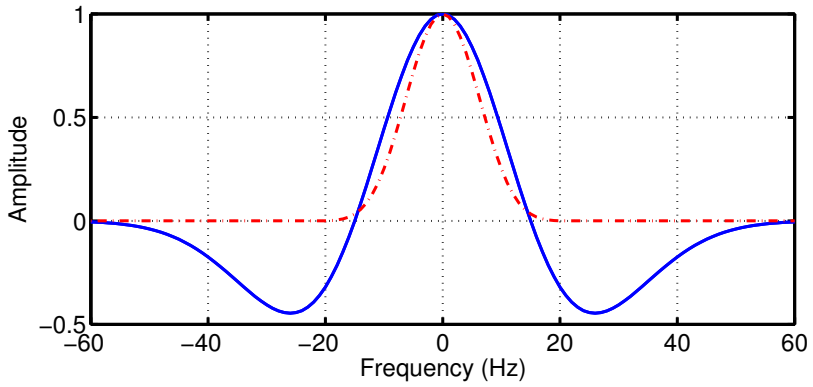


Figure 4.2: Mexican hat wavelet for $\sigma = 15$ (solid line) and PSD of a Hamming window of length equal to 90 ms (dash-dot line).

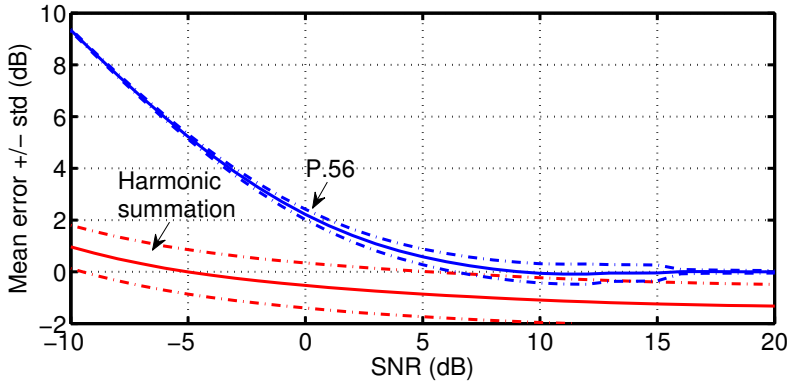


Figure 4.3: Variation of the harmonic summation (red) and P.56 (blue) mean error (solid line) plus and minus the standard deviation (dash-dot line) with SNR for white noise on 1000 utterances from the training set of the TIMIT database [37].

greatly attenuated.

The energy, E_t, of the first K harmonics in a voiced time frame t, is estimated as

$$E_t = \sum_{k=1}^{K} \max\left(0, \int R_t(f) h_a(f - k f_0) df\right) \tag{4.3}$$

The maximum function is included in (4.3) since the integral can be negative when the SNR is poor. The active speech level can now be estimated as

$$\hat{l}_h = \frac{1}{|V|} \sum_{t \in V} E_t \tag{4.4}$$

where V represents the subset of frames which are classified as voiced by the pitch detector.

Figure 4.3 shows the mean and standard deviation of the estimation error as a function of SNR both for ITU-T P.56 and for the harmonic summation algorithm described above. While ITU-T P.56 obtains very good results at high SNRs, its performance degrades rapidly for negative SNRs. On the other hand, the reliability of the harmonic summation method is more constant across all SNRs but its standard deviation is higher and it underestimates the speech level at high SNRs.

To compensate for the unvoiced speech energy and the underestimation of the

harmonic energy we introduced an offset, β, such that

$$l_h = 10\log_{10}\left(\hat{l}_h\right) + \beta \tag{4.5}$$

The value of β is determined from a training set by minimizing the cost function $J = \sum_{u=1}^{U}\left(l^u - 10\log_{10}\left(\hat{l}_h^u\right)\right)^2$ with respect to β. This gives

$$\beta = \frac{\sum_{u=1}^{U}\left(l^u - 10\log_{10}\left(\hat{l}_h^u\right)\right)}{U} \tag{4.6}$$

where l^u is the speech active level ground truth in dB for the u^{th} utterance and U is the number of utterances used for the training.

4.3 Composite algorithm

As Fig. 4.3 illustrates, the P.56 active level estimate is more accurate at high SNRs but the harmonic summation method provides better results at negative SNRs. Accordingly, we combine the results from both algorithms into a new estimate that will provide reliable estimation over a larger SNR range.

In order to be able to combine the methods, we need to find a measure which identifies the transition point at which the performance of the harmonic summation method starts to be more reliable than that of ITU-T P.56. This is achieved by

$$\gamma = 10\log_{10}\frac{\hat{l}_h}{P_N} \tag{4.7}$$

where $\hat{l}_h$ is defined in (4.4) and P_N represents the noise power estimated using the algorithm described in [39]. Although it could be considered an SNR estimation, we are not aiming to estimate the SNR and consequently we are not directly concerned with the accuracy of the SNR estimate. Figure 4.4 shows the root mean squared error of ITU-T P.56 and the harmonic summation method for different values of γ. Three different noises were used at SNRs from -10 dB to 20 dB: white noise, car noise and babble noise. As we can observe in Fig. 4.4, γ provides a good way of

identifying the point at which ITU-T P.56 performance starts to degrade and the harmonic summation method becomes the most reliable.

The final speech active level estimate, l_c, is calculated as a linear combination of the ITU-T P.56 estimate, l_p, and the harmonic summation method estimate, l_h,

$$l_c = \rho l_p + (1 - \rho)\, l_h \tag{4.8}$$

where ρ defines the contribution of each algorithm.

To determine the optimum mapping function $\rho(\gamma)$, we minimize the cost function $J = \sum_{u=1}^{U} (l - l_c)^2$ with respect to ρ and we obtain

$$\rho(\gamma) = \frac{\sum_{u \in G(\gamma)} \left(l^u - l_h^u\right)\left(l_p^u - l_h^u\right)}{\sum_{u=1}^{U} \left(l_h^u - l_p^u\right)^2} \tag{4.9}$$

where $G(\gamma)$ is the set of utterances having a particular value of γ.

From training data, we determined the optimal ρ for selected values of γ as shown in Table 4.1. We perform linear interpolation on this table for intermediate values of γ.

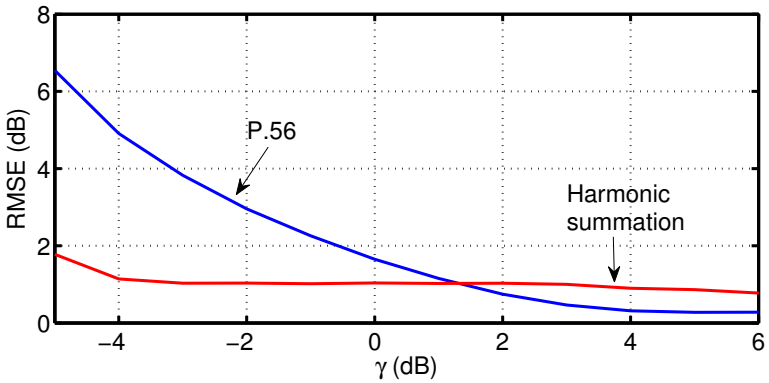


Figure 4.4: Variation of the root mean squared error of P.56 and harmonic summation method with γ on 1000 utterances from the training set of the TIMIT database for white noise, car noise and babble noise.

4.4 Experiments

The test set and a subset of the training set from the TIMIT database [37] were respectively used for testing and training the algorithm. The sampling frequency of the speech material is 16 kHz. To determine the ground truth for the speech active level, ITU-T P.56 was applied on the clean speech signal.

For training and testing, noise from the RSG-10 database [131] was added to the speech files to generate the noisy signals. The calculation of SNR used ITU-T P.56 [68, 15] for the speech level and unweighted power for the noise.

The STFT used a Hamming analysis window of 90 ms duration and the inter-frame time increment was 10 ms. This frame duration is long enough to resolve the pitch harmonics even for low values of f_0 but short enough to limit the pitch variation within a frame.

The speech active level estimation described in this chapter includes a number of algorithm parameters whose values were determined empirically using the training set from the TIMIT database. The β parameter was calculated from equation (4.6) using 1000 utterances from the training set. Three types of noise were used at different SNRs ranging from -5 to $+5$ dB: white noise, car noise and babble noise. These three noises have different spectral characteristics and were chosen to make the results relatively independent of the noise type. The final value was set to $\beta = 0.85$.

The linear combination of ITU-T P.56 and the harmonic summation method was determined by the optimization of ρ for different values of γ. For the calculation of the noise power, P_N, use to calculate γ in (4.7) we use the implementation of the algorithm in [39] provided in [15]. The range of γ used for the estimation was from -2 dB to 4 dB every 0.5 dB. Below $\gamma = -2$ dB, the error from the harmonic

Table 4.1: Optimized ρ values for different γ values

γ (dB)	-2	-1	0	1	2	3	4
$\rho(\gamma)$	0	0.16	0.28	0.44	0.68	0.89	1

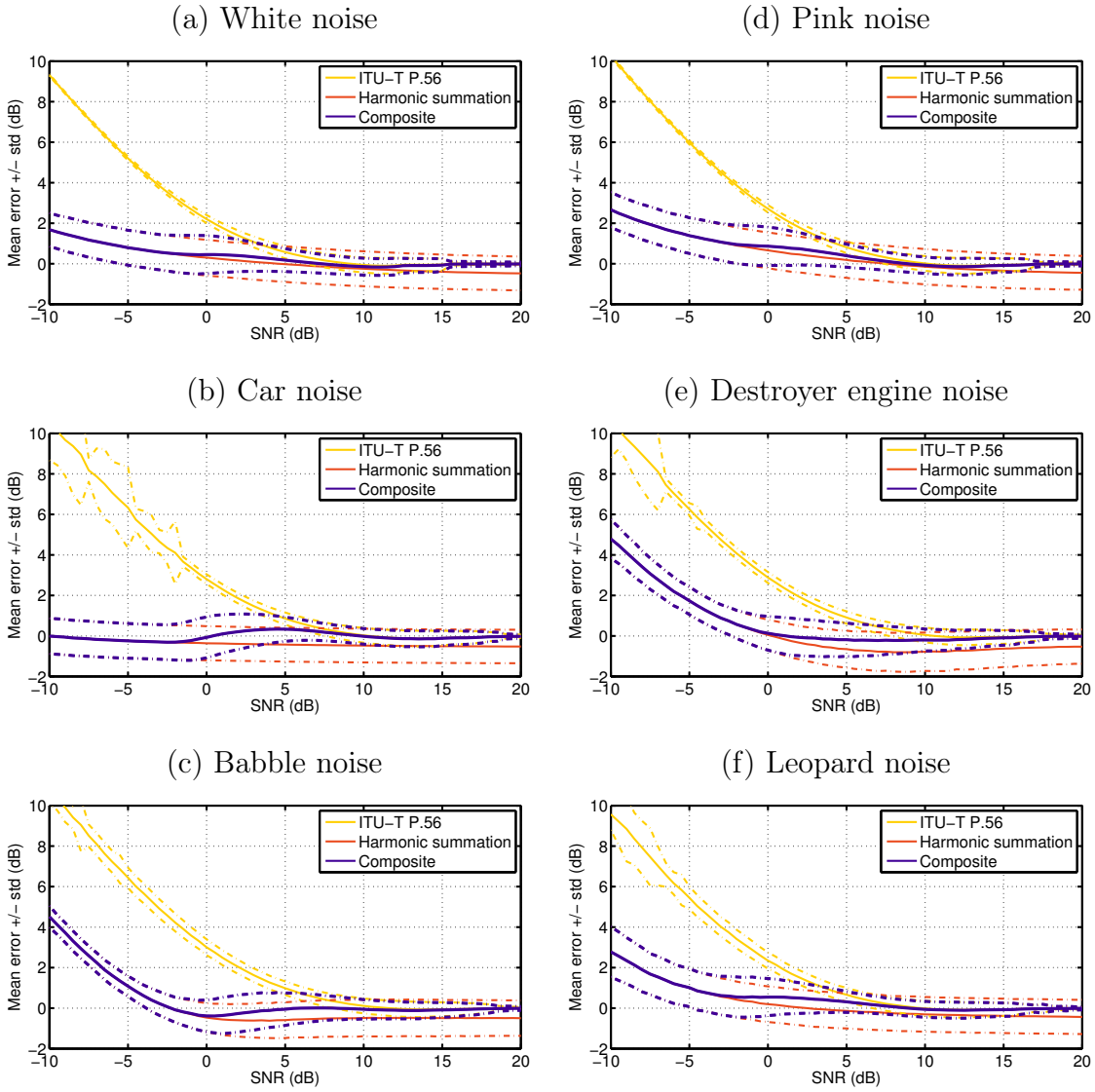


Figure 4.5: Variation of speech active level estimation accuracy on the test set of the TIMIT database with SNR for (a) white noise, (b) car noise, (c) babble noise, (d) pink noise, (e) destroyer engine noise and (f) leopard tank noise. The solid lines show the mean error of the estimation and the dashed lines the mean error plus/minus the standard deviation for each of the algorithms.

summation algorithm is much lower than that of ITU-T P.56 and $\rho = 0$ and above $\gamma = 4\,\text{dB}$, the superiority of the ITU-T P.56 algorithm is clear, $\rho = 1$. Table 4.1 shows how, as expected, the optimum calculated value of ρ smoothly increases with γ.

4.5 Results

In this section, the performance of the proposed speech active level estimator is evaluated on the test set of the TIMIT database [37]. Six types of noise from the RSG-10 database [131] were evaluated at different SNRs from -10 to $+20\,\text{dB}$: white, car, babble, pink, destroyer engine and leopard tank noise. While the first three kinds of noises were used in the training, the last three were new kinds of noises to the algorithm. This allows the performance evaluation of the proposed method on untrained conditions.

For each of the six noise types, Fig. 4.5 shows the mean and standard deviation of the estimation error for three algorithms: ITU-T P.56, the harmonic summation algorithm from Sec. 4.2 and the composite algorithm from Sec. 4.3. We observe how the combined method is able to select the best estimate at different SNRs, both on noises used for the training and on new noises. Babble and destroyer engine noise have the worst performances, with a mean error of approximately $4.5\,\text{dB}$ at $-10\,\text{dB}$ SNR, and car noise have best performance, with a mean error close to $0\,\text{dB}$ even at $-10\,\text{dB}$ SNR. Overall, the proposed method is able to provide a good estimation at both high and low SNRs for all the tested noise types. Spectrograms of all the noise types used in training and testing are included in Appendix A.

4.6 Summary

In this chapter we have presented a new method for estimating the speech active level which combines the ITU-T Recommendation P.56 with novel harmonic summation approach. The harmonic summation method extracts the energy of the speech har-

monics and provides a reliable estimation of the speech active level even at low SNRs. A fixed offset determined from training data compensates for any unvoiced speech power and for the underestimation of voiced speech power. The final speech active level estimate is calculated as a linear combination of the ITU-T P.56 estimate and the harmonic summation method estimate. The algorithm has been evaluated on the TIMIT test set with a range of noise types and extends by more than 7 dB the range of SNRs for which reliable estimation is possible.

Chapter 5

Sibilant speech detection in noise

Our goal for binary mask estimation is to identify the time-frequency regions that contain significant speech energy. In voiced speech regions, this energy is concentrated in the pitch harmonics and in Chapter 3 we showed that it was possible, even at poor SNRs, to identify these regions and to estimate the pitch. In this chapter, we address the problem of detecting unvoiced speech energy.

Recent work has illustrated the significance of unvoiced speech detection for several applications. In [127], for instance, it was shown that enhancing noisy unvoiced speech plays a greater role in achieving accurate speech recognition than enhancing voiced speech. Detecting unvoiced speech in noise is especially important for hearing-impaired listeners, who typically have severe high frequency hearing loss, as well as for speech enhancement algorithms, which can benefit from adaptivity to different phoneme classes. An increasing interest in unvoiced speech detection has specifically emerged for binary mask estimation [59, 61], where most previous approaches have focused on voiced speech segregation [56, 60], as seen in Section 2.3.2.

Aperiodic speech energy at high frequencies is mainly contained in stops, fricatives and affricatives (a sequence of a stop followed by a fricative [94]). Sibilant phones, a subset of fricatives and affricative sounds, have more energy than their non-sibilant counterparts and most of their energy is concentrated at higher frequencies. Therefore, sibilant speech sounds accounts for a large fraction of aperiodic high frequency speech energy. In English, they comprises the fricatives /s/, /ʃ/, /z/ and /ʒ/ and the

affricatives /tʃ/ and /dʒ/.

In this chapter, we present a sibilant detection algorithm robust to high levels of noise for wide-band speech that operates in the frequency domain and that does not rely on voicing detection. Rather than identifying explicit sibilant onsets and offsets, a sustained increase in energy during the sibilant is instead detected. Under the hypothesis of a sibilant presence within a time-frame, its mean power in each frequency band is estimated using a maximum likelihood approach. This information is sent to a classifier which discriminates sibilant from non-sibilant time frames. As far as we are aware, there is no other sibilant detector in the literature for noisy conditions.

5.1 Proposed method

Following [31], we assume that the short-time Fourier transform (STFT) coefficients of speech and noise can be modelled as statistically independent complex Gaussian random variables. Given a noisy speech signal, the power, $Y_{t,f}$, in a time-frequency STFT bin is therefore distributed as

$$p(y_{t,f}) = \frac{1}{\mu_{t,f}} \exp\left(-\frac{y_{t,f}}{\mu_{t,f}}\right) \tag{5.1}$$

where t and f are the time-frame and frequency indices and $\mu_{t,f}$ is the mean power.

Fig. 5.1 shows the time-variation of power at 5 kHz for a noisy speech example corrupted with white noise at 5 dB SNR containing the phone /ʃ/. We can divide the time interval into three segments as indicated above the waveform: a central segment S that encompasses the sibilant and two surrounding intervals, $N1$ and $N2$, that contain no sibilant energy. We assume the mean power of the speech to be constant over S and that of the noise to be constant over the entire interval $N1 + S + N2$, giving $\mu_{N1,f} = \mu_{N2,f} = a_f$ and $\mu_{S,f} = a_f + b_f$. From (5.1), the log-likelihood of the

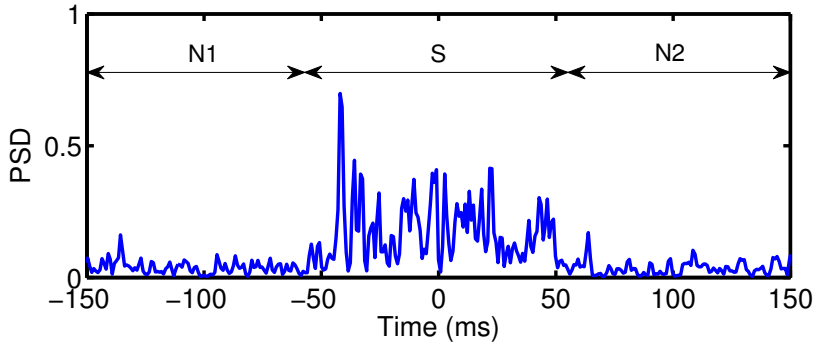


Figure 5.1: Power spectral density (PSD) at 5 kHz versus time of a speech segment containing the sibilant phone /ʃ/ using a Hamming analysis window of 3.6 ms duration with 75% overlap. The speech has been corrupted with white noise at 5 dB SNR. The time origin represents the centre of the sibilant phone.

observed signal can then be expressed as

$$\begin{aligned} L_f &= \sum_{t \in S} \left(-\ln(a_f + b_f) - \frac{y_{t,f}}{a_f + b_f} \right) \\ &\quad + \sum_{t \in N1,N2} \left(-\ln(a_f) - \frac{y_{t,f}}{a_f} \right). \end{aligned} \tag{5.2}$$

5.1.1 Sibilant speech energy estimation

By maximizing the log-likelihood in (5.2), the sibilant mean, b_f, and the noise mean, a_f, can be estimated if the exact time and duration of the sibilant phone are known. However, the duration of an actual sibilant is unknown and varies in each case. Fig. 5.2 shows the sibilant duration distribution in the TIMIT training set [37]. We observe that 74% of sibilant durations lie within 60 and 130 ms. Therefore, if $t_s = 0$ represents the centre of a sibilant $|t_s| < 30$ ms has a high probability of lying within the sibilant while the region $|t|_s > 65$ ms has a high probability of lying outside the sibilant. To account for this, we apply a weighting function, w_t, to the time frames when calculating the log-likelihood that reduces the contribution of the transition region 30 ms $< |t_s| <$ 65 ms as shown in Fig. 5.4. The weighted log-likelihood can now be

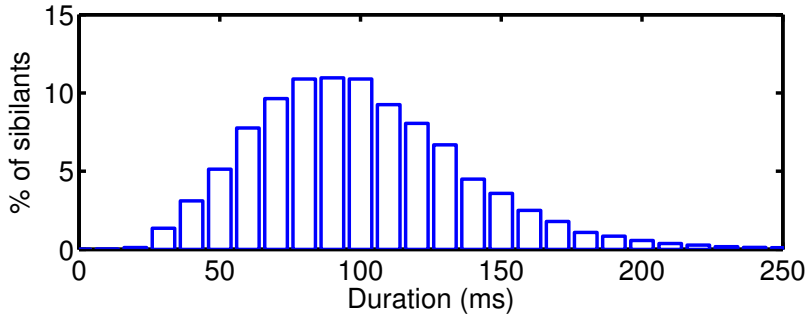


Figure 5.2: Sibilant duration distribution in the TIMIT training set.

expressed as

$$\begin{aligned}\tilde{L}_f &= \sum_{t\in S} w_t \left(-\ln(a_f+b_f) - \frac{y_{t,f}}{a_f+b_f}\right) \\ &\quad + \sum_{t\in N1,N2} w_t \left(-\ln(a_f) - \frac{y_{t,f}}{a_f}\right). \end{aligned} \tag{5.3}$$

We maximise the value of the log-probability with respect to a_f and b_f by setting the partial derivatives to zero

$$\begin{aligned} 0 = \frac{\partial \tilde{L}_f}{\partial a_f} &= \sum_{t\in S} w_t \left(-\frac{1}{a_f+b_f} - \frac{y_{t,f}}{(a_f+b_f)^2}\right) \\ &\quad + \sum_{t\in N1,N2} w_t \left(-\frac{1}{a_f} + \frac{y_{t,f}}{a_f^2}\right) \end{aligned} \tag{5.4}$$

$$0 = \frac{\partial \tilde{L}_f}{\partial b_f} = \sum_{t\in S} w_t \left(-\frac{1}{a_f+b_f} - \frac{y_{t,f}}{(a_f+b_f)^2}\right) \tag{5.5}$$

from which we can estimate the mean noise energy, a_f, and the mean sibilant energy, b_f, as

$$\hat{a}_f = \frac{\sum_{t\in N1,N2} w_t y_{t,f}}{\sum_{t\in N1,N2} w_t} \tag{5.6}$$

$$\hat{b}_f = \frac{\sum_{t\in S} w_t y_{t,f}}{\sum_{t\in S} w_t} - \hat{a}_f \tag{5.7}$$

Under the hypothesis that time-frame t lies at the centre of a fixed-length sibilant phone, we can estimate the mean sibilant power in frequency bin f using (5.7). We

denote this estimate as $\hat{b}_{t,f}$, where the index t represents the time-frame considered to be the centre of segment S. Fig. 5.3(a) shows the PSD waveform of $\hat{b}_{t,f}$ for the /ʃ/ sibilant example shown in Fig. 5.1. We see that it reaches a maximum when t lies near the centre of the phone and becomes negative either side of the phone when region $N1$ or $N2$ overlaps significantly with the true sibilant.

5.1.2 Maximum filter and normalization

The quantity $\hat{b}_{t,f}$ from (5.7) will give a reliable estimate of sibilant power near the centre of a sibilant phone and also in signal regions where no sibilant is present. However, as can be seen in Fig. 5.3(a), the estimate of sibilant power is less accurate in frames near the sibilant boundary. To counter this effect, we apply a maximum filter to the sibilant power estimate

$$\tilde{b}_{t,f} = \max_{|m-t|<W/2} \hat{b}_{m,f} \tag{5.8}$$

where W, the filter support, represents the minimum sibilant duration. Fig. 5.3(b) shows the filter output, $\tilde{b}_{t,f}$, using $W = 30\,\text{ms}$ and we observe that the estimated $\tilde{b}_{t,f}$ remains at a high level for most of the sibilant duration.

To make the estimate independent of the overall speech level, the estimated sibilant mean power within each frame is normalized to give

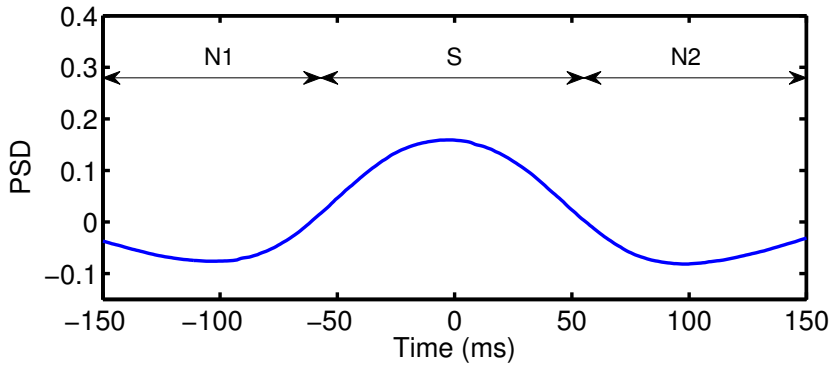
$$\bar{b}_{t,f} = \frac{\tilde{b}_{t,f}}{\frac{1}{N_f}\sum_{f=1}^{N_f}\left|\tilde{b}_{t,f}\right|} \tag{5.9}$$

The absolute value is used because as seen in Fig. 5.3(b), $\tilde{b}_{t,f}$ can be negative when the sibilant occupies a region that was assumed to be noise.

5.1.3 Gaussian mixture model

For each frame, the normalized sibilant power spectrum, $\bar{b}_{t,f}$ for $f \in [1, N_f]$, forms the input to two GMMs: one trained on non-sibilant speech and the other on sibilant

(a) Estimated sibilant mean power, $\hat{b}_{t,f}$



(b) Estimated sibilant mean power after maximum filter, $\tilde{b}_{t,f}$

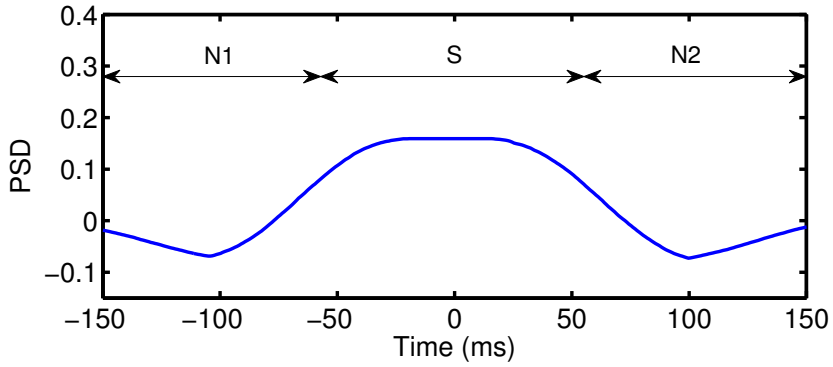


Figure 5.3: Estimated sibilant PSD for the segment of speech shown in Fig. 5.1. Plot (a) shows the raw estimate, $\hat{b}_{t,f}$, from (5.7) and plot (b) shows the output of the maximum filter (5.8), $\tilde{b}_{t,f}$.

speech. The probability that a time frame contains a sibilant phone is calculated from the likelihood ratio of the two GMMs.

5.2 Experiments

The sibilant detector described in this chapter includes a number of algorithm parameters whose values were determined using the training set of the TIMIT database [37], which includes phonetic transcription. The STFT used a Hamming analysis window of 3.6 ms duration with 75% overlap. The relatively short analysis window provides a high time resolution and a frequency resolution that is able to characterize the sibilant power spectrum without resolving pitch harmonics. The power spectrum

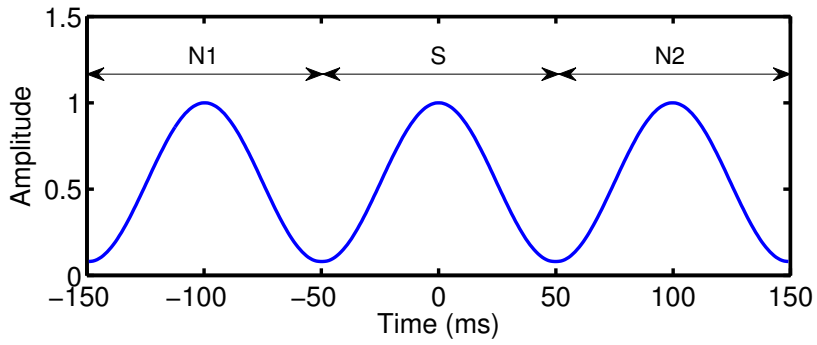


Figure 5.4: Weighting function, w_i, used in (3) to accommodate variations in sibilant duration.

of each frame was interpolated using triangular filters to give 14 frequency bins whose centres are uniformly spaced from 1.5 kHz to 8 kHz.

The sibilant duration, S, as well as the duration of $N1$ and $N2$ need to be fixed in order to estimate the mean sibilant energy, b_f, from equation (5.7). We evaluated a range of fixed widths for S as well as a variable width approach in which (5.3) was maximized with respect to the phone boundaries in addition to the powers a_f and b_f. We found that a fixed S, $N1$ and $N2$ duration of 100 ms gave the highest performance on a training set. The weighting function used in (5.3) was the concatenation of three Hamming windows shown in Fig. 5.4 and the length of the maximum filter in (5.8) was set to $W = 30$ ms.

The input for the GMMs was a 14-component vector containing the estimated sibilant power spectrum from 1.5 kHz to 8 kHz every 500 Hz. The GMMs for sibilant and non-sibilant speech respectively used 14 and 28 full-covariance mixtures and were trained on the training subset of TIMIT. Sibilant phones and phones that sometimes include sibilant-like characteristics, such as stop consonants and non-sibilant fricatives, were excluded when training the non-sibilant GMM. To avoid problems caused by transcription alignment errors, phone transitions were omitted from the training. The SNR used for training was 0 dB in order to make the algorithm robust to noise, as it represents the lowest SNR at which sibilants/non-sibilants discrimination is practicable.

5.3 Results

In this section, the performance of the proposed sibilant speech detector is evaluated. The results were calculated using the test set from the TIMIT database, which contains a total of 168 speakers and 1344 utterances. For evaluation purposes all non-sibilant phones were taken into account including stops and non-sibilant fricatives previously excluded for the training. Every time-frame was evaluated, without the removal of phone transitions.

White Gaussian noise, babble noise and car noise from the RSG-10 database [131] were added to the speech files to generate the noisy test signals. The measurement of SNR used ITU-T P.56 [68, 15] for the speech level and unweighted power for the noise. Spectrograms of all the noise types used in training and testing are included in Appendix A.

The results obtained for -5 dB, 0 dB, 5 dB and 10 dB SNR as well as for clean speech are shown in Fig. 5.5 for the three types of noise: white noise, babble noise and car noise. The DET curves [105] in Fig. 5.5 shows the miss probability, P_{miss}, versus the false alarm probability, P_{fa}, as the likelihood ratio threshold is varied between 0.05 and 19.0. Because of the noise-like nature of sibilant phones at high frequencies, we observe that it is more difficult to detect sibilants in white noise, Fig. 5.5(a), than in other typical stationary noise sources where lower frequencies often dominate, such as babble noise, Fig. 5.5(b), or car noise, Fig. 5.5(c). The results for car noise, Fig. 5.5(c), show that the algorithm performance is very similar for all noise levels, as car noise does not mask the sibilant power. The performance on babble noise is illustrated in Fig. 5.5(b), where we observe that, although the performance degrades as the SNR decreases, the results for positive SNR are better than for white noise.

The equal error rates, where $P_{miss} = P_{fa}$, are listed in Table 5.1. and we see that at 0 dB SNR the highest equal error rate, 16.5%, occurs with white noise; this means that even in the worst tested case 83.5% of frames are correctly classified.

The circle on each line in Fig. 5.5 corresponds to a likelihood ratio threshold of unity corresponding to an estimated sibilant probability of 0.5. The values of P_{miss}

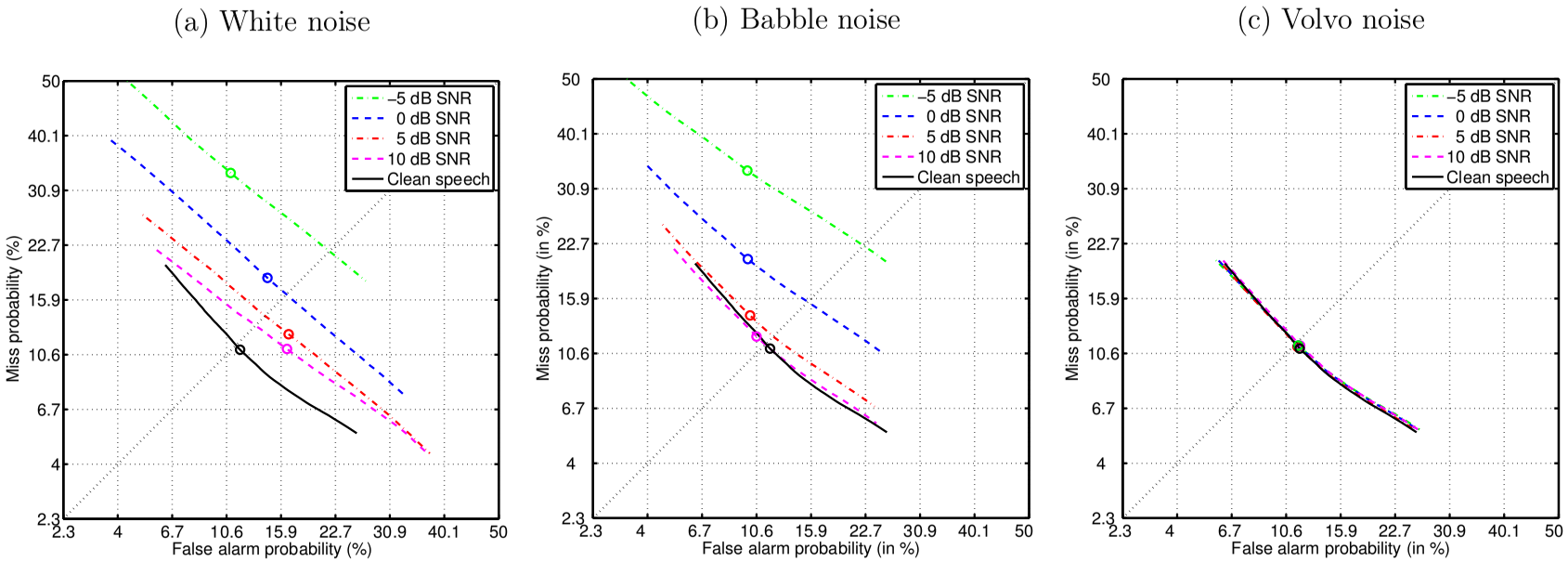


Figure 5.5: DET curve of the sibilant detection algorithm obtained for −5 dB, 0 dB, 5 dB and 10 dB SNR as well as for clean speech. The circles represent the results for a likelihood ratio threshold of unity.

Table 5.1: Classifier equal error rates as a function of SNR.

Equal error rate, $P_{er}(\%)$					
SNR (dB)	∞	+10	5	0	-5
White noise	11.4	13.2	14.2	16.5	21.8
Babble noise	11.4	11.3	12.0	13.5	22.4
Car noise	11.4	11.5	11.4	11.4	11.4

Table 5.2: Unity-threshold classification performance as a function of SNR.

Likelihood ratio of unity						
	SNR (dB)	∞	+10	5	0	-5
White noise	$P_{miss}(\%)$	11.0	11.1	12.4	18.4	33.7
	$P_{fa}(\%)$	11.7	16.5	16.7	14.4	10.9
Babble noise	$P_{miss}(\%)$	11.0	12.0	14.1	20.6	33.8
	$P_{fa}(\%)$	11.7	10.6	10.0	9.9	9.8
Car noise	$P_{miss}(\%)$	11.0	11.2	11.1	11.2	11.2
	$P_{fa}(\%)$	11.7	11.8	11.5	11.6	11.6

and P_{fa} when using this threshold are listed in Table 5.2. For clean speech both P_{miss} and P_{fa} is approximately 11%. Manual inspection of the missed sibilant frames indicates that most of them correspond either to sibilant boundaries or to phones with very low energy, whereas false alarms usually correspond to non-sibilant fricatives or stops. Moderate levels of white noise cause an increase in P_{fa}, while, in contrast, moderate levels of babble noise cause an increase in P_{miss}. The reason behind this is that while white noise adds energy at high frequencies which the algorithm may identify as sibilant energy, babble noise distorts the normalized sibilant power of the frame, therefore increasing P_{miss}.

5.4 Conclusions

In this chapter we have presented a sibilant detection algorithm robust to high levels of white Gaussian noise. The algorithm comprises a sibilant mean power estimation stage, which is based on a maximum likelihood approach, followed by a classification stage in which the likelihood ratio of two GMMs, one for sibilant speech and one for non-sibilant speech, is used. The algorithm has been evaluated on the TIMIT test set over a range of noise types and SNRs and consistently achieved over 80 % classification accuracy for positive SNRs.

Chapter 6

Mask estimation

In the binary mask approach to speech enhancement, a binary-valued gain mask is applied to the speech in the time-frequency domain and the signal is then transformed back into the time-domain. This procedure is similar to that used in conventional approaches such as spectral subtraction or MMSE estimators except that, in the latter cases, a continuously variable gain function is applied. The principal advantage of the binary mask approach over other state-of-the-art algorithms operating in the time-frequency domain is that the problem of enhancement is changed from one of gain estimation to one of classification.

A detailed review of the goals of binary masks enhancement systems was given in Chapter 2. The most common binary mask is the Ideal Binary Mask (IBM), based on the SNR at each time-frequency bin. The Target Binary Mask (TBM), more recently proposed and with the same intelligibility performance as the IBM [87], removes dependency on the noise by comparing the clean speech to the LTASS of the speaker. In Chapter 2 we proposed a variation of the TBM, the Universal Target Binary Mask (UTBM), and we showed it has a similar performance to that of the TBM while also removing dependency on the speaker by using a universal LTASS.

Our aim in this chapter is to estimate the UTBM, which selects time-frequency regions whose speech energy is above a frequency-dependent threshold. Accordingly, in the previous chapters, we have been exploring approaches that aim to identify time-frequency regions that contain high speech energy. We have proposed algorithms for

detecting voiced speech and identifying its pitch, estimating the speech active level and localizing sibilant phones. In this chapter, we focus on the estimation of the binary mask by exploiting the information extracted with the algorithms developed in previous chapters.

6.1 System overview

A block diagram of the binary mask estimation system is shown in Fig. 6.1, which illustrates the steps of training and binary mask estimation. The purpose of the estimation system is to determine binary-valued mask gain, $\hat{M}(t, f_e)$, for each time frame, t, and each frequency bin, f_e. In the training step (shown in the upper portion of Fig. 6.1), the inputs to the classifier training block for each time frame consists of a set of 145 features derived from the noisy training signal, $y(\tau)$, together with the corresponding binary-valued mask target, $M(t, f_e)$, derived from the clean speech, $s(\tau)$. In the mask estimation phase (shown in the lower portion of Fig. 6.1), the input consists only of the 145 features and the mask, $\hat{M}(t, f_e)$, is estimated by the classifier.

6.1.1 Feature estimation

The UTBM, whose definition was given in (2.3), preserves time-frequency regions whose energy is above a set threshold such that

$$\text{UTBM}(t, f) = \begin{cases} 1 & \text{if } S_{dB}(t, f) > L_{dB}(f) + \alpha + \text{LC}, \\ 0 & \text{otherwise.} \end{cases} \tag{6.1}$$

where α is a variable to adjust the power of the threshold function to the speech active level.

The selected feature set aims to provide information about the energy distribution of the speech. The feature set, as explained below in detail, contains information

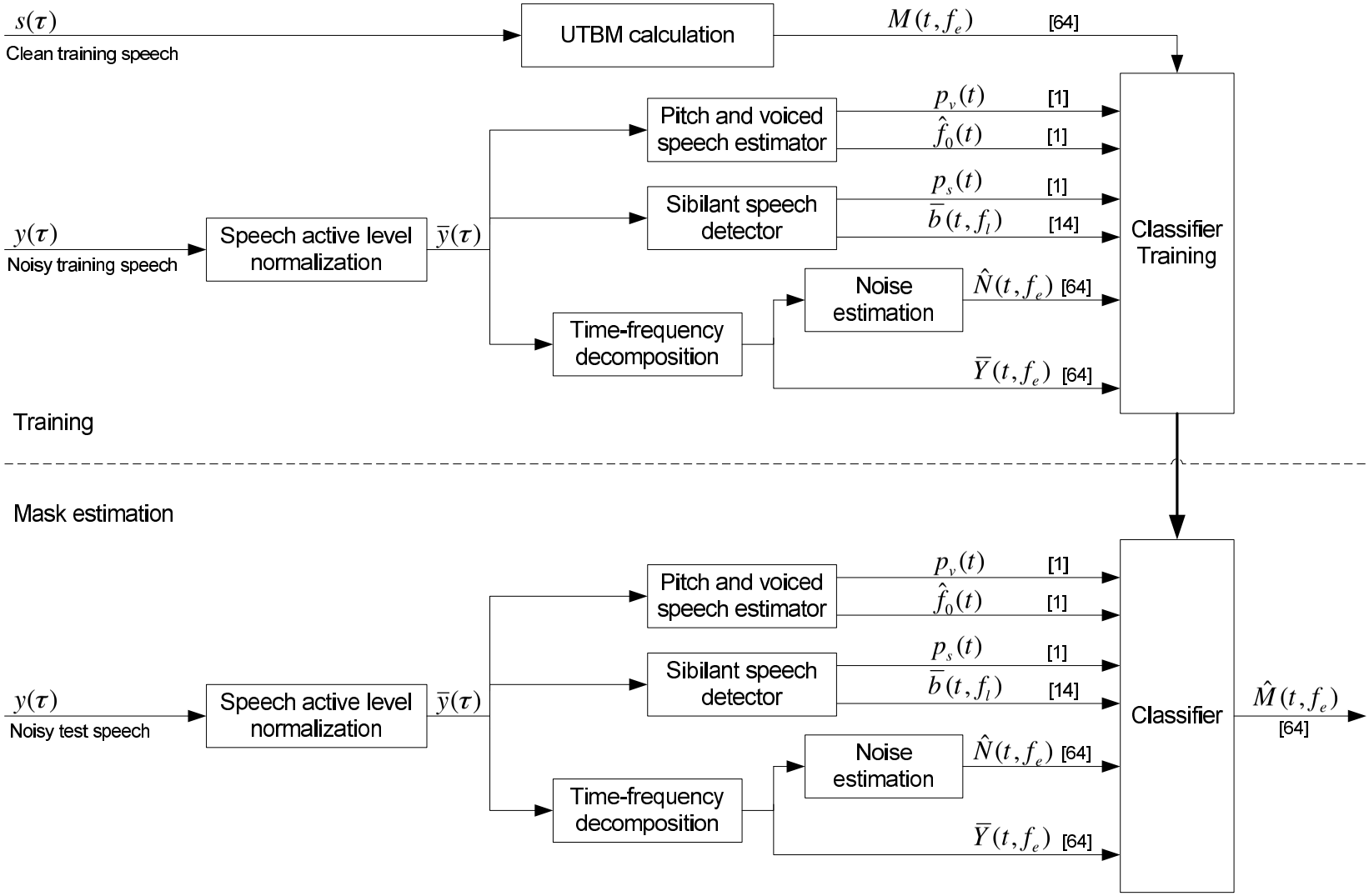


Figure 6.1: Block diagram of the mask estimation system proposed. Signal vector dimensions are indicated in brackets.

about the presence of voiced speech and its fundamental frequency and also about the presence of sibilant speech. Moreover, the feature set also includes the normalized noisy speech and a noise estimate, which provides information about the SNR energy at each time-frequency bin.

In the next subsections, we explain the various processing blocks in Fig. 6.1 which are used to extract the system parameters.

6.1.1.1 Level normalization

To ensure that classification is independent of the signal input level, the first step of the system is the power normalization of the speech component of the noisy speech signal, $y(\tau)$. The speech active level is estimated using the algorithm described in Chapter 4 and the normalization is performed such that:

$$\overline{y}(\tau) = 10^{-l_c/20} y(\tau) \tag{6.2}$$

where l_c is the estimated active speech level in dB and $\overline{y}(\tau)$ the normalized signal. In our experiments, the power normalization is performed over the entire duration of the utterance. If the input signal was long enough to include changes in the speech active level, the signal could be divided up into segments to perform this stage.

6.1.1.2 Pitch and voiced speech estimator

Most voiced speech energy is concentrated within the fundamental frequency and its harmonics. Therefore, identifying voiced speech segments and estimating their fundamental frequency makes it possible to locate high speech energy regions. In Chapter 3, we have described a robust method to identify voiced frames and estimate pitch in high levels of noise, the PEFAC algorithm. The PEFAC algorithm provides a fundamental frequency estimate at every time-frame, together with a probability of each time-frame containing voiced speech. Both features are used as inputs to the classifier:

$p_v(t)$ voiced speech probability for frame t.

$\hat{f}_0(t)$ estimated fundamental frequency for frame t.

6.1.1.3 Sibilant speech detector

Identifying time-frames which contain sibilant phones is important for the preservation of aperiodic speech energy at high frequencies. Furthermore, an estimation of the power spectrum of the sibilant phone would also help identifying the frequency bands containing most of the sibilant speech energy. In Chapter 5, we have proposed an algorithm for locating sibilant phones, which is used to extract:

$p_s(t)$ sibilant speech probability for frame t.

$\bar{b}(t, f_l)$ a 14-component vector for each time-frame, t, containing the normalized sibilant power spectrum estimate in 500 Hz bands from 1.5 kHz to 8 kHz.

6.1.1.4 Time-frequency decomposition

The inclusion of the normalized noisy speech periodogram and the noise estimation as parameters aids the mask estimation algorithm by providing information about the energy distribution across frequency of both speech and noise. The normalized input signal, $\bar{y}(\tau)$, is transformed into the time-frequency domain using the STFT. The spectrum of each frame is interpolated onto 64 ERB spaced frequency bands ranging from 40 Hz to 8 kHz. By using the ERB frequency scale, which is based on the equivalent rectangular bandwidths of the human ear, the frequency bands have a closer correspondence with the spectral resolution of the ear. The output of the time-frequency transformation, $\overline{Y}(t, f_e)$, is used as a parameter for the classifier together with a noise estimation, $\hat{N}(t, f_e)$:

$\overline{Y}(t, f_e)$ normalized periodogram of the noisy speech at time-frame t.

$\hat{N}(t, f_e)$ noise periodogram estimated at time-frame t using the algorithm described in [39] and the implementation provided in [15].

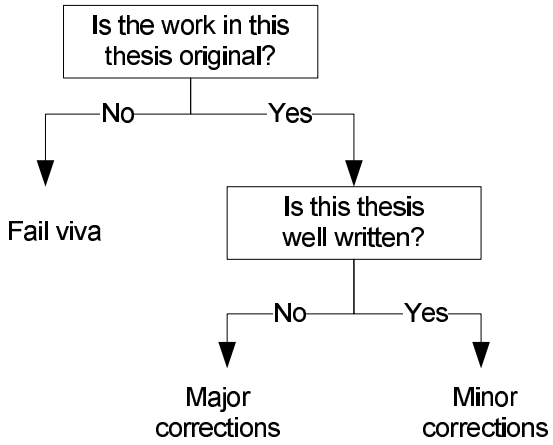


Figure 6.2: Binary tree example.

6.2 Classifier

The non-parametric CART approach [13] has been used to generate the mask. The CART approach is convenient to handle the heterogeneous nature of the mask estimation algorithm parameters and the complex relationship between them and the target mask. CART is a procedure that constructs a binary decision tree for predicting the output response or class from a set of input parameters taking discrete or continuous values. Each internal node compares one of the input parameters to a threshold and continues to a sub-branch of the tree according to the binary output. This process continues until a terminal node is reached, where prediction is performed by aggregating or averaging all the training data points which reach that node. A visual example of how a binary tree operates is shown in Fig. 6.2.

The CART approach can either be used for classification or regression. Classification trees provide a categorical value at each terminal node while regression trees provide a continuous output. For each internal node of the tree, the training process selects a feature to test and a threshold against which it is compared. These choices are made in order to minimize the average value of a misclassification function, $R(d)$. In the case of classification trees, the misclassification rate is estimated as

$$R(d) = \frac{1}{N_t} \sum_{n=1}^{N_t} \chi(d(\mathbf{p}_n) \neq c_n) \tag{6.3}$$

where N_t is the number of samples in the training set, $d(\cdot)$ is the binary-valued prediction function, $\mathbf{p}_n$ contains the input parameters for sample n, c_n is the sample class and $\chi(\cdot)$ is the indicator function, defined to be 1 if the statement is true and 0 otherwise. For regression trees, however, the goal of the training is to minimize the mean square error of the prediction, estimated using the training set as

$$R(d) = \frac{1}{N_t} \sum_{n=1}^{N_t} (x_n - d(\mathbf{p}_n))^2 \tag{6.4}$$

where x_n is the ground truth value. The predicted value at each terminal node u, $x(u)$, that minimizes $R(d)$ is the average of x_n for all cases within u

$$\bar{x}(u) = \frac{1}{N_u} \sum_{\mathbf{p}_n \in \mathrm{u}} x_n \tag{6.5}$$

Although in our case the ground truth provides binary values, $M(t, f_e)$, it is not necessary for the CART output to be binary. We train, therefore, a regression tree, whose output can later be converted to binary values. As we have seen in (6.5), the continuous output of the regression tree is the average of the ground truth values within each terminal node. The binary ground truth values in our application are 0 and 1, which means that the output of the regression tree can be interpreted as the probability that the corresponding time-frequency bin energy lies above the UTBM energy threshold. The estimated probability can than be converted to a binary value by setting a threshold.

6.3 Experiments

The training set and the test set from the TIMIT database [37] were respectively used for training and testing the algorithm. The training and testing sets of the TIMIT database contains different speakers. Most of the sentence texts are also different

between the two sets, with only 20% overlap. The sampling frequency of the speech material is 16 kHz. To determine the ground truth for the binary mask, the UTBM was calculated for each utterance on the clean speech signal. The LC parameter was set to -5 dB, which, as was shown in Fig. 2.5, provides the best intelligibility results.

The STFT used a Hamming analysis window of 90 ms duration and an inter-frame time increment of 22.5 ms. The length of the window was chosen so that speech harmonics could be resolved for all f_0 values. The inter-frame time increment was set to achieve perfect signal reconstruction when the Hamming window was used for both analysis and synthesis. The spectrum of each frame was interpolated onto 64 ERB spaced frequency bands ranging from 40 Hz to 8 kHz. We expect this frequency resolution to provide good intelligibility performance since, as was shown in Fig. 2.3, high intelligibility is obtained above 16 frequency bands.

To train the regression tree we used 300 TIMIT utterances from the training set mixed with 12 noises from the RSG-10 database [131]. The noise types included: factory, babble, buccaneer and F16 fighter jets, engine room, operation room, HF radio channel, leopard and M109 tank, pink, car and white. The power spectrogram of these noise types is provided in Appendix A. The calculation of the SNR used ITU-T P.56 [68, 15] for the speech level and unweighed power for the noise. SNRs from –5 to +9 dB in 2 dB steps were used. A separate regression tree was trained for each of the 64 frequency bands. The input to each regression tree contained the entire feature vector, rather than just its local frequency components.

6.4 Results

The performance of the mask estimation was evaluated using 100 utterances from the test set of the TIMIT database mixed with noises from both the RSG-10 database [131] and the ITU-T P.501 standard [69]. SNRs from -5 to $+10$ dB were used for evaluation. This range was chosen because at SNRs above $+10$ dB, speech is fully intelligible whereas below -5 dB SNR the speech signal is so degraded that reliable feature extraction is not possible.

A visual example of the performance of the algorithm can be found in Fig. 6.3. A speech utterance containing the sentence "She had your dark suit in greasy wash water all year" corrupted with white noise at -5 dB SNR is shown in Fig. 6.3(a). By applying the proposed method, we estimated the mask shown in Fig. 6.3(b). The ground truth of the algorithm, the UTBM, is illustrated in Fig. 6.3(c). Figure 6.3(d), (e) and (f) correspond to the clean speech, segregated speech with the estimated mask and segregated speech with the UTBM respectively. We can observe how in Fig. 6.3(e) we are able to extract most speech power, Fig. 6.3(d), while greatly reducing the background noise. The classifier has accurately identified the major features of the UTBM with the exception of the relatively weak sibilant at $t = 2.7$ s. However, some noise has been introduced in the segregated speech, which is especially visible at high frequencies. The energy in the low frequencies, concentrated in the fundamental frequency and its harmonics, is well-preserved with little distortion.

Intelligibility evaluation of the results was achieved using the intrusive measure STOI [134], which has shown good intelligibility correlation for binary masks. This objective algorithm provides a value between 0 and 1 which has been shown to have a monotonic relation with the subjective speech-intelligibility as discussed in Section 1.4.2 [134].

6.4.1 Continuous versus binary-valued masks

First of all, we evaluated the performance of the continuous versus the binary gain mask. For that, we set a probability threshold, p_b, above which the mask is set to 1

$$\hat{M}_B(t,f) = \begin{cases} 1 & \text{if } \hat{M}_C(t,f) > p_b, \\ 0 & \text{otherwise.} \end{cases} \tag{6.6}$$

where $\hat{M}_B(t,f)$ and $\hat{M}_C(t,f)$ represent the binary and continuous gain mask respectively. We evaluated the results for different p_b on 100 utterances from the test set on the same noise types used for training. It was found that the highest STOI values

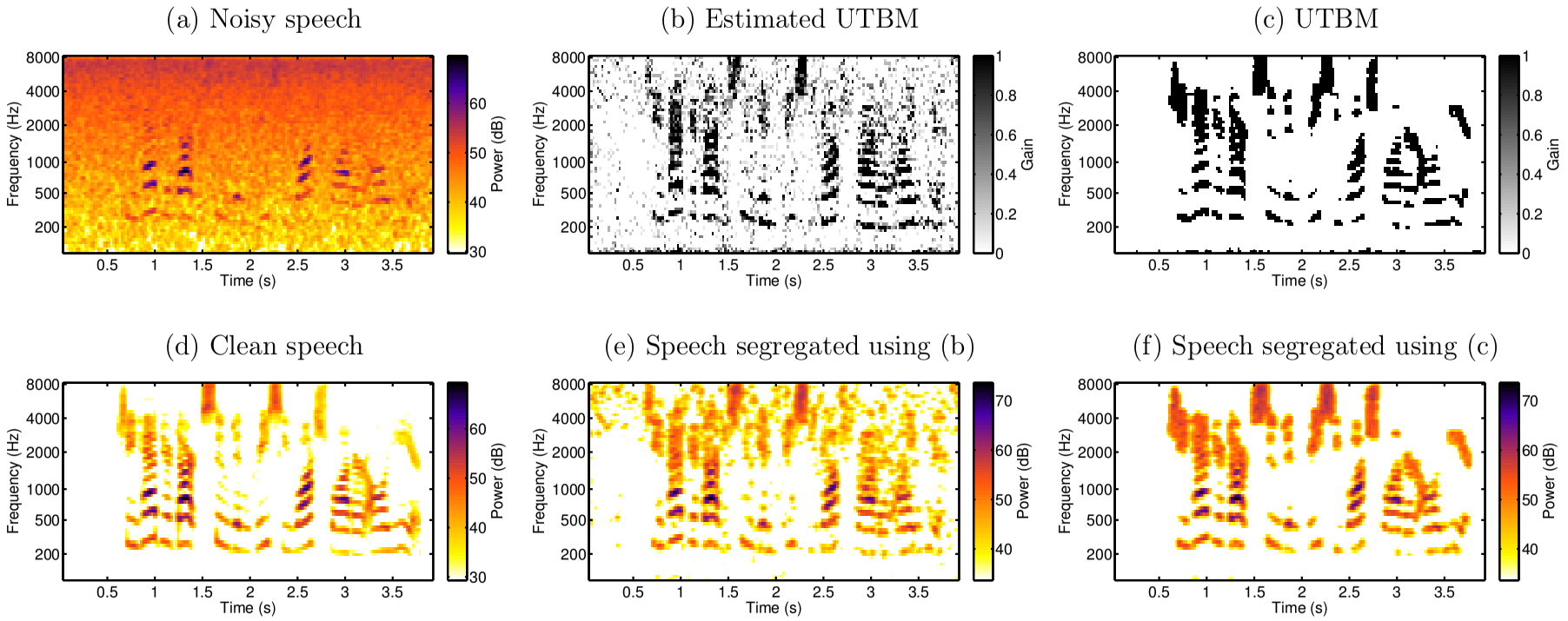


Figure 6.3: (a) Speech utterance from the test set of the TIMIT database containing the sentence "She had your dark suit in greasy wash water all year" corrupted with white noise at −5 dB SNR; (b) estimated mask using the proposed algorithms from the noisy speech in (a); (c) ground truth mask – the UTBM; (d) clean speech utterance, (e) segregated speech from the noisy speech in (a) by using the estimated UTBM shown in (b); and (f) segregated speech using the ground truth mask from (c).

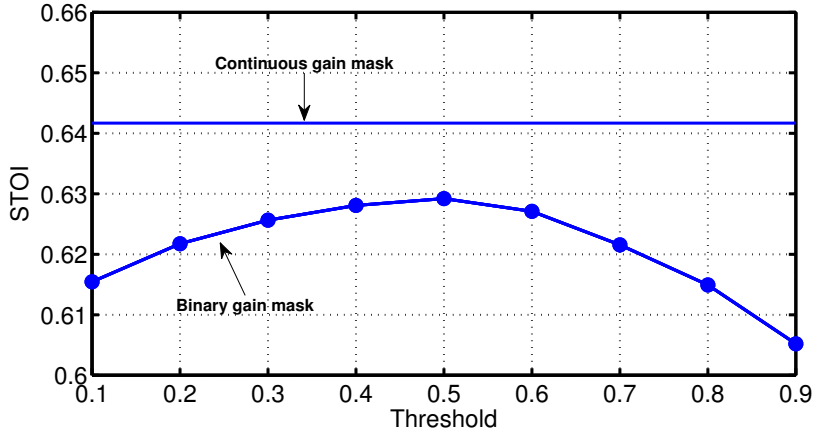


Figure 6.4: STOI values for the continuous gain mask and the different binary masks for factory noise at -5 dB SNR. The STOI values are the average over 100 utterances.

were achieved when using the continuous gain mask. An example for factory noise at -5 dB SNR is shown in Fig. (6.4), where the STOI value for the continuous gain mask outperforms the binary mask estimated for any tested threshold.

6.4.2 Evaluation on seen noise types

For performance comparison, the log-MMSE algorithm [32], and the spectral subtraction [11] speech enhancement algorithm were used. In both cases, the noise was estimated using the algorithm described in [39], the same one used for the proposed binary mask estimation. The results for the noise types used in training are shown in Table 6.1 averaged over 100 utterances from the test set. Note that although the noise types were the same as in training, the actual noise samples used were different in every test. The STOI performance of the oracle binary mask, the UTBM, is also shown. In the table we observe how the STOI values for both the MMSE and the spectral subtraction methods are very similar to that of the noisy speech. This is consistent with the results shown in previous studies [5, 66, 96] where it was found that none of the evaluated algorithms was able to increase speech intelligibility significantly.

The proposed mask-based algorithm, as seen in Table 6.1, is able to increase the STOI values at low SNR while preserving the high STOI values at high SNRs, where

Table 6.1: STOI results for different speech enhancement algorithms on the noise types used for training the proposed algorithm. MMSE corresponds to the log-spectral amplitude MMSE approach [32], SS corresponds to spectral subtraction [11]. Each entry gives the average STOI over 100 utterances from the TIMIT test set.

		STOI values			
SNR (dB)		-5	0	5	10
Babble noise	Noisy	0.50	0.62	0.74	0.83
	MMSE	0.47	0.60	0.73	0.83
	SS	0.46	0.60	0.73	0.84
	Proposed	**0.62**	**0.71**	**0.79**	**0.85**
	Oracle mask	*0.77*	*0.81*	*0.85*	*0.87*
Factory noise	Noisy	0.51	0.63	0.75	0.85
	MMSE	0.50	0.63	0.75	0.84
	SS	0.49	0.63	0.75	0.85
	Proposed	**0.64**	**0.74**	**0.82**	**0.87**
	Oracle mask	*0.78*	*0.82*	*0.86*	*0.88*
Pink noise	Noisy	0.54	0.66	0.77	0.87
	MMSE	0.56	0.67	0.78	0.87
	SS	0.54	0.67	0.78	0.87
	Proposed	**0.67**	**0.76**	**0.84**	**0.88**
	Oracle mask	*0.79*	*0.83*	*0.86*	*0.89*
Engine room noise	Noisy	0.55	0.67	0.78	0.88
	MMSE	0.60	0.72	0.82	**0.90**
	SS	0.60	0.72	0.83	**0.90**
	Proposed	**0.69**	**0.78**	**0.85**	0.88
	Oracle mask	*0.79*	*0.83*	*0.87*	*0.89*
HF radio channel noise	Noisy	0.55	0.67	0.79	0.88
	MMSE	0.57	0.69	0.80	0.89
	SS	0.55	0.69	0.81	**0.90**
	Proposed	**0.70**	**0.78**	**0.84**	0.88
	Oracle mask	*0.80*	*0.85*	*0.88*	*0.90*
White noise	Noisy	0.59	0.71	0.82	0.90
	MMSE	0.60	0.72	0.82	0.90
	SS	0.58	0.71	0.83	**0.91**
	Proposed	**0.70**	**0.79**	**0.85**	0.89
	Oracle mask	*0.82*	*0.86*	*0.88*	*0.90*
Leopard tank noise	Noisy	0.76	0.80	**0.84**	0.87
	MMSE	0.75	0.80	**0.84**	**0.88**
	SS	0.75	0.80	**0.84**	**0.88**
	Proposed	**0.78**	**0.82**	**0.84**	0.87
	Oracle mask	*0.82*	*0.83*	*0.85*	*0.87*
Volvo noise	Noisy	**0.88**	**0.92**	0.94	**0.97**
	MMSE	0.87	0.91	**0.95**	**0.97**
	SS	0.87	0.91	0.94	**0.97**
	Proposed	0.86	0.88	0.89	0.90
	Oracle mask	*0.87*	*0.88*	*0.89*	*0.90*
Overall	Noisy	0.61	0.71	0.80	0.88
	MMSE	0.61	0.72	0.81	0.88
	SS	0.60	0.72	0.81	**0.89**
	Proposed	**0.71**	**0.78**	**0.84**	0.88
	Oracle mask	*0.80*	*0.85*	*0.87*	*0.89*

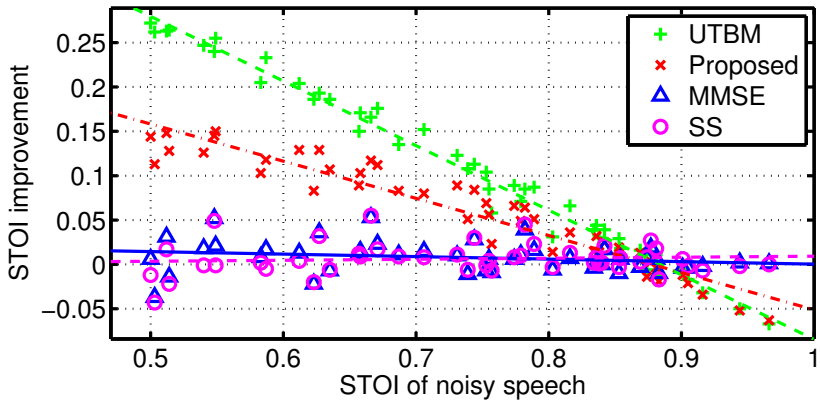


Figure 6.5: STOI improvement using the proposed algorithm versus the STOI of the noisy signal for seen noise types. The STOI values are the average over 100 utterances. The straight lines in the figure are least-squares linear fits to the data points.

the speech is already intelligible. In terms of intelligibility, the most damaging noise types are those whose energy is distributed across the same frequencies as the speech signal, such as babble noise (see Appendix A). In this situation, the estimated binary mask is able to improve the noisy STOI value substantially, and, for example, for HF radio channel noise, the proposed algorithm can increase STOI by as much as 0.15 at -5 dB SNR. On the other hand, noise types whose energy is concentrated within a relatively narrow band of frequencies, such as volvo noise or leopard noise, have less effect on intelligibility and the STOI value remain high even at low SNRs; the proposed algorithm, therefore, does not change substantially the STOI value. On average, for seen noises, the STOI value is increased by 0.10 at -5 dB SNR with the proposed algorithm while the increment using the oracle mask is 0.19. At $+10$ dB SNR, when the speech intelligibility is high, both the estimated and the oracle mask have almost no impact on the STOI value.

It can be observed in Table 6.1 that the higher improvements come when the STOI value is low. Therefore, it is instructive to plot the STOI improvement versus the STOI of the noisy speech. The results obtained for the three evaluated algorithms and the oracle mask are shown in Fig. 6.5 for all 12 seen noise types used for the training. The different markers on the figure correspond to the average STOI improvement over 100 test utterances and the straight lines represent the least-squares linear fit to the

data points for each speech enhancement method. On average, both the MMSE (blue triangles, solid blue line) and spectral subtraction (pink circles, pink dashed line) algorithms do not change substantially the input STOI value. However, in Fig. 6.5 we can observe how the proposed mask is consistently able to improve the STOI of noisy speech for values below 0.8. It is worth noting that when the STOI value is above 0.7, the speech intelligibility is very high [134] (see Fig. 1.9) and the the impact on intelligibility of small changes to the STOI score will be insignificant. In particular, for noisy speech STOI values above 0.9, the small decreases in STOI introduced by our proposed algorithm will not significantly affect intelligibility. The oracle mask has similar performance to the estimated mask for high STOI values while providing a STOI improvement of approximately 0.25 for an initial STOI of 0.5 versus the 0.15 improvement of the estimated mask. When the noisy speech STOI value is below 0.5, the original speech is too corrupted to extract reliable information and the proposed algorithm will not improve the predicted intelligibility.

6.4.3 Evaluation on unseen noise types

The performance of the proposed algorithm on six unseen noise types is shown in Table 6.2 together with the results obtained for the log-MMSE algorithm [32], and the spectral subtraction [11] algorithm. In Table 6.2 we can observe how the estimated mask does not increase the STOI values as much as in Table 6.1, with an average STOI increment of 0.04 at -5 dB SNR. For higher SNRs, the proposed algorithm can slightly degrade the STOI, which changes from an average STOI value of 0.89 for noisy speech to an average STOI value of 0.87 for the processed speech using the proposed algorithm at 10 dB SNR. Overall, the STOI value for unseen noise types at -5 dB SNR is higher than for seen noise types. The reason behind this is that the majority of the unseen noise types belong to the database from the ITU-T P.501 standard [69], and, as we can observe in the spectrograms provided in Appendix A, most of their energy is concentrated at low frequencies and speech information is preserved. As expected, the MMSE and spectral subtraction algorithms so not significantly change the STOI value at any input SNR.

Table 6.2: STOI results for different speech enhancement algorithms on unseen noise types. MMSE corresponds to the log-spectral amplitude MMSE approach [32], SS corresponds to spectral subtraction [11]. Each entry gives the average STOI over 100 utterances from the TIMIT test set.

		STOI values			
SNR (dB)		-5	0	5	10
Cafeteria noise	Noisy	**0.53**	0.64	0.75	**0.84**
	MMSE	0.49	0.61	0.73	0.83
	SS	0.49	0.62	0.74	0.83
	Proposed	**0.53**	**0.66**	**0.77**	0.83
	Oracle mask	*0.77*	*0.81*	*0.85*	*0.87*
Car production hall noise	Noisy	0.67	0.77	0.85	**0.91**
	MMSE	0.68	0.77	0.85	**0.91**
	SS	0.67	0.77	0.85	**0.91**
	Proposed	**0.75**	**0.81**	**0.86**	0.89
	Oracle mask	*0.82*	*0.85*	*0.87*	*0.89*
Restaurant noise	Noisy	0.70	0.78	**0.84**	**0.89**
	MMSE	0.68	0.76	0.83	0.88
	SS	0.68	0.77	0.83	0.89
	Proposed	**0.73**	**0.80**	**0.84**	0.87
	Oracle mask	*0.81*	*0.84*	*0.86*	*0.89*
Office noise	Noisy	0.70	0.78	**0.84**	**0.90**
	MMSE	0.70	0.78	**0.84**	**0.90**
	SS	0.70	0.78	**0.84**	**0.90**
	Proposed	**0.74**	**0.80**	**0.84**	0.88
	Oracle mask	*0.81*	*0.84*	*0.86*	*0.88*
Street noise	Noisy	0.70	0.79	**0.85**	0.90
	MMSE	0.71	0.79	**0.85**	**0.91**
	SS	0.71	0.79	**0.85**	**0.91**
	Proposed	**0.75**	**0.81**	**0.85**	0.88
	Oracle mask	*0.82*	*0.85*	*0.87*	*0.88*
Railway station noise	Noisy	0.72	0.78	**0.84**	**0.88**
	MMSE	0.71	0.78	0.83	**0.88**
	SS	0.71	0.78	0.83	**0.88**
	Proposed	**0.75**	**0.80**	**0.84**	0.87
	Oracle mask	*0.81*	*0.83*	*0.86*	*0.88*
Overall	Noisy	0.67	0.76	**0.83**	**0.89**
	MMSE	0.66	0.75	0.82	0.88
	SS	0.66	0.75	0.82	**0.89**
	Proposed	**0.71**	**0.78**	**0.83**	0.87
	Oracle mask	*0.81*	*0.83*	*0.86*	*0.88*

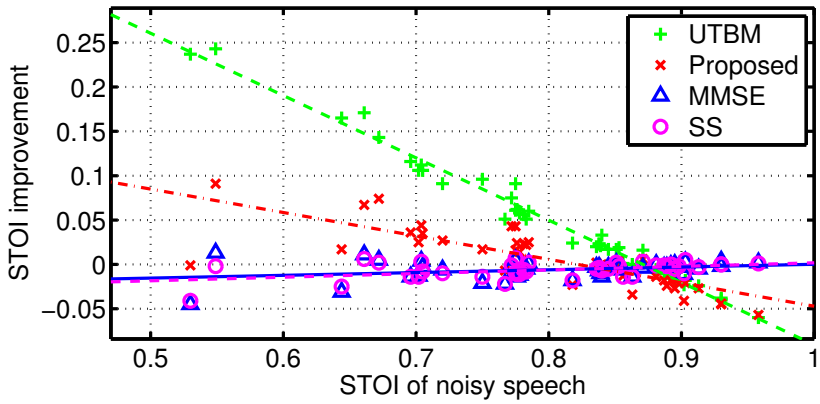


Figure 6.6: STOI improvement using the proposed algorithm versus the STOI of the noisy signal for unseen noise types. The STOI values are the average over 100 utterances. The straight lines in the figure are least-squares linear fits to the data points.

The STOI improvement versus its input value is shown in Fig. 6.6 calculated over the 10 noise types of the ITU-T P.501 standard [69] and 3 noise types from RSG-10 database [131] which were not used for training. We see that the oracle mask results (indicated by green +) are very similar to those shown in Fig. 6.5 for the noise types used in training. Due to the limited number of noises used for the training, our algorithm does not generalise well on all types of unseen noise and the results (indicated by red x) are not as consistent as for the seen noises. However, the proposed algorithm trend, indicated by the linear fit to the data (red dash-dot line), is to increase the STOI value when the input STOI is low, although the average increase is about half that obtained on seen noise types. The MMSE and SS algorithms do not substantially modify the input STOI value at any SNR, which is consistent with the results found in [65].

6.5 Summary

In this chapter we have presented a mask-based algorithm that is able to increase the predicted intelligibility calculated using the objective STOI measure. We extracted 145 features per frame from the noisy speech using previously developed algorithms and trained a regression tree for each frequency band using the Universal Target Bin-

ary Mask (UTBM) as a target. Utterances from the TIMIT training set and noise types from the RSG-10 database [131] were used to train the regression trees. The proposed mask estimation algorithm was evaluated on the TIMIT test set with a variety of noise types, some of which had been previously used in the training stage. We conclude that the proposed algorithm is able to increase the predicted intelligibility for noises seen in the training while maintaining or increasing the predicted intelligibility on unseen noise types.

Chapter 7

Conclusions

7.1 Thesis summary

Speech signals can be degraded in many ways during their acquisition in noisy environments and they can also be further degraded in the electronic domain. Serious signal degradation, however, is most commonly caused by noise from unwanted acoustic sources in the environment, which may affect the speech quality and/or intelligibility of the wanted signal. In this thesis, we have focused on the enhancement of single-channel speech signals that have been corrupted by levels of additive noise that are high enough to affect the intelligibility of the speech.

Numerous approaches for single-channel speech enhancement, mainly driven by telecommunications companies and hearing aid manufacturers, have been developed over many years. The majority of algorithms perform the enhancement in a transform domain in which both speech and noise signals are sparse. The time-frequency domain is the dominant domain for speech enhancement procedures. There are several approaches which enhance the signal using time-frequency gain modification, such as spectral subtraction or MMSE-based algorithms. Although most approaches aim to estimate the clean speech by applying a continuous gain, the more recently proposed time-frequency binary mask approach aims to retain important speech information by using binary gain values.

Several studies [5, 66, 96] have evaluated the impact on quality and intelligibility

of state-of-the-art speech enhancement algorithms. The results show that in most cases intelligibility gets worse although perceived quality may improve. Although no current approach has been able to improve speech intelligibility, several studies [18, 146] have shown the potential of time-frequency binary masks in this task.

7.1.1 Time-frequency binary masks

Time-frequency binary masks aim to identify regions of the time-frequency plane that contain information from the target sound. The original goal of binary mask estimation was to identify the regions where the SNR was higher than 0 dB [144, 98]. Later research [146, 86], however, shows that the optimum SNR threshold in terms of intelligibility depends on the input SNR. In recent years, an alternative goal has been proposed [86], which aims at preserving time-frequency regions with high speech energy.

In Chapter 2, we provided a detailed explanation of the different binary mask targets, the Ideal Binary Mask (IBM) and the Target Binary Mask (TBM). The IBM bases its decision on the SNR while the TBM bases its decision on the LTASS of the speaker. We proposed a variation of the TBM, the Universal Target Binary Mask (UTBM) and we have shown a similar predicted intelligibility performance to that of the TBM while removing dependency on the speaker.

Based on the idea that a binary mask based only on the speech is possible, our approach to the binary mask estimation problem aims to preserve high speech energy independently of the noise present. Accordingly, we have in this thesis developed methods for detecting the presence of voiced and sibilant speech components in a noisy speech signal and for characterizing them in the time frequency domain. In addition we have developed an algorithm for estimating the active level of a speech signal even when high levels of noise are present.

7.1.2 Voicing and pitch detection

The PEFAC algorithm, described in Chapter 3, is both a fundamental frequency estimator and a voiced speech detector which has robust performance at low SNRs. The first stage of the algorithm is a spectral normalization designed (i) to remove the dependency on the overall speech level, (ii) to compensate for the channel response and (iii) to attenuate narrowband noise components. The second stage is the convolution in the frequency domain with a pitch estimation filter that rejects broadband noise that has a smooth power spectrum. Dynamic programming is then used to impose soft temporal continuity constraints by selecting between pitch candidates in each frame. For voiced speech detection, two GMMs are trained on voiced and unvoiced frames respectively and the likelihood ratio of the two models is used to classify each frame.

The PEFAC algorithm was evaluated on different speech corpuses with a variety of noise types and consistently outperformed other widely used algorithms. It was also evaluated on reverberant speech without a degradation in performance. The voiced activity detector is able to discriminate between voiced and unvoiced with a lower overall error rate than the detectors implemented by other competing algorithms.

7.1.3 Speech active level estimation

In Chapter 4, we proposed a new method for estimating the speech active level in high levels of noise. The method combines the ITU-T Recommendation P.56 [68] with novel harmonic summation approach. The harmonic summation approach extracts the energy contained at the fundamental frequency and its harmonics in order to estimate the speech energy. The final speech active level estimate is calculated as a linear combination of the ITU-T P.56 estimate, which is more accurate at high SNRs, and the harmonic summation method estimate, which provides a reliable estimation of the speech active level even at poor SNRs. The algorithm has been evaluated on the TIMIT test set with a range of noise types and extends by more than 7 dB the range of SNRs for which reliable estimation is possible.

7.1.4 Sibilant speech detection

In order to locate the presence of aperiodic speech energy at high frequencies we presented in Chapter 5 a sibilant speech detection algorithm robust to high levels noise. Rather than identifying explicit onsets and offsets, a sustained increase in energy during the sibilant is instead detected. The algorithm, which does not rely on voicing detection, comprises a sibilant mean power estimation stage based on a maximum likelihood approach followed by a classification stage in which the likelihood ratio of two GMMs, one for sibilant speech and one for non-sibilant speech, is used. The algorithm has been evaluated on the TIMIT test set over a range of noise types and SNRs and consistently achieved over 80% classification accuracy for positive SNRs.

7.1.5 Mask estimation

In Chapter 6, we used a machine learning approach to estimate the UTBM. The parameters used for the estimation are extracted from the noisy speech using the previously developed algorithms together with a noise estimate. A regression tree is trained for each frequency band on a range of noise types. The proposed mask estimation algorithm was evaluated on the TIMIT test set with a variety of noise types, some of which had been previously used in the training stage and the predicted intelligibility was calculated using the objective algorithm STOI. While no other evaluated speech enhancement technique was able to considerably improve the predicted intelligibility; our algorithm, for seen noise types, can improve substantially the STOI values for low SNRs while maintaining them at high SNRs. On average, for unseen noise types, the estimated binary mask still gave an improvement, although it was smaller than for noise types included in the training data.

7.2 Future work

There are several directions in which further work can be approached. We can either focus on the improvement of each of the proposed algorithms, on the development of new ones to extract more speech information or on the enhancement of the mask estimate.

7.2.1 Voicing and pitch detection

There are different ways in which the PEFAC algorithm performance could be further improved. As the active level estimation algorithm performance depends on the accuracy on both voicing detection and pitch estimation, any improvement to the PEFAC algorithm would also benefit its performance. Future work to improve the PEFAC algorithm could include the application of temporal continuity constraints to the voicing probability estimate. The voiced/unvoiced classifier provides a probability estimate for each time-frame independently of neighbouring information. We could take advantage of the knowledge about the average duration of voiced speech segments and the separation between them to improve the final probability estimate.

Recent research [40] has shown the valuable information the speech phase contains. Within the PEFAC algorithm, it would be possible to use phase consistency to distinguish between true harmonic peaks and spurious peaks.

7.2.2 Speech active level estimation

The speech active level method identifies the voiced speech segments of a speech signal and calculates the speech active level from the energy in the fundamental frequency harmonics. However, for a practical speech active level estimation operating on continuous speech the algorithm would need to be modified. There is a need to determine a window length to calculate the speech active level over and also to ensure that the system works properly when no speech is present.

7.2.3 Unvoiced speech detection

The identification of aperiodic noise components could also benefit from further research. The sibilant detector described in this thesis classifies each frame individually; it is possible that its classification accuracy could be further improved by applying temporal constraints to the classification decisions.

An important class of speech sound that we do not currently detect explicitly is stop consonants and, in particular, plosive stops. The sibilant detection algorithm could be adapted to estimate the presence of stops by accommodating the duration of the sustained increase in energy to that of stop consonants and by retraining the classifier. It is worth noting that some of the false alarms of the sibilant detection algorithm were caused by this type of consonants.

7.2.4 Mask estimation

The classification features for mask estimation include information about voiced speech, sibilant speech and the energy distribution in frequency of the noisy speech and the estimated noise. The inclusion of new classification features containing information about types of phonemes such as stops or non-sibilant fricatives could further improve the performance of the mask estimation algorithm.

Although the CART approach has shown to provide a good performance, other appropriate machine learning techniques could be investigated, such as SVMs. Furthermore, in order to improve the mask generalization to unseen noise conditions, more noise types may be used in the training stage.

The output of the machine learning could be further improved by taking advantage of neighbouring time-frequency information. It can be seen from Fig. 6.3 that the estimated mask includes isolated false positive cells. The occurrence of these could be reduced by applying continuity constraints in the time and/or frequency directions or by including in the parameter vector the classifier outputs from nearby time-frequency bins.

Despite the possible improvements in the mask estimation stage, we believe that

one of the major limitations of the proposed mask estimation method is not the machine learning technique or the input parameters, but rather the UTBM that we have used as the ground truth when training the classifier. There is a need to better understand what are the key elements of the speech signal which makes it intelligible. Only by understanding this process can we set an appropriate target for the mask estimation problem.

References

[1] J. Allen and L. Radiner, "A unified approach to short-time Fourier analysis and synthesis," *Proc. IEEE*, vol. 65, no. 11, pp. 1558–1564, 1977.

[2] J. B. Allen, "Short term spectral analysis, synthesis, and modification by discrete Fourier transform," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 25, no. 3, pp. 235–238, Jun. 1977.

[3] ANSI, "Methods for the calculation of the articulation index," American National Standards Institute, New York, ANSI Standard ANSI S3.5–1969, 1969.

[4] ——, "Methods for the calculation of the speech intelligibility index," American National Standards Institute, ANSI Standard S3.5–1997 (R2007), 1997.

[5] M. Anzalone, L. Calandruccio, K. Doherty, and L. Carney, "Determination of the potential benefit of time-frequency gain manipulation," *Ear & Hearing*, vol. 27, pp. 480–492, 2006.

[6] C. M. Ayer, M. J. Hunt, and M. Brookes, "A discriminatively derived linear transform for improved speech recognition," in *Proc. European Signal Processing Conf. (EUSIPCO)*, vol. 1, Berlin, Sep. 1993, pp. 583–586.

[7] W. Bauer and W. Blankenship, "Dyptrack–a noise-tolerant pitch tracker," Dept. of Defence (NSA), USA, Unclassified Report NASL-S-210, 1974.

[8] M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 4, 1979, pp. 208–211.

[9] R. W. Berry, "Speech-volume measurements on telephone circuits," *Proc. Institution of Electrical Engineers*, vol. 118, no. 2, pp. 335–338, Feb. 1971.

[10] P. Boersma and D. Weenink, "Praat: doing phonetics by computer (Version 5.3.23)," http://www.praat.org/, 2012.

[11] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-27, no. 2, pp. 113–120, Apr. 1979.

[12] M. S. Brandstein and D. B. Ward, Eds., *Microphone Arrays: Signal Processing Techniques and Applications*. Berlin, Germany: Springer-Verlag, 2001.

[13] L. Breiman, J. Friedman, R. A. Olshen, and C. J. Stone, *Classification and Regression Trees*. Chapman and Hall, Jan. 1984.

[14] I. Brons, R. Houben, and W. A. Dreschler, "Perceptual effects of noise reduction by time-frequency masking of noisy speech," vol. 132, no. 4, pp. 2690–2699, Jun. 2012.

[15] M. Brookes, "VOICEBOX: A speech processing toolbox for MATLAB," http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html, 1997–2013.

[16] M. Brookes, N. D. Gaubitch, M. Huckvale, and P. A. Naylor, "Speech cleaning literature review," Tech. Rep. CTR-2, Feb. 2008. [Online]. Available: www.clear-labs.com/Tutorial-LitReview/index.html

[17] J. C. Brown, "Musical fundamental frequency tracking using a pattern recognition method," *J. Acoust. Soc. Am.*, vol. 92, no. 3, pp. 1394–1402, Sep. 1992.

[18] D. S. Brungart, P. S. Chang, B. D. Simpson, and D. L. Wang, "Isolating the energetic component of speech-on-speech masking with ideal time-frequency segregation," *J. Acoust. Soc. Am.*, vol. 120, pp. 4007–4018, 2006.

[19] D. Byrne, H. Dillon, K. Tran, S. Arlinger, K. Wilbraham, R. Cox, B. Hayerman, R. Hetu, J. Kei, C. Lui, J. Kiessling, M. N. Kotby, N. H. A. Nasser, W. A. H. E.

Kholy, Y. Nakanishi, H. Oyer, R. Powell, D. Stephens, , T. Sirimanna, G. Tavartkiladze, G. I. Frolenkov, S. Westerman, and C. Ludvigsen, "An international comparison of long-term average speech spectra," *J. Acoust. Soc. Am.*, vol. 96, no. 4, pp. 2108–2120, Oct. 1994.

[20] D. Chan, A. Fourcin, D. Gibbon, B. Granstrom, M. Huckvale, G. Kokkinakis, K. Kvale, L. Lamel, B. Lindberg, A. Moreno, J. Mouropoulos, F. Senia, I. Trancoso, C. Veld, and J. Zeiliger, "EUROM - a spoken language resource for the EU," in *Proc. European Conf. on Speech Communication and Technology*, Sep. 1995, pp. 867–870.

[21] M. G. Christensen and A. Jakobsson, *Multi-Pitch Estimation*, ser. Synthesis Lectures on Speech and Audio Processing. Morgan & Claypool, 2009.

[22] M. G. Christensen, A. Jakobsson, and S. H. Jensen, "Joint high-resolution fundamental frequency and order estimation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 5, pp. 1635–1644, 2007.

[23] M. Cooke, P. Green, L. Josifovski, and A. Vizinho, "Robust automatic speech recognition with missing and unreliable acoustic data," *Speech Communication*, vol. 34, pp. 267–285, 2001.

[24] M. Cooke, "Modelling auditory processing and organisation," Ph.D. dissertation, University of Sheffield, 1993.

[25] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continously spoken sentences," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 28, no. 4, pp. 357–366, Aug. 1980.

[26] A. de Cheveigne and H. Kawahara, "YIN, a fundamental frequency estimator for speech and music," *J. Acoust. Soc. Am.*, vol. 111, no. 4, pp. 1917–1930, Apr. 2002.

[27] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from

incomplete data via the EM algorithm," *Journal Royal Statistical Society, Series B*, vol. 39, no. 1, pp. 1–38, 1977.

[28] M. Dendrinos, S. Bakamidis, and G. Carayannis, "Speech enhancement from noise: a regenerative approach," *Speech Communication*, vol. 10, no. 1, pp. 45–67, Feb. 1991.

[29] L. Dolansky and P. Tjernlund, "On certain irregularities of voiced-speech waveforms," *IEEE Trans. Audio Electroacoust.*, vol. 16, no. 1, pp. 51–56, Mar. 1968.

[30] J. Droppo and A. Acero, "Maximum a posteriori pitch tracking," in *Proc. Intl. Conf. on Spoken Lang. Processing (ICSLP)*, 1998.

[31] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 32, no. 6, pp. 1109–1121, Dec. 1984.

[32] ——, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 33, no. 2, pp. 443–445, 1985.

[33] Y. Ephraim and H. L. Van Trees, "A signal subspace approach for speech enhancement," *IEEE Trans. Speech Audio Process.*, vol. 3, no. 4, pp. 251–266, Jul. 1995.

[34] N. R. French and J. C. Steinberg, "Factors governing the intelligibility of speech sounds," *J. Acoust. Soc. Am.*, vol. 19, no. 1, pp. 90–119, 1947.

[35] S. Furui, "Cepstral analyisis technique for automatic speaker verification," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 29, no. 2, pp. 254–272, Apr. 1981.

[36] M. J. F. Gales and S. J. Young, "Robust continuous speech recognition using parallel model combination," *IEEE Trans. Speech Audio Process.*, vol. 4, pp. 352–359, Sep. 1996.

[37] J. S. Garofolo, "Getting started with the DARPA TIMIT CD-ROM: An acoustic phonetic continuous speech database," National Institute of Standards and Technology (NIST), Gaithersburg, Maryland, Technical Report, Dec. 1988.

[38] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, N. L. Dahlgren, and V. Zue, "TIMIT acoustic-phonetic continuous speech corpus," Linguistic Data Consortium, Philadelphia, Corpus LDC93S1, 1993.

[39] T. Gerkmann and R. C. Hendriks, "Unbiased MMSE-based noise power estimation with low complexity and low tracking delay," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 4, pp. 1383 –1393, May 2012.

[40] T. Gerkmann and M. Krawczyk, "MMSE-optimal spectral amplitude estimation given the STFT-phase," *IEEE Signal Process. Lett.*, vol. 20, no. 2, pp. 129–132, 2013.

[41] J. D. Gibson, B. Koo, and S. D. Gray, "Filtering of colored noise for speech enhancement and coding," *IEEE Trans. Signal Process.*, vol. 39, no. 8, pp. 1732–1742, 1991.

[42] R. L. Goldsworthy and J. E. Greenberg, "Analysis of speech-based speech transmission index methods with implications for nonlinear operations," *J. Acoust. Soc. Am.*, vol. 116, no. 6, pp. 3679–3689, Dec. 2004.

[43] S. Gonzalez and M. Brookes, "A pitch estimation filter robust to high levels of noise (PEFAC)," in *Proc. European Signal Processing Conf. (EUSIPCO)*, Barcelona, Aug. 2011.

[44] M. Goto, "A real-time music-scene-description system: predominant-f0 estimation for detecting melody and bass lines in real-world audio signals," *Speech Communication*, vol. 43, no. 4, pp. 311–329, 2004.

[45] K. Han and D. L. Wang, "An SVM based classification approach to speech separation," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2011, pp. 4632–4635.

[46] K. Han and D. Wang, "Towards generalizing classification based speech separation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 1, pp. 168–177, Jan. 2013.

[47] J. H. L. Hansen, V. Radhakrishnan, and K. H. Arehart, "Speech enhancement based on generalized minimum mean square error estimators and masking properties of the auditory system," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 6, pp. 2049–2063, Nov. 2006.

[48] R. Hendriks, R. Heusdens, and J. Jensen, "MMSE based noise PSD tracking with low complexity," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Mar. 2010, pp. 4266–4269.

[49] D. J. Hermes, "Measurement of pitch by subharmonic summation," *J. Acoust. Soc. Am.*, vol. 83, no. 1, pp. 257–264, Jan. 1988.

[50] G. Hilkhuysen, N. Gaubitch, M. Brookes, and M. Huckvale, "Effects of noise suppression on intelligibility ii: a validation of physical metrics," *J. Acoust. Soc. Am.*, submitted.

[51] V. Hohmann, "Frequency analysis and synthesis using a gammatone filterbank," *Acta Acustica United with Acustica*, vol. 88, no. 3, pp. 433–442, 2002.

[52] T. Houtgast and F. Dubbelboer, "The relation between the speech-envelope spectrum and speech intelligibility revisited," VU University medical center, Tech. Rep., 2007. [Online]. Available: http://www.silicon-speech.com/Media/TemporalDynamics/PDF/Houtgast_TemporalDynamics.pdf

[53] T. Houtgast and H. J. M. Steeneken, "A review of the MTF concept in room acoustics and its use for estimating speech intelligibility in auditoria," *J. Acoust. Soc. Am.*, vol. 77, no. 3, pp. 1069–1077, 1985.

[54] G. Hu and D. L. Wang, "Segregation of stop consonants from acoustic interference," in *IEEE XIII Workshop on Neural Networks for Signal Processing*, 2003.

[55] ——, "Separation of stop consonants," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 2, Apr. 2003.

[56] ——, "Monaural speech segregation based on pitch tracking and amplitude modulation," *IEEE Trans. Neural Netw.*, vol. 15, no. 5, pp. 1135–1150, Sep. 2004.

[57] ——, "An auditory scene analysis approach to monaural speech segregation," in *Topics in Acoustic Echo and Noise Control*, E. Hänsler and G. Schmidt, Eds. Springer Berlin Heidelberg, 2006.

[58] ——, "Auditory segmentation based on onset and offset analysis," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 2, pp. 396–405, Aug. 2007.

[59] ——, "Segregation of unvoiced speech from nonspeech interference," *J. Acoust. Soc. Am.*, vol. 124, pp. 1306–1319, Aug. 2008.

[60] ——, "A tandem algorithm for pitch estimation and voiced speech segregation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 8, pp. 2067–2079, Nov. 2010.

[61] K. Hu and D. L. Wang, "Unvoiced speech segregation from nonspeech interference via CASA and spectral subtraction," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 6, pp. 1600–1609, Aug. 2011.

[62] Y. Hu and P. C. Loizou, "Evaluation of objective measures for speech enhancement," in *Proc. Interspeech Conf.*, 2006, pp. 1447–1450.

[63] ——, "Techniques for estimating the ideal binary mask," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2008.

[64] ——, "A subspace approach for enhancing speech corrupted by colored noise," *IEEE Signal Process. Lett.*, vol. 9, no. 7, pp. 204–206, Jul. 2002.

[65] ——, "Subjective comparison and evaluation of speech enhancement algorithms," *Speech Communication*, vol. 49, no. 7–8, pp. 588–601, Jul. 2007.

[66] ——, "A comparative intelligibility study of single-microphone noise reduction algorithms," *J. Acoust. Soc. Am.*, vol. 122, pp. 1777–1786, 2007.

[67] T. Inoue, H. Saruwatari, Y. Takahashi, K. Shikano, and K. Kondo, "Theoretical analysis of musical noise in generalized spectral subtraction based on higher order statistics," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 6, pp. 1770–1779, Aug. 2011.

[68] ITU-T, *Objective Measurement of Active Speech Level*, International Telecommunications Union (ITU-T) Recommendation P.56, Mar. 1993.

[69] ——, *Test signals for use in telephonometry*, International Telecommunications Union (ITU-T) Recommendation P.501, Aug. 1996.

[70] ——, *Subjective performance evaluation of telephone band and wideband codecs*, International Telecommunications Union (ITU-T) Recommendation P.830, 1998.

[71] ——, *Artificial Voices*, International Telecommunications Union (ITU-T) Standard P.50, Sep. 1999.

[72] ——, *Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs*, International Telecommunications Union (ITU-T) Recommendation P.862, Feb. 2001.

[73] ——, *Mapping function for transforming P.862 raw result scores to MOS-LQ*, International Telecommunications Union (ITU-T) Recommendation P.862.1, 2003.

[74] ——, *Perceptual Objective Listening Quality Assessment: An advanced objective perceptual method for end-to-end listening speech quality evaluation of fixed, mobile, and IP-based networks and speech codecs covering narrowband, wideband, and super-wideband signals*, International Telecommunications Union (ITU-T) Standard P.863, Jan. 2011.

[75] J. Jensen and R. C. Hendriks, "Spectral magnitude minimum mean-square error estimation using binary and continuous gain functions," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 1, pp. 92–102, 2012.

[76] Z. Jin and D. L. Wang, "HMM-based multipitch tracking for noisy and reverberant speech," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 5, pp. 1091–1102, Jul. 2011.

[77] S. Jørgensen and T. Dau, "Predicting speech intelligibility based on the signal-to-noise envelope power ratio after modulation-frequency selective processing," vol. 130, no. 3, pp. 1475–1487, Sep. 2011.

[78] B. Juang and S. Katagiri, "Discriminative learning for minimum error classification," *IEEE Trans. Signal Process.*, vol. 40, no. 12, pp. 3043–3054, Dec. 1992.

[79] A. Kain, "CSLU: Voices," Linguistic Data Consortium, Philadelphia, http://www.ldc.upenn.edu/Catalog/catalogEntry.jsp?catalogId=LDC2006S01, 2006.

[80] J. M. Kates and K. H. Arehart, "Coherence and the speech intelligibility index," *J. Acoust. Soc. Am.*, vol. 117, no. 4, pp. 2224–2237, 2005.

[81] D. S. Kim and A. Tarraf, "Anique+: A new american national standard for non-intrusive estimation of narrowband speech quality," *Bell Labs Tech. J.*, vol. 12, pp. 221–236, 2007.

[82] G. Kim and P. Loizou, "Improving speech intelligibility in noise using environment-optimized algorithms," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 8, pp. 2080–2090, Nov. 2010.

[83] G. Kim, Y. Lu, Y. Hu, and P. C. Loizou, "An algorithm that improves speech intelligibility in noise for normal-hearing listeners," *J. Acoust. Soc. Am.*, vol. 126, no. 3, pp. 1486–1494, Sep. 2009.

[84] W. Kim and R. M. Stern, "Band-independent mask estimation for missing-feature reconstruction in the presence of unknown background noise," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 1, May 2006, pp. 305–308.

[85] N. Kitawaki and T. Yamada, "Subjective and objective quality assessment for noise reduced speech," in *ETSI Workshop on Speech and Noise in Wideband Communication*, Sophia Antipolis, France, May 2007.

[86] U. Kjems, J. B. Boldt, M. S. Pedersen, T. Lunner, and D. L. Wang, "Role of mask pattern in intelligibility of ideal binary-masked noisy speech," *J. Acoust. Soc. Am.*, vol. 126, no. 3, pp. 1415–1426, Sep. 2009.

[87] U. Kjems, M. S. Pedersen, J. B. Boldt, T. Lunner, and D. L. Wang, "Speech intelligibility of ideal binary masked mixtures," in *Proc. European Signal Processing Conf. (EUSIPCO)*, Aalborg, Denmark, Aug. 2010, pp. 1909–1913.

[88] A. Klapuri, "Multiple fundamental frequency estimation by summing harmonic amplitudes," in *Proc Intl Conf Music Inf. Retrieval*, vol. 6, 2006, pp. 216–221.

[89] ——, "Multipitch analysis of polyphonic music and speech signals using an auditory model," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 2, pp. 255–266, Feb. 2008.

[90] B. Kollmeier and R. Koch, "Speech enhancement based on physiological and psychoacoustical models of modulation perception and binaural interaction," *J. Acoust. Soc. Am.*, vol. 95, no. 3, pp. 1593–1602, Mar. 1994.

[91] A. A. Kressner, D. V. Anderson, and C. J. Rozell, "A novel binary mask estimator based on sparse approximation," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2013.

[92] K. Kryter, "Methods for the calculation and use of the articulation index," *J. Acoust. Soc. Am.*, vol. 34, no. 11, pp. 1689–1697, 1962.

[93] P. Ladefoged, *Preliminaries to Linguistic Phonetics*. Chicago: University of Chicago Press, 1971.

[94] ——, *A Course in Phonetics*, 2nd ed. Harcourt Brace Jovanovich, 1982, iSBN 0155151789.

[95] J. Le Roux, H. Kameoka, N. Ono, A. de Cheveigne, and S. Sagayama, "Single and multiple f0 contour estimation through parametric spectrogram modeling of speech in noisy environments," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 4, pp. 1135–1145, 2007.

[96] J. Li, L. Yang, J. Zhang, Y. Yan, Y. Hu, M. Akagi, and P. C. Loizou, "Comparative intelligibility investigation of single-channel noise-reduction algorithms for Chinese, Japanese, and English," *J. Acoust. Soc. Am.*, vol. 129, no. 5, pp. 3291–3301, May 2011.

[97] N. Li and P. C. Loizou, "Factors influencing intelligibility of ideal binary-masked speech: Implications for noise reduction," *J. Acoust. Soc. Am.*, vol. 123, no. 3, pp. 1673–1682, Mar. 2008.

[98] Y. Li and D. L. Wang, "On the optimality of ideal binary time-frequency masks," *Speech Communication*, vol. 51, no. 3, pp. 230–239, 2009.

[99] J. S. Lim and A. V. Oppenheim, "Enhancement and bandwidth compression of noisy speech," *Proc. IEEE*, vol. 67, no. 12, pp. 1586–1604, Dec. 1979.

[100] P. C. Loizou, "Speech enhancement based on perceptually motivated Bayesian estimators of the magnitude spectrum," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 5, pp. 857–869, 2005.

[101] T. Lotter, C. Benien, and P. Vary, "Multichannel speech enhancement using Bayesian spectral amplitude estimation," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 1, Apr. 2003.

[102] C. Ludvigsen, C. Elberling, and G. Keidser, "Evaluation of a noise reduction method–comparison between observed scores and scores predicted from STI," *Scandinavian audiology. Supplementum*, vol. 38, pp. 50–55, 1993.

[103] J. Ma, Y. Hu, and P. C. Loizou, "Objective measures for predicting speech intelligibility in noisy conditions based on new band-importance functions," *J. Acoust. Soc. Am.*, vol. 125, no. 5, pp. 3387–3405, 2009.

[104] N. Madhu, A. Spriet, S. Jansen, R. Koning, and J. Wouters, "The potential for speech intelligibility improvement using the ideal binary mask and the ideal wiener filter in single channel noise reduction systems: Application to auditory prostheses," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 1, pp. 63–72, Jan. 2013.

[105] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki, "The DET curve in assessment of detection task performance," in *Proc. European Conf. on Speech Communication and Technology*, 1997, pp. 1895–1898.

[106] P. Martin, "Comparison of pitch detection by cepstrum and spectral comb analysis," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 7, May 1982, pp. 180–183.

[107] R. Martin, "Spectral subtraction based on minimum statistics," in *Proc. European Signal Processing Conf*, 1994, pp. 1182–1185.

[108] ——, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. Speech Audio Process.*, vol. 9, pp. 504–512, Jul. 2001.

[109] ——, "Speech enhancement based on minimum mean-square error estimation and supergaussian priors," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 5, pp. 845–856, Sep. 2005.

[110] R. McAulay and M. Malpass, "Speech enhancement using a soft-decision noise

suppression filter," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 28, no. 2, pp. 137–145, Apr. 1980.

[111] Y. Medan, E. Yair, and D. Chazan, "Super resolution pitch determination of speech signals," *IEEE Trans. Signal Process.*, vol. 39, no. 1, pp. 40–48, 1991.

[112] R. Meddis, "Simulation of auditory–neural transduction: Further studies," vol. 83, no. 3, pp. 1056–1063, 1988.

[113] N. Mesgarani, M. Slaney, and S. Shamma, "Discrimination of speech from non-speech based on multiscale spectro-temporal modulations," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 3, pp. 920–930, May 2006.

[114] U. Mittal and N. Phamdo, "Signal/noise KLT based approach for enhancing speech degraded by colored noise," *IEEE Trans. Speech Audio Process.*, vol. 8, no. 2, pp. 159–167, Mar. 2000.

[115] B. C. J. Moore and B. R. Glasberg, "Suggested formulae for calculating auditory-filter bandwidths and excitation patterns," *J. Acoust. Soc. Am.*, vol. 74, pp. 750–753, 1983.

[116] H. Ney, "A dynamic programming technique for nonlinear smoothing," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 1981, pp. 62–65.

[117] ——, "Dynamic programming algorithm for optimal estimation of speech parameter contours," *IEEE Trans Syst, Man and Cybernetics*, vol. 13, pp. 208–214, 1983.

[118] A. Noll, "Pitch determination of human speech by the harmonic product spectrum, the harmonic sum spectrum, and a maximum likelihood estimate," in *Proceedings of the Symposium on Computer Processing in Communications*, 1969, pp. 779–797, vol. XIX, Polytechnic Press: Brooklyn, New York, (1970).

[119] T. W. Parsons, "Separation of speech from interfering speech by means of harmonic selection," *J. Acoust. Soc. Am.*, vol. 60, no. 4, pp. 911–918, Oct. 1976.

[120] R. Patterson, I. Nimmo-Smith, J. Holdsworth, and P. Rice, "An efficient auditory filterbank based on the gammatone function," MRC Applied Physiology Unit, Cambridge, Tech. Rep., Dec. 1987.

[121] K. Pearson, "On lines and planes of closest fit to systems of points in space," *Philosophical Magazine*, vol. 2, no. 11, pp. 559–572, 1901.

[122] N. Roman, D. Wang, and G. J. Brown, "Speech segregation based on sound localization," *J. Acoust. Soc. Am.*, vol. 114, no. 4, pp. 2236–2252, Oct. 2003.

[123] J. Rouat, Y. C. Liu, and D. Morissette, "A pitch determination and voiced/unvoiced decision algorithm for noisy speech," *Speech Communication*, vol. 21, no. 3, pp. 191–207, 1997.

[124] M. Sambur, "LMS adaptive filtering for enhancing the quality of noisy speech," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 3, Apr. 1978, pp. 610–613.

[125] N. Sasaoka, M. Watanabe, Y. Itoh, and K. Fujii, "Noise reduction system based on LPEF and system identification with variable step size," in *Proc. Intl. Symp. on Circuits and Systems*, 2007, pp. 2311–2314.

[126] M. R. Schroeder, "Period histogram and product spectrum: New methods for fundamental-frequency measurement," *J. Acoust. Soc. Am.*, vol. 43, no. 4, pp. 829–834, Apr. 1968.

[127] M. Seltzer, B. Raj, and R. Stern., "A Bayesian classifier for spectrographic mask estimation for missing feature speech recognition," *Speech Communication*, vol. 43, pp. 379–393, Sep. 2004.

[128] D. Sharma and P. A. Naylor, "Evaluation of pitch estimation in noisy speech for application in non-intrusive speech quality assessment," in *Proc. European Signal Processing Conf. (EUSIPCO)*, Glasgow, Aug. 2009.

[129] E. C. Smith and M. S. Lewicki, "Efficient auditory coding," *Nature*, vol. 439, pp. 978–982, Feb. 2006.

[130] S. So, K. K. Wocicki, J. G. Lyons, A. P. Stark, and K. K. Paliwal, "Kalman fitler with phase spectrum compensation algorithm for speech enhancement," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2009, pp. 4405–4408.

[131] H. J. M. Steeneken and F. W. M. Geurtsen, "Description of the RSG.10 noise data-base," TNO Institute for perception, Tech. Rep. IZF 1988–3, 1988.

[132] H. J. M. Steeneken and T. Houtgast, "A physical method for measuring speech-transmission quality," *J. Acoust. Soc. Am.*, vol. 67, no. 1, pp. 318–326, Jan. 1980.

[133] ——, "Mutual dependence of the octave-band weights in predicting speech intelligibility," *Speech Communication*, vol. 28, pp. 109–123, 1999.

[134] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time-frequency weighted noisy speech," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 7, pp. 2125–2136, Sep. 2011.

[135] J. Tabrikian, S. Dubnov, and Y. Dickalov, "Maximum a-posteriori probability pitch tracking in noisy environments using harmonic model," *IEEE Trans. Speech Audio Process.*, vol. 12, no. 1, pp. 76–87, Jan. 2004.

[136] J. Taghia, N. Mohammadiha, J. Sang, V. Bouse, and R. Martin, "An evaluation of noise power spectral density estimation algorithms in adverse acoustic environments," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, May 2011, pp. 4640 –4643.

[137] D. Talkin, "A robust algorithm for pitch tracking (RAPT)," in *Speech Coding and Synthesis*, W. B. Kleijn and K. K. Paliwal, Eds. Amsterdam: Elsevier, 1995, pp. 495–518.

[138] J. Tribolet, P. Noll, B. McDermott, and R. Crochiere, "A study of complexity and quality of speech waveform coders," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 3, 1978, pp. 586–590.

[139] G. Turin, "An introduction to matched filters," *IRE Transactions on Information Theory*, vol. 6, no. 3, pp. 311–329, Jun. 1960.

[140] A. P. Varga and R. K. Moore, "Hidden Markov model decomposition of speech and noise," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Apr. 1990, pp. 845–848.

[141] N. Virag, "Single channel speech enhancement based on masking properties of the human auditory system," *IEEE Trans. Speech Audio Process.*, vol. 7, no. 2, pp. 126–137, Mar. 1999.

[142] T. Virtanen and A. Klapuri, "Separation of harmonic sound sources using sinusoidal modeling," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing ICASSP '00*, vol. 2, Jun. 2000, pp. II765–II768.

[143] K. Wagener, J. Josvassen, and R. Ardenkjær, "Design, optimization and evaluation of a danish sentence test in noise," *International journal of audiology*, vol. 42, no. 1, pp. 10–17, 2003.

[144] D. L. Wang, "On ideal binary mask as the computational goal of auditory scene analysis," in *Speech Separation by Humans and Machines*, P. Divenyi, Ed. Kluwer Academic, 2005, pp. 181–197.

[145] D. L. Wang and G. J. Brown, "Separation of speech from interfering sounds based on oscillatory correlation," *IEEE Trans. Neural Netw.*, vol. 10, no. 3, pp. 684–697, May 1999.

[146] D. L. Wang, U. Kjems, M. S. Pedersen, J. B. Boldt, and T. Lunner, "Speech intelligibility in background noise with ideal binary time-frequency masking," *J. Acoust. Soc. Am.*, vol. 125, pp. 2336–2347, 2009.

[147] D. L. Wang, U. Kjems, M. S. Pedersen, J. B. Bolt, and T. Lunner, "Speech perception of noise with binary gains," *J. Acoust. Soc. Am.*, vol. 124, no. 4, pp. 2303–2307, Oct. 2008.

[148] Y. Wang, K. Han, and D. L. Wang, "Exploring monaural features for classification-based speech segregation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 2, pp. 270–279, Feb. 2013.

[149] Y. Wang and D. L. Wang, "Boosting classification based speech separation using temporal dynamics." in *Proc. Interspeech Conf.*, 2012.

[150] ——, "Towards scaling up classification-based speech separation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 7, pp. 1381–1390, Jul. 2013.

[151] J. Wen, N. D. Gaubitch, E. Habets, T. Myatt, and P. A. Naylor, "Evaluation of speech dereverberation algorithms using the MARDY database," in *Proc. Intl. Workshop Acoust. Echo Noise Control (IWAENC)*, Paris, France, Sep. 2006.

[152] M. Wohlmayr, M. Stark, and F. Pernkopf, "A probabilistic interaction model for multipitch tracking with factorial hidden Markov models," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 4, pp. 799–810, 2011.

[153] M. Wu, D. L. Wang, and G. J. Brown, "A multipitch tracking algorithm for noisy speech," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 3, pp. 229–241, May 2003.

[154] A. Yasmin, P. Fieguth, and L. Deng, "Speech enhancement using voice source models," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 2, Mar. 1999, pp. 797–800.

[155] E. Zwicker, "Subdivision of audible frequency range into critical bands," *J. Acoust. Soc. Am.*, vol. 33, p. 248, 1961.

Appendix A

Noise databases

Two different noise databases have been used in this thesis: the RSG-10 database [131] and the noise database from the ITU-T P.501 standard [69]. In this appendix, we present further details about the noise types present in each database together with their power spectrogram.

A.1 RSG-10 database

All the descriptions provided in this section have been extracted from [131].

Babble noise: The source of this babble noise is 100 people speaking in a canteen. The room radius is over two meters; therefore, individual voices are slightly audible.

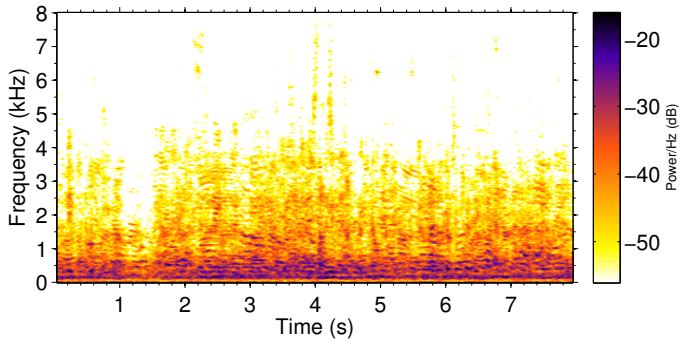


Figure A.1: Babble noise power spectrogram.

Buccaneer noise 1: The Buccaneer jet was moving at a speed of 190 knots, and an altitude of 1000 feet, with airbrakes out.

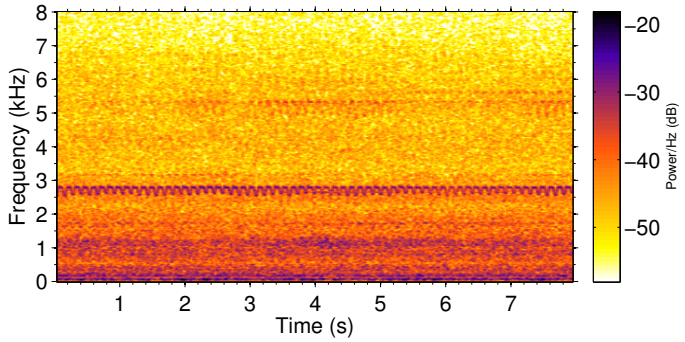


Figure A.2: Buccaneer noise 1 power spectrogram.

Buccaneer noise 2: The Buccaneer was moving at a speed of 450 knots, and an altitude of 300 feet.

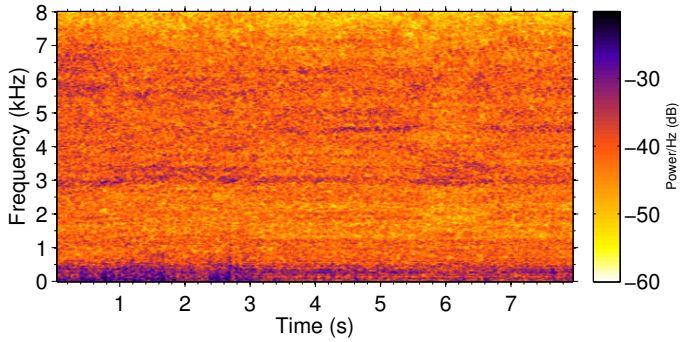


Figure A.3: Buccaneer noise 2 power spectrogram.

Destroyer engine noise: Engine Room noise.

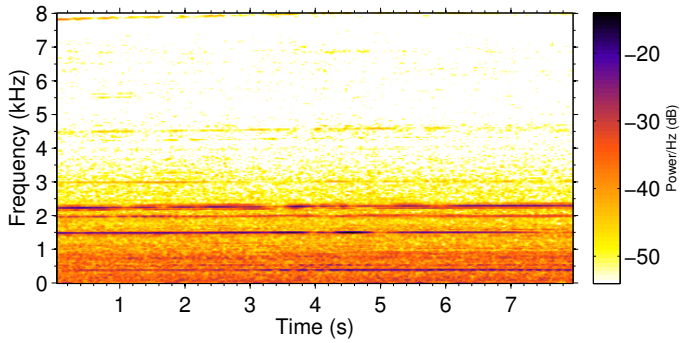


Figure A.4: Destroyer engine noise power spectrogram.

Destroyer operations noise: Operations Room noise.

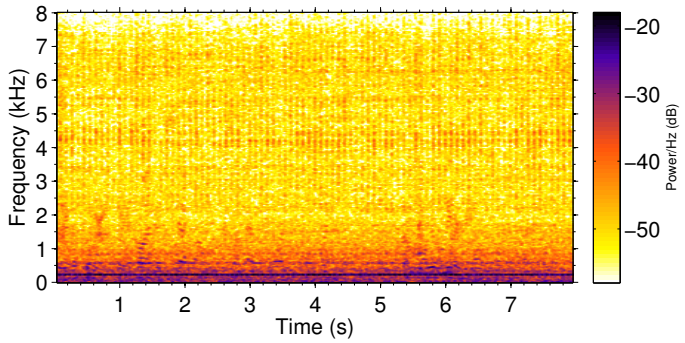


Figure A.5: Destroyer operations room noise power spectrogram.

F16 noise: The noise was recorded at the co-pilot's seat in a two-seat F-16, travelling at a speed of 500 knots, and an altitude of $300 - 600$ feet. It was found that the flight condition had only a minor effect on the noise. The reproduced noise can therefore be considered to be representative.

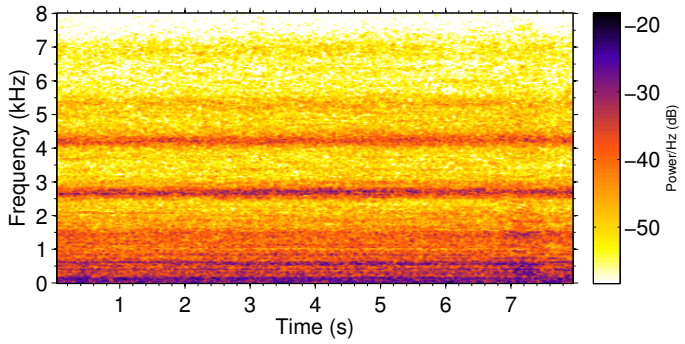


Figure A.6: F16 noise power spectrogram.

Factory noise 1: This noise was recorded near plate-cutting and electrical welding equipment.

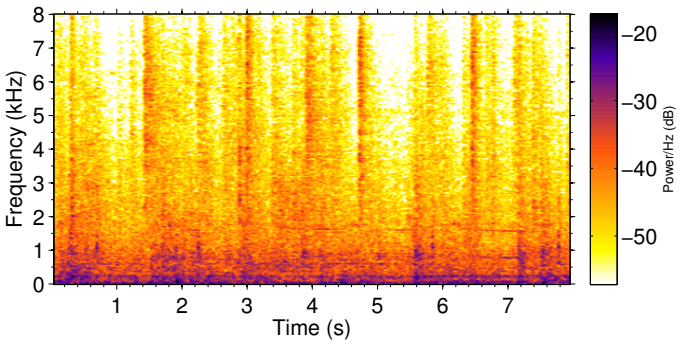


Figure A.7: Factory noise 1 power spectrogram.

Factory noise 2: This noise was recorded in a car production hall.

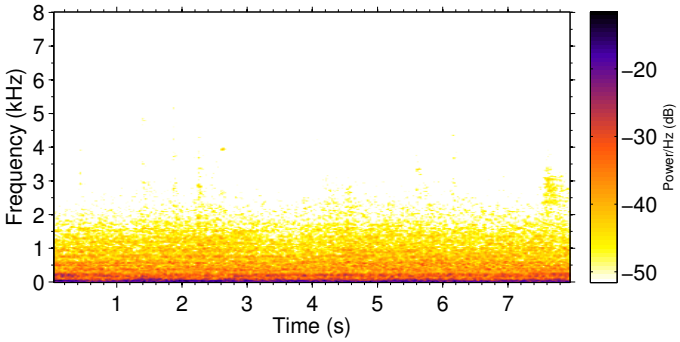


Figure A.8: Factory noise 2 power spectrogram.

HF radio noise: Recording of noise in an HF radio channel after demodulation.

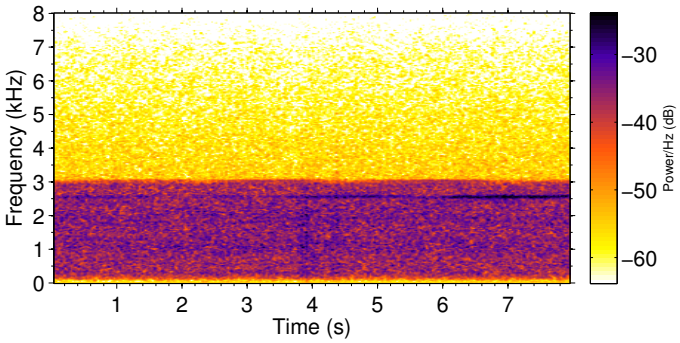


Figure A.9: HF radio noise power spectrogram.

Leopard tank noise: The Leopard vehicle was moving at a speed of 70 km/h.

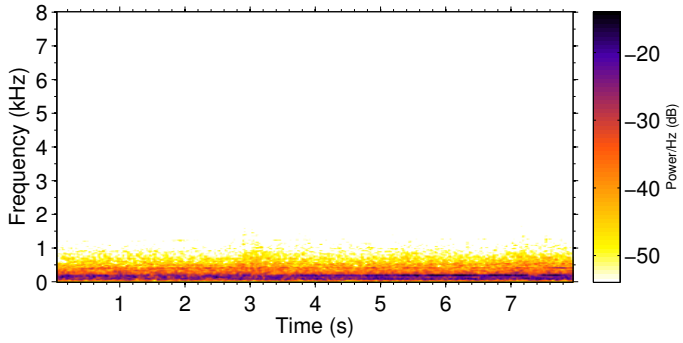


Figure A.10: Leopard tank noise power spectrogram.

M109 tank noise: The M109 tank was moving at a speed of 30 km/h.

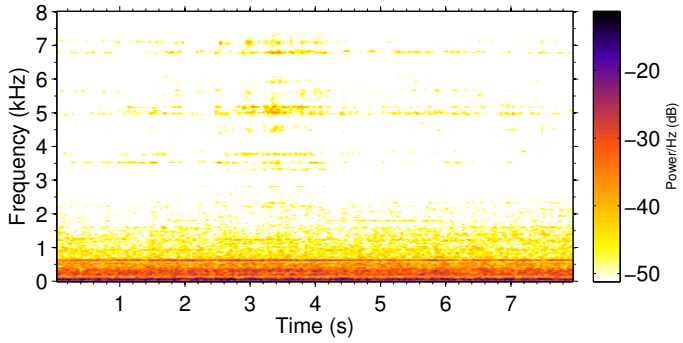


Figure A.11: M109 tank noise power spectrogram.

Machine gun noise: The weapon used was a .50 calibre gun fired repeatedly.

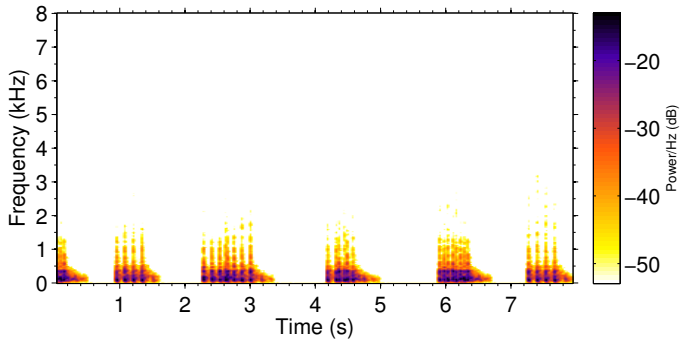


Figure A.12: Machine gun noise power spectrogram.

Pink noise: Noise acquired by sampling high-quality analog noise generator. Exhibits equal energy per 1/3 octave.

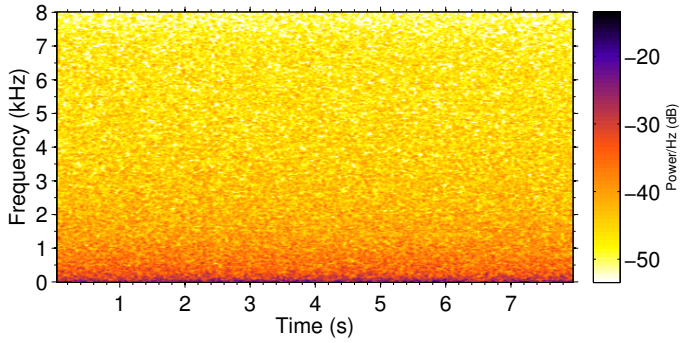


Figure A.13: Pink noise power spectrogram.

Volvo noise Volvo 340 noise acquired at 120 km/h, in 4th gear, on an asphalt road, in rainy conditions.

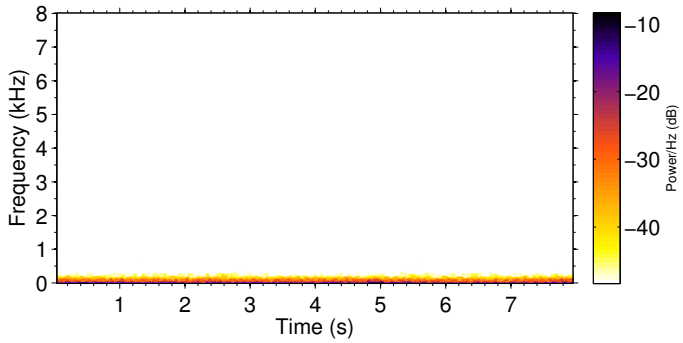


Figure A.14: Volvo car noise power spectrogram.

White noise: White noise acquired by sampling high-quality analog noise generator. Exhibits equal energy per Hz bandwidth.

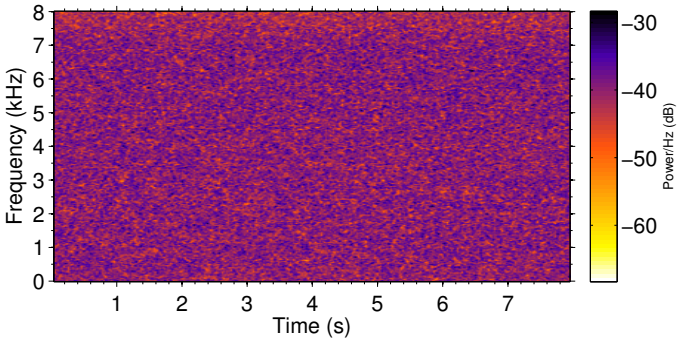


Figure A.15: White noise power spectrogram.

A.2 Noise database from the ITU-T P.501 standard

All the descriptions provided in this section have been extracted from [69].

Cafeteria noise Typical cafeteria noise

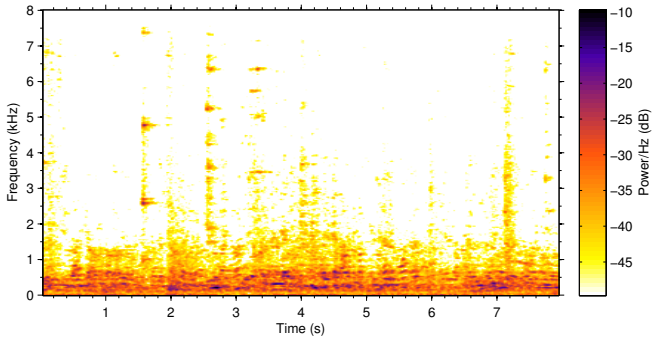


Figure A.16: Cafeteria noise power spectrogram.

In car noise Noise inside a typical medium size car.

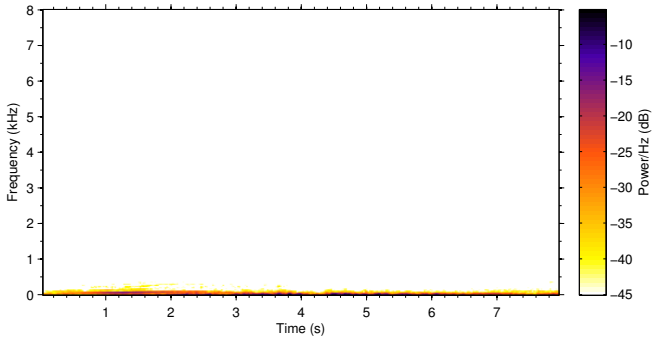


Figure A.17: In car noise power spectrogram.

Street noise Typical street noise

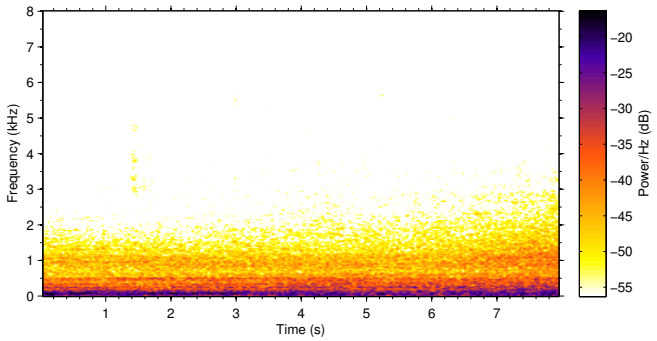


Figure A.18: Street power spectrogram.

Car noise Car interior noise, car driving, radio on (speech programme).

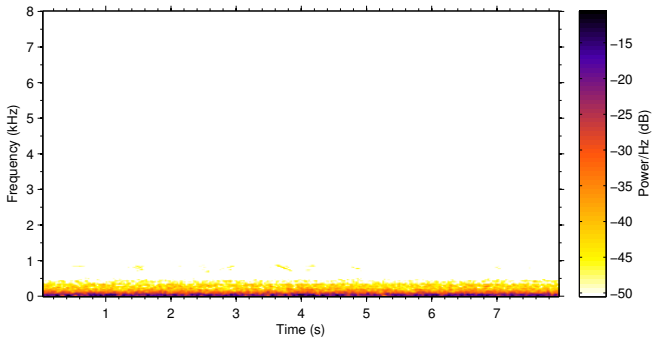


Figure A.19: Car noise power spectrogram.

Construction noise: Construction noise, impulse type noise (hammering), sawing noise.

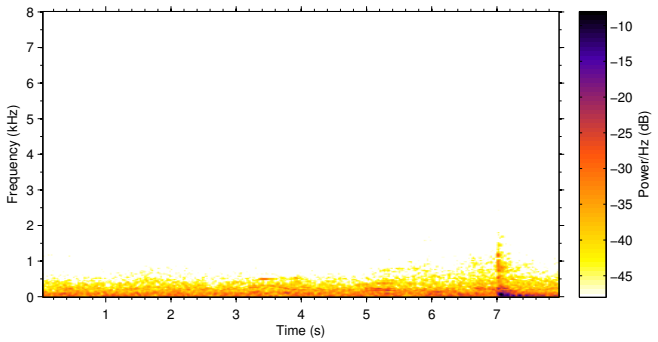


Figure A.20: Construction noise power spectrogram.

Metro noise: Metro train arriving to the station.

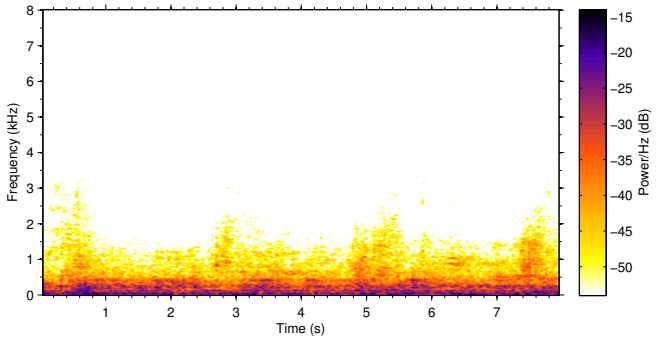


Figure A.21: Metro noise power spectrogram.

Office noise: Office noise, fans, typing, phone ringing, noise from chair.

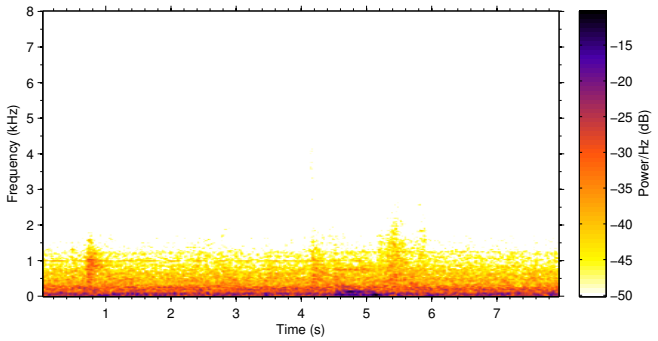


Figure A.22: Office noise power spectrogram.

Railway station noise: Railway station, echoing surroundings, speech, shoes clacking.

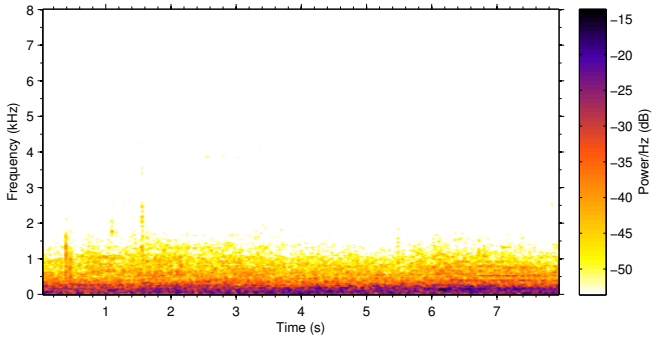


Figure A.23: Railway station noise power spectrogram.

Restaurant noise: Restaurant, babble, water, dishes.

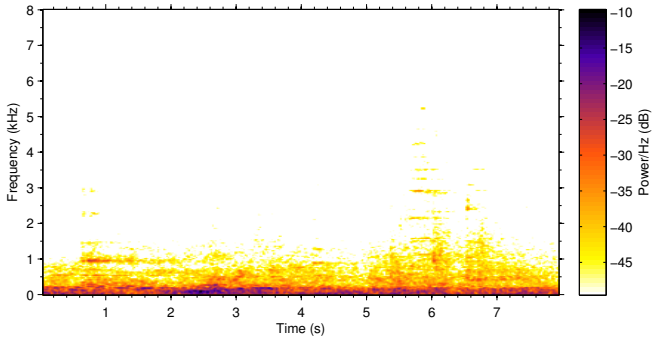


Figure A.24: Restaurant noise power spectrogram.