University of Wisconsin Milwaukee
UWM Digital Commons

Theses and Dissertations

December 2016

Improving the Speech Intelligibility By Cochlear Implant Users

Behnam Azimi
University of Wisconsin-Milwaukee

Follow this and additional works at: https://dc.uwm.edu/etd

Part of the Electrical and Electronics Commons, and the Science and Mathematics Education Commons

Recommended Citation

Azimi, Behnam, "Improving the Speech Intelligibility By Cochlear Implant Users" (2016). *Theses and Dissertations*. 1345.
https://dc.uwm.edu/etd/1345

This Dissertation is brought to you for free and open access by UWM Digital Commons. It has been accepted for inclusion in Theses and Dissertations by an authorized administrator of UWM Digital Commons. For more information, please contact open-access@uwm.edu.

IMPROVING THE SPEECH INTELLIGIBILITY BY COCHLEAR IMPLANT USERS

by

Behnam Azimi

A Dissertation Submitted in

Partial Fulfillment of the

Requirements for the Degree of

Doctor of Philosophy

in Engineering

at

The University of Wisconsin-Milwaukee

December 2016

ABSTRACT

IMPROVING THE SPEECH INTELLIGIBILITY BY COCHLEAR IMPLANT USERS

by

Behnam Azimi

The University of Wisconsin-Milwaukee, 2016
Under the Supervision of Professor Yi Hu

In this thesis, we focus on improving the intelligibility of speech for cochlear implants (CI) users. As an auditory prosthetic device, CI can restore hearing sensations for most patients with profound hearing loss in both ears in a quiet background. However, CI users still have serious problems in understanding speech in noisy and reverberant environments. Also, bandwidth limitation, missing temporal fine structures, and reduced spectral resolution due to a limited number of electrodes are other factors that raise the difficulty of hearing in noisy conditions for CI users, regardless of the type of noise. To mitigate these difficulties for CI listener, we investigate several contributing factors such as the effects of low harmonics on tone identification in natural and vocoded speech, the contribution of matched envelope dynamic range to the binaural benefits and contribution of low-frequency harmonics to tone identification in quiet and six-talker babble background. These results revealed several promising methods for improving speech intelligibility for CI patients. In addition, we investigate the benefits of voice conversion in improving speech intelligibility for CI users, which was motivated by an earlier study showing that familiarity with a talker's voice can improve understanding of the conversation. Research has shown that when adults are familiar with someone's voice, they can more

accurately – and even more quickly – process and understand what the person is saying. This theory identified as the “familiar talker advantage” was our motivation to examine its effect on CI patients using voice conversion technique. In the present research, we propose a new method based on multi-channel voice conversion to improve the intelligibility of transformed speeches for CI patients.

© Copyright by Behnam Azimi, 2016
All Rights Reserved

TABLE OF CONTENTS

LIST OF FIGURES vii

LIST OF TABLES x

Chapter One – INTRODUCTION 1

- What is Speech? 1
- Normal Hearing 2
- Cochlear Implants and its structure 3
- Patient difficulties and limitations 11

Chapter Two - Objectives and Motivation 12

Chapter Three - Introduction to Voice Conversion and algorithm 17

- Voice Conversion Techniques 20
- Linear algebra techniques 21
- Signal processing techniques 23
- Cognitive techniques 28
- Statistical Techniques 30

Chapter Four – Feature Enhancement 37

- Effects of low harmonics on tone identification in natural and vocoded speech 37
    - 4-1-1 Introduction 37
    - 4-1-2-Method 39
    - 4-1-3- Results 42
    - 4-1-4- Discussion 46
- The Contribution of Matched Envelope Dynamic Range to the Binaural Benefits in Simulated Bilateral Electric Hearing 49
    - 4-2-1-Introduction 49
    - 4-2-2-Method 54
    - 4-2-3-Procedure 61
    - 4-2-4-Results 62
    - 4-2-5-Discussion and Conclusions 64
- Contribution of low-frequency harmonics to tone identification in quiet and six-talker babble background 69
    - 4-3-1-INTRODUCTION 69

    - 4-3-2 Method 72
    - 4-3-3-RESULTS 76
    - 4-3-4- DISCUSSION 83
    - 4-3-5-SUMMARY 93
- Chapter Five – Methods 94
  - Multi-channel Voice Conversion algorithm 94
  - Noise reduction strategies in Cochlear Implants 95
    - Noise reduction on noisy acoustic inputs 97
    - Noise Reduction on Noisy Electrical Envelopes 100
  - Feature Extraction 107
  - Multi-channel GMM Voice conversion algorithm 116
    - Training Stage - Multi Channel process 117
    - Feature Extraction 121
    - Speech Alignment 122
    - GMM model 123
  - Multi-channel ANN Voice conversion algorithm 124
- Chapter Six - Results and Conclusion 126
- References 128
- CURRICULUM VITAE 134

LIST OF FIGURES

Figure 1.1 shows voiced "ASA.wav" 1
Figure 1.2- Anatomy of the Ear 2
Figure 1.3, Cochlear implant 3
Figure 1.4: Cochlear implant device 4
Figure 1.5: Cochlear implant Signal processing 5
Figure 1.6: Band Pass Filters 6
Figure 1.7: Channels 1 to 4 signals and envelop 8
Figure 1.8: Channels 5 to 8 signals and envelop 9
Figure 1.9: Channels 9 to 12 signals and envelop 9
Figure 1.10: Channels 13 to 16 signals and envelop 10
Figure 2.1: Percentage of correct word recognition is plotted at each signal-to-noise ratio. [11]. 13
Figure 2.2: Talker identification accuracy during the six training sessions for both German-trained and English-trained listeners separated by good and poor voice learners. Two training sessions were completed on each day of training. by Levi, Winters and Pisoni [8]. 14
Figure 2.3: Proportion phonemes correct by SNR for English-trained and German-trained learners divided by learning ability (good and poor). Levi, Winters and Pisoni [8]. 15
Figure 3.1: Concept of Voice conversion. 17
Figure 3.2 shows two stages of voice conversion system. 19
Figure 3.3 Shows LPC model 22
Figure 3.4 Dynamic Time Warping (DTW) 24
Figure 3.5 (Top) Learning Stage (Bottom) Converting Stage 25
Figure 3.6 shows increasing (a) and decreasing (b) of pitch with PSOLA method. 27
Figure 3.7 Shows Artificial Neural Network 28
Figure 3.8 Shows Artificial Neural Network learning stage 29
Figure 3.9 a) Shows GMM classification. b) Shows HMM model 30
Figure 3.10 Probability density of the Gaussian distribution N(0,1.5), N(-1,2) and N(1,3). 31
Figure 3.11 Gaussian Mixture Model 32
Figure 3.12 Gaussian Mixture Model using Expectation-Maximization algorithm 33
Figure 3.13 Block diagram of the learning procedure [28]. 35
Figure 3.14 Block diagram of the voice conversion procedure [28]. 36
Figure 4.1: Tone identification as a function of the four tone categories in the quiet condition (a) and noisy condition (b). Listening condition for natural (left panels) and vocoded (right panels) speech and for the female (upper panels) and male (lower panel) speakers (All: all harmonics included; H1: the first harmonic; H2: the second harmonic; H3: the third harmonic). 43
Figure 4.2: Linear regression function of average tone identification scores over the ten listeners vs. amplitude-F0 correlation index (Pearson r¼0.66, p<0.01). 45
Figure 4.3- The plot of envelope dynamic range (DR) value as a function of compression factor α, which takes the values from 0.05 to 1 corresponding to different DR values. 57
Figure 4.4- The histograms for the DR values of the eight temporal envelopes in the 240 vocoded Mandarin Hearing in Noise Test (MHINT; Wong et al., 2007) sentences as a function of compression

factor a, which takes the values of 1/13 (DR = 5 dB; Panel A), 1/5 (DR = 10 dB; Panel B), 1/3 (DR = 15 dB; Panel C), and 1 (Panel D). The dotted line in each subplot fits the histogram of the envelope DR values using a normal distribution function with parameter values of mean (μ) and standard deviation (σ)...... 60
Figure 4.5- Mean recognition scores for (A) Mandarin sentence and (B) Mandarin tone at signal-to-noise ratio (SNR) levels of 5 and 0 dB by the simulated bilateral and unilateral electric hearing. 63
Figure 4.6- The amplitude contours of a recording of the Mandarin Chinese vowel /a / in four tones when the DR of the envelope is the original value (Panel A) and compressed to 15 dB (Panel B) and 10 dB (Panel C). 65
Figure 4.7: Tone identification scores and standard errors averaged over the three vowels as a function of listening conditions (Quiet, and SNRs of 5 and 0 dB) for the female (top) and male (bottom) speaker. 77
Figure 4.8: Tone identification scores and standard errors as a function of the four tone categories in quiet for the female (left panels) and male (right panels) speakers with the three vowels /i/ (top panels), /ɣ/ (middle panels), and /ɑ/ (bottom panels), for six harmonic conditions (All: all harmonics included; High: high harmonics; Low: low harmonics; H1: the first harmonic; H2: the second harmonic; H3: the third harmonic). 79
Figure 4.9: Tone identification scores and standard errors as a function of the four tone categories in the noisy listening condition of SNR at 5 dB. The layouts and symbols are the same with Figure 4.8. 80
Figure 4.10: Tone identification scores and standard errors as a function of the four tone categories in the noisy listening condition of SNR at 0 dB. The layouts and symbols are the same with Figure 4.8 82
Figure 4.11: The LPC long-term spectrum for the 10-s six-talker Mandarin babble using a sampling frequency of 12207 Hz with the RMS level at 70 dB SPL 88
Figure 4.12: The sigmoidal fitting function of tone identification as a function of local SNRs for clustered H1 and H2 (top; R=0.86, p<0.05) and for H3 (bottom; R=0.65, p<0.05). 88
Figure 4.13- The sigmoidal fitting function of tone identification as a function of SNRs for the low (R=0.77, p<0.05) and high (R=0.52, p<0.05) harmonics. 89
Figure 5.1 Top side shows position of preprocessing technique in signal processing of cochlear implants; Bottom side shows noise reduction on noisy Electrical envelopes. 96
Figure 5.2 Block diagram of single-microphone noise reduction methods based on preprocessing the noisy acoustic input signals. 97
Figure 5.3 describes the envelope-weighting strategy block diagram. 101
Figure 5.4 Block diagram of single-microphone noise reduction methods based on attenuating the noisy electrical envelopes through envelope-selection 105
Figure 5.5 the block diagram of traditional MFCC features extraction. 108
Figure 5.6 shows the frame with 400 samples length and frame step 160 samples............................... 109
Figure 5.7 shows 10 Mel filter-banks start from 300Hz and ends at 8000Hz 112
Figure 5.8: Comparison of the traditional MFCC (left) with the integrated approach (right) [82]........... 115
Figure 5.9 shows Training and transformation stage for Multi-GMM Voice conversion. 116
Figure 5.10 block diagram of submodules in Multi-Channel process 117
Figure 5.11 Phoneme Segmentation with Praat 118
Figure 5.12 Epoch detection in Praat 120
Figure 5.13 channeling Block Diagram 121

Figure 5.14: multi-channel VC with artificial neural network 124
Figure 5.15: Best validation performance is 0.05669 at epoch 4, Blue line shows the decreasing error on training data, Green shows error on Validation data and red shows the error on test data. 125
Figure 6.1: compares the quality of Single band and multiband with GMM and ANN methods. 126
Figure 6.2: compares the quality of a various number of channels. 127

LIST OF TABLES

Table 1.1: lower and upper frequencies in the 16 bandpass filters 7

Table 4.1- Filter cutoff (–3 dB) frequencies used for the channel allocation. 59

Table 4.2: Fundamental frequency (F0), F1, and F2 frequency (Hz) of the three vowels with four tones. 74

Table 4.3: Results of four-factor (F0 value × vowel category × tone category × stimulus harmonic) repeated-measures ANOVAs and Tukey post hoc tests in quiet, and the SNR of 5 and 0 dB. DF: degree of freedom; F: Female; M: Male; Hi: High; Lo: Low. 84

Table 4.4: Results of three-factor (F0 value × vowel category× stimulus harmonic) repeated-measures ANOVAs for each of the four tones in the SNR of 5 dB. DF: degree of freedom; F: female; M: male; Hi: high; Lo: low. 85

Table 4.5: Results of three-factor (F0 value × vowel category × stimulus harmonic) repeated-measures ANOVAs for each of the four tones in the SNR of 0 dB. DF: degree of freedom; F: female; M: male; Hi: high; Lo: low. 86

Table 4.6: Confusion matrix of tone identification in quiet for the six harmonic conditions and two talkers (female: left; male: right). 87

Table 4.7: Confusion matrix of tone identification at the SNR of 5 dB for the six harmonic conditions and two talkers (female: left; male: right). 87

Table 4.8- Confusion matrix of tone identification at the SNR of 0 dB for the six harmonic conditions and two talkers (female: left; male: right). 88

Table 4.9: Average amplitude (dB SPL) over the four tones for individual low-frequency harmonics (H1, H2, and H3) for the three vowels and for the female and male speaker. 92

Table 5.1 shows frequency and corresponding Mel values for 10 filter-banks. 111

Table 5.2 result of phoneme segmentation section 119

Chapter One – INTRODUCTION

What is Speech?

Speech is the vocalized form of communicating. Speech signal, produced by muscle actions in the head, neck and using the lungs and the vocal folds in the larynx. In this process air is forced from the lungs through the vocal cords and vocal tract then produced at speaker's mouth. Speech consists of the following:

Articulation: is the movement of the tongue, lips and other speech organs in order to make speech sound.

Voice: The use of the vocal folds and breathing to produce sound.

Fluency: The rhythm, intonation, stress, and related attributes of speech.

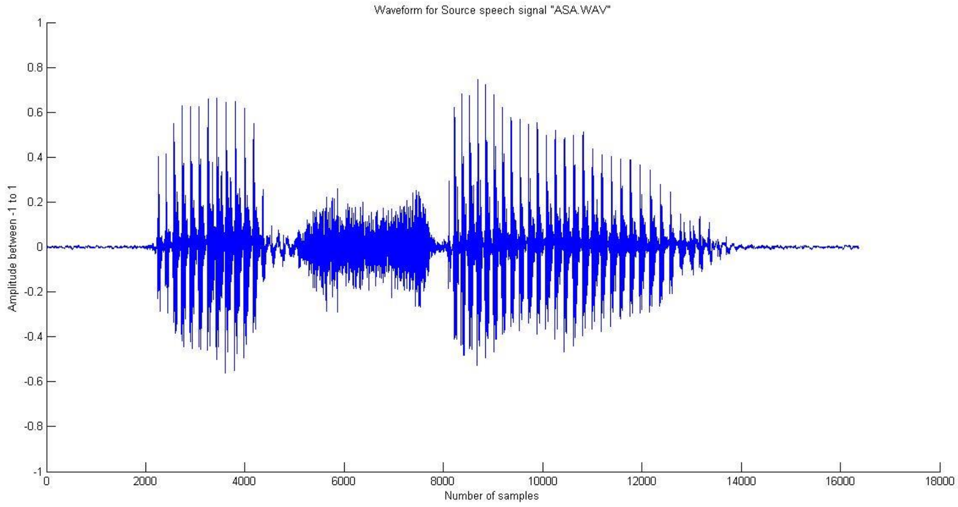


Figure 1.1 shows voiced "ASA.wav"

Normal Hearing

The hearing process starts with catching waves by outer area of the ear. Outer ear helps us to determine the direction of a sound. Next sound travels through 10 mm wide ear canal and reaches to the eardrum and causes it to vibrate. Drum vibrates and shakes three small bones in the middle ear that cause amplify signal 22 times and transfer it to Cochlear (figure 1.2).

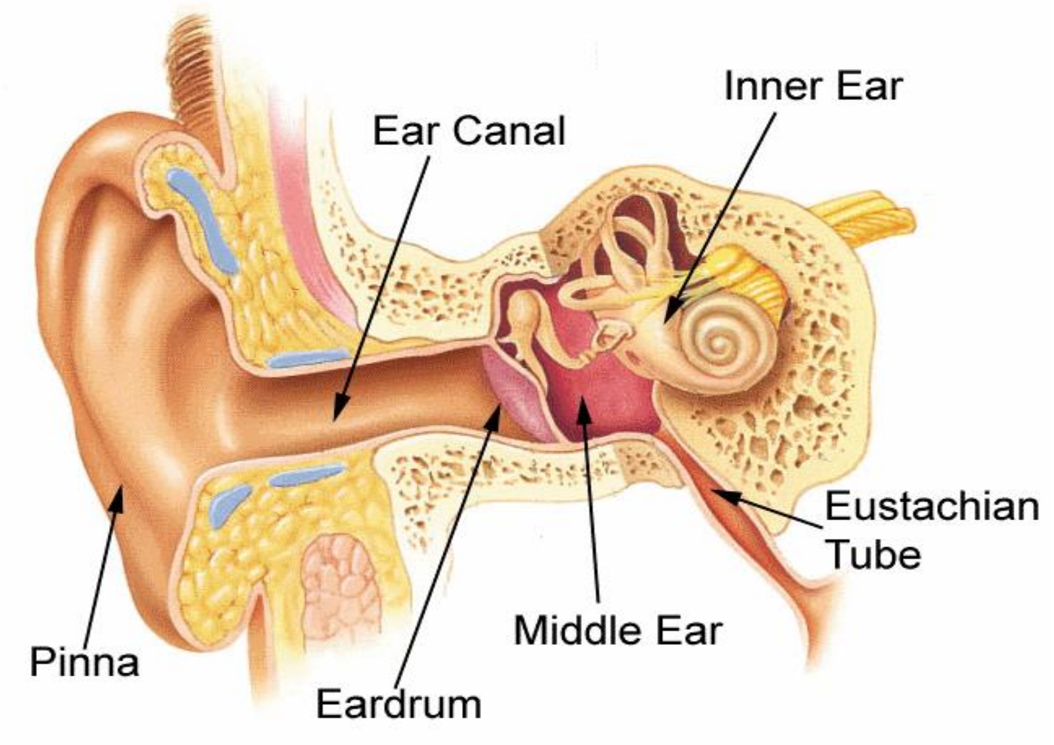


Figure 1.2- Anatomy of the Ear

Cochlear Implants and its structure

People have severe to profound hearing loss in a different age; this may happen by diseases, accident, age or Innate. Hearing loss can be categorized by which part of the auditory system is damaged. There are three basic types of hearing loss: conductive hearing loss, sensorineural hearing loss, and mixed hearing loss. For many years, researchers could not find a solution to restore hearing to profoundly deafened individuals. But starting from the seventies, scientists have been able to make progress in restoring hearing sensation to profoundly deafened individuals by electrically stimulating the auditory neurons.

Fortunately these days, cochlear implant, can be surgically implanted into the inner ear and provide a hearing sensation to a person who is profoundly deaf or severely hard of hearing. It's often referred to as a bionic ear. Cochlear implants can restore hearing in patients suffering deafness due to loss of sensory hair cells in their cochlea. Figure 1.3 shows the cochlear implants. Some people have cochlear implants in both ears (bilateral), and some just have in one ear (unilateral). Cochlear implants often represent speech formants quite well and enable patients to have telephone conversations unaided.

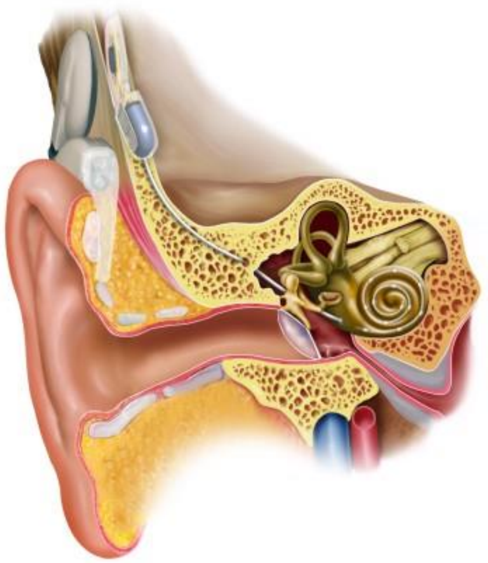


Figure 1.3, Cochlear implant

Not all patients with hearing problems are candidates for cochlear implantation. They need to meet certain audiological criteria. First, the hearing deficiency has to be very severe in one or both ears and receive little or no benefit from hearing aids. Profound deafness [1], generally defined as a hearing loss of more than 90 dB, and hearing loss is typically measured as the average of pure tone hearing thresholds at octave frequencies. Second, the candidate has to score 65% or less on sentence recognition tests done by hearing professional in the ear to be implanted. Since 2000, cochlear implants have been FDA-approved for use in eligible children beginning at 12 months of age.

Over decades, several cochlear implant devices have been developed, and all of them have the following components:

1. The microphone that converts sounds into electrical signals.
2. The sound processor that modifies acoustic signals for the purpose of auditory stimulation.
3. The radio-frequency link that transmits the electrical signals to the implanted electrodes.
4. An electrode or an electrode array (consisting of multiple electrodes) that is inserted into the cochlear by a surgeon during a surgery.

Figure 1.4 describes different section of cochlear implant device.

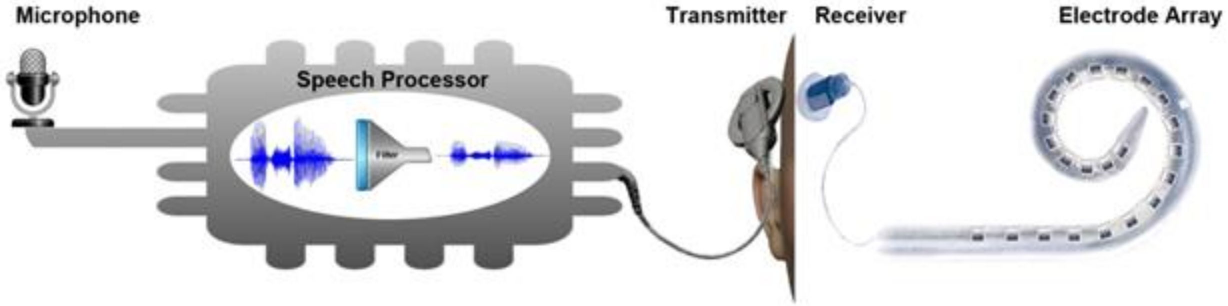


Figure 1.4: Cochlear implant device

In single-channel implants, only one electrode is used. However, in multi-channel cochlear implants, an electrode array that typically consists of 16-22 electrodes is inserted into the cochlea. Thus more auditory nerve fibers can be stimulated at different places in the cochlea(e.g., Shannon et al. [2], Dorman et al. [3]). Different electrodes are stimulated depending on the frequency range of the acoustic signal. Electrodes near the base of the cochlea are stimulated with high-frequency signals while electrodes near the apex are stimulated with low-frequency signals. The primary function of the signal processor is filtering the input signal into different frequency bands or channels and delivering the filtered signals to the assigned electrodes. The sound processor is used to analyze the acoustic signal into different frequency components, similar to the way healthy cochlear processes acoustic signals.

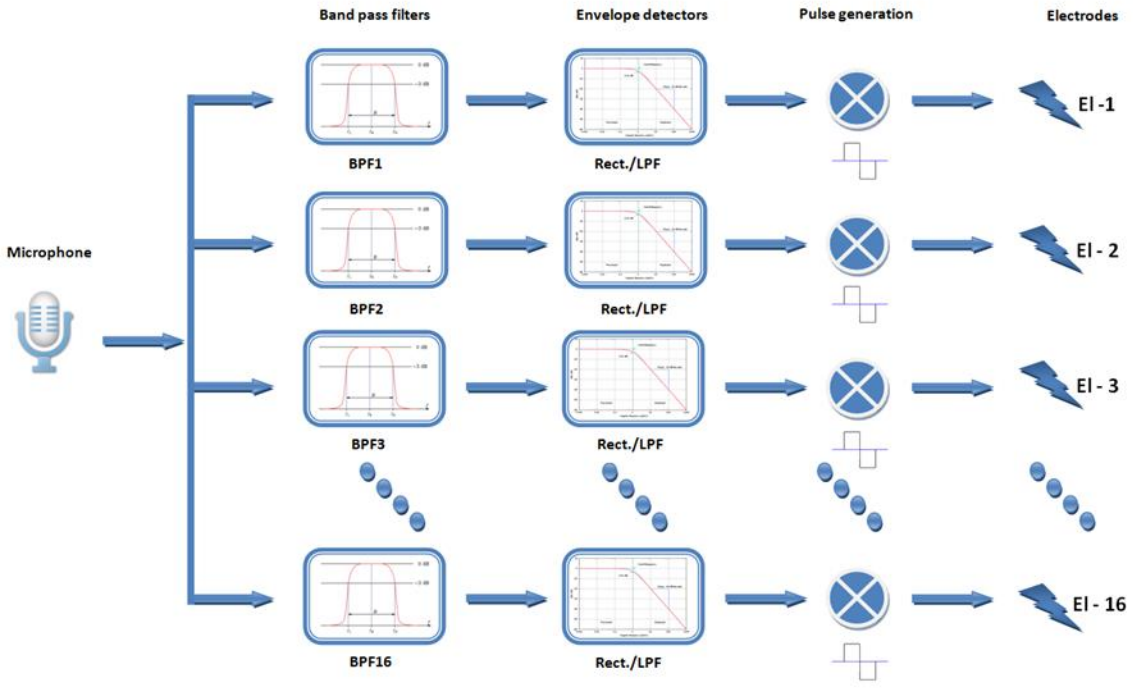


Figure 1.5: Cochlear implant Signal processing

Multi-channel cochlear implants consist of multiple bandpass filters. The total number of bandpass filters is denoted, as the total number of channels. For example, if we have 16 channels we need to have 16 bandpass filters. The Frequency bandwidth for each channel is different from the others and defined in logarithmic scale. The range of frequency is arbitrarily set between 350Hz to 5500Hz for the simulation purposes. Figure 1.6 shows 16 band pass filter.

Band boundaries can be calculated as below:

$$Range=log\ \left(\frac{Upper\_Frequency}{Lower\_Frequency}\right) \quad (1-1)$$

$$Interval=\frac{Range}{Number\_of\_Channels} \quad (1-2)$$

$$Upper\_Bandi = Lower\_Frequency * 10^{Interval*i} \quad (1-3)$$

$$Lower\_Bandi = Lower\_Frequency * 10^{Interval*(i-1)} \quad (1-4)$$

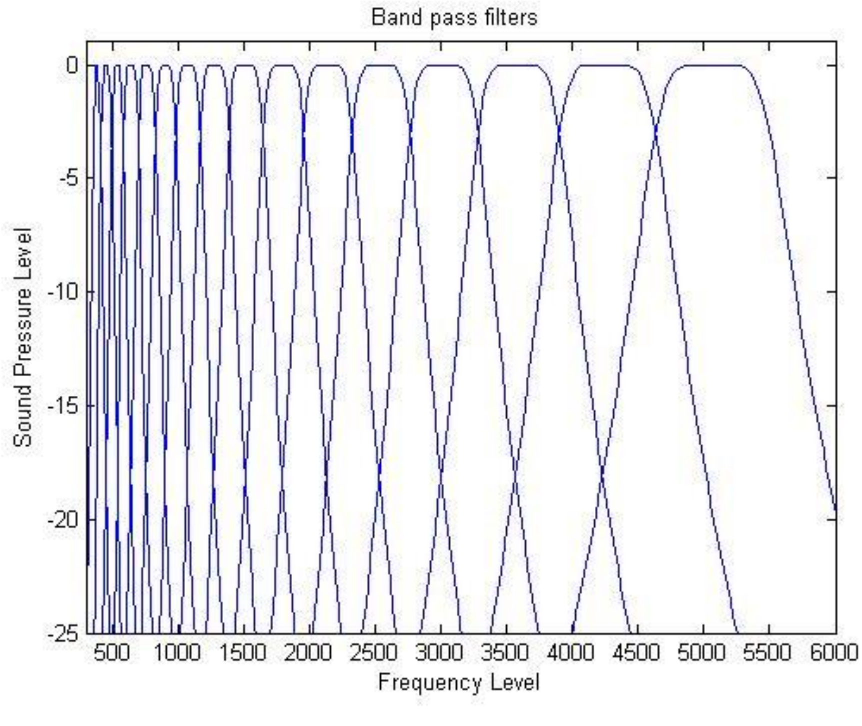


Figure 1.6: Band Pass Filters

16 channels bandpass filters shows in table 1.1

	Lower Band	**Center**	**Upper Band**
Channel 1	350 Hz	382.87 Hz	415.75 Hz
Channel 2	415.75 Hz	454.80 Hz	493.86 Hz
Channel 3	493.86 Hz	540.25 Hz	586.64 Hz
Channel 4	586.64 Hz	641.74 Hz	696.85 Hz
Channel 5	696.85 Hz	762.31 Hz	827.77 Hz
Channel 6	827.77 Hz	905.52 Hz	983.28 Hz
Channel 7	983.28 Hz	1075.64 Hz	1168.01 Hz
Channel 8	1168.01 Hz	1277.72 Hz	1387.44 Hz
Channel 9	1387.44 Hz	1517.77 Hz	1648.10 Hz
Channel 10	1648.10 Hz	1802.91 Hz	1957.72 Hz
Channel 11	1957.72 Hz	2141.62 Hz	2325.52 Hz
Channel 12	2325.52 Hz	2543.96 Hz	2762.41 Hz
Channel 13	2762.41 Hz	3021.90 Hz	3281.38 Hz
Channel 14	3281.38 Hz	3589.62 Hz	3897.85 Hz
Channel 15	3897.85 Hz	4263.99 Hz	4630.14 Hz
Channel 16	4630.14 Hz	5065.07 Hz	5500 Hz

Table 1.1: lower and upper frequencies in the 16 bandpass filters

In cochlear implants, only temporal envelope information is delivered to the auditory neurons. We use envelope detection to generate the output signal for each channel. A common and efficient technique for envelope detection is based on the Hilbert Transform.

In our implementation after bandpass filtering, we use the low pass filter to detect envelop. Our cutoff frequency in discrete-time domain is calculated as below:

$$Lpf = 400\ Hz, \tag{1-5}$$

$$Fs = sampling\ frequency \tag{1-6}$$

$$W0 = Lpf/Fs \tag{1-7}$$

After calculating filter parameters, we use them to generate filtered signal from absolute values of bandpass filtered signals. Ultimately, we modulate the envelope signals to biphasic pulse carriers and deliver them to electrodes. Figures 1.7, 1.8, 1.9 and 1.10, shows different channels for signal “ASA” and envelope detection for each of them.

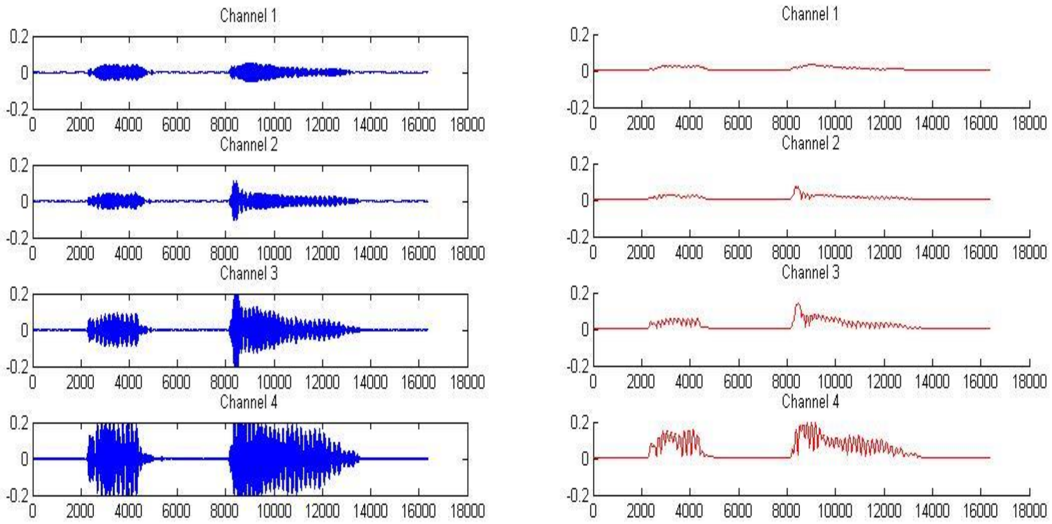


Figure 1.7: Channels 1 to 4 signals and envelop

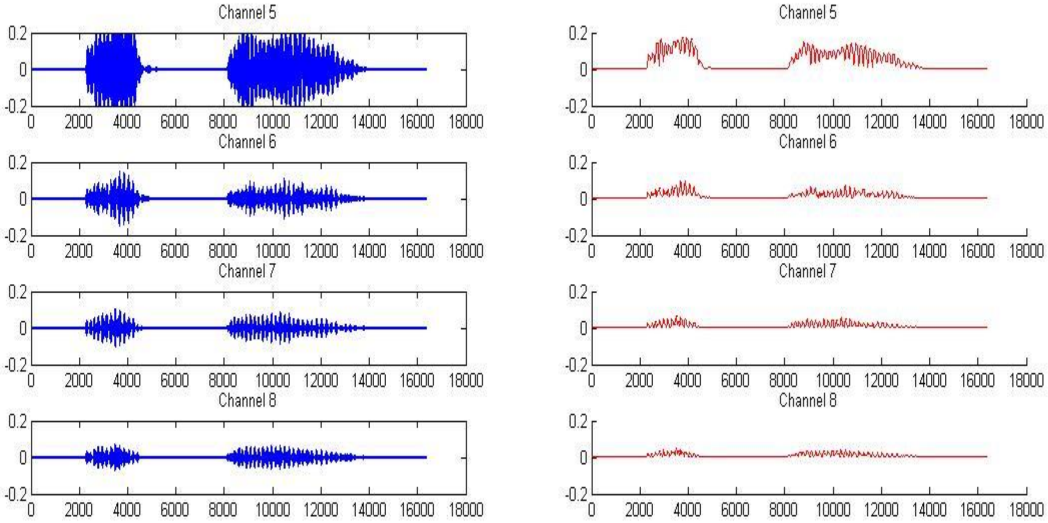


Figure 1.8: Channels 5 to 8 signals and envelop

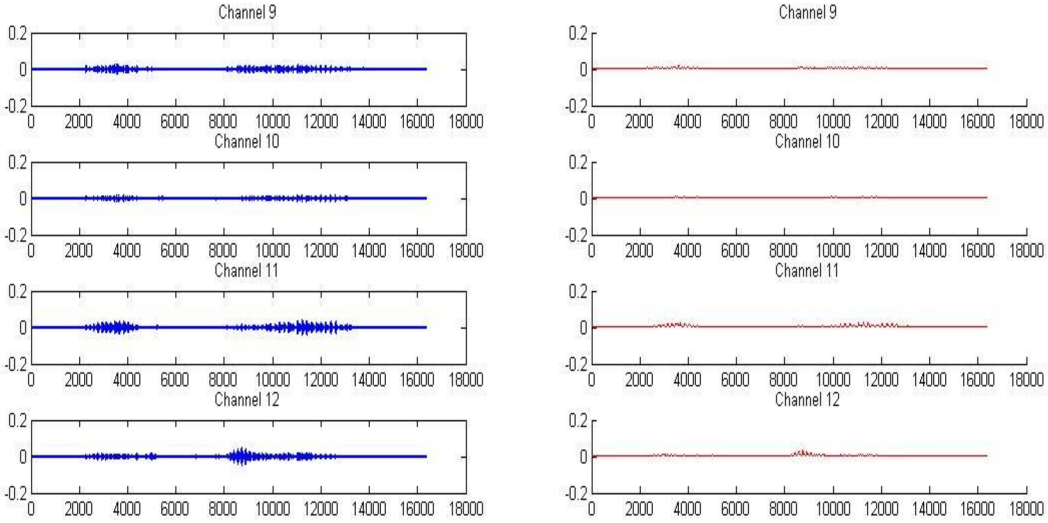


Figure 1.9: Channels 9 to 12 signals and envelop

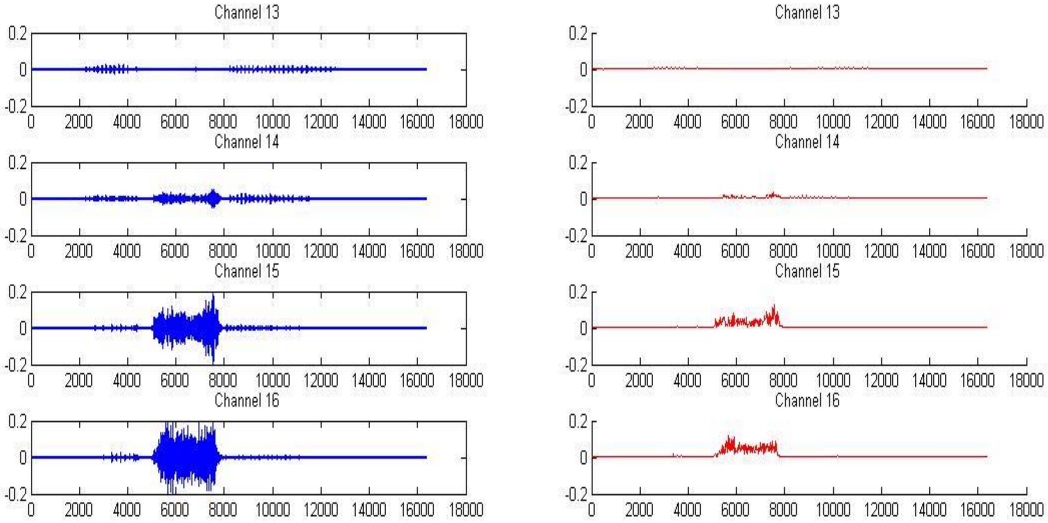


Figure 1.10: Channels 13 to 16 signals and envelop

As can be seen, envelope signal is quite different from the original signal. The problem at hand is how to use normal hearing subjects to test the stimuli heard by the CI user? For simulation purposes, we modulate “white noise” carrier to the envelope signals.

Patient difficulties and limitations

Cochlear Implants can restore sufficient hearing for patients who has profound hearing loss in both ears and allow them to understand of speech in a quiet background. However, they still have serious problems in understanding speech in noisy environments, such as restaurant, airport, and classroom. Several studies show in the presence of background noise, the speech recognition of CI listeners is more sensitive to background noise than that of normal-hearing (NH) listeners. This phenomenon is most likely because of the limited frequency, temporal, and amplitude resolution that can be transmitted by the implant device (Qin & Oxenham, 2003[4]). In fact, a combination of weak frequency and temporal resolution and channel interaction (or current spread) in the stimulating electrodes are some of the reasons implant users have difficulty in recognizing speech in noisy environments. These factors raise the difficulty of hearing in noisy conditions for CI users, regardless of the type of noise (e.g., steady-state or modulated) present (e.g., see Fu & Nogaki, 2004[5]). In the study by Firszt et al. (2004)[6], speech recognition was assessed using the Hearing in Noise Test (HINT) sentences (Nilsson, Soli, & Sullivan, 1994 [7]). Results revealed that CI recipients' performance on sentence recognition tasks in a present of noise was significantly lower than listening at a soft conversational level in quiet.

Present implants are unable to deliver complete temporal and spectral fine structure information to allow implanters to enjoy pitch and musical melody, or to localize sound sources accurately. These limitations make it hard for patients to follow conversations in environments with high background noise.

Chapter Two - Objectives and Motivation

There is no doubt that human life depends on relation and communication with other peoples. Interaction is fundamental to mankind society. “People’s participation is becoming the central issue of our time,” says UNDP in its Human Development Report 1993. Thousands of research and reports show benefits of communication in our daily life. Communication is the root of human sociality since beginning till now, and visiting a relative or friend is essential in social activities. Talking and listening is a core of communication and most people believe that talking to friends and family is very delightful. Some peoples consider the comfortable conversation with family is the primary factor and make it pleasant. However, better perception and understanding can be another reason to attract people to communicate with a familiar person.

An earlier study has revealed that familiarity with a talker’s voice can improve understanding of the conversation. Research has shown that when adults are familiar with someone’s voice, they can more accurately – and even more quickly – process and understand what the person is saying. This theory, identified as the “familiar talker advantage,” becomes into action in locations where it is difficult to hear. For example, in a loud or crowded place, adults can better understand those whose voices they already know [8].

In some cases, manipulation of the talker dimension been shown to improve linguistic performance. Individual research by Magnuson(1995) [9], Nygaard and Pisoni(1998) [10] and Sommers(1994) [11] shows, familiar talkers, improved word recognition in adult speech perception. In those researches, they trained listeners with a set of speakers once subjects are familiar with talkers' voices; they perform a linguistic processing task such as word recognition or sentence transcription. These studies have shown that normal hearing listeners' performance on the linguistic tasks with familiar talkers is consistently better than unfamiliar talkers. Figure 2.1 shows the result of research conducted by Nygaard L. C., Sommers M. S., and Pisoni D. B. (1994) [11]. Mean intelligibility of words presented in noise for trained and control subjects in four SNR levels (-5dB, 0dB, +5dB and +10dB). Trained, or experimental, subjects were trained with one set of talkers and tested with words produced by these familiar talkers. Control subjects were trained with one set of talkers and tested with words produced by a novel set of talkers.

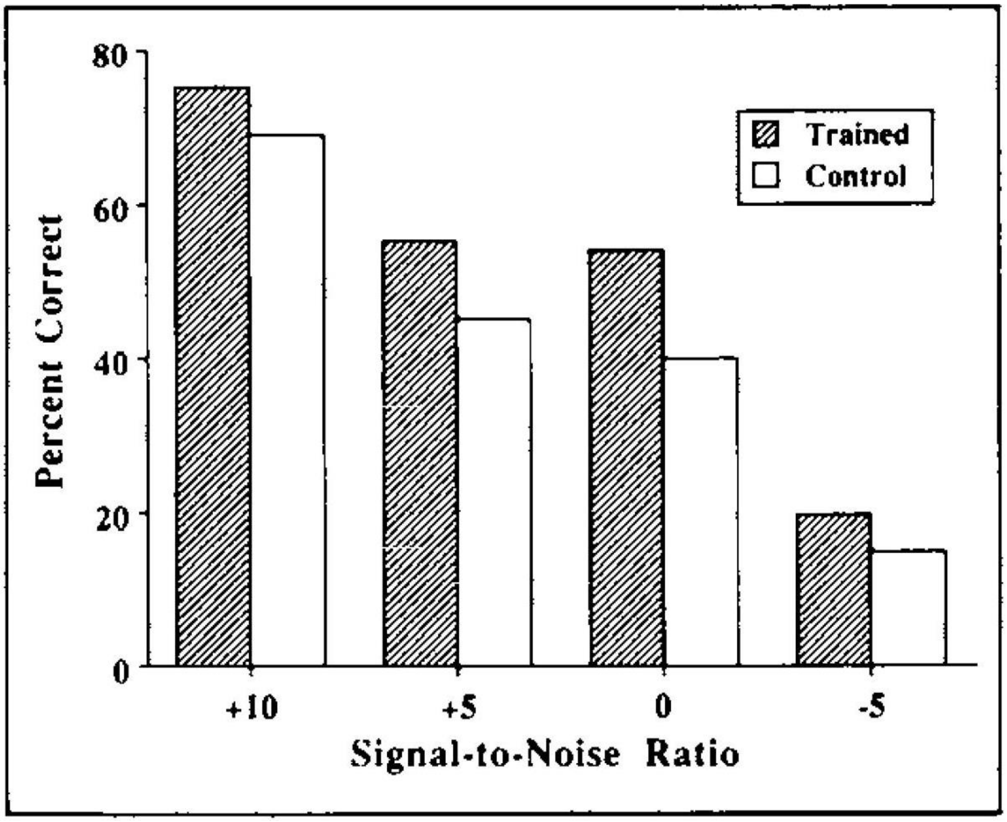


Figure 2.1: Percentage of correct word recognition is plotted at each signal-to-noise ratio. [11].

In research was done by Levi, Winters and Pisoni [8]; Native talkers in two languages (English and German) who were unfamiliar to all listeners selected. Also, listeners divided into four groups based on language and their ability to voice learning. Then listeners trained in six sessions for three days. They assessed the performance of each session by percentage of speaker identification. These results are shown in Figure 2.2.

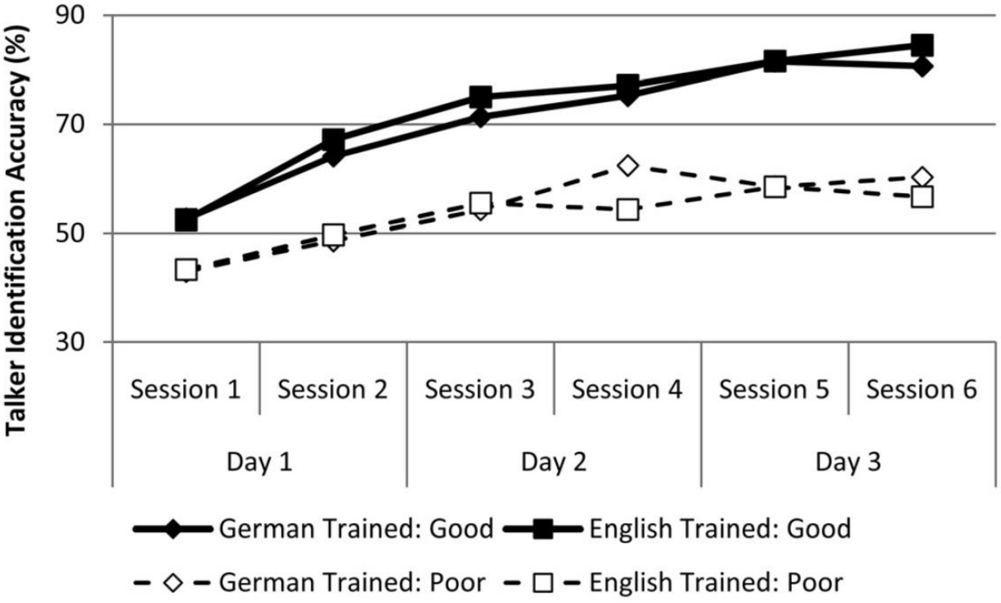


Figure 2.2: Talker identification accuracy during the six training sessions for both German-trained and English-trained listeners separated by good and poor voice learners. Two training sessions were completed on each day of training. by Levi, Winters and Pisoni [8].

Response obtains based on proportion of correct phonemes in three SNR levels (0dB, +5dB and +10dB) to familiar and unfamiliar talker in both languages. Figure 2.3 shows their results.

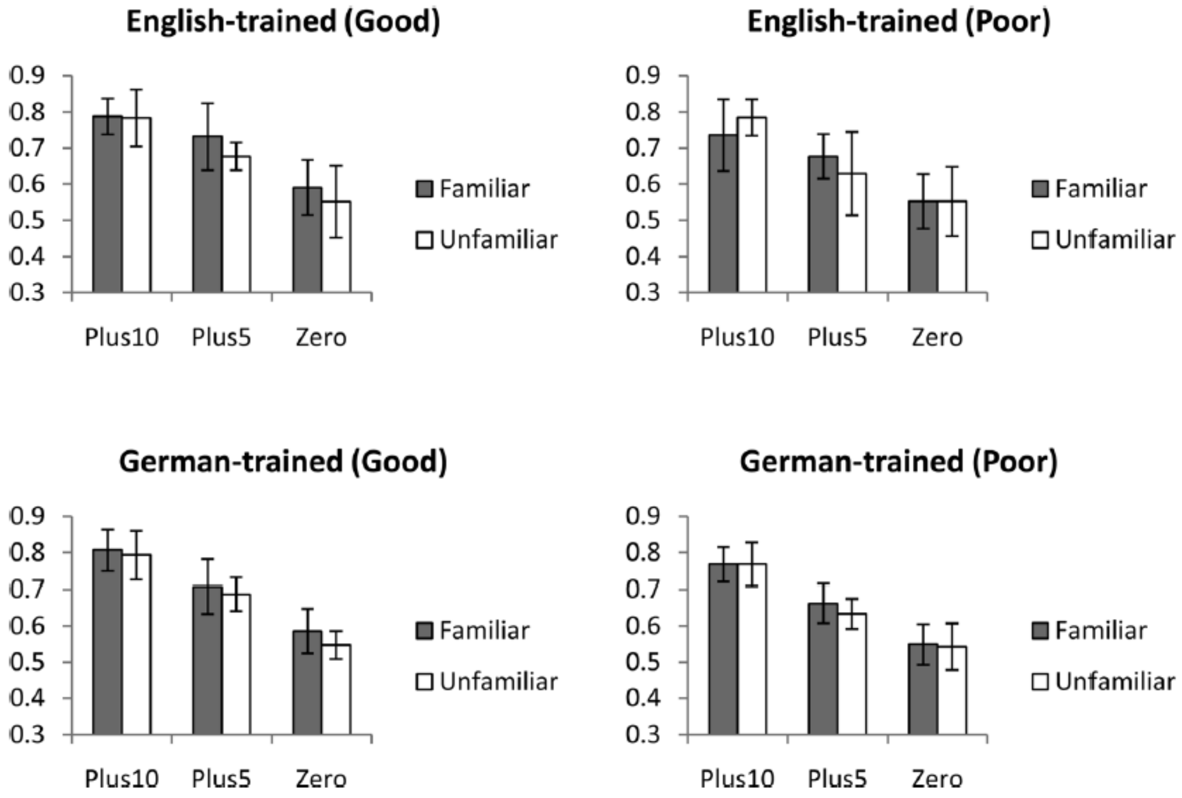


Figure 2.3: Proportion phonemes correct by SNR for English-trained and German-trained learners divided by learning ability (good and poor). Levi, Winters and Pisoni [8].

In a study issued by Hazan and Markham (2004) [12], conducted single word materials (124 words) from 45 talkers (from a homogeneous accent group) and presented to 135 adult and child listeners in low-level background noise. It appears that word intelligibility was considerably related to the total energy in the 1- to 3-kHz region and word duration. Also shows that the relativistic intelligibility of different talkers was very consistent across listener age groups. That implies the acoustic-phonetic characteristics of a talker's speech are the primary factor in determining talker understandability.

Several reports have shown that listeners are highly sensitive to the talker dimension, and it’s been proven that knowledge about a talker alters segmental perception. (Ladefoged and

Broadbent, 1957[13]; Ladefoged, 1978[14]; Johnson, 1990[15]; Johnson et al. 1999[16]; Allen and Miller, 2004[17]; Eisner and McQueen, 2005[18]; Kraljic and Samuel, 2005, 2006, 2007[19][20][21]; Kraljic et al., 2008[22]). Also similar studies for cochlear implant users show CI users are sensitive to different talkers. For example Green et al. (2007) [23] investigated the effects of cross-talker on speech intelligibility in CI users, and normal hearing listeners were listening to acoustic CI simulations. They select two groups of talkers each group consisted of one male adult, one female adult, and one female child. These groups were divided consider to previous data collected with NH listeners according to mean word error rates (high or low intelligibility talkers). Results explained differences in understandability between the two talker groups, for different conditions

The aim of this research is to develop algorithms which can convert speech from one speaker to another one. Based on previews studies Voice conversion algorithms can use to generate familiar talker for cochlear implant (CI) users. This voice can be a voice of family member or friends who have the highest performance of intelligibility for CI user.

Chapter Three - Introduction to Voice Conversion and algorithm

Voice Conversion, which is also mentioned as voice transformation and voice morphing, is a technique to alter a source speaker's speech statement to sound as target speaker. In other words, it refers to a system of changing a person's voice to either make them sound like someone else or to pretend their voice. In voice conversion technique we may change the tone or pitch, add distortion to the user's voice, or a combination of all of the above to reshape the structure of sound and have similar characteristics to another speaker.

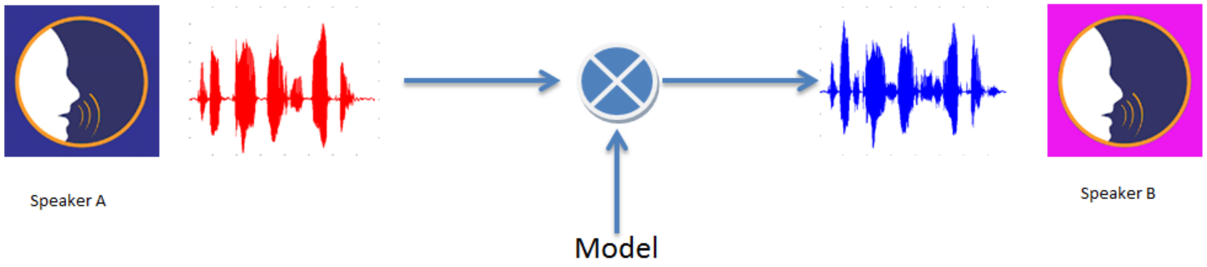


Figure 3.1: Concept of Voice conversion.

There are many applications which may benefit from this sort of technology. For example, a Text To Speech system with voice morphing technology combined can generate various voices. In TTS systems, synthesized speech can be formed by concatenating pieces of recorded speech that are filed in a database. With Voice Conversion technique we can extend these databases and create numerous voices. Another case can be anywhere that the speaker character can change, such as dubbing movies and TV-shows, computer games and cartoons the availability of high-quality voice morphing technology will be very worthy because allowing the fitting voice to be generated without the real actors being present. The other application of using this technique is voice restoration system. People who have lost or damage their speech system can use this device to communicate.

All of the applications mentioned above are assume that both source and target speakers, talk in the same language. However, cross-language voice conversion system assumes that source and target speakers use different languages. In this method timbre or the vocal identity of speaker replace in recorded sentences. [24]

Based on those applications we can divide voice conversion technique to two categories:

- Text-dependent or parallel recording

- Text-independent or non-parallel recording

In Text-dependent systems, the set of utterances recorded by both source and target speakers use as training material to create conversion model. On the other hand, in the text-independent application such as the cross-lingual voice conversion, speakers talk in different languages so it focuses on timbre and acoustic feature of speech.

The process of voice conversion contained two stages:

- Training stage
- Transforming stage

In the training phase(First stage), a set of parallel materials usually refers to a list of sentences covered all phonetic content of language enter as input then aligned for synchronous feature extraction and conversion model creation. Acoustic feature extraction usually down frame-by-frame and 15ms or 25ms commonly select for each frame length. In order to have high-quality data, the sample rate of 8 kHz and 16 kHz with 16 bits or more sample required. Figure 3.2 shows two stages of voice conversion system.

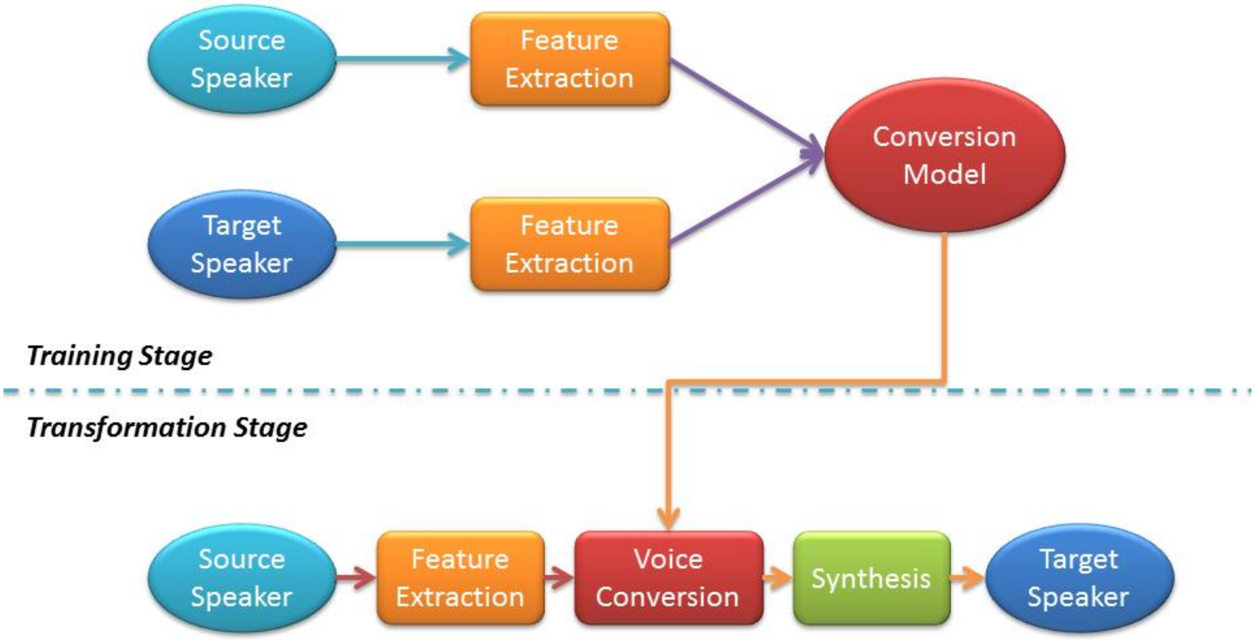


Figure 3.2 shows two stages of voice conversion system.

In the second stage or transformation step, source speech converts to target speech by the model created in the training stage.

Voice Conversion Techniques

For the first time Childers et al.[25] introduced Voice conversion (VC) in 1985. In his paper, he proposed a method that includes mapping acoustical features of the source speaker to the target speaker. In 1986, Shikano[26] offered to use vector quantization(VQ) techniques, and codebook sentences. In 1992, Valbert[27] proposed personalized Text to Speech using Dynamic Frequency Warping(DFW). Seven years later Stylianou[28] introduced Gaussian Mixture Models(GMMs) merged with Mel-Frequency Cepstral Coefficients(MFCCs) for Speaker transformation algorithm. Also, Rentzosin in 2003[29], Ye in 2006[30], Rao in 2006[31] and Zhang in 2009[32] have concentrated on probabilistic techniques, such as GMMs, ANN and codebook sentences.

The transformation stage in voice conversion systems is concerned all acoustic feature applied in the voice signal. Such as pitch shifting and energy compensation. A. F. Machado and M. Queiroz. In Voice conversion: A critical survey [33] categories all transformation techniques in four groups:

- Linear Algebra Techniques
- Signal Processing Techniques
- Cognitive Techniques
- Statistical Techniques

Linear algebra techniques

Linear algebra techniques such as Bilinear Models, Linear predictive coding (LPC), Singular Value Decomposition (SVD), Weighted Linear Interpolations (WLI) and Perceptually Weighted Linear Transformations, and Linear Regression (LRE, LMR, MLLR) are based on geometrical interpretations of data. For example in finding simplified models by orthogonal projection (linear regression), in getting linear combinations of input data (weighted interpolations), or in decomposing transformations into orthogonal components (SVD).

Linear predictive coding (LPC) is a mechanism frequently adopted in audio signal processing, and speech processing such as speech synthesis, speech recognition, and voice conversion. This method used the information of a linear predictive model in the compressed form of a digital signal of speech for representing the spectral envelope. It is one of the most useful methods for encoding good quality speech at a low bit rate and provides extremely accurate estimates of speech parameters and one of the simplest methods of predicting projected values of a data set. It is using previous data in series to predict the next value in a sequence. In this technique, we assume we can learn the behavior of the sequence from a chunk of training information, and then we can apply our learning to places where the next point is unknown. The number of previous points we use to predict the next item in a sequence known as the order of LPC model. For example, if 4 data points are used to predict a fifth one, then we call it an order 4 model. The typical vocal track model an all-pole LTI system.

$$H(z) = \frac{1}{1 + a_1 z^{-1} + a_2 z^{-2} + a_3 z^{-3} + \cdots + a_p z^{-p}}$$

Or

$$H(z) = \frac{1}{1 + \sum_{k=1}^{p} a_k z^{-k}}$$

Where p is the prediction order and the typical value of p are from 8 to 12 (two linear prediction coefficients for each dominant frequencies "formants"). For the signal y(n) an autoregressive(AR) model shown below:

$y(n) = e(n) - a_1y(n-1) - a_2y(n-2) - a_3y(n-3) - \ldots - a_py(n-p)$

Which e(n) is excitation signal.

$$e(n) \longrightarrow \boxed{H(z)} \longrightarrow y(n)$$

The goal is for each given speech signal s(n), find the excitation signal e(n) and coefficients a_k to generate y(n) close to signal s(n). Thus, we can store coefficients a_k and excitation signal e(n) to represent signal s(n). Steps to create LPC model shown in figure below:

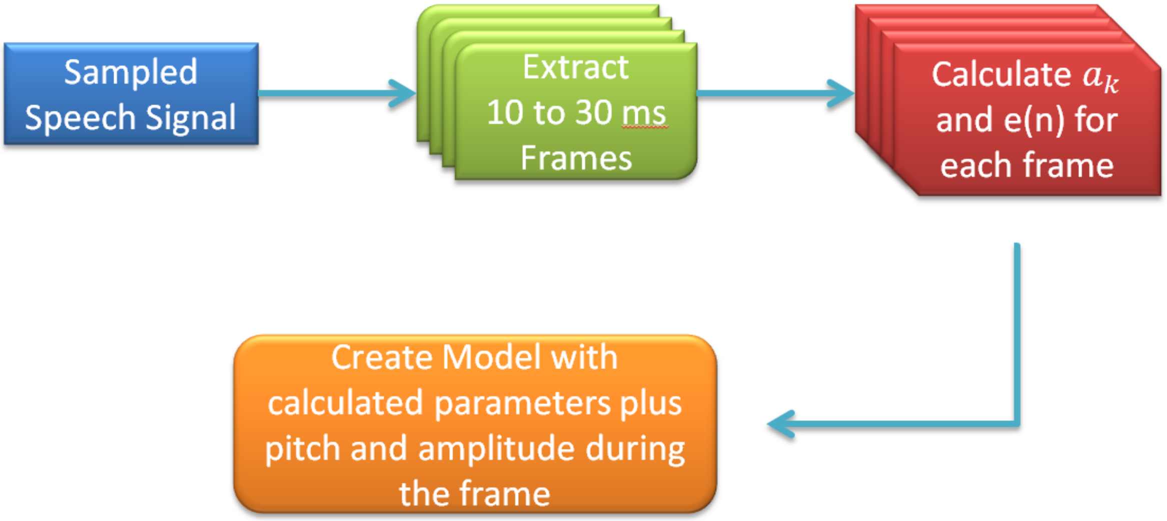

Sampled Speech Signal
Extract 10 to 30 ms Frames
Calculate a_k and e(n) for each frame
Create Model with calculated parameters plus pitch and amplitude during the frame

Figure 3.3 Shows LPC model

In order to calculate LPC coefficients for each frame f(n), first we need to calculate the autocorrelation of f(n):

$$\text{corr}(n) = f(n)*f(n-1)$$

Then solve the Yule-Walker equation [34]. Example below shows Yule-Walker equation for p=4

$$\begin{bmatrix} corr(0) & corr(1) & corr(2) & corr(3) \\ corr(1) & corr(0) & corr(1) & corr(2) \\ corr(2) & corr(1) & corr(0) & corr(1) \\ corr(3) & corr(2) & corr(1) & corr(0) \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ a_4 \end{bmatrix} = - \begin{bmatrix} corr(1) \\ corr(2) \\ corr(3) \\ corr(4) \end{bmatrix}$$

Next step after generating H(z) for each frame is computing the excitation signal.

$$y(n) \longrightarrow 1/H(z) \longrightarrow e(n)$$

1/H(z) is an FIR filter thus excitation signal calculate by filtering s(n) with FIR filter 1/H(z).

We can perform voice conversion by replacing the excitation component from the given speaker with a new one. Since we are still using the same transfer function H(z), the resulting speech sample will have the same voice quality as the original. However, since we are using a different excitation component, the resulting speech sample will have the same sounds as the new speaker.

Signal processing techniques

Signal processing techniques refer to methods such as Vector Quantization (VQ) and Codebook Sentences, Speaker Transformation Algorithm using Segmental Codebooks (STASC), and Frequency Warping (FW, DFW and WFW) which represent transformations based on time-domain or frequency-domain representations of the signal. This group of methods encodes a signal using libraries of signal segments or code words. The other member of this family transforms timbre-related voice features by altering frequency scale representations.

Before we review voice conversion in this group, we need to describe some of the algorithms has an important impact on all techniques. Dynamic time warping (DTW) is an algorithm for measuring the relationship between two temporal series which may vary in time or speed. For example, Finding walking pattern between to person even one of them walks faster than another one. In fact, any type of data such as Audio and Video can be transformed into a linear sequence then analyzed by DTW. In definition, DTW is an algorithm that estimates an optimal similarity between two given sequences (e.g. time sequences) with specific limitations. To measure a correlation between sequences in the time domain, the sequences are "warped" non-linearly. Dynamic frequency warping (DFW) is an exact analog of dynamic time warping (DTW) which is used to reduce the difference in frequency scale of speech and normalize the frequency correctly.

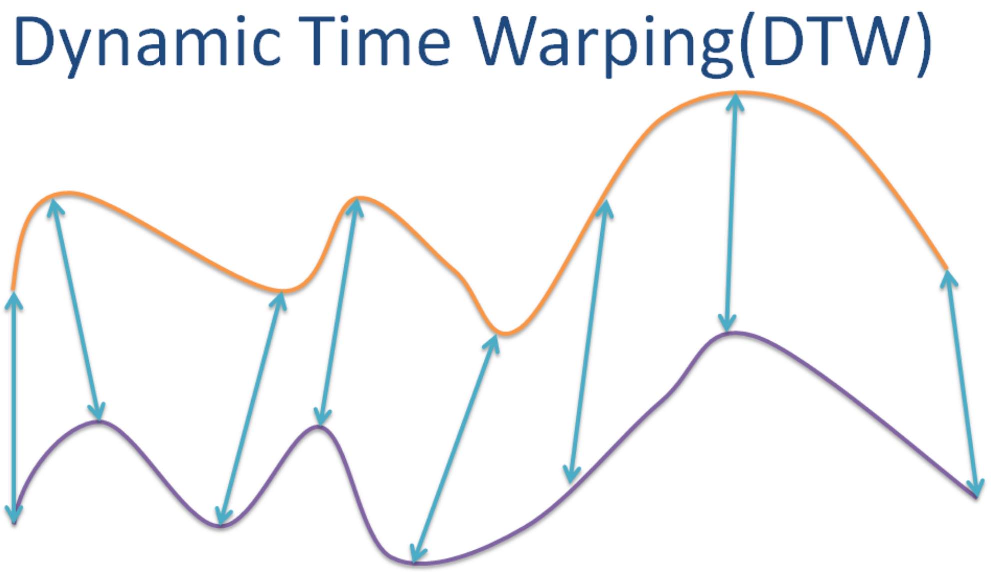


Figure 3.4 Dynamic Time Warping (DTW)

Vector quantization (VQ) is one of the classic signal processing techniques that used for data compression. This method uses the distribution of prototype vectors for modeling of probability density functions. In this technique, large vectors (set of points) separated into groups owning almost the same number of points nearest to them. Each group is expressed by its centroid point same as k-means and some other classification algorithms. In 1986, Shikano[26] introduced the speaker adaptation algorithms which use VQ codebooks of two representatives(input speaker and reference speaker). In this method, vectors in the codebook of a reference speaker exchange with vectors of the input speaker's codebook. Abe et al.(1988)[36] present voice conversion method based on a speaker adaptation method by Shikano et al. (1986). This method relies on producing a discrete mapping between source and target spectral envelopes. This algorithm contains two stages: learning stage and converting stage.

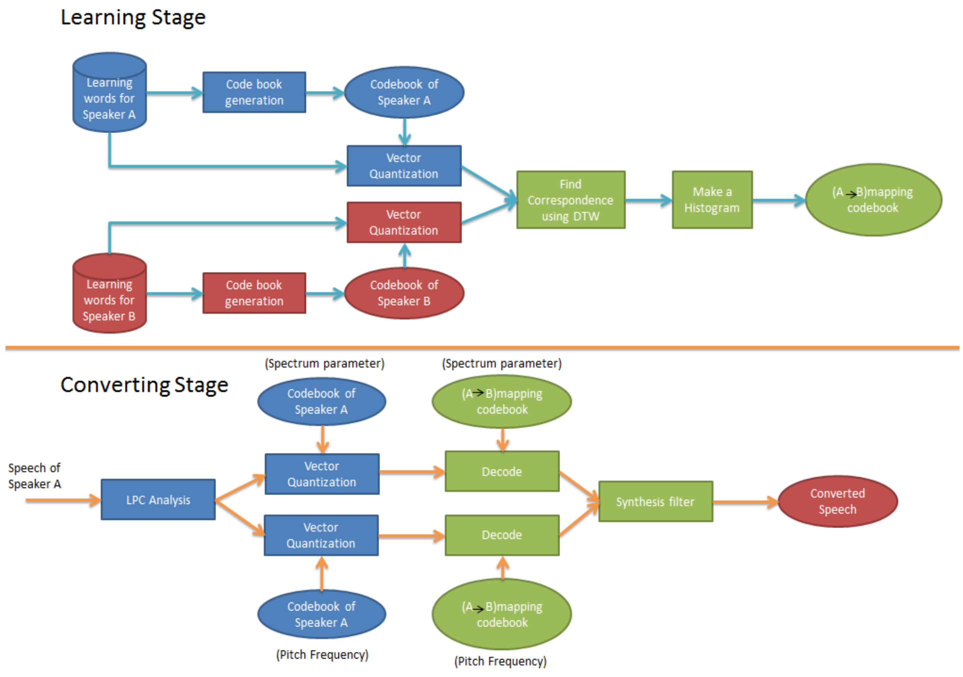
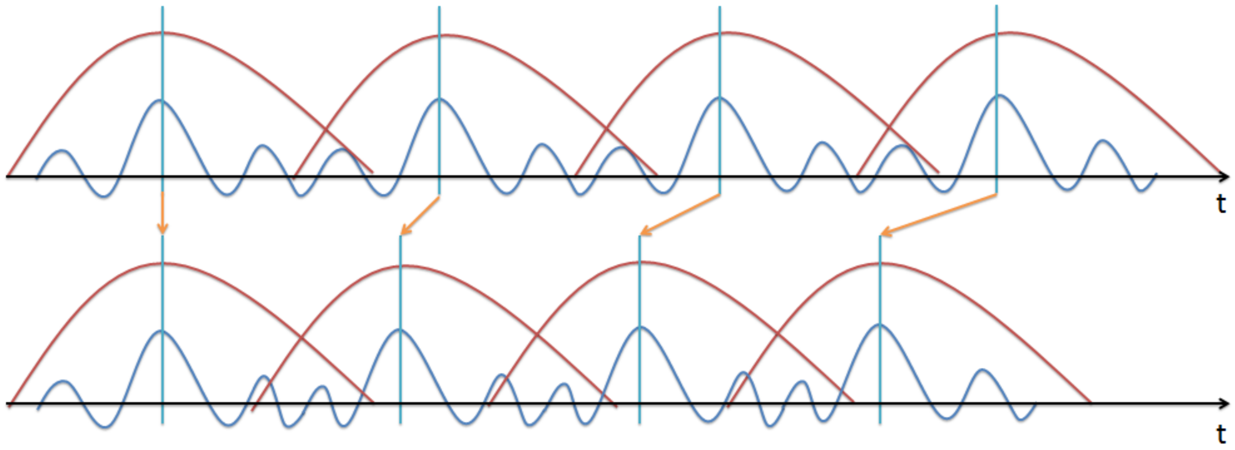


Figure 3.5 (Top) Learning Stage (Bottom) Converting Stage

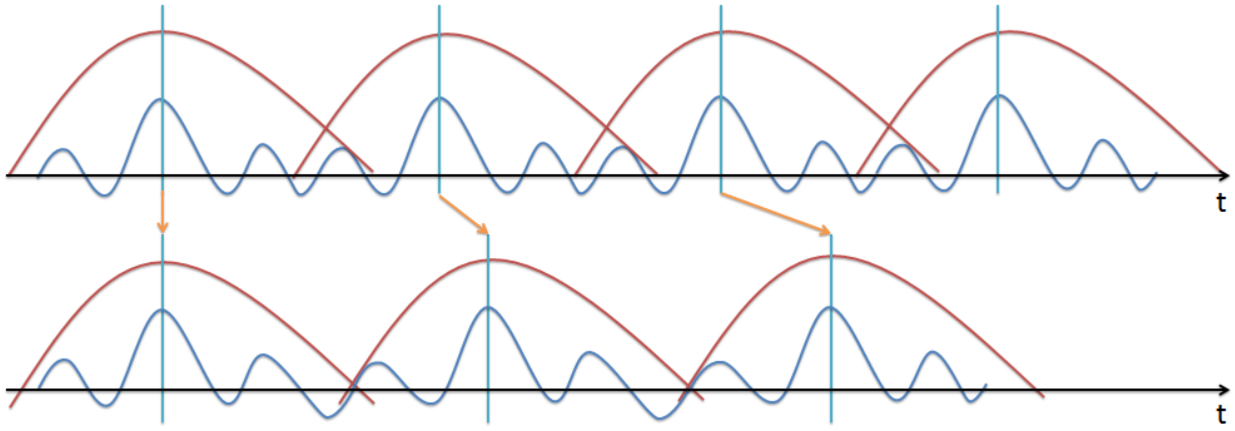
Figure 3.5 shows each step of this method. This method requires parallel recordings from both speakers (source and target) as input to learning stage. The first step is generating spectral vectors for all recordings (in this case Mel-frequency cepstral coefficients).The second step, creating vector quantization (VQ) of each speaker. In next step frames of each sentence align respect to another speaker by using dynamic time warping (DTW) then similarity vector between two speakers store as a 2D histogram. And in the last step of learning stage, it generates mapping codebook for both pitch and energy of speaker B (a linear combination of speaker's B vectors) by using each histogram as a weighting function.

In Transform stage, First step is, Create VQ of input speech with respect to spectrum and pitch. Next use mapping codebook to decode (convert) parameters of speaker A to the speaker B. and at last synthesis (reconstruct) the new speech.

Another famous technique in this group is Pitch Synchronous Overlap and Add (or PSOLA). PSOLA is digital signal processing method used for speech synthesis, and it can be used to modify the pitch and duration of a speech signal. In PSOLA algorithm speech waveform divided into the small overlapping sections. To increasing the pitch of signal, the segments must move closer to each other and to decreasing the pitch of signal distance between segments has to increment. In the same way, to increase the duration of the signal, the segments has to copy multiple times and to decrease the duration, some of the segments have to eliminate. After that, remaining segments merged again using the overlap-add technique.



a) Increasing the pitch



b) Decreasing the pitch

Figure 3.6 shows increasing (a) and decreasing (b) of pitch with PSOLA method.

In 1992 Valbert et al.[27] introduced voice conversion method using all above methods together. Key steps in this approach listed below:

- Divided parallel recording of both speakers (source and target) to frames.
- Align frames using DTW and correlate pairs.
- Extracting pitch marks for each set of speech.
- Vector quantize of source frames
- Build a dynamic frequency warp (DFW) for each cluster between source and target spectral envelope

- adjusting pitch and duration using LP-PSOLA

Cognitive techniques

Cognitive techniques cover all methods that using abstract neuronal structures, and usually depend on a training phase such as Artificial Neural Networks (ANN), Radial Basis Function Neural Networks (RBFNN), Classification and Regression Trees (CART), Topological Feature Mapping, and Generative Topographic Mapping. It is necessary to both inputs, and outputs are available. Usually, they are used for cases that only two possible output values are available. One of the good examples of these decision problems is speech recognition. For solving this problem we need separate network (model), trained for each specific phoneme or word or sentence that is going to be recognized.

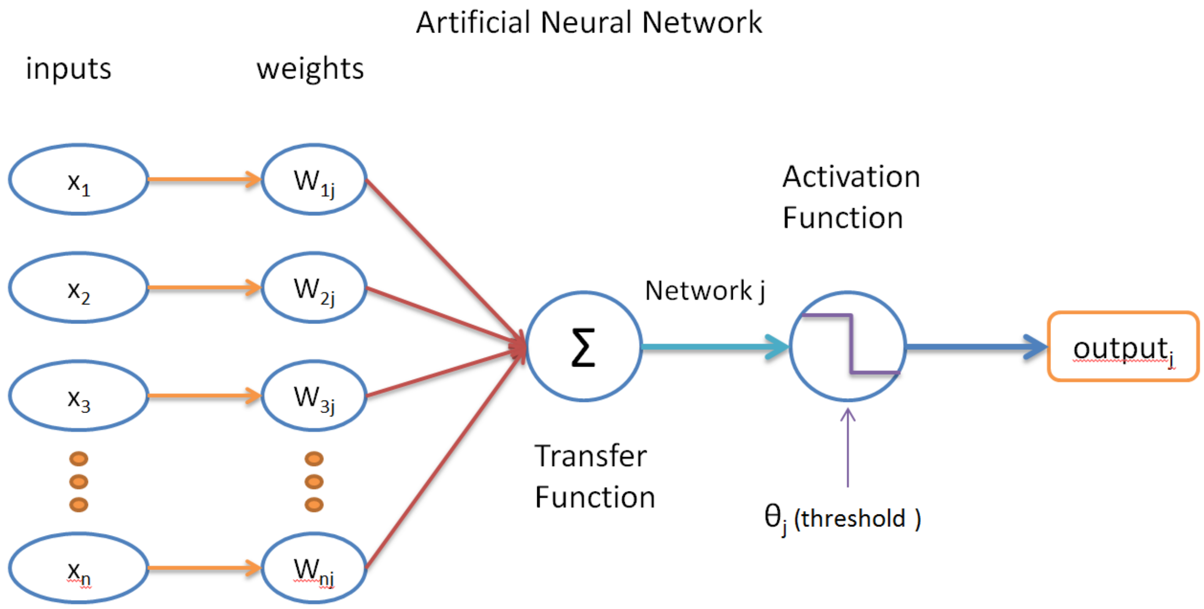


Figure 3.7 Shows Artificial Neural Network

Artificial Neural Network (ANN) algorithms are machine learning and cognitive science technique which are used to compute functions that can depend on a lot of inputs and are regularly unknown. These models include interconnected "neurons" which exchange data between each other. The connection between two nodes has a weight associated with it that can be tuned based on learning. In 2009 Srinivas et al. [35] offer to use this technique(ANN) for Voice conversion. The ANN is trained to convert Mel-cepstral coefficients (MCEPs) of the source speaker to the target speaker's MCEP's. That approach used a parallel set of sentences from source and target speakers. In feature extraction step, MCEPs and fundamental frequency extract as filter parameters and excitation feature. In next step, dynamic time warping is used to align MCEP vectors between the source and target speakers. The output of this stage is set of paired feature vector X and Y which used to train ANN model to perform the transforming from X to Y.

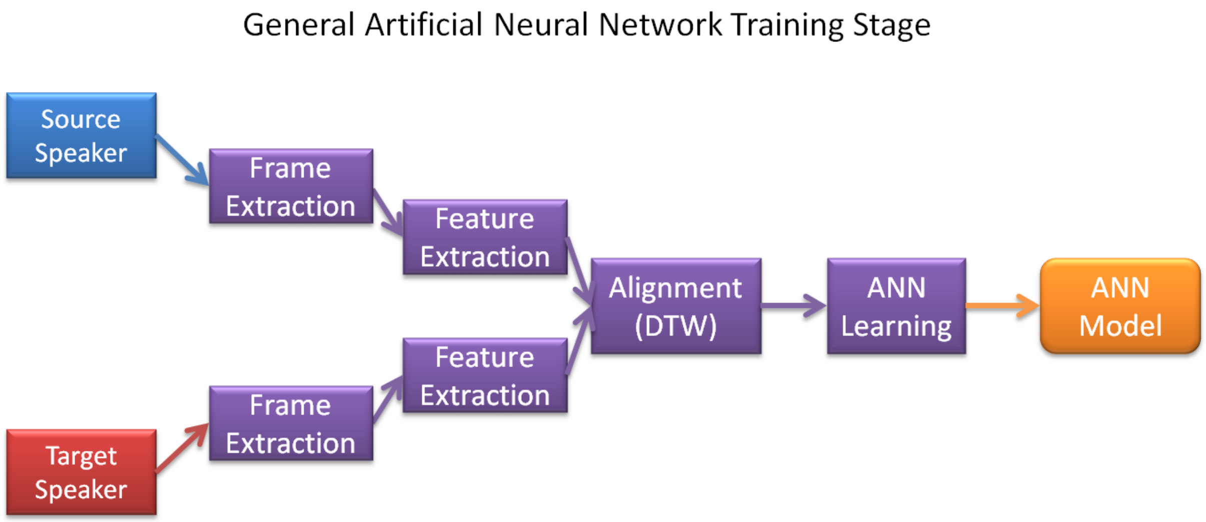


Figure 3.8 Shows Artificial Neural Network learning stage

Statistical Techniques

Statistical Techniques include Gaussian Mixture Models (GMM), Hidden Markov Models (HMM), Multi Space Probability Distributions, Maximum Likelihood Estimators (MLE), Principal Component Analysis (PCA), Unit Selection (US), Frame Selection (FS), K-means and K-histograms. In this group, some of the techniques such as Gaussian model assume feature vectors or vocal parameters have a random component and may be expressed by means and standard deviations. The other group such as Markov models develops over time according to simple rules based on the recent past.

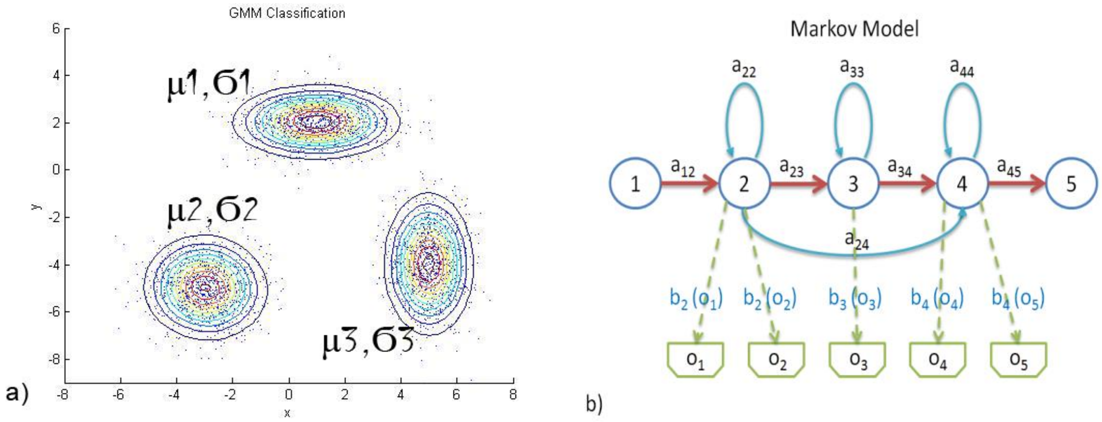


Figure 3.9 a) Shows GMM classification. b) Shows HMM model

A Gaussian Mixture Model (GMM) is a parametric probability density function represented as multiple Gaussian distributions (Distribution based on population mean and the variance). GMM is one of the famous signal processing techniques that use for speaker recognition and in voice conversion system known as the state of the art technology because it has the best quality of transformed speech. GMM parameters are estimated from well-trained datasets using the iterative Expectation-Maximization (EM) algorithm.

The probability density of the Gaussian distribution shown in equation below:

$$f(x|\mu,\sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Where

- x is data point.

- μ is mean or expectation of the distribution:

$$\mu = \frac{1}{M}\sum_{i}^{M} X_i$$

- σ is Standard Deviation and $\sigma^2$is Variance

$$\sigma^2 = \frac{1}{M}\sum_{i}^{M} (X_i - \mu)^2$$

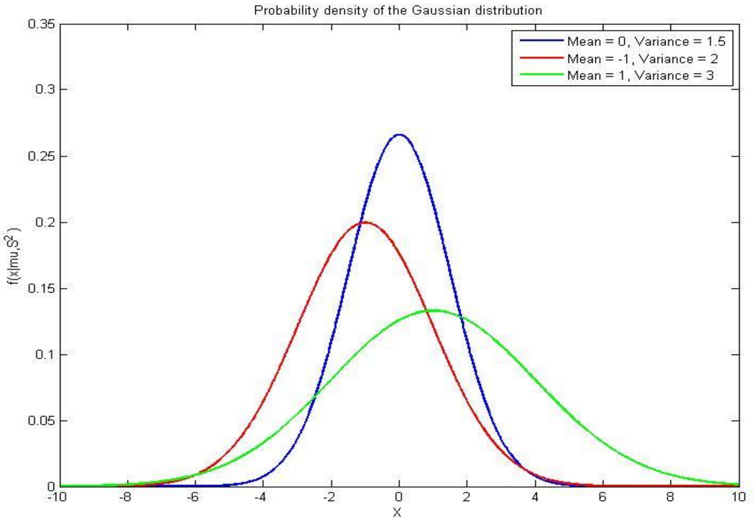


Figure 3.10 Probability density of the Gaussian distribution N(0,1.5), N(-1,2) and N(1,3)

To demonstrate GMM we can use a set of data shown in figure 3.11 (a). That data doesn't like one Gaussian, and it looks like we have three groups of data. Figure 3.11(b) shows simple Gaussian Mixture Model that involves three Gaussian distributions.

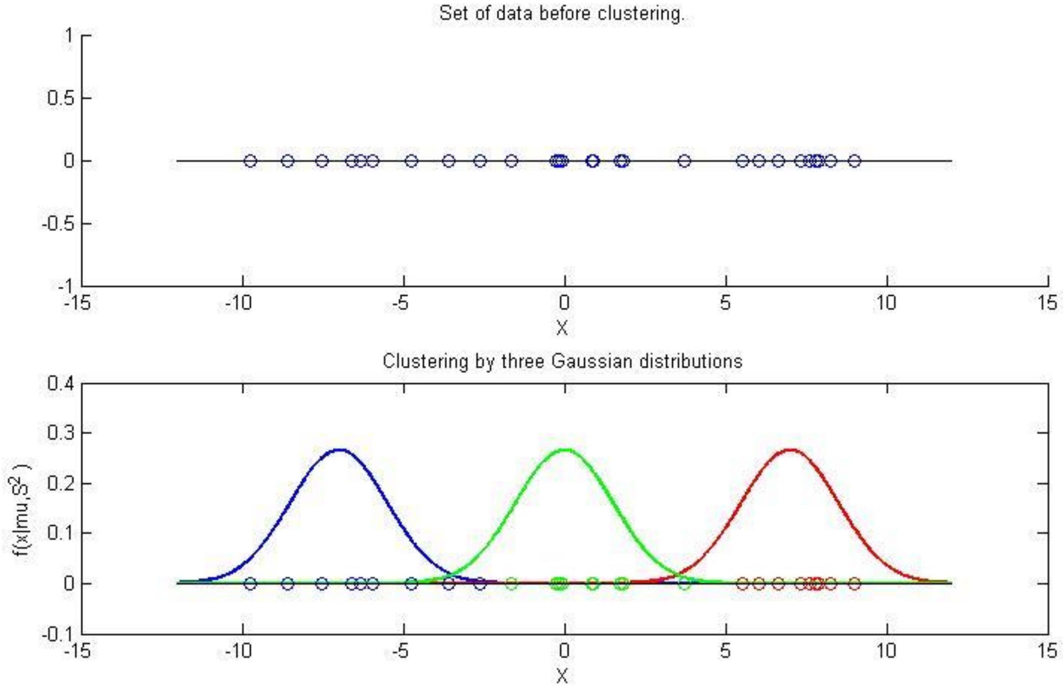


Figure 3.11 Gaussian Mixture Model

GMM has a probability distribution that indicates the probability that each point belongs to the cluster. There are various techniques available for determining the parameters of a GMM. A General Gaussian mixture model is the linear combination of several Gaussian functions given by the equation below:

$$p(x_i|\gamma) = \sum_{i=1}^{k} \omega_i f(x|\mu_i, {\sigma_i}^2)$$

Where x is continuous-valued data vector (features), k is the number Gaussian distribution, ω_i $(i = 1..K)$ are prior probabilities or the mixture weights, γ is set of mixture model with k components $(\{\mu_1, \sigma_1{}^2, \omega_1\}, \{\ldots\}, \{\mu_K, \sigma_K{}^2, \omega_K\})$ and $f(x|\mu_i, \sigma_i{}^2)$ are Gaussian densities given in equation below:

$$f(x|\mu_i, \sigma_i{}^2) = \frac{1}{\sigma_i\sqrt{2\pi}} e^{-\frac{(x-\mu_i)^2}{2\sigma_i{}^2}}$$

One of the most popular methods for Parameter Estimation and unsupervised learning is Expectation-Maximization (EM). Expectation-Maximization (EM) is a parameter estimation algorithm for GMMs that will determine an optimal setting for all of the GMM parameters, using a set of data points. EM algorithm start with initializing each Gaussian model randomly (μ, σ *and* ω), then estimate a new model base on initial one. After that, the new model becomes the initial model for the next iteration, and it will continue until convergence. Figure 3.12 shows parameter estimation for GMM using EM algorithm.

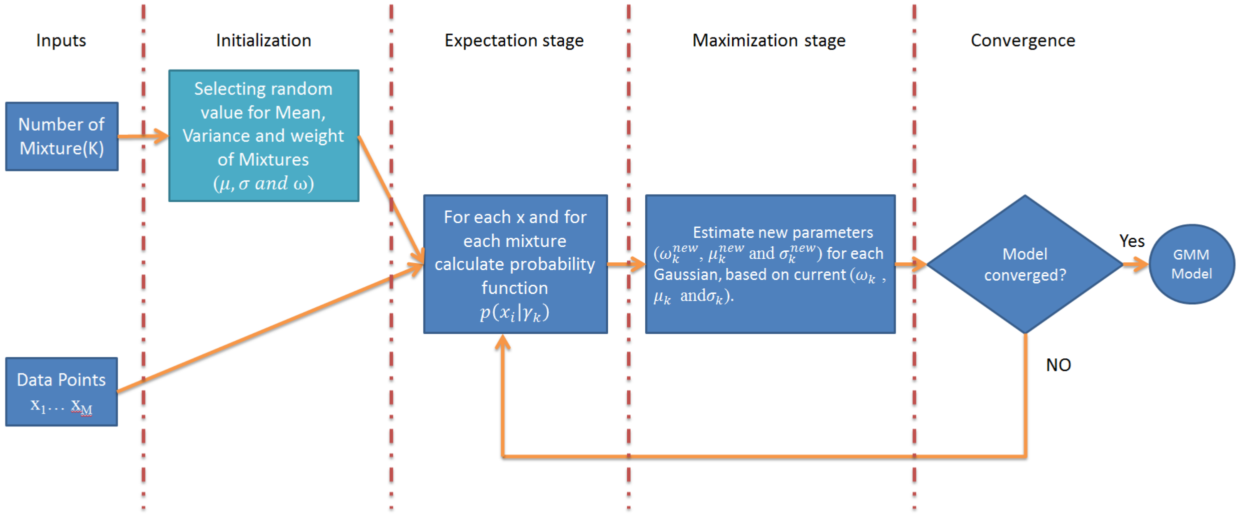

Inputs
Initialization
Expectation stage
Maximization stage
Convergence
Number of Mixture(K)
Selecting random value for Mean, Variance and weight of Mixtures (μ, σ and ω)
Data Points $x_1 \ldots x_M$
For each x and for each mixture calculate probability function $p(x_i|\gamma_k)$
Estimate new parameters (ω_k^{new}, μ_k^{new} and σ_k^{new}) for each Gaussian, based on current (ω_k, μ_k and σ_k).
Model converged?
Yes
NO
GMM Model

Figure 3.12 Gaussian Mixture Model using Expectation-Maximization algorithm

In Expectation stage, for each input data x_i and Gaussian mixture $k^{th}(m_1 \dots m_K)$, probability density needs to compute. The formula below used to derive posterior probability or membership weight:

$$p(x_i|\gamma_k) = \frac{\omega_k f(x|\mu_k, {\sigma_k}^2)}{\sum_{m=1}^{K} \omega_m f(x|\mu_m, {\sigma_m}^2)}$$

Where x_i is from training vector x={x$_1$… x$_M$} and ω_k is weights for mixture k^{th}. In Maximization stage, a new value (ω_k , μ_k andσ_k) for each mixture needs to re-estimate (update). The equations in the Maximization stage required to be calculated in this order, first compute the M new Mixture weight, then the M new means, and finally the M new Variances. Equations below show estimation formula for ω_k , μ_k , and σ_k.

Prior probabilities or Mixture weight:

$$\omega_k = \frac{1}{M} \sum_{i=1}^{M} p(x_i|\gamma_k)$$

Mean (calculate μ_k^{new} for all mixtures):

$$\mu_k^{new} = \frac{\sum_{i=1}^{M} p(x_i|\gamma_k) * x_i}{\sum_{i=1}^{M} p(x_i|\gamma_k)}$$

Variance (used new μ calculated in last step):

$$\sigma_k = \frac{\sum_{i=1}^{M} p(x_i|\gamma_k) * (x_i - \mu_i^{new})^2}{\sum_{i=1}^{M} p(x_i|\gamma_k)}$$

Convergence is commonly recognized by measuring the value of the log-likelihood at the end of each iteration and declares finish when there is no significant change between current and last iteration. We can compute the log-likelihood as defined below:

$$l(\gamma) = \sum_{i=1}^{M} log\left(\sum_{m=1}^{K} p(x_i\,|\,\gamma_m)\right) = \sum_{i=1}^{M}\left(log\sum_{m=1}^{K} \omega_m f(x_i|\mu_m, {\sigma_m}^2)\right)$$

Where M is a number of data points, K is a number of mixtures (Gaussian component), and p is the Gaussian density for the mth mixture component.

In 1998 Stylianou et al. [28] proposed the Gaussian Mixture Model (GMM) for Voice Conversion. Their approach assumes two sets of parallel speech from both speakers are present. Then from each set, spectral envelope (e.g., MFCC) extracted with a fixed 10ms frame rate and aligned using Dynamic Time Warping (DTW). In next step GMM model generate for source data by using Expectation-Maximization (EM) algorithm. And in the final step of learning procedure conversion function generate by applying least squares (LS) Optimization. After a transformation function has been computed, the process can be iterated back to re-estimating the time-alignment step between the transformed envelopes and the objective envelopes. Figure 3.13 shows the block diagram of the learning procedure.

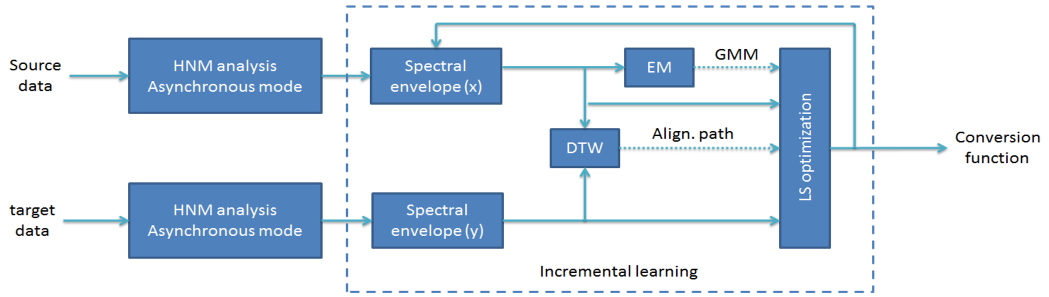


Figure 3.13 Block diagram of the learning procedure [28].

In next stage (conversion stage) use PSOLA like modifications to re-computing the pitch-synchronous synthesis. After applying conversion signal to spectral envelopes, results need to filter to eliminate noises. The noise portion is adjusted with two different fixed filters (so-called corrective filters) for voiced and unvoiced frames. Figure 3.14 shows the block diagram of the voice conversion procedure.

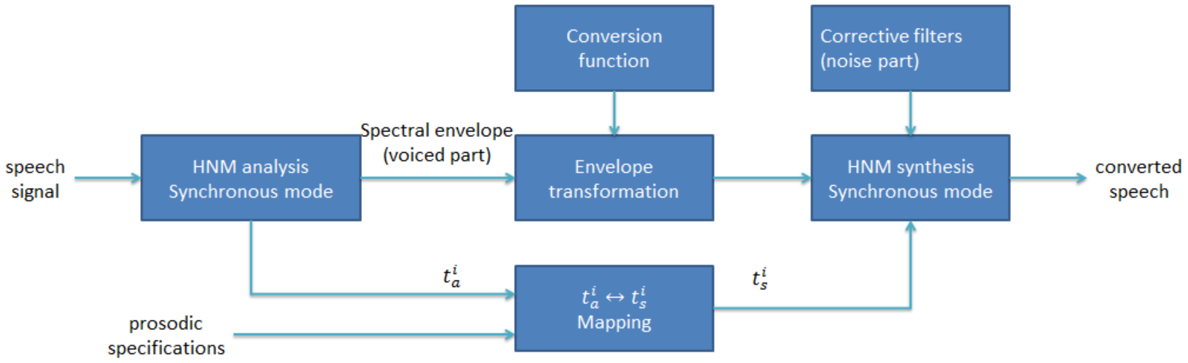

Conversion function
Corrective filters (noise part)
speech signal
HNM analysis Synchronous mode
Spectral envelope (voiced part)
Envelope transformation
HNM synthesis Synchronous mode
converted speech
t_a^i
t_s^i
$t_a^i \leftrightarrow t_s^i$ Mapping
prosodic specifications

Figure 3.14 Block diagram of the voice conversion procedure [28].

Chapter Four – Feature Enhancement

Effects of low harmonics on tone identification in natural and vocoded speech

4-1-1 Introduction

In order to have adequate linguistic communication, for audiences in tone languages same as Mandarin Chinese, the perception of lexical tones is important to understanding word meanings [37]. For example, the syllable [ma] in the Mandarin language can pronounce in four different fundamental frequency (F0) height and contour shapes. And each tone has a different meaning (high-level as tone 1, high-rising as tone 2, low-fall-rise as tone 3, and high-falling as tone 4, means “mother, hemp, horse, and scold,” respectively).[38] F0 height or contour primarily contains the tone information. However, the other cues such as amplitude contour and syllable duration also make contributions. [39][40][41]

When F0 leads are not available such as whispering speech or signal-correlated-noise stimuli, Mandarin-native normal-hearing (NH) audiences were able to recognize Mandarin tones by using temporal envelope cues with an accuracy of 60%–80%[39][40][41]. As a matter of fact, these studies also shows that amplitude contours significantly associated with the F0 contours of natural speech, although such correlations may depend on vowel context, tone category, and speakers. Also, higher correlations between amplitude contours and F0 contours can increase tone identification accuracy.[41] Another study shows that adjusting the amplitude contour to more closely resemble the F0 contour improved tone identification for NH listeners listening to cochlear implant (CI) simulations. This result shows the influence of amplitude contours on tone identification. [42]

In a later study, Luo and Fu[43] measured Mandarin tone, phoneme, and sentence recognition in steady-state speech-shaped noise for Mandarin Chinese-native speakers listening to an acoustic simulation of binaurally combined electric and acoustic hearing (i.e., low-pass filtered speech in one ear and a six-channel CI simulation in the other ear). Results revealed that frequency information below 500 Hz mostly presented to tone recognition, while frequency information above 500 Hz was important to phoneme recognition.

In this study, we investigate the effect of the low harmonics below or near 500 Hz of Mandarin speech on tone identification for natural and vocoded speech. CI patient's auditory perception (e.g., temporal modulation detection and pitch perception) depends on a temporal envelope (e.g., amplitude contour) cues because of lack of resolved harmonics and temporal fine structure ([44],[45]). Besides, the cues of F0, harmonic structures, and temporal fine structures above 500 Hz (Refs. [42] and [43]) are not well maintained in CI processing. Therefore, it is important to recognize the contribution of the temporal envelope of the low harmonics (e.g., the first three harmonics, the frequencies of which are below or near 500 Hz) to Mandarin tone recognition with CI processing.

As we know, speech sounds are processed into a number of continuous frequency bands in CIs, a noise-vocoded speech of each of the three low harmonics was used to examine the effects of their amplitude contours on tone recognition. The effect of background noise was also measured, considering that noise may interrupt the temporal properties of speech signals and low-frequency harmonics. It is assumed that a closer correlation between amplitude contour and F0 contour leads to better tone recognition.

4-1-2-Method

4-1-2-1-Listeners

Ten young Mandarin Chinese-native listeners ranging in age from 20 to 25 years old participated in this study. They had normal hearing sensitivity with pure-tone thresholds less or equal to 15 dB hearing level [46] at octave intervals between 250 and 8000 Hz in both ears. Listeners were paid for their participation. All participants, undergraduate or graduate students in Beijing, China, were from northern China and spoke standard Mandarin. None of the listeners had a formal musical education of more than five years, and no listeners had received any musical training in the past three years.

4-1-2-2-Stimuli

4-1-2-2-1-Speech signal

Four tones of the Chinese vowel /Ç/ were recorded from two young speakers (one female and one male Mandarin native speaker) in isolation forms. The F0 for the female speaker ranged between 160 to 324 Hz and for the male speaker varied between 95 to 215 Hz across the four tones. The average F1 and F2 frequencies of the vowels with four tones were 531 and 1492 Hz for the female speaker and were 430 and 1102 Hz for the male speaker. The duration of each tone was leveled at 210ms, and the vowel signals were normalized to the same root-mean-square (rms) value.

The steps to generate stimuli with individual low-frequency harmonics described as follows. First, vowel signals were segmented into 30ms frames with 50% overlap, and a pitch-detection algorithm based on an autocorrelation function was used to obtain F0 values in each voiced frame. Second, harmonic analysis and synthesis of the vowel signals were conveyed in the frequency domain using a 2048 point fast Fourier transform. The harmonics identified as magnitude spectrum peaks around the integer multiple of F0. For example, The second harmonic classified as the magnitude spectrum peak around 2*F0. There were four harmonic synthesis conditions: first, three using the three individual harmonics (H1, H2, and H3) and the condition four using all harmonics (i.e., all). Third, the synthesis was implemented by using the overlap-and-add approach for the corresponding harmonic condition. The level of stimuli with all harmonics was set at 70 dB sound pressure level (SPL) for both quiet and noise (i.e., 10 dB signal-to-noise ratio (SNR) for the noise condition). Also, the level of stimuli with individual harmonics classified from 56.7 to 66.7 dB SPL, depending on the rms level of each harmonic relative to the rms level of natural speech with all harmonics.

The stimuli in both quiet and noise conditions vocoded with eight-channel noise-vocoder and described as follows. First, Both clean harmonic signals (i.e., H1, H2, H3, and all) and noisy harmonic signals (i.e., signal and noise were mixed before the vocoded processing) were divided into eight frequency bands between 80 and 6000 Hz using sixth-order Butterworth filters. The cutoff frequencies of the eight channels were set as 80, 221, 426, 724, 1158, 1790, 2710, 4050, and 6000 Hz, respectively. Also, the equivalent rectangular bandwidth scale [47] was used to allocate the eight channels.

In the second step, the temporal envelope was extracted by full-wave rectification and low-pass filtering using a second-order Butterworth filter with a 160 Hz cutoff frequency in each frequency band. White Gaussian noise was modulated by the temporal envelope of each frequency band, followed by band-limiting using the sixth-order Butterworth band-pass filters.

Third, the envelope-modulated noises of each band were added together, and the level of the synthesized vocoded speech was normalized to produce the same rms value as the original harmonic signal.

4-1-2-2-2- Noise

400ms of long term speech shaped (LTSS) noise used as the noise due to its comparison to the spectra of speech signals. Gaussian noise that was shaped by a filter with an average spectrum of Mandarin six-talker babble used to generate the LTSS noise at 60 dB SPL[48]. In noise conditions, the 210ms signal was temporally presented at the center of the 400ms noise.

4-1-2-2-3- Stimulus condition

In this study, five factors considered for generating a total of 128 stimulus conditions:

1. Two talkers (one male and one female)
2. Four harmonic conditions (H1, H2, H3, and all)
3. Two types of speech (natural and vocoded speech)
4. Two listening conditions (quiet and noise with a 10 dB SNR)
5. Four tones (tones 1–4)

4-1-2-3- Stimulus presentation

Digital stimuli, sampled at 12207 Hz, were presented to the listeners’ right ears through MDR-7506 headphones. Listeners were seated in the Psychological Behavioral Test rooms of the National Key Laboratory of Cognitive Science and Learning at Beijing Normal University. A

Tucker-Davis Technologies (TDT) mobile processor (RM1) was used for the stimulus presentation. The SPLs of acoustic stimuli were calibrated in a NBS 6 cm3 coupler using a Larson-Davis (Depew, NY) sound-level meter (Model 2800) to set the linear weighting band.

4-1-2-4- Procedure

The experimental procedure was handled by TDT SYKOFIZX VR v2.0 (Alachua, FL). After each stimulus presentation, the listeners' asked to identify and select the tone of the signal from four-choice close-set buttons (tones 1, 2, 3, and 4). Each signal was presented 15 times to each listener, resulting in 1920 cases (128 stimulus conditions * 15 repetitions). For a given block, the tone of the stimulus was randomly presented with the four remaining factors (talker, harmonic, listening condition, and type of speech) fixed. To familiarize listeners with the vocoded speech and experimental procedure, listeners provided with 30 min training session using six blocks of vocoded speech (vowel /i/) before the test session.

4-1-3- Results

Figure 4.1 illustrated the tone identification as a function of the four tone categories for the female speaker (upper panels) and male speaker (lower panels), and for natural (left panels) and vocoded (right panels) in quiet [Fig. 4.1(a)] and noisy [(Fig. 4.1(b)] conditions, respectively. In quiet condition, tone identification was 96% for natural speech and 69% for a vocoded speech on average over the speakers, harmonics, and tone categories; although in noisy condition, the average tone identification became 89% and 38% for natural and vocoded speech. For statistical objectives, the intelligibility scores were converted from percentage correct to rationalized arcsine transformed units (RAU), extending the upper and lower ends of score ranges. [49]

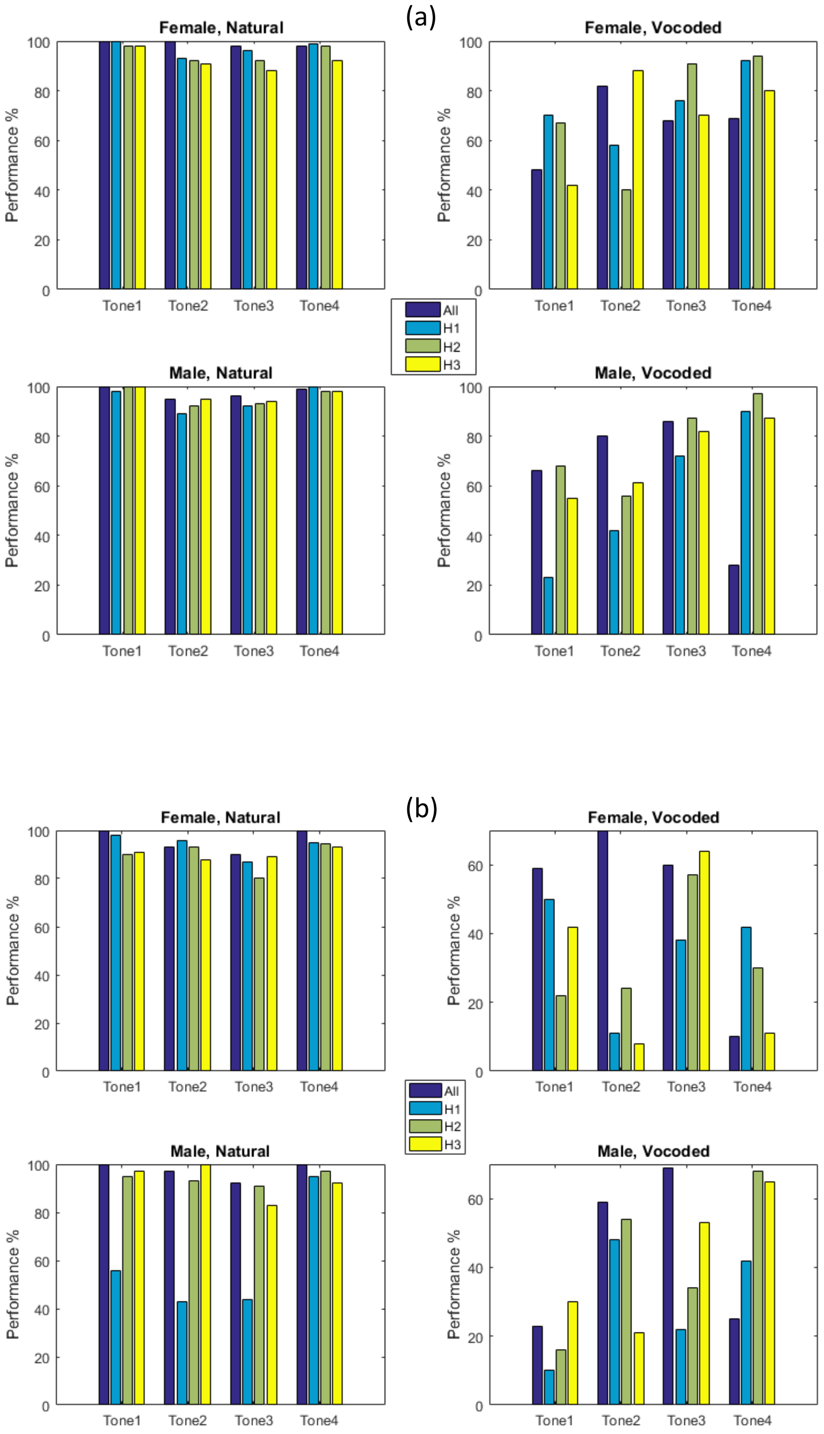


Figure 4.1: Tone identification as a function of the four tone categories in the quiet condition (a) and noisy condition (b). Listening condition for natural (left panels) and vocoded (right panels) speech and for the female (upper panels) and male (lower panel) speakers (All: all harmonics included; H1: the first harmonic; H2: the second harmonic; H3: the third harmonic).

A five-factor (talker * speech type * harmonics * listening condition * tone category) repeated-measures analysis of variance (ANOVA) with tone identification scores in RAU as the dependent variable was produced. Results showed that tone identification was significantly influenced by speech type, listening condition, tone category, and stimulus harmonics, but not by speakers. All the two-factor interaction effects were significant, besides the interactions of talker * speech type, talker * listening condition, speech type * stimulus harmonic, and listening condition * tone category. All the three-factor interaction effects were significant, except the interactions of talker * speech type * stimulus harmonic, and talker * listening condition * stimulus harmonic. In addition, all the four-and five-factor interaction effects were significant.

Results showed that the effect of three-factor interaction (speech type, listening condition, and stimulus harmonic) was notable, to expose the main effect of stimulus harmonic. Therefore four three-factor (talker * stimulus harmonics * tone category) repeated measures ANOVAs with tone identification in RAU were conducted for natural and vocoded speech in the quiet and noisy conditions. In quiet with natural speech, there was no significant effect of any of the three factors, nor of the two and three-factor interactions, showing that single low-frequency harmonics of natural speech carried enough tone information in quiet. When the vocoded speech was performed in quiet, tone identification was significantly affected by stimulus harmonic, tone category, and all the two-factor and three-factor interactions, but not by a speaker. For natural speech in the noise condition, there was a significant effect of stimulus harmonic and tone category, but not by the talker. Also, all the two-factor and three-factor interaction effects were significant, except the two-factor interaction of speaker * tone category. As vocoded speech was presented in noisy condition, all the one-factor, two-factor and three-factor effects were significant except speaker.

In fact, as indicated in the statistical results above, the interaction effects of stimulus harmonic and the other factors were significant except for the natural speech in quiet. Thus, two-factor (stimulus harmonics tone category) repeated-measures ANOVAs were conducted for the male and female talker, separately, for natural speech in noise and vocoded speech in quiet and noisy conditions. Results are shown in Figures 4.1(a) and 4.1(b); a significant difference in tone identification between individual-harmonic conditions and the all-harmonic condition was designated with an asterisk (*).

To explain the complex pattern of the effects of stimulus harmonics, a linear regression used between tone identification scores averaged over the ten listeners and the amplitude-F0 contour index for vocoded speech across the quiet and noisy conditions, following Fu and Zeng.[41] The amplitude-F0 contour index was calculated as the correlation between the F0 contour of the corresponding harmonic signal and the amplitude contour of a vocoded stimulus.

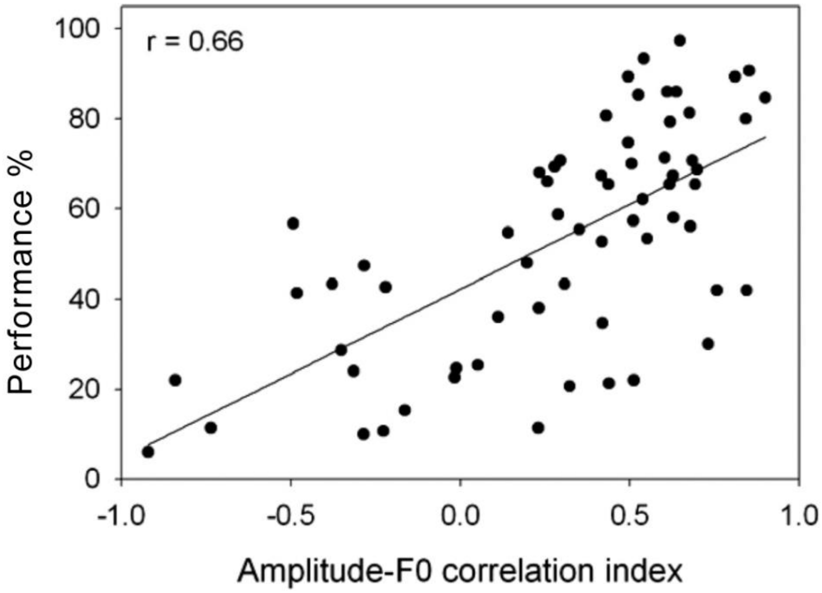


Figure 4.2: Linear regression function of average tone identification scores over the ten listeners vs. amplitude-F0 correlation index (Pearson r¼0.66, p<0.01).

As shown in Figure 4.2, there was a significant correlation between tone accuracy and the amplitude-F0 correlation index for vocoded speech, implying that as the amplitude contours were closer to the F0 contour, tone identification scores were raised.

4-1-4- Discussion

The purpose of this study was to investigate the contribution of low-frequency harmonics to Mandarin tone identification in vocoded speech simulating CI processing. Overall, the effects of low-frequency harmonics on tone identification were complicated, depending on the other four factors. For natural speech presented in quiet and noisy conditions of a 10 dB SNR, the three low harmonics of both male and female speakers led to high tone identification accuracy (>90%), except H1 of the male speaker in noisy condition. The result of this condition might be due to relatively low audibility at the low-frequency (H1) of the male speaker (e.g., ~161 Hz). For vocoded speech in a quiet condition, tone identification with the three low harmonics was comparable to that with all harmonics, with average scores reaching about 60%–70%. However, for vocoded speech in noise, average tone identification over the talkers and tone categories was reduced to 30%–40% and was better for all harmonics than for individual harmonics, especially H1. It should also be mentioned that the contribution of low-frequency harmonics changed with the tone categories for vocoded speech. For instance, for tones 1, 2, and 3, the stimuli with all harmonics had better or similar tone identification compared with the stimuli with individual harmonics, whereas for tone 4, the stimuli with all harmonics had substantially lower tone accuracy than those with individual harmonics. As shown in Figure 4.2, the complicated pattern of tone identification for the vocoded speech was partially considered by the amplitude-F0 contour index. Furthermore, tone identification was significantly better for tone 3 and tone 4 than for tone 1 and tone 2, especially for vocoded speech in quiet, consistent with previous reports.[40]

These results designate that the frequency contour of F0 and its harmonics played a dominating role in determining tone identification in quiet and in relatively high-SNR conditions, even though the frequency contour and amplitude contour of natural speech sometimes were not well correlated (e.g., the low amplitude-F0 contour index for tone 4). In a natural speech with phonation, frequency contours of low-frequency harmonics appeared to coincide with F0 contours such that high tone accuracy is reached. However, when low harmonics like H1 were not well audible in noise, tone identification was significantly reduced for tones 1, 2, and 3 of the male speaker (see the left-lower panel of Figure 4.1(b)). Interestingly, the identification score of tone 4 for the male H1 was quite high (96%), also possibly due to the relatively better audibility of the male H1 for tone 4 (e.g., the frequency ranged from 224 to 109 Hz). On the other hand, for vocoded speech, which removed F0 and tonal duration cues in the present study, listeners appeared to rely on temporal envelope cues like amplitude contour to identify tones [41], although tone identification was significantly degraded. Moreover, tone identification of vocoded speech was significantly lower in noise than in quiet, partially due to the disruption of the amplitude contour of speech signals by noise.

The effect of amplitude contour on tone identification seemed to be frequency independent. That is, regardless of broad frequency ranges for the stimuli with all harmonics or narrow low-frequency ranges for the individual harmonics. Tone identity was significantly connected with the amplitude contour of the stimulus. For example, tone 4 had significantly lower identification scores for the stimulus with all harmonics than for stimuli with individual harmonics, mainly because the amplitude contour was slightly rising or flat for the stimulus with all harmonics (i.e., the amplitude-F0 contour index ranged from -0.23 to +0.23), whereas H1, H2, and H3 had falling amplitude contours (i.e., most of the amplitude-F0 contour index ranged from 0.54 to 0.85). This finding implies that for vocoded speech, “more” frequency information in vowel signals (i.e., more harmonics) did not necessarily produce better tone recognition when the amplitude contour of vowel signals did not follow the F0 contour. This result is likely due to the complex interaction pattern among the amplitude contours generated by individual harmonics. However, the amplitude contour only partially considered for tone identification with vocoded speech, and other factors may contribute to listeners’ performance. For example, to be consistent with earlier studies[50], a cutoff frequency of 160 Hz was used in the low-pass filter for envelope extraction, which affected the periodicity fluctuations important to the temporal pitch cues available to CI users, especially under noisy conditions. Furthermore, the intensities of individual harmonics were observed to be quite varied in different listening conditions, and this might contribute to the audibility differences of the temporal envelopes.

The results obtained with NH listeners of the present study suggest that listeners relied on the overall temporal envelope cues of the stimulus, although the temporal envelope of each frequency band may differ from the overall temporal envelope. Considering that speech sounds are processed into several frequency bands in CI processing, the amplitude contours of each frequency band may carry tone information; however, the amplitude contours of individual frequency bands may be integrated into the central auditory system for CI users.

The Contribution of Matched Envelope Dynamic Range to the Binaural Benefits in Simulated Bilateral Electric Hearing

4-2-1-Introduction

A significant amount of proof exists to supporting the hearing benefits of binaural cochlear implants (CIs) regarding better speech recognition in challenging environments [51], sound localization [52], language acquisition and learning [53], and quality of life [51]. The binaural benefits for CI speech perception in noise have been receiving an increased level of interest, with researchers examining the head-shadow effect, the binaural summation effect, and the binaural squelch effect [54]. The head-shadow effect has probably the highest influence on binaural CI hearing, and the difference in speech recognition thresholds (SRTs) for the two monaural listening conditions (one condition with the CI closer to the noise sources and the other condition with the CI on the shadowed side) ranges between 4 and 7 dB[55]. The benefit of the binaural summation effect (i.e., presenting the identical stimulus bilaterally) is essentially associated to the redundancy of information in the stimuli presented at the two ears. The summation benefits observed from listeners with bilateral CIs are relatively larger [56] than those seen from normal-hearing (NH) listeners (it was suggested to be approximately 1 dB in Bronkhorst & Plomp, [57]). Although the benefit of the binaural squelch effect is rarer and is sensitive to the test conditions, a few studies suggested that some bilateral CI users obtained this benefit [55].

Lately, several studies in which scientists focused on how the unilateral performance affected binaural benefits [58] showed that significant binaural benefits in bilateral CIs depend on comparable performance across the two implanted ears. In general, subjects having very different unilateral speech understanding performance, earn little regarding binaural advantages. However, only a small number of studies examined the possible origins of the performance mismatch between ears. Garadat, Litovsky, Yu, and Zeng (2010) studied the effects of “dead regions” for the performance difference between the two implanted ears and recommended that customized programming for bilateral CI processors based on knowledge about dead areas could enhance performance in difficult listening situations. Yoon, Liu, and Fu (2011) confirmed that the CI insertion-depth difference between ears might contribute to mismatched unilateral performance, thereby decreasing binaural benefits. Chen, Wong, Tahmina, Azimi, and Hu (2012)[59] examined the impact of mismatched spectral resolution between implanted ears on the recognition of Mandarin tones and sentences and concluded that matched spectral resolution maximized binaural summation benefits for Mandarin speech perception by simulated bilateral electric hearing. In the present study, we investigated another likely source of unilateral performance discrepancy—the dynamic range (DR) of the temporal envelope—in the context of Mandarin speech perception in noise by simulated bilateral electric hearing. Our reason for the present work was that, because of the two different maps used to transform acoustic signal units to electric current units, electric current signals have different DRs across the two implanted ears, which may contribute to the unilateral CI performance difference.

Researchers have extensively used acoustic simulations [3] involving NH subjects in CI research to study the effects of various factors on speech perception by unilateral and bilateral CIs. They have done so because it is beneficial to avoid the influence of patient-specific confounding factors (e.g., insertion depth, neural surviving pattern, etc.) that exist in clinical communities. In general, the vocoding processing used in the CI simulations mimics the way sound is processed in CI devices, with the exception that the final mapping stage is usually ignored. In consequence, without the logarithmic compression, the DRs of the temporal envelopes in acoustic simulations could be much larger than those generated by the clinical devices. In several studies[60], researchers examined how DR compression could affect the performances of the vocoding-processing in CI simulations and real CI listening. Fu and Shannon (1999) used peak or center clipping to reduce the DR of the input speech signals, and they measured the effects of reduced DR on phoneme recognition by CI users in quiet and in noise. Zeng and Galvin (1999) showed that in quiet, reduced DR had little impact on phoneme recognition. However, in a noisy area, phoneme recognition was significantly degraded by the reduction of the DR. Zeng et al. (2002) examined how to best convert the acoustic signals to optimize the DR of electric currents in CI listeners. Van Hoesel et al. (2005) assessed the effects of amplitude-mapping adjustment on speech intelligibility with unilateral and bilateral CI patients by reducing the stimulation level (also referred to as the “DR”) of standard monaural amplitude-mapping function when used bilaterally. They found a modest but statistically significant decrease in performance when stimulation level was lowered in quiet and in noise. In their CI simulation study, Loizou et al. (2000) used a six-channel sinusoid vocoder and compressed the amplitudes of the sine waves in a systematic fashion to simulate the effects of a reduced DR between 6 and 24 dB on speech understanding. The major drawback of their approach is that the minimum envelope amplitude both before and after compression across all spectral channels was always set to the quantization noise floor value (i.e., the value of 1 in the pulse code modulation wave file format often used in CI simulations). As a result, (a) in most cases, the DR of the input acoustic signals was clearly overestimated and (b) the arbitrarily set minimum value of the output envelope amplitude was not accurate and could lead to undesirable signal distortions. To address these issues, in our present study, we used an approach that dynamically set the minimum envelope amplitude both before and after compression across all spectral channels.

Among the three types of binaural benefits (i.e., head shadow, binaural summation, and binaural squelch), the head-shadow effect is mainly from the physical aspect that the head would attenuate some frequency components of the interfering noise signals and prevent these elements from reaching the other side of the head. In other words, the head would produce a shadow; for this reason, the head-shadow effect does not involve central auditory processing. Yoon, Shin, and Fu (2012) showed that binaural spectral mismatch did not appear to have an adverse impact on the head-shadow effect. Also, Garadat et al. (2010) showed that spectral holes in the higher frequency range would have a negative impact on the head-shadow effect because these holes undermine the spectral cues underlying the head-shadow effect. The hypothesized result of the envelope DR mismatch is that the mismatch might have an adverse impact on the head-shadow effect. For example, in the listening scenario where speech is presented from the front of the head and noise from the right, usually, the left ear will have a signal-to-noise ratio (SNR) advantage over the right ear. This benefit happened because the head shadow attenuates the low-frequency sound by 3–6 dB and attenuates the high-frequency sound by about 20 dB (Tyler et al., 2003). There are three scenarios for the envelope DR between the two ears: The left is better than the right, the left is similar to the right, or the left is worse than the right. In the scenario that the left ear takes advantage of the improved SNR, the envelope DR of the left ear cannot be extremely worse than right ear, because the reduced DR undermines the spectral contrast, which needs to meet a certain threshold for phoneme recognition (see, e.g., Loizou et al., 2000). In other words, if the DR of the left ear is noticeably worse than the right ear, the head-shadow effect would be severely reduced. The binaural summation and squelch effects rely on the redundancy and integration of the speech information from the two ears and involve central processing. The binaural summation effect occurs when both ears are presented with similar speech information (i.e., diotic listening), leading to the increment of the perceived loudness. The hypothesized impact of the envelope DR mismatch on the summation effect is that the mismatch would have a negative impact on the summation effect because the summation effect requires similar speech information from the two ears for the central processing. The binaural squelch effect refers to the advantage associated with bilateral listening when compared to monaural listening with the shadowed ear. By the results from Yoon et al. (2012), the hypothesized impact of the envelope DR mismatch on the squelch effect is that the mismatch would have a negative impact on the squelch effect because the squelch effect might require similar speech information from the two ears for central processing. To accurately measure the effects of the envelope DR mismatch on the head-shadow and squelch effects, researchers need to conduct spatial-hearing experiments such as those involving head-related transfer functions (HRTFs; e.g., Garadat et al., 2010; Yoon et al., 2012). In the present study, we focused on the binaural summation benefits.

We assessed the effects of envelope DR mismatch in the present study by using vocoded Mandarin speech to simulate unilateral and bilateral CIs. The binaural listening experiments were designed to answer the question of how envelope DR mismatch would affect the intelligibility of Mandarin speech in steady-state noise by bilateral electric hearing. From the results of previous studies in which the authors used English speech materials[58], we hypothesized that compared to the conditions with unmatched DR across the two implanted ears; those states with matched DR would yield more precise binaural summation benefits for Mandarin speech perception in noise. As to the binaural Mandarin tone identification and its relationship with sentence recognition, on the basis of our earlier studies on spectral resolution mismatch in Chen et al. (2012) as well as those results on tone recognition in quiet by Fu, Zeng, Shannon, and Soli (1998), we hypothesized that (a) DR mismatch may have less effect on tone identification than on sentence recognition and (b) factors other than tone identification might account more for the variance of Mandarin sentence recognition performance in noise.

4-2-2-Method

4-2-2-1-Subjects and Materials

Ten (five male and five female) NH native Mandarin-Chinese listeners participated in the experiments. The subjects' ages ranged from 18 to 28 years, and the majority of subjects were graduate students at The University of Hong Kong. Subjects were compensated for their participation in the study.

The Mandarin tone materials included six single-vowel syllables (/a/, /o/, /e/, /i/, /u/, and /ü/) in each of the four Mandarin tones (1: flat and high, 2: rising, 3: falling and rising, and 4: falling) produced by two adult native Mandarin- Chinese (one male and one female) speakers in a sound-treated booth, resulting in a total of 48 vowel tokens (i.e., two speakers × six vowels

× four tones) for Mandarin tone identification. The Mandarin sentence materials included sentence lists taken from the database of Mandarin Hearing in Noise Test (MHINT; see Wong, Soli, Liu, Han, & Huang, 2007, for more details). All sentences were produced by a native Mandarin-Chinese male speaker, with fundamental frequency (F0) ranging from 75 to 180 Hz. Both Mandarin sentences and vowels were recorded at a sampling rate of 16 kHz, and their waveforms were then adjusted to have the same root-mean-square (rms) values. The duration of the vowel tokens was normalized (Fu & Zeng, 2000). We used steady-state, speech-shaped noise (SSN) to corrupt the target speech materials at 5 and 0 dB SNR levels before the vocoding processing, and we chose the SNR levels to avoid the ceiling effect.

4-2-2-2-Amplitude Envelope Dynamic-Range Compression

The amplitude values of the temporal envelopes were extracted in the vocoding processing (see the Signal Processing section, below) and then were linearly compressed within a predetermined DR. The basic mechanism for amplitude envelope DR compression in the present study is similar to that developed in Loizou et al. (2000)—that is, experimenters use a linear transform to convert the range of the input amplitude envelope to a smaller range of the output amplitude envelope. However, the main differences lie in (a) how the minimum envelope amplitude of the input signal is determined and (b) how the linear transform is designed.

To determine the minimum and maximum amplitude values of the temporal envelopes in the spectral channels (denoted as X_{min} and X_{max}, respectively), we computed the histograms for the amplitudes of the temporal envelopes using the 240 sentences in the MHINT database.

X_{max} was selected to include 99% of all sorted amplitude counts, and X_{min} was chosen to include 1% of all sorted amplitude counts in ascending order. The amplitude envelope compression was conducted as follows: Let x and y denote the input and output amplitude envelopes, respectively. The compressed output amplitude envelope y was computed as:

$$y = \alpha \times (x - \bar{x}) + \bar{x} \qquad (4\text{-}1)$$

Where $\bar{x}$ is the mean of the input amplitude envelope x, and α is a constant (i.e., the compression factor) chosen in order for the output amplitude envelope to fall within a certain DR—that is,

$$Y_{max} = Y_{min} \times 10^{\frac{DR}{20}} \qquad (4\text{-}2)$$

Where Y_{max} and Y_{min} indicate the maximum and minimum output amplitude values, respectively. It is clear that the mean value of the output amplitude envelope equals the mean value of the input amplitude envelope (i.e.,$\bar{y} = \bar{x}$), regardless of the value of DR selected; in contrast, the approach in Loizou et al. (2000) yielded different $\bar{y}s$ for various DR values. Combining Equations (1) and (2), the compression factor a can be obtained as

$$\alpha = \frac{\left(10^{\frac{DR}{20}} - 1\right) \times \bar{x}}{X_{max} - \bar{x} - 10^{\frac{DR}{20}} \times (X_{min} - \bar{x})} \qquad (4\text{-}3)$$

To summarize, we performed the compression of the amplitude envelope by first computing the α value given the DR value using Equation (4-3), and then α was used in Equation (4-1) to map the input amplitude envelope x to the output amplitude envelope y. Figure 4.3 shows the relationship between the values of the averaged targeted DR and the values of the compression factor α, computed from the 240 sentences taken from the MHINT database. Equation (4-1) shows that when α equals 0, the compressed amplitude envelope becomes a direct current signal with a constant value of $\bar{x}$ (i.e.,$\bar{y} = \bar{x}$), and the DR is 0 dB. Alternatively, when α equals 1 in Equation (4-1), the output amplitude envelope keeps the original DR of the input amplitude envelope.

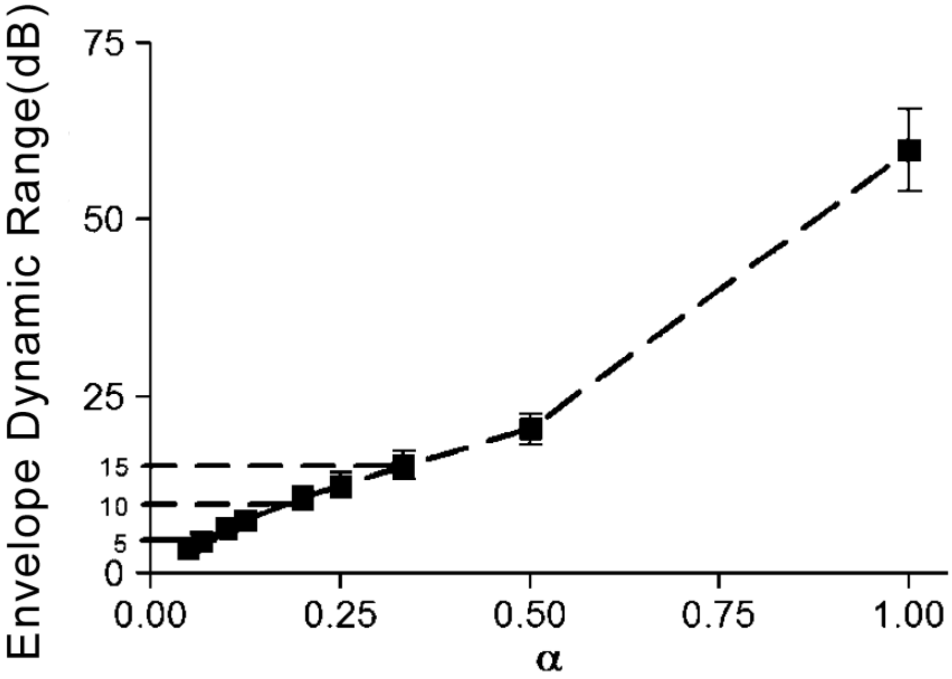


Figure 4.3- The plot of envelope dynamic range (DR) value as a function of compression factor α, which takes the values from 0.05 to 1 corresponding to different DR values.

Figure 4.4 shows the histograms for the DR values of the amplitude envelopes of the eight spectral channels computed from the 240 vocoded MHINT sentences for different values of α, which takes the value of 1/13, 1/5, and 1/3 in Figures 4.4A, 4.4B, 4.4C, and 4.4D, respectively. Figure 4.4D shows that the input amplitude envelopes across the eight spectral channels have a DR between around 40 to 80 dB, indicating a nearly normal distribution with a mean value around 60 dB. Similarly, the output amplitude envelopes have a mean DR value of around 5, 10, and 15 dB in Figures 4.4A, 4.4B, and 4.4C, respectively. Figure 4.4 also shows the plots (i.e., dotted lines) of fitting the histograms of the envelope DR values using normal-distribution functions with parameter values of mean (μ) and standard deviation (σ).

4-2-2-3-Signal Processing

An eight-channel noise vocoder was used in the vocoding processing in this study. We did this because Friesen, Shannon, Baskent, and Wang (2001) showed that, despite the relatively large number (16–22) of electrodes available, most CI users receive only a limited number (i.e., up to eight) of channels of spectral information. In implementing the eight channel noise vocoder, the corrupted Mandarin speech materials were first processed through a pre-emphasis high-pass filter (2000-Hz cutoff) with a 3-dB/octave roll off and then were band-pass filtered into eight frequency bands between 80 and 6000 Hz using sixth-order Butterworth filters. We used the equivalent rectangular bandwidth scale (Glasberg &Moore, 1990) to allocate the eight channels with the specific bandwidth (see Table 4.1). The temporal envelope of each spectral channel was extracted by full-wave rectification followed by low-pass filtering using a second-order Butterworth filter with a cutoff frequency of 160 Hz. The envelope of each frequency band was then compressed to the pre-set DR (e.g., α = 1/5 for compressing the envelope DR to 10 dB; see Figure 4.4B) and was used to modulate a white-noise signal, followed by band-limiting using the same Butterworth band-pass filters. The envelope-modulated noises of each band were finally summed up, and the level of the synthesized speech was adjusted to yield the same rms value as that of the original speech.

Channel	Low(Hz)	High(Hz)
1	**80**	**221**
2	**221**	**426**
3	**426**	**724**
4	**724**	**1158**
5	**1158**	**1790**
6	**1790**	**2710**
7	**2710**	**4550**
8	**4050**	**6000**

Table 4.1- Filter cutoff (–3 dB) frequencies used for the channel allocation.

For simulate DR mismatch across the two implanted ears, the right ear was presented with the eight-channel noise-vocoded stimuli that were compressed into 15-dB DR, whereas the left ear was given with the eight-channel noise-vocoded stimuli that were compressed into 5-, 10-, and 15-dB DR, respectively. Note that the 15-dB DR was used because researchers have found that the electrical DR was about 15 dB for CI listeners (Zeng et al., 2002). The three processing conditions described above are referred to as R15_L5, R15_L10, and R15_L15, respectively, where R and L indicate the right and left ears, respectively. As the control condition, we implemented the condition R15_L0, which presented only to the right ear the eight-channel noise vocoded stimuli that were compressed into 15-dB DR.

Note that, because the amplitude of the white noise was not constant, using white noise as the carrier signal might include an undesirable change to the DR of the envelope modulated noise in each frequency band. Although the envelope DR before modulating the white-noise signal (i.e., the DR of envelope y) was controlled following Equations (4-1) and (4-3).

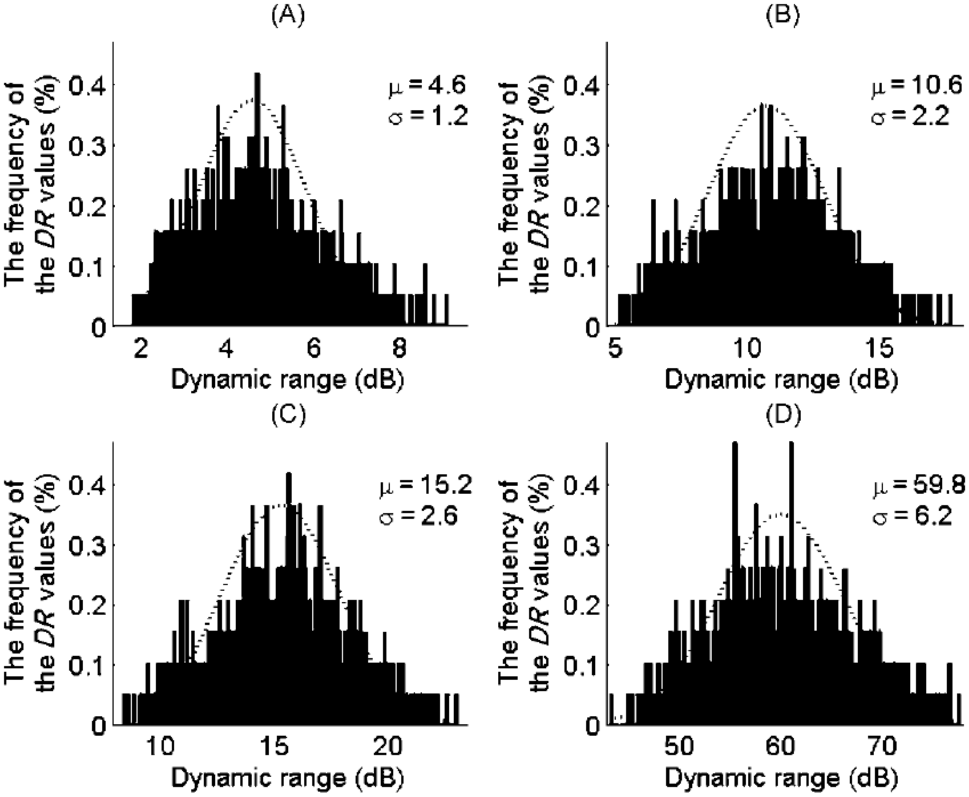


Figure 4.4- The histograms for the DR values of the eight temporal envelopes in the 240 vocoded Mandarin Hearing in Noise Test (MHINT; Wong et al., 2007) sentences as a function of compression factor a, which takes the values of 1/13 (DR = 5 dB; Panel A), 1/5 (DR = 10 dB; Panel B), 1/3 (DR = 15 dB; Panel C), and 1 (Panel D). The dotted line in each subplot fits the histogram of the envelope DR values using a normal distribution function with parameter values of mean (μ) and standard deviation (σ).

However, we did not expect the simulations in this study to accurately predict CI listeners' speech perception performance with various DRs; rather, we wanted to explore the effects of DR mismatch on binaural summation benefits of speech intelligibility in simulated bilateral electric hearing. In contrast, because of the constant amplitude of the sinusoidal carrier signal, a sinusoid vocoder may be selected in future studies. Further investigation is warranted, in which researchers compare the effects of envelope DRs on speech intelligibility between noise-vocoded and sinusoid-vocoded speech.

4-2-3-Procedure

The operation was made in a soundproof booth, and stimuli were played to listeners through a set of Sennheiser HD 250 Linear II circumaural headphones at a comfortable listening level. Before the test, each subject participated in a 20-min practice session and listened to the noise-vocoded speech with both matched and unmatched DRs in the temporal envelopes across the two ears, to become familiar with the vocoded stimuli and the testing procedure and conditions. Each subject participated in eight testing conditions (i.e., two SNR levels × four processing conditions) in both Mandarin sentence recognition and tone identification tests. The order of the eight testing conditions was randomized across subjects. For the Mandarin sentence test, a total of 20 Mandarin sentences were used per condition, and none of the sentences were repeated across various testing conditions. The subjects were instructed to write what they heard on the response sheets, and they were allowed to repeat the stimuli only once. The subjects were also instructed to guess when they were unsure about what they heard. For the Mandarin tone test, each condition consisted of two presentations of each vowel stimulus spoken by a male talker and a female talker, and each subject listened to a total of 96 randomized vowel stimuli (i.e., two repetitions × two speakers × six vowels × four tones) per condition. The Mandarin tone responses were collected with custom software through the use of a computer display of response alternatives and a mouse as a response key. The subjects were allowed to use a repeat key as many times as they wished to repeat the presentations of the test stimuli during the tone test. We calculated the percent correct score for each condition by dividing the number of words or tones correctly identified by the total number of words or vowel stimuli presented, respectively. Subjects were given a 5-min break every 30 min during the test.

4-2-4-Results

Figure 4.5A shows the mean percent correct scores of Mandarin sentence recognition for all conditions at different SNRs. We determined statistical significance by using the percent correct score as the dependent variable and using SNR levels and processing conditions as the two within-subjects factors. Two-way analysis of variance (ANOVA) with repeated measures indicated a significant effect of SNR level, F (1, 9) = 601.7, p < .0005; a significant effect of DR, F (3, 27) = 12.0, p < .0005; and a nonsignificant SNR Level × Dynamic Range interaction, F (3, 27) = 1.63, p = .206. Significant improvement in intelligibility was observed when listeners had access to the binaural vocoded stimuli with the same envelope DR (i.e., condition R15_L15) across the two ears, compared with the binaural stimulation with unmatched envelope DR (i.e., conditions R15_L10 and R15_L5). For instance, the improvement with the R15_L15 stimuli relative to the R15_L5 stimuli ranged from 15 percentage points at 5 dB SNR to 9 percentage points at 0 dB SNR. Furthermore, post hoc pairwise comparisons revealed a significant binaural benefit only for the condition R15_L15 (p < .05), compared with the condition R15_L0, in which the vocoded stimuli were presented unilaterally to the right ear. When the envelope DRs were not matched between the ears (i.e., conditions R15_L10 and R15_L5), there was no performance improvement with bilateral stimulation over condition R15_L0.

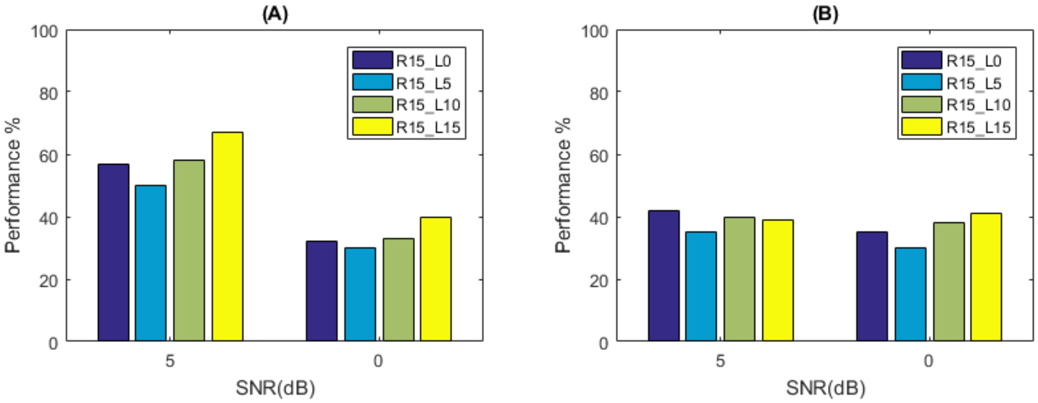


Figure 4.5- Mean recognition scores for (A) Mandarin sentence and (B) Mandarin tone at signal-to-noise ratio (SNR) levels of 5 and 0 dB by the simulated bilateral and unilateral electric hearing.

The results from the tone identification are shown in Figure 4.5B and are not quite similar to those observed from sentence recognition in Figure 4.5A. A two-way ANOVA with repeated measures indicated a nonsignificant effect of SNR level, F(1, 9) = 1.51, p = .255; a (weakly) significant effect of envelope DR, F(3, 27) = 6.04, p = .04; and a nonsignificant SNR Level × Envelope Dynamic Range interaction, F(3, 27) = 0.50, p = .685. Furthermore, post hoc pairwise comparisons revealed a significant binaural benefit only for the condition R15_L15 (p < .05) at 0 dB SNR, compared with the condition R15_L0, in which the vocoded stimuli were presented unilaterally to the right ear. Again, when the envelope DRs were not matched across ears (i.e., conditions R15_L10 and R15_L5), there were no significant binaural benefits over unilateral hearing (i.e., condition R15_L0) at both 5 and 0 dB SNR.

4-2-5-Discussion and Conclusions

A noticeable feature of the proposed compression scheme is that the probability distribution of the input-amplitude DR is more or less maintained in the output-amplitude DR. As can be seen in Figure 4.4, the nearly normal distribution shown in subplot D for the original DR is mostly kept intact in subplots B and C (for the compressed DRs of 10 and 15 dB, respectively), besides that the mean and standard deviation values are reduced accordingly due to the compression. Subplot A shows that the DR distribution for the compressed DR of 5 dB is skewed toward lower values and slightly deviates from the normal distribution. In contrast, the compression scheme in Loizou et al. (2000) used arbitrarily set minimum values of input and output amplitudes. As a result, as shown in Figure 3 of that article, the DR values of the compressed envelope amplitudes heavily tilted toward lower numbers, which might lead to artificially introduced distortion in the test stimuli, thereby underestimating the speech recognition performance. It is important to point out that both compression schemes used in the present study and Loizou et al. (2000) are based on linear transforms, which differ from the logarithmic compression typically applied in CIs. In order to match the loudness growth between electric stimulation and acoustic stimulation (Zeng & Shannon, 1992), CI processing typically uses logarithmic compression, as the loudness growth in acoustic stimulation of the auditory system features a power function of the physical intensity of the sound. Because our simulation study used acoustic stimulations involving the typical peripheral auditory system, the perceived loudness of the stimuli by NH listeners had a logarithmic relationship with the rms values of the stimuli. In other words, a linear transform would be more suited for our acoustic simulation study, although there would be differences between the loudness growth in our study and that in electric hearing. To our understanding, a logarithmic transformation would skew the DR distribution toward higher values, and as a consequence, the DR distribution would deviate away from the nearly normal distribution in the original envelopes. In the acute testing setup, a logarithmic transform might lead to more practice time, but with that exception; we would expect results to be similar to those using a linear transform. Certainly, in future studies, nonlinear compression schemes warrant further investigation.

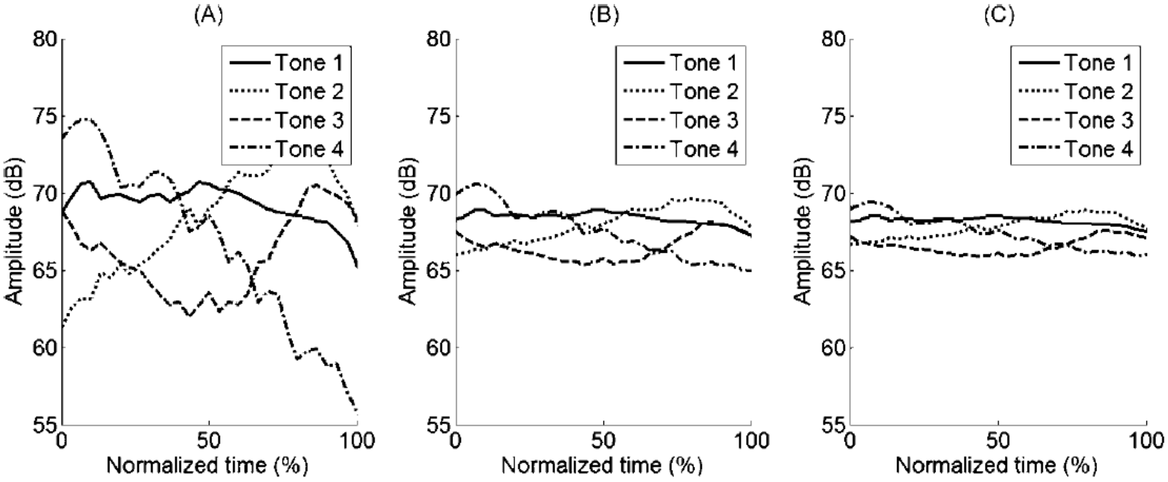


Figure 4.6- The amplitude contours of a recording of the Mandarin Chinese vowel /a / in four tones when the DR of the envelope is the original value (Panel A) and compressed to 15 dB (Panel B) and 10 dB (Panel C).

The findings of this study suggest that Mandarin tone identification may be robust to noise interference (i.e., at least for steady-state noise (SSN) maskers of 5- to 0-dB SNR levels) when the amplitude envelopes are compressed to a smaller DR (e.g., DR = 15 dB). Outcomes from other studies have reported the relatively small effect of spectral manipulations on Mandarin tone identification performance. For example, Fu et al. (1998) found that when the spectral resolution in the noise-vocoder simulation was altered from n = 1 to n = 4, the identification of Mandarin tones was almost consistent. The result is around 60% and 80% with envelope filter cutoff frequencies of 50 and 500 Hz, respectively. Zhou and Xu (2008) simulated the effect of mismatched spectral distribution of envelopes on Mandarin tone recognition using a noise vocoder; they found that spectral shift might not pose a severe problem for tone recognition when the carrier bands are shifted basally relative to the analysis bands by 1–7mm in the cochlea. To some extent, the present result is consistent with previous findings on the relative robustness of Mandarin tone recognition when compared to performances recognizing other speech components (e.g., phonemes and sentences). Chen et al. (2012) proposed that the smaller effects of noise on Mandarin tone recognition than on speech recognition might be partially due to (a) the favorable local SNRs of vowel segments where lexical tones were located and (b) a higher chance level for tone identification (i.e., 25%) than for consonant or vowel recognition. In addition to these factors, we believe that another factor for the relatively unchanged performance of Mandarin tone recognition across the three DR conditions may be the severe degradation of the temporal amplitude contour cues used for tone recognition, such that there was very little difference in the availability of this cue across the three reduced DRs tested. Many studies suggested that CI listeners used temporal amplitude contour and periodicity cues for tone identification (Fu & Zeng, 2000; Fu et al., 1998; Yuan et al., 2009). The envelope filter cutoff frequency was set to 160Hz in the present study because researchers have used this setting in other acoustic simulation studies to investigate the performance of Mandarin tone identification (see, e.g.,Chen et al., 2012; Zhou&Xu, 2008).Nevertheless, using a relatively low cutoff frequency to extract temporal amplitude envelope means that listeners relied primarily on the limited temporal cues (e.g., tonal amplitude envelope/contour; Fu et al.,

1998) for tone identification. Further analysis unveiled that when the amplitude envelope was compressed to a smaller DR(e.g., 15dB), the distinction among the amplitude contours of four Mandarin tones was reduced, which made it difficult for listeners to obtain sufficient amplitude contour information for reliable tone identification. Figure 4.6 illustrates the amplitude contours of a recording of the Mandarin Chinese vowel /a/ in quiet and at three DRs: the original range, 15 dB, and 10 dB. It is seen in Figure 4.6A that the amplitude contours of four lexical tones are notably different, with each representing the F0 contours of four Mandarin tones accordingly. However, when the DR of amplitude envelope was compressed to 15 dB and 10 dB in Figures 4.6B and 4.6C, respectively, it is clear that the difference among the four amplitude contours is not as salient as that observed in Figure 4.6A (i.e., with the original DR). This diminished distinction among amplitude contours at a smaller DR might also partially account for the lower and relatively consistent recognition scores of Mandarin tones in Figure 4.5B. Note that further study is needed, in which researchers investigate how the periodicity cues and DR would interact with one another to influence the performance of tone recognition.

Moreover, the decreased DR might also damage the formant representation in vowels and other temporal envelope cues beneficial for phoneme recognition (Fu & Zeng, 2000; Fu et al., 1998; Loizou et al., 2000). Therefore, comparing the identification scores in Figure 4.5 of the current article with those in Figure 1 of Chen et al. (2012), it is not surprising to see that, for the same eight-channel noise-vocoding processing and the same testing materials, the proposed compression scheme reduced the unilateral sentence and tone recognition performance by about 27%–36% and 14%–24%, respectively, when the amplitude envelope DR was compressed from (the original) 60 dB to 15 dB.

In conclusion, in the present study, we valued the effects of envelope DR mismatch on the binaural summation benefits for Mandarin speech intelligibility in a simulated bilateral electric hearing with an improved DR compression scheme. It is important to note that because this is an NH based acoustic simulation study that involved “ideal” assumptions unlikely to hold in actual patients with bilateral CIs, caution should be used in interpreting the present findings, and a follow-up study with bilateral CI users is needed to verify these results. The testing conveyed was acute and was not performed after long-term experience with the mismatched DRs, and we do not yet understand how CI users adapt their DR mapping after long-term use. Also, there are many confounding factors in electric hearing such as the depth of electrode insertion, neuron survival patterns, and current spread. With this in mind, in this article, we examine the implications of the present study. Consistent with previous findings on the binaural summation benefits for English materials[58], as well as those on the effects of spectral resolution mismatch for Mandarin speech materials (Chen et al., 2012[59]), the present results suggest that the sentence perception performance of binaural CI listening in noise is affected by the difference of envelope DRs between the two ears. Also, binaural summation benefits are maximized with matched DRs in the two implanted ears. In addition, the present study shows that in the context of compressed envelope amplitudes, tone identification in steady-state SSN is steadier than sentence recognition, suggesting that tone identification performance in noise might not predict sentence recognition performance in unilateral and bilateral CIs.

Contribution of low-frequency harmonics to tone identification in quiet and six-talker babble background

4-3-1-INTRODUCTION

As a tonal language, Mandarin Chinese has four normal tones to carry the lexical meaning of each word (Chao, 1968; Howie, 1976). Acoustically, the four lexical tones are mostly defined by the fundamental frequency (F0) height and contour shapes: tone 1, 2, 3, and four are associated with a high-level, mid-rising, falling-then-rising, and high falling F0 contour, respectively. For instance, the syllable [ma] means "mother" with tone 1, "hemp" with tone 2, "horse" with tone 3, and "scold" with tone 4 (Chao, 1968).

Native Mandarin Chinese listeners with normal hearing are able to identify Mandarin tones with nearly perfect scores (Kong and Zeng, 2006; Lee et al., 2010), in quiet listening conditions. However, when Mandarin tones are presented in noisy conditions (e.g., speech-shaped noise or white noise), tone identification is expectedly reduced as signal-to-noise ratio (SNR) decreases (Kong and Zeng, 2006; Dees et al., 2007; Lee et al., 2010). For example, native Mandarin listeners identify tones above 90% for the SNR of 5 dB, and tone identification accuracy drops to 85% for the SNR of 15 dB (Lee et al., 2010). Although many studies have focused on tone recognition in quiet conditions for listeners with normal hearing and users of cochlear implants (see, e.g., Fu and Zeng, 2000; Luo and Fu, 2004, 2006; Liu et al., 2012) and some other studies have investigated tone identification in speech-shaped noise (Kong and Zeng, 2006; Lee et al., 2010). However, normal-hearing listeners' tone identification in multi-talker babble is limited. Dees et al. (2007) found that tone identification of Mandarin Chinese listeners in six-talker babble was 61% and 93% for SNR of 25 and 15 dB, respectively. Compared to speech-shaped noise, low-N multi-talker babble (e.g., two- and four-talker babble) contains greater temporal modulations (Rosen et al., 2013) and potentially more informational masking (Brungart et al., 2001). Thus, one goal of this study was to examine tone identification in multi-talker babble further.

Various acoustic cues contribute to Mandarin tone identification such as F0 height or contour, amplitude contours, and syllable duration (Ho, 1976; Whalen and Xu, 1992; Fu et al., 1998; Fu and Zeng, 2000; Gao, 2002; Liu et al., 2012). Among these cues, F0 height and contour play a primary role for tone identity for normal-hearing listeners (Fu and Zeng, 2000; Liu et al., 2012). That is, even when amplitude contours of speech signals were not matched with F0 contours, tone identification was determined by the F0 contour (Liu et al., 2012). While relying on F0 contours to identify Mandarin tones, listeners needed to resolve harmonics, especially low- and middle-frequency harmonics (Stagray et al., 1992; Luo and Fu, 2006; Oxenham, 2008). Using a methodology based on filtering, Stagray et al. (1992) measured thresholds for detection and identification of Mandarin tones for stimuli with different types of harmonic structures such as F0 only, resolved harmonics between 315 and 1000 Hz, unresolved harmonics above 2000 Hz, and all harmonics. Although the F0-only stimulus showed the highest detection threshold, for identification, its threshold was the lowest, when expressed as sensation level above the detection limit. Also, resolved harmonics had relatively low detection thresholds and low identification thresholds, when shown in sensation level. Stagray et al. (1992) concluded that F0 and low-frequency (below 1000 Hz) resolved harmonics were critical cues for tone identity, although unresolved high-frequency harmonics also cued Mandarin tones.

Luo and Fu (2006) examined Mandarin speech recognition with an acoustic simulation of bilaterally combined electric and acoustic hearing. They reported that low-frequency acoustic information below 500 Hz made significant contributions to tone identity, while low-frequency acoustic information above 500 Hz was necessary for phonemic recognition. In addition, compared to low-frequency harmonics, from a signal processing perspective, high frequency harmonics are much harder to regenerate (for the scenario where harmonics are mostly missing) or enhance (for the situation where harmonics are mostly distorted by noise and high-frequency harmonics usually have much lower SNRs than low-frequency harmonics in multi-talker babble and speech-shaped noise). For these reasons, almost all existing algorithms of harmonic regeneration and harmonic enhancement focus on low-frequency harmonics (e.g., Zavarehei et al., 2007; Jin et al., 2010). Therefore, another goal of the present study was to investigate the role of low-frequency harmonics on tone recognition in babble.

The studies above investigated tone identification with different types of stimuli (e.g., individual, partial, and all harmonics) in a variety of listening condition (e.g., quiet and speech-shaped noise), indicating that resolved harmonics (e.g., low harmonics) primarily carry tone information. However, research questions on tone identification still remain, such as the effects of vowel category and F0. In general, the F0s of Mandarin speech produced by native Mandarin Chinese male speakers range from about 90 to 150 Hz, while the F0s by native Mandarin Chinese female speakers range from about 180 to 310 Hz (Xu, 1997). Given the marked difference in F0 between male and female speakers, the low harmonics (e.g., the first three harmonics) may differ in frequency by several hundred Hz, possibly resulting in different patterns of tone identification in noise across this broad range. Also, formant structures of vowels may interact with vowel harmonics, i.e., when vowel harmonics coincide with formant frequency, the harmonic amplitude becomes greater; therefore, it is reasonable to predict that vowel category with different formant frequencies may affect tone identification.

In summary, the aims of the present study were twofold. The first goal was to investigate the contribution of individual and combined low harmonics (i.e., below and near 500 Hz) of Mandarin speech to tone identification in quiet and multi-talker babble for normal-hearing listeners. The second goal of this study was to examine the interaction effects of F0 and vowel category with harmonic structures on tone identification in noise.

4-3-2 Method

4-3-2-1- Listeners

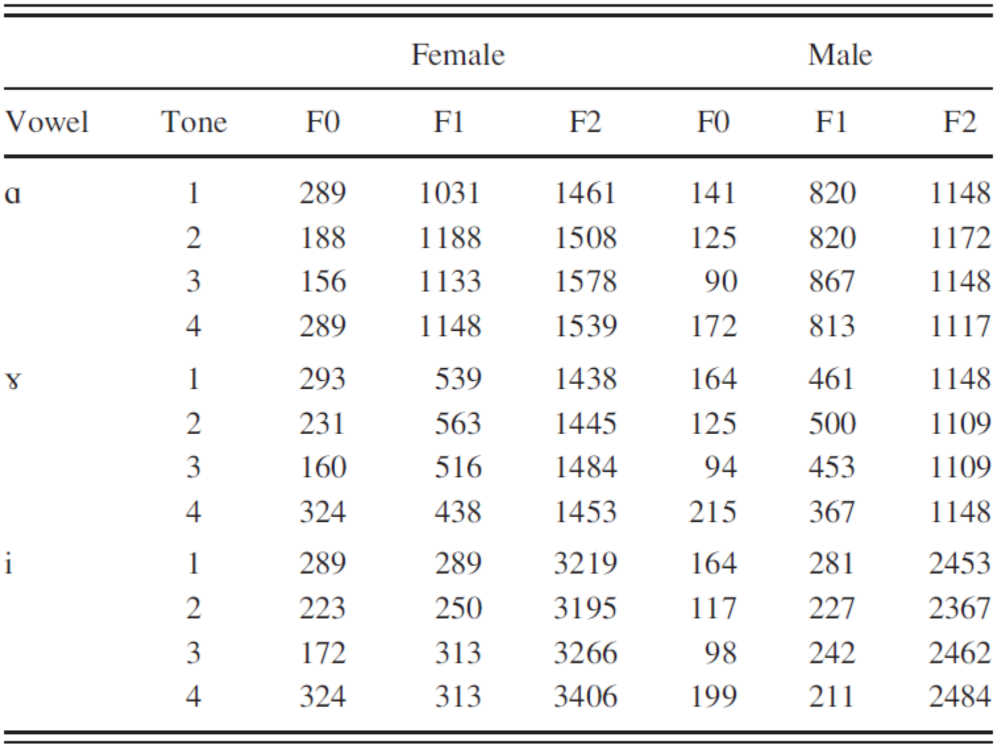
Ten young Mandarin Chinese-native listeners from 20 to 25 years old participated in this study. They had normal hearing sensitivity with pure-tone thresholds<15 dB hearing level (HL) (ANSI, 2010) at octave intervals between 250 and 8000 Hz in both ears. All participants, undergraduate or graduate students in Beijing, China, were from northern China and spoke standard Mandarin. None of the listeners had a formal musical education of more than five years, and no listeners had received any musical training in the past three years.

4-3-2-2-Stimuli

4-3-2-2-1- Speech signal

Three Mandarin vowels, /ɑ, ɤ, i/ with four tones were recorded from a young female and a young male Mandarin native speaker as isolated tokens. The tone duration ranged from 238 to 466 ms across the three vowels and four tones. Because the primary goal of this study was to examine tone identification with various harmonic structures, the duration of the four tones was equalized at 210 ms by removing the onset and offset segments with the steady-state vowel nucleus remaining to eliminate the duration effect on tone identification. Then the vowel signals were normalized to the same root-mean-square (RMS) value. The F0, F1, and F2 values are shown in Table 4.2. Stimuli with individual and combined low-frequency harmonics were generated as follows. First, vowel signals were segmented into 366-sample frames (around 30 ms) with 50% overlap, and a pitch detection algorithm based on an autocorrelation function was used to estimate F0 values in each voiced frame (Dubnowski et al., 1976). The 30-ms speech segment was zero-padded to generate a frame of 2048 samples before the application of the fast Fourier transform (FFT), and this was done to increase the frequency-bin resolution for the FFT.

Second, harmonic extraction and guidance of the vowel signals were made in the frequency domain using a 2048-point FFT implemented to the 2048-sample frame windowed by a Hamming window. The magnitude spectrum peaks around the integer multiple of F0 were identified as the harmonics; for example, the magnitude spectrum peak around 3xF0 was identified as the third harmonic. The F0 values, as well as the magnitude values of the surrounding harmonics in each frame, were manually verified to ensure their accuracy, and this was done to make certain of their preservation in the re-synthesis. The harmonic-synthesis conditions using the first three individual harmonics were labeled as H1, H2, and H3, respectively. The condition using the lowest three harmonics (e.g., H1+H1+H3) was labeled as "Low," the corresponding condition using all harmonics except the lowest three (i.e., the "higher" harmonics) was labeled as "High," and the condition using all harmonics were labeled as "All." After the analysis and synthesis had been done in the frequency domain, the inverse FFT was applied to generate a frame of 2048 samples in the time domain, which was truncated to 366 samples to maintain the 30-ms window.



		Female			Male		
Vowel	Tone	F0	F1	F2	F0	F1	F2
ɑ	1	289	1031	1461	141	820	1148
	2	188	1188	1508	125	820	1172
	3	156	1133	1578	90	867	1148
	4	289	1148	1539	172	813	1117
ɤ	1	293	539	1438	164	461	1148
	2	231	563	1445	125	500	1109
	3	160	516	1484	94	453	1109
	4	324	438	1453	215	367	1148
i	1	289	289	3219	164	281	2453
	2	223	250	3195	117	227	2367
	3	172	313	3266	98	242	2462
	4	324	313	3406	199	211	2484

Table 4.2: Fundamental frequency (F0), F1, and F2 frequency (Hz) of the three vowels with four tones.

Third, the synthesis was performed by applying the overlap and add approach for the corresponding harmonic condition. The level of stimuli with All harmonics was set at 70 dB sound pressure level (SPL) for both quiet and noisy. Also the level of noise was set at 65 dB SPL and 70 dB SPL [i.e., 5 and 0 dB SNR for the noisy conditions], while the level of stimuli with individual, Low, and High harmonics ranged from 56.7 to 66.7 dB SPL, depending on the RMS level of each harmonic relative to the RMS level of the speech with all harmonics.

4-3-2-2-2- Noise

A random segment of 400 ms duration from a 10-s sample of six-talker Mandarin Chinese babble (Van Engen, 2010), was selected to serve as the masker. For each trial, the 210-ms signal was presented at the temporal center of the masker.

4-3-2-3-Stimulus presentation

Signal and noise stimuli, sampled at 12,207 Hz, were presented to the right ears of listeners via SONY MDR-7506 headphones. Listeners were seated in the Psychological Behavioral Test rooms of the National Key Laboratory of Cognitive Science and Learning at Beijing Normal University. The stimulus presentation was manipulated through a Tucker-Davis Technologies (TDT) mobile processor (RM1). The sound-pressure levels of signal and noise were calibrated in an AEC201-A IEC 60318-1 ear simulator by a Larson-Davis sound-level meter (model 2800) with linear weighting. The procedure was manipulated by the software Sykofizx v2.0.

4-3-2-4-Experimental condition

There were five experimental factors and a total of 432 experimental conditions in the present study:

- Two F0s (one for male and one for female)
- Six harmonic configurations (H1, H2, H3, Low, High, and All)
- Three vowel categories (/ɑ, ɣ, i/)
- Three listening conditions (quiet and noise with 5 dB and 0 dB SNRs)
- Four tones (tone 1-4).

4-3-2-5-Procedure

After each stimulus performance, the listener’s task was to identify the tone of the signal (i.e., tone 1, 2, 3, or 4). Listeners made responses in front of an LCD monitor on which a subject interface displayed four-choice close-set buttons (tone 1, 2, 3, and 4). Each signal was presented 15 times for each listener, resulting in a total of 6480 trials (432 stimulus conditions × 15 repetitions). The 432 experimental conditions were divided into 36 blocks (3 listening conditions × 6 harmonic configurations × 2 F0 values), each of which contained the four tones and three vowels. Thus, for a given block, three factors (F0 value, harmonic configuration, and listening condition) were fixed and the other two factors (vowel category and tone) were mixed and presented randomly trial by trial. The performance of the 36 blocks was randomized. Before the test session, a 5-min training session was provided to listeners to familiarize them with the experimental procedure. Natural Mandarin tones, recorded from a native Mandarin Chinese speaker who was not included with the two speakers above, were used as stimuli in the training session. Feedback was provided to indicate the correct response on each trial for the training session, but not for the test sessions.

4-3-3-RESULTS

4-3-3-1-Overall results

Figure 4.7 shows average tone identification scores over the three vowels of the female (top) and male (bottom) talker as a function of listening conditions for the six harmonic conditions. Commonly, in quiet, tones with All, Low, High, and individual harmonics were identified with high accuracy (95%). When presented in the six-talker babble, tone identification scores were still quite high (90%) for signals with All harmonics, but dropped significantly for Low, High, and individual harmonics; for instance, tone identification was less than 40% at an SNR of 0 dB for all three individual-harmonic signals. In other words, the negative effect of noise was clearly observed for Low, High, and individual harmonics, but not for signals with All harmonics.

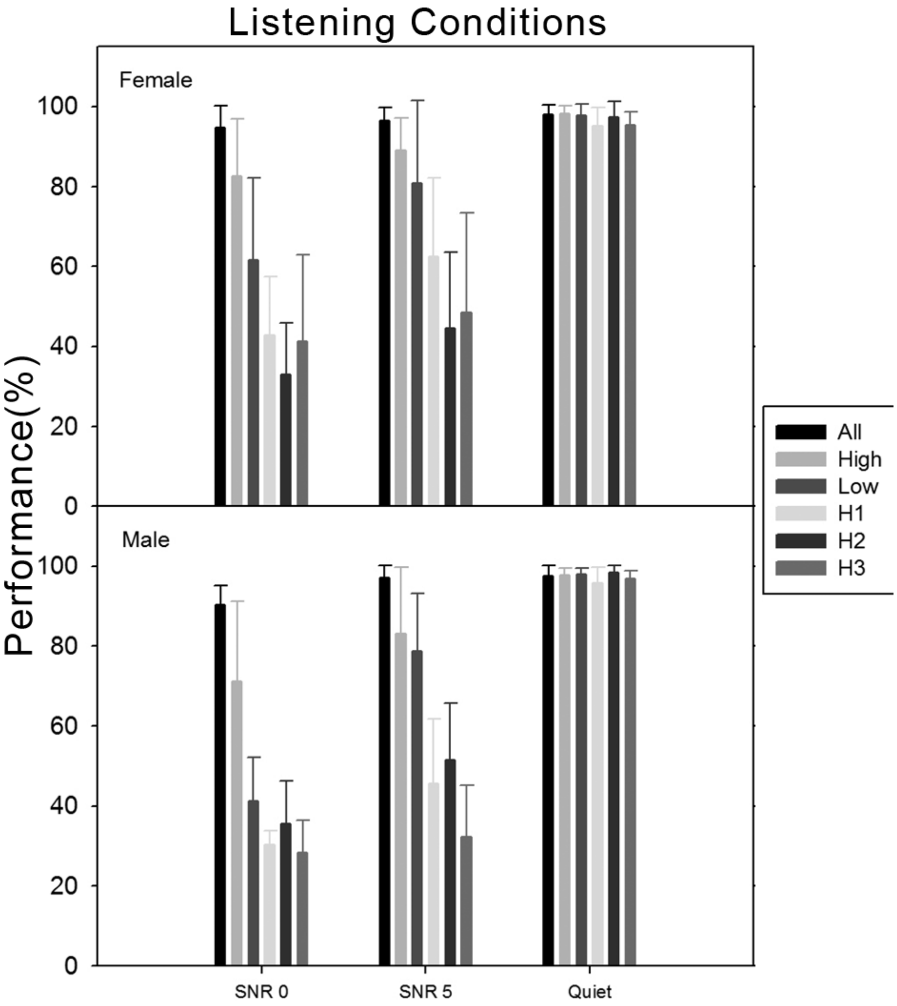


Figure 4.7: Tone identification scores and standard errors averaged over the three vowels as a function of listening conditions (Quiet, and SNRs of 5 and 0 dB) for the female (top) and male (bottom) speaker.

Tone identification as a function of the four tone categories for the high F0 (female, left panels) and low F0 (male, right panels) is shown in Figures 4.8 (quiet), 4.9 (SNR at 5 dB), and 4.10 (SNR at 0 dB), respectively. These figures suggest that the effect of stimulus harmonics on tone identification differed across the three listening conditions. Therefore, data analysis was conducted separately for the quiet and noisy conditions below. To overcome the compressive scores near the ceiling performance of tone identification, identification scores in percent for individual listeners and experimental conditions were converted to rationalized arcsine units (RAU; Studebaker, 1985). The RAUs were then used in the data analysis throughout the manuscript.

4-3-3-1-1-Tone identification in quiet

For the quiet condition, a four-factor (vowel category × tone category × F0 × stimulus harmonic), repeated measures analysis of variance (ANOVA) was conducted. To minimize the type I error, the significant level was set at 0.001. Results revealed that there was a significant effect of tone category ($F_{3,27}$=18.038, p<0.001), but not by stimulus harmonic ($F_{5,45}$=3.778, p=0.006), F0 ($F_{1,9}$=0.983, p=0.347), and vowel category ($F_{2,18}$=4.360, p=0.029). Between the two-, three- and four-factor interactions, only the two-factor interaction effects of stimulus harmonic and tone category ($F_{15,135}$=2.747, p<0.001), and tone category and vowel category ($F_{6,54}$=6.583, p<0.001), and the three-factor interaction of F0, vowel category, and tone category ($F_{6,54}$=9.373, p<0.001) were significant.

Due to the significant interaction effect of tone category and stimulus harmonic, a three-factor (F0 × stimulus harmonics × vowel category) analysis of variance (ANOVA) was run to examine the simple main effect of stimulus harmonic under each tone category. Identification of tone 1 was not significantly affected by any of the three factors and their multi-factor interactions (all p>0.001). For tone 2, none of the three factors and the two- and three-factor interaction had significant effects on tone identification (all p>0.001), except vowel category ($F_{2,18}$=15.134, p<0.001). Furthermore, Tukey post hoc test suggested that identification scores of tone 2 for vowel /ɣ/ were significantly higher than those for vowel /ɑ/ (p<0.001). For tone 3, tone identification was significantly affected by stimulus harmonic ($F_{5,45}$=4.474, p<0.001) and the interaction of vowel category and F0 ($F_{2,18}$=15.803, p<0.001), but not by the other two factors and other multi-factor interactions (all p>0.001). For tone 4, none of the three factors and multi-factor interactions was significant (all p>0.001).

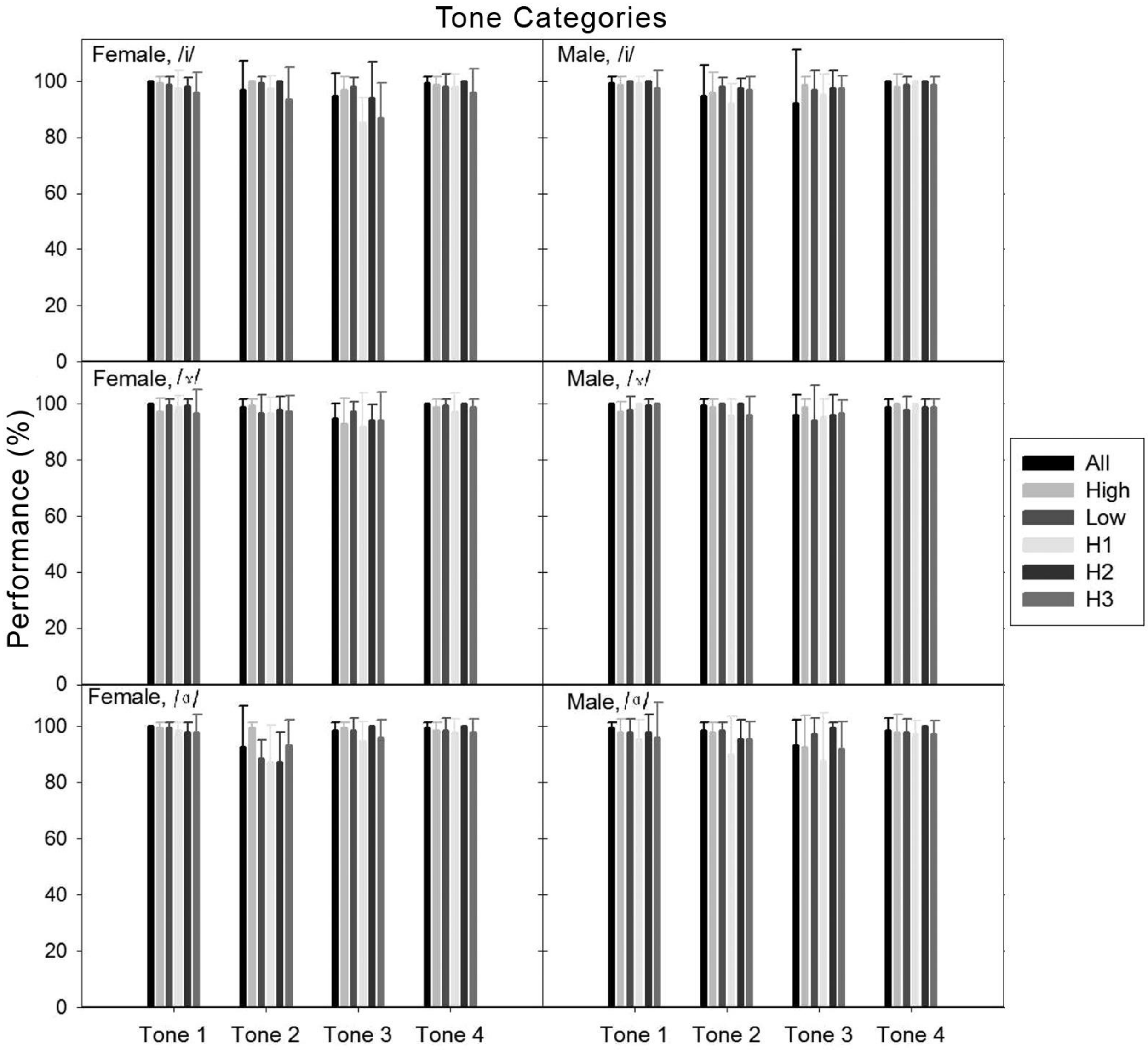


Figure 4.8: Tone identification scores and standard errors as a function of the four tone categories in quiet for the female (left panels) and male (right panels) speakers with the three vowels /i/ (top panels), /ɣ/ (middle panels), and /ɑ/ (bottom panels), for six harmonic conditions (All: all harmonics included; High: high harmonics; Low: low harmonics; H1: the first harmonic; H2: the second harmonic; H3: the third harmonic).

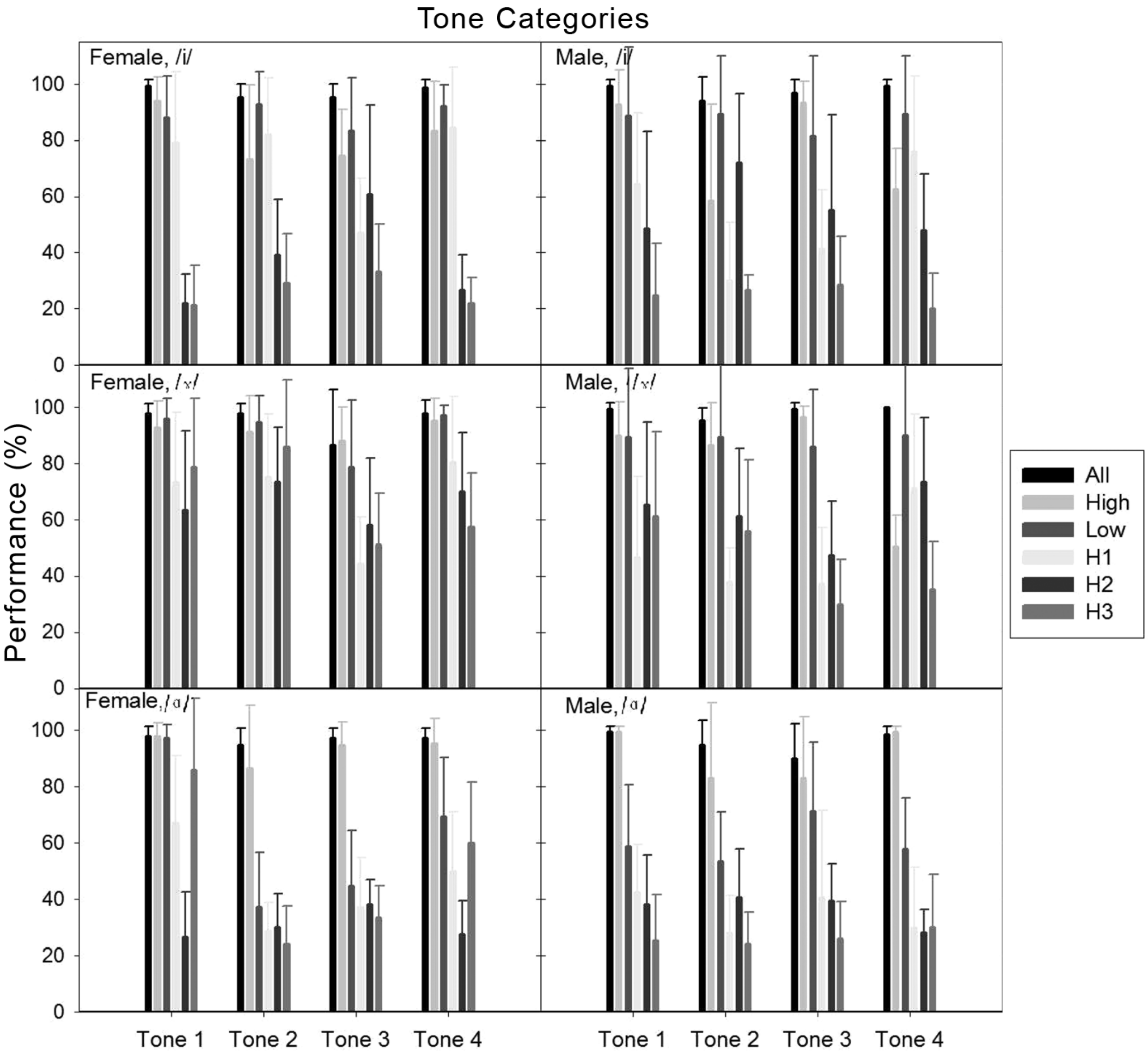


Figure 4.9: Tone identification scores and standard errors as a function of the four tone categories in the noisy listening condition of SNR at 5 dB. The layouts and symbols are the same with Figure 4.8.

4-3-3-1-2-Tone identification in multi-talker babble

As the speech stimuli were presented in babble, a five factor (vowel category × tone category × F0 × stimulus harmonic × SNR) suggested that tone identification was significantly affected by each of the five factors: stimulus harmonic ($F_{5,45}$=181.897, p<0.001), tone category ($F_{3,27}$=8.340, p<0.001), F0 ($F_{1,9}$=27.391, p<0.001), vowel category ($F_{2,18}$=37.996, p<0.001), and SNR ($F_{1,9}$=142.155, p<0.001). All the two-factor interactions were significant (all p<0.001) except the interactions of SNR and tone category, and SNR and F0 (both p>0.001). Tukey post hoc tests indicated that stimuli with All harmonics had the highest tone identification scores, followed by stimuli with High and Low harmonics, then the H1, and the H2 and H3 had the lowest scores (all p<0.001). Also, vowel /ɣ/ had the greatest tone identification scores, followed by vowel /i/ and then /ɑ/ (all p<0.001). The high F0 had significantly better tone identification scores than the low F0 (p<0.001). Tone identification comparisons across the four tone categories indicated that tone 1 had significantly higher identification scores than tone 3 (p<0.001).

Because of the significant interaction effect of stimulus harmonic and SNR, a four-factor (vowel category × tone category × F0 × stimulus harmonic) repeated-measures ANOVA was run for the SNR of 5 and 0 dB, respectively, to examine the simple main effect of stimulus harmonic. Results of the two ANOVAs and the post hoc tests are shown in Table 4.3. Most of the single-factor effects and multiple-factor interactions were significant. In particular, similar to the data analysis in a quiet, there was a significant interaction effect of stimulus harmonic and tone category for each SNR condition. Thus, for each SNR condition, three factors (stimulus harmonic × F0 × vowel category) repeated measures ANOVAs were conducted for each of the four tone categories, respectively, to reveal the simple main effect of stimulus harmonic. Results of these analyses are illustrated in Table 4.4 for the SNR of 5 dB and in Table 4.5 for the SNR of 0 dB. Across all the tones in noisy conditions, the effect of stimulus harmonic was significant. Tukey post hoc tests indicated that tone identification was greater for the multiple harmonic signals than for the individual harmonic signals.

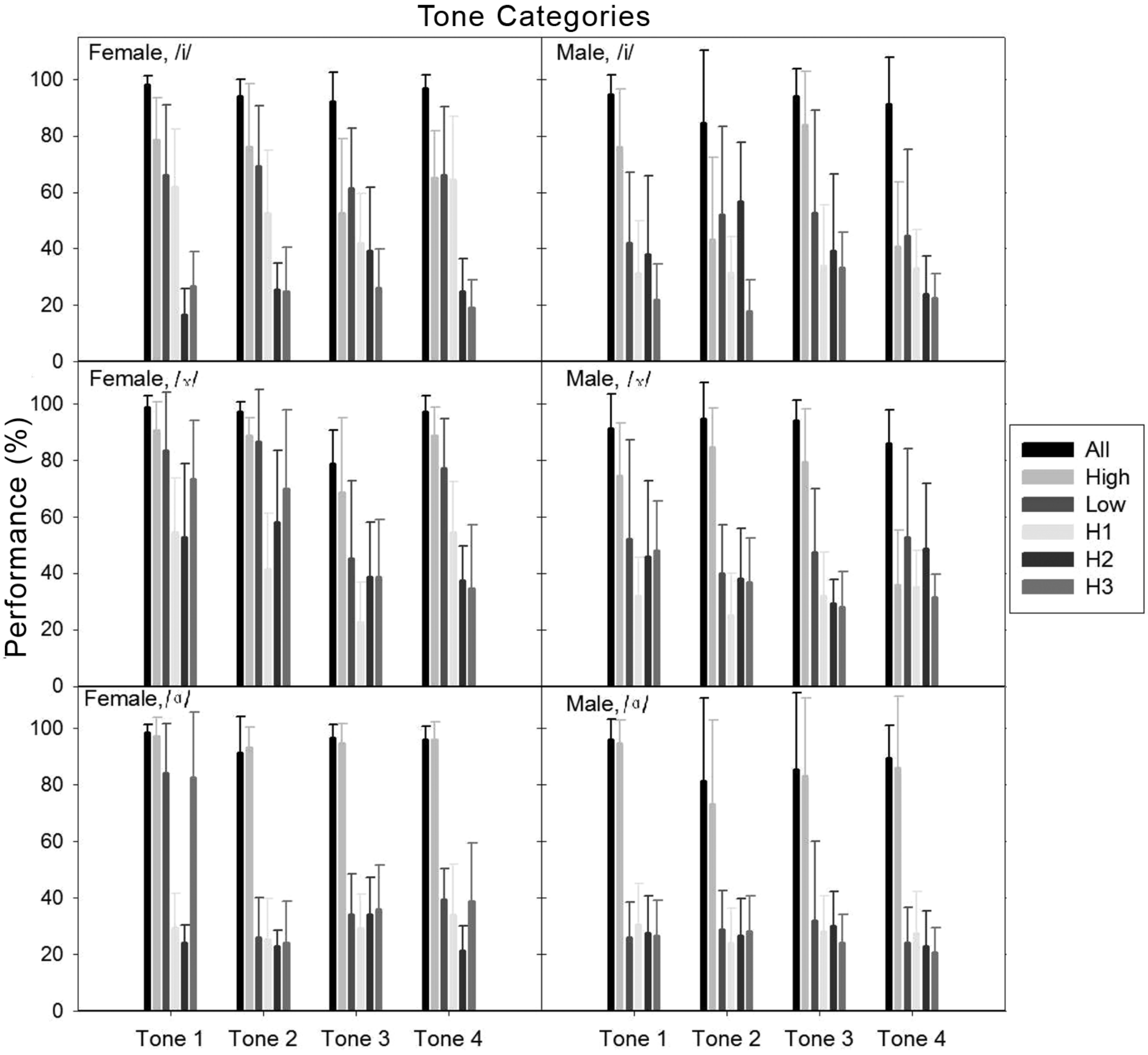


Figure 4.10: Tone identification scores and standard errors as a function of the four tone categories in the noisy listening condition of SNR at 0 dB. The layouts and symbols are the same with Figure 4.8

Furthermore, the all harmonics had the greatest tone identification scores, followed by the High harmonics and then the Low harmonics, while there were no significant differences among the three individual harmonics. For more than half of the tone conditions in noise, tone identification was significantly affected by vowel category. Also, vowel /ɤ/ had higher scores than vowels /ɑ/ and /i/.

4-3-3-2- Error pattern analysis: confusion matrix of tone identification

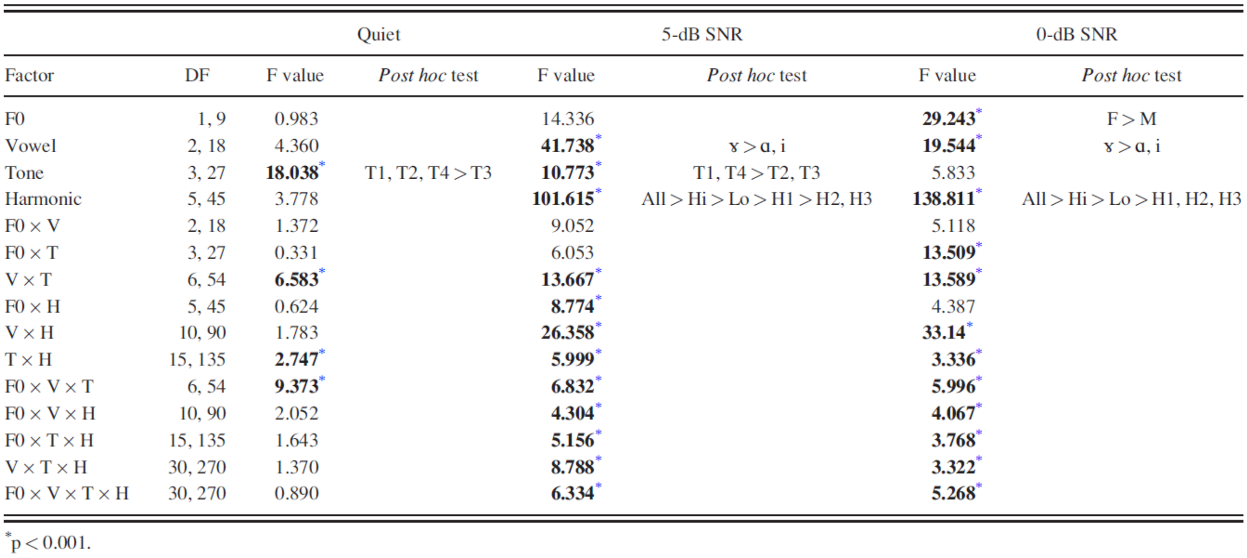
Tone confusion matrices are illustrated in Tables 4.6, 4.7, and 4.8 for quiet, and the SNR of 5 and 0 dB, respectively. In general, for specific harmonics, each of the four tones was confused equally with the other three tones. These confusion matrixes were somewhat different from other studies (Lee et al., 2010). For instance, in long-term speech-shaped noise, the identification of tone 2 was mainly confused by tone 1 and 3, while the determination of tone 3 was primarily confused by tone 2 (Lee et al., 2010). The differences in confusion matrix between these studies could be due to the different types of maskers. That is, the six-talker babble in the present study may carry plenty of tonal information that distracted listeners' tone identification.

4-3-4- DISCUSSION

4-3-4-1- Correlations between tone identification and signal-to-noise ratios of speech harmonics

The goal of this study was to investigate the contribution of low harmonics to Mandarin tone identification in quiet and babble listening conditions by normal-hearing listeners. As the statistical results above showed, tone identification patterns were quite complicated, depending on vowel category, stimulus harmonics, F0, and listening condition. Correlations between tone identification and SNRs of speech harmonics were examined to interpret the tone identification with the effects of the experimental factors. The SNRs of speech harmonics were computed as follows:

- First, a long-term average spectrum of the six-talker babble (see Fig. 4.11) was collected by the linear-predictive coding (LPC), using 16 LPC coefficients with a modified version of Colea MATLAB VR code (Loizou, 2000);
- Second, for individual harmonics such as H1, H2, and H3, noise level was computed within one auditory filter band (equivalent rectangular bandwidth-ERB; Glasberg and Moore, 1990), while for Low and High harmonics, noise level was computed within the corresponding frequency range (e.g., noise at frequencies below 3.5 × F0 for Low harmonics and noise at frequencies above 3.5 × F0 for High harmonics).
- Third, harmonic SNRs were calculated by subtracting noise level from signal level (both expressed in dB) for the conditions with Low and High harmonics and with individual harmonics.



		Quiet		5-dB SNR		0-dB SNR	
Factor	DF	F value	*Post hoc* test	F value	*Post hoc* test	F value	*Post hoc* test
F0	1, 9	0.983		14.336		**29.243***	F > M
Vowel	2, 18	4.360		**41.738***	ɤ > ɑ, i	**19.544***	ɤ > ɑ, i
Tone	3, 27	**18.038***	T1, T2, T4 > T3	**10.773***	T1, T4 > T2, T3	5.833	
Harmonic	5, 45	3.778		**101.615***	All > Hi > Lo > H1 > H2, H3	**138.811***	All > Hi > Lo > H1, H2, H3
F0 × V	2, 18	1.372		9.052		5.118	
F0 × T	3, 27	0.331		6.053		**13.509***	
V × T	6, 54	**6.583***		**13.667***		**13.589***	
F0 × H	5, 45	0.624		**8.774***		4.387	
V × H	10, 90	1.783		**26.358***		**33.14***	
T × H	15, 135	**2.747***		**5.999***		**3.336***	
F0 × V × T	6, 54	**9.373***		**6.832***		**5.996***	
F0 × V × H	10, 90	2.052		**4.304***		**4.067***	
F0 × T × H	15, 135	1.643		**5.156***		**3.768***	
V × T × H	30, 270	1.370		**8.788***		**3.322***	
F0 × V × T × H	30, 270	0.890		**6.334***		**5.268***	

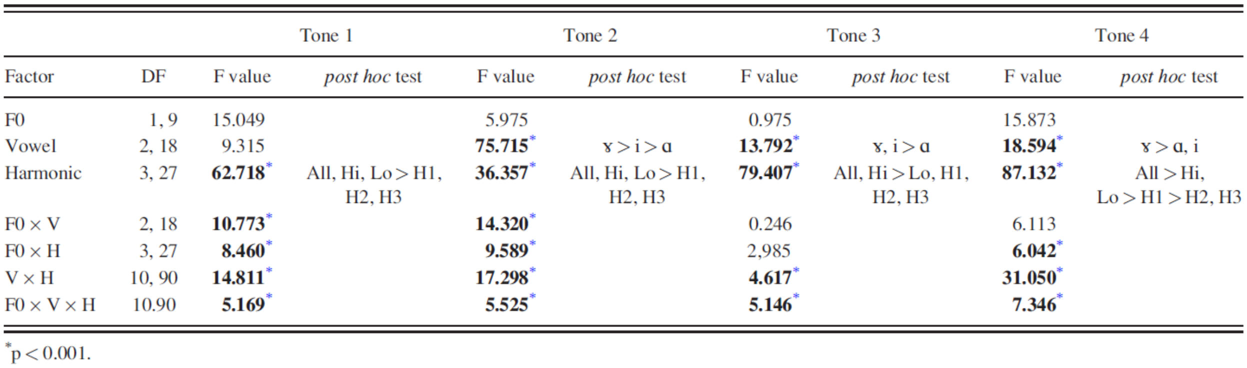
$^*p < 0.001$.

Table 4.3: Results of four-factor (F0 value × vowel category × tone category × stimulus harmonic) repeated-measures ANOVAs and Tukey post hoc tests in quiet, and the SNR of 5 and 0 dB. DF: degree of freedom; F: Female; M: Male; Hi: High; Lo: Low.

Due to the binomial distribution of the tone identification scores, a sigmoidal model was used to fit the tone identification data as a function of SNRs as shown in the following equation:

$$p(c) = \alpha + \left[(1.0 - \alpha)/\left(1.0 + e^{-(x-x_0)/b}\right)\right]$$

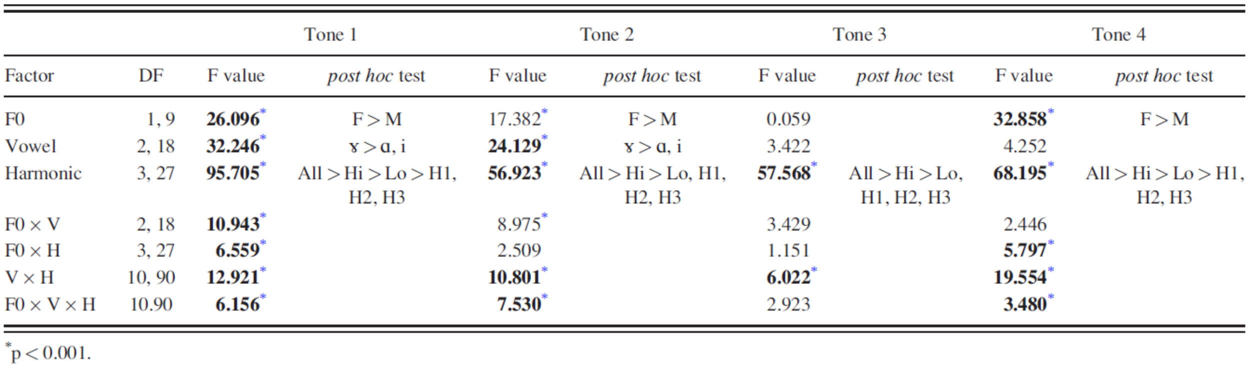
Where α =25% (chance performance or 1/4), b is the steepness of the function, x_0 is the midpoint (62.5% correct) of the function, and x is the SNR value. The sigmoidal functions were fit for individual harmonics and for Low and High harmonics, respectively, using the Marquardt-Levenberg algorithm (Marquardt, 1963) implemented in SigmaPlot v10.0. For particular harmonics, the tone identification data were clustered across the two individual harmonics (H1 and H2), while the tone identification data were somewhat separate for H3 (see Fig. 4.12), due to the greater possibility of formant interaction for H3 than for H1 and H2. For multiple harmonics, tone identification differed between the conditions of Low and High harmonics (see Fig. 4.13). Thus, the sigmoidal model was fit for H1+H2, H3, Low harmonics, and High harmonics, respectively. The result showed that a significant goodness of fit was observed for the conditions of H1+H2 (R=0.86, p<0.05), H3 (R=0.65, p<0.05), Low harmonics (R=0.77, p<0.05), and High harmonics (R=0.52, p<0.05). These results suggest that audibility of the harmonics (e.g., local SNRs) may play a fundamental role in tone identification in multi-talker babble.



		Tone 1		Tone 2		Tone 3		Tone 4	
Factor	DF	F value	*post hoc* test	F value	*post hoc* test	F value	*post hoc* test	F value	*post hoc* test
F0	1, 9	15.049		5.975		0.975		15.873	
Vowel	2, 18	9.315		**75.715***	ɤ > i > ɑ	**13.792***	ɤ, i > ɑ	**18.594***	ɤ > ɑ, i
Harmonic	3, 27	**62.718***	All, Hi, Lo > H1, H2, H3	**36.357***	All, Hi, Lo > H1, H2, H3	**79.407***	All, Hi > Lo, H1, H2, H3	**87.132***	All > Hi, Lo > H1 > H2, H3
F0 × V	2, 18	**10.773***		**14.320***		0.246		6.113	
F0 × H	3, 27	**8.460***		**9.589***		2,985		**6.042***	
V × H	10, 90	**14.811***		**17.298***		**4.617***		**31.050***	
F0 × V × H	10.90	**5.169***		**5.525***		**5.146***		**7.346***	

*$p < 0.001$.

Table 4.4: Results of three-factor (F0 value × vowel category× stimulus harmonic) repeated-measures ANOVAs for each of the four tones in the SNR of 5 dB. DF: degree of freedom; F: female; M: male; Hi: high; Lo: low.



		Tone 1		Tone 2		Tone 3		Tone 4	
Factor	DF	F value	*post hoc* test	F value	*post hoc* test	F value	*post hoc* test	F value	*post hoc* test
F0	1, 9	**26.096***	F > M	17.382*	F > M	0.059		**32.858***	F > M
Vowel	2, 18	**32.246***	ɤ > ɑ, i	**24.129***	ɤ > ɑ, i	3.422		4.252	
Harmonic	3, 27	**95.705***	All > Hi > Lo > H1, H2, H3	**56.923***	All > Hi > Lo, H1, H2, H3	**57.568***	All > Hi > Lo, H1, H2, H3	**68.195***	All > Hi > Lo > H1, H2, H3
F0 × V	2, 18	**10.943***		8.975*		3.429		2.446	
F0 × H	3, 27	**6.559***		2.509		1.151		**5.797***	
V × H	10, 90	**12.921***		**10.801***		**6.022***		**19.554***	
F0 × V × H	10.90	**6.156***		**7.530***		2.923		**3.480***	

$^*p < 0.001$.

Table 4.5: Results of three-factor (F0 value × vowel category × stimulus harmonic) repeated-measures ANOVAs for each of the four tones in the SNR of 0 dB. DF: degree of freedom; F: female; M: male; Hi: high; Lo: low.

4-3-4-2- Mechanisms of tone recognition in quiet and in babble

In quiet, the stimuli with three individual harmonics reached high tone identification accuracy (average scores >95% over vowel and tone categories) for both male and female speakers, implying that listeners were able to recognize Mandarin tones accurately as long as the individual harmonic was well audible. In the six-talker babble, signals with all harmonics showed high accuracy of tone recognition (e.g., >90%). The small noise effect on tone identification for the signals with all harmonics was likely because the two SNRs in the present study were not low enough. When Mandarin speech signals were presented in speech-shaped noise with the SNRs at 10, 0, and -5 dB, tone identification accuracy was greater than 95% (Kong and Zeng, 2006; Lee et al., 2010; Liu et al., 2012). Dees et al. (2007) also reported that tone identification in six-talker babble reached above 90% at 15 dB SNR. According to the participants' informal reports, the six-talker babble contained plenty of tonal information, implying that informational masking may significantly affect tone identification.

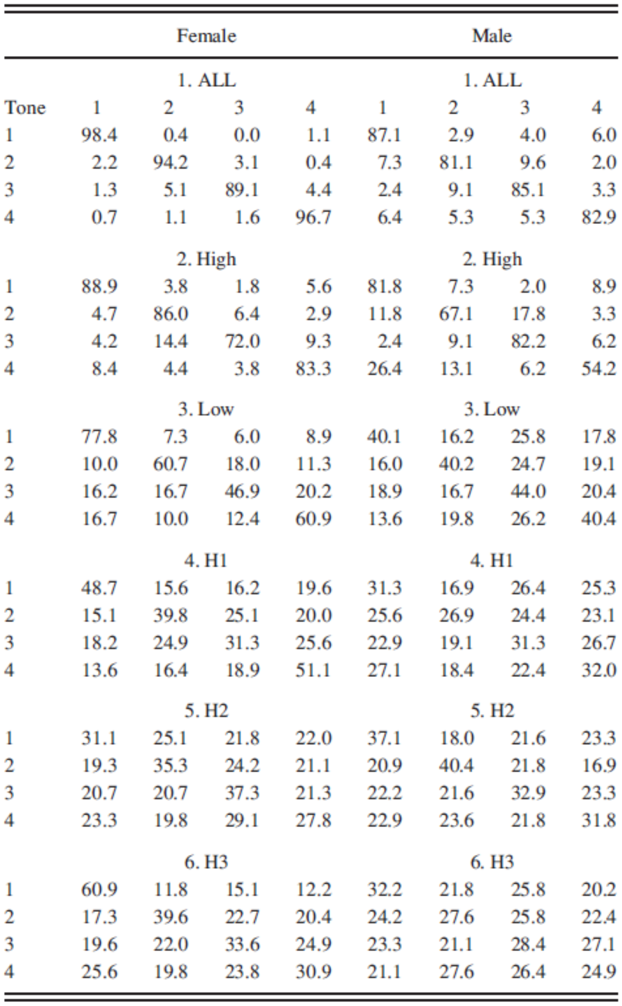
	Female				Male			
	1. ALL				1. ALL			
Tone	1	2	3	4	1	2	3	4
1	100				99.6	0.4		
2	0.7	96.0	3.1	0.2	1.3	97.6	1.1	
3	0.9	4.4	94.4	0.2	3.1	2.9	93.8	0.2
4			0.4	99.6		0.7	0.2	99.1
	2. High				2. High			
1	98.7	1.3			97.8	1.3	0.2	0.4
2		99.6	0.4		1.1	97.6	1.3	
3		3.3	96.2	0.4	0.2	2.7	96.7	0.4
4	0.4	0.7	0.2	98.7		0.4	0.9	98.7
	3. Low				3. Low			
1	99.1	0.4	0.4		98.7	1.1	0.2	
2	1.1	94.9	4.0		0.4	98.9	0.4	0.2
3	0.9	0.9	98.0	0.2	3.1	0.9	96.0	0.0
4	0.4		0.9	98.7	0.7		1.1	98.2
	4. H1				4. H1			
1	98.2	0.4	0.7	0.7	98.2	1.1	0.2	0.2
2	0.7	93.8	4.7	0.7	1.1	92.7	6.0	0.2
3	2.7	5.3	90.7	1.3	2.9	3.1	92.9	1.1
4	1.3	0.2	0.7	97.8	0.2		0.7	99.1
	5. H2				5. H2			
1	98.4	1.1	0.2	0.2	99.1	0.7	0.2	
2	0.2	95.1	4.4	0.2	0.7	97.6	1.3	0.2
3	0.7	2.7	96.0	0.7	1.1	0.9	97.6	0.2
4				100.0		0.2	0.2	99.6
	6. H3				6. H3			
1	96.9	0.9	1.3	0.9	97.8	0.2	1.6	0.4
2	0.4	94.7	4.2	0.7	1.6	96.0	2.0	0.4
3	1.1	5.8	92.2	0.9	1.8	2.0	95.3	0.7
4	0.2	1.3	0.9	97.6	0.2	0.4	1.1	98.2

Table 4.6: Confusion matrix of tone identification in quiet for the six harmonic conditions and two talkers (female: left; male: right).

	Female				Male			
	1. ALL				1. ALL			
Tone	1	2	3	4	1	2	3	4
1	98.4	1.3	0.0	0.2	99.3	0.2	0.2	0.2
2	2.0	95.3	2.0	0.7	2.7	95.1	2.0	0.2
3	0.2	5.3	92.9	1.6	0.7	4.0	94.7	0.7
4	0.2	0.4	1.1	98.2	0.2	0.2	0.2	99.3
	2. High				2. High			
1	94.9	2.0	0.4	2.7	94.0	1.6	0.7	3.8
2	5.1	83.8	9.8	1.3	5.6	76.2	16.7	1.6
3	2.7	8.4	85.8	3.1	0.4	6.7	91.1	1.8
4	4.7	1.8	2.2	91.3	15.8	10.0	3.3	70.9
	3. Low				3. Low			
1	93.8	3.3	1.3	1.6	78.9	8.9	5.6	6.7
2	3.1	74.9	17.1	4.9	10.4	77.3	5.6	6.7
3	10.4	10.7	68.9	10.0	5.6	7.8	79.6	7.1
4	6.7	2.7	4.4	86.2	7.3	6.7	6.9	79.1
	4. H1				4. H1			
1	73.3	7.6	7.8	11.3	51.3	16.2	14.9	17.6
2	9.3	62.0	16.9	11.8	21.8	32.0	23.6	22.7
3	18.9	17.3	43.1	20.7	20.0	21.8	39.8	18.4
4	10.9	8.7	8.7	71.8	16.0	12.4	12.4	59.1
	5. H2				5. H2			
1	37.3	19.1	24.7	18.9	50.4	15.6	19.1	14.9
2	13.8	47.6	23.6	15.1	16.9	58.0	12.9	12.2
3	16.7	15.8	52.2	15.3	18.2	18.4	46.0	17.3
4	13.3	20.9	24.4	41.3	18.2	16.4	16.2	49.1
	6. H3				6. H3			
1	62.0	13.1	13.6	11.3	37.3	23.3	22.9	16.4
2	18.0	46.4	20.4	15.1	17.6	37.1	24.0	21.3
3	16.4	26.0	39.3	18.2	23.6	21.8	28.7	26.0
4	17.6	13.3	22.7	46.4	23.1	24.4	25.3	27.1

Table 4.7: Confusion matrix of tone identification at the SNR of 5 dB for the six harmonic conditions and two talkers (female: left; male: right).

Also, the confusion matrix of the present study showed equal distribution (see Table 4.6, 4.7, and 4.8), while the confusion matrix in speech-shaped noise showed more specificity (Lee et al., 2010; e.g., tone 3 was primarily confused by tone 2), showing different masking effects of babble and speech-shaped noise on tone recognition. In general, greater temporal variation and informational masking were present for the multi-talker babble than for the speech-shaped noise (Brungart et al., 2001; Mi et al., 2013; Rosen et al., 2013).



	Female				Male			
	1. ALL				1. ALL			
Tone	1	2	3	4	1	2	3	4
1	98.4	0.4	0.0	1.1	87.1	2.9	4.0	6.0
2	2.2	94.2	3.1	0.4	7.3	81.1	9.6	2.0
3	1.3	5.1	89.1	4.4	2.4	9.1	85.1	3.3
4	0.7	1.1	1.6	96.7	6.4	5.3	5.3	82.9
	2. High				2. High			
1	88.9	3.8	1.8	5.6	81.8	7.3	2.0	8.9
2	4.7	86.0	6.4	2.9	11.8	67.1	17.8	3.3
3	4.2	14.4	72.0	9.3	2.4	9.1	82.2	6.2
4	8.4	4.4	3.8	83.3	26.4	13.1	6.2	54.2
	3. Low				3. Low			
1	77.8	7.3	6.0	8.9	40.1	16.2	25.8	17.8
2	10.0	60.7	18.0	11.3	16.0	40.2	24.7	19.1
3	16.2	16.7	46.9	20.2	18.9	16.7	44.0	20.4
4	16.7	10.0	12.4	60.9	13.6	19.8	26.2	40.4
	4. H1				4. H1			
1	48.7	15.6	16.2	19.6	31.3	16.9	26.4	25.3
2	15.1	39.8	25.1	20.0	25.6	26.9	24.4	23.1
3	18.2	24.9	31.3	25.6	22.9	19.1	31.3	26.7
4	13.6	16.4	18.9	51.1	27.1	18.4	22.4	32.0
	5. H2				5. H2			
1	31.1	25.1	21.8	22.0	37.1	18.0	21.6	23.3
2	19.3	35.3	24.2	21.1	20.9	40.4	21.8	16.9
3	20.7	20.7	37.3	21.3	22.2	21.6	32.9	23.3
4	23.3	19.8	29.1	27.8	22.9	23.6	21.8	31.8
	6. H3				6. H3			
1	60.9	11.8	15.1	12.2	32.2	21.8	25.8	20.2
2	17.3	39.6	22.7	20.4	24.2	27.6	25.8	22.4
3	19.6	22.0	33.6	24.9	23.3	21.1	28.4	27.1
4	25.6	19.8	23.8	30.9	21.1	27.6	26.4	24.9

Table 4.8- Confusion matrix of tone identification at the SNR of 0 dB for the six harmonic conditions and two talkers (female: left; male: right).

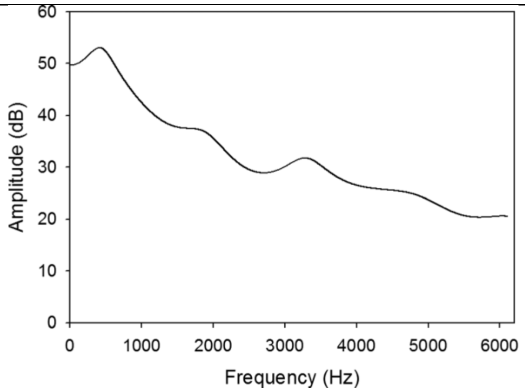


Figure 4.11: The LPC long-term spectrum for the 10-s six-talker Mandarin babble using a sampling frequency of 12207 Hz with the RMS level at 70 dB SPL

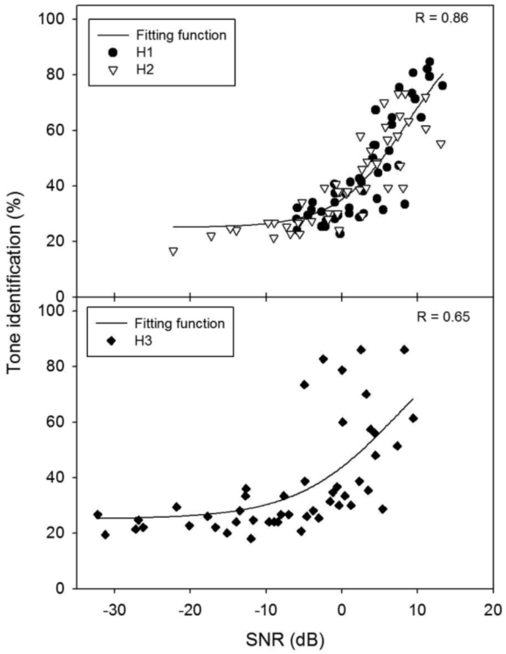


Figure 4.12: The sigmoidal fitting function of tone identification as a function of local SNRs for clustered H1 and H2 (top; R=0.86, p<0.05) and for H3 (bottom; R=0.65, p<0.05).

However, compared to the speech-shaped noise used in the previous studies (Kong and Zeng, 2006; Lee et al., 2010; Liu et al., 2012), the babble of this study had different SNRs or different spectral and temporal features. Thus, the role of temporal variation and informational masking in the babble was difficult to study. More research is needed to reveal their roles by systematically manipulating SNRs and spectral and temporal properties of noise.

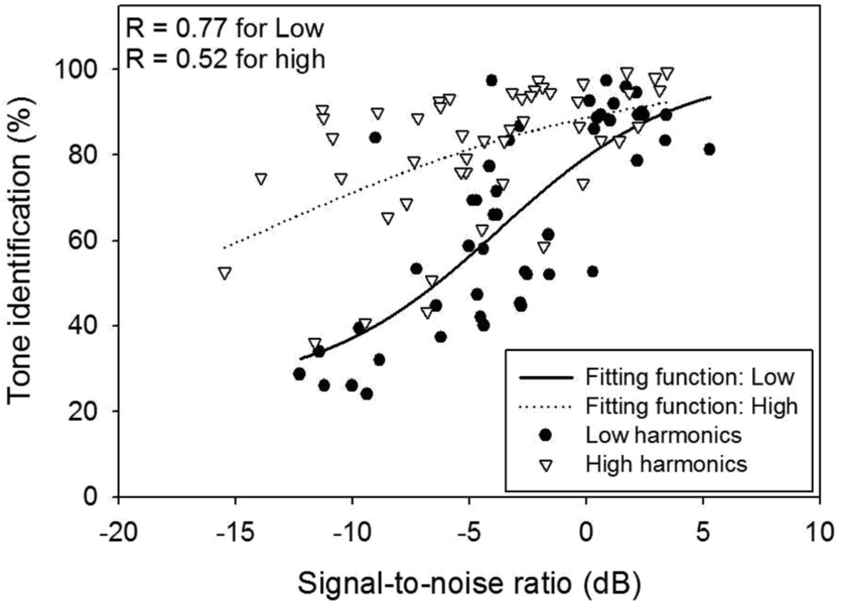


Figure 4.13- The sigmoidal fitting function of tone identification as a function of SNRs for the low (R=0.77, p<0.05) and high (R=0.52, p<0.05) harmonics.

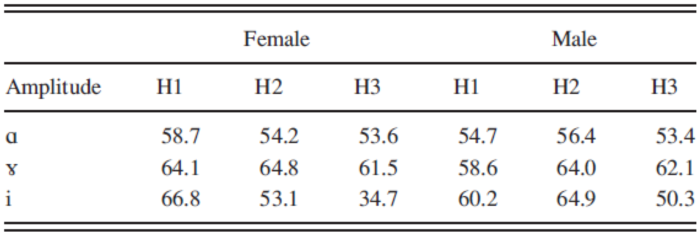
As individual, Low, and High harmonics were presented in babble, tone identification was significantly affected by SNR, indicating the important effect of noise masking; specifically, as shown in Figs. 4.12 and 4.13, tone identification accuracy had a significant relationship with local SNRs for an individual, Low, and High harmonics. These results suggest that when an individual harmonic was performed, audibility might be a primary factor accounting for tone identification in noise. This is consistent with the finding of Liu et al. (2012) that tone identification scores were quite low (e.g., 40%-50%) for H1 in speech-shaped noise due to its low audibility [see their Fig. 4.7(b)]. The Low and High harmonics showed significantly better tone identification compared with individual harmonics, suggesting that listeners were able to integrate tone information from different frequency bands (i.e., auditory channels) to identify Mandarin tones (Oxenham et al., 2004). Moreover, compared with Low harmonics, High harmonics had significantly better tone identification in noise (see Figs. 4.9 and 4.10), likely due to their relatively better audibility in noise (see Fig. 4.13). The High harmonics in the present study were composed of frequency components above 500 Hz for the low F0 and above 900 Hz for the high F0, indicating that several harmonics in the middle-frequency range might be resolved and used for tone perception in noise. Also, for High harmonics, the nonlinear processing of the cochlea may generate a frequency component at F0 (e.g., otoacoustic emission products; Pressnitzer and Patterson, 2001; Yost, 2009). The distortion spectrum (e.g., low-frequency harmonics) was produced with the highest amplitude at the F0 component, and the level of these distortion products was well above audible thresholds, possibly contributing to tone recognition (Pressnitzer and Patterson, 2001).

Studies of pitch perception with complex tones suggested that resolved harmonics (e.g., low and middle frequencies) may be coded by their excitation pattern in the cochlea, by phase locking of the temporal pattern of auditory neuron firing, or by both (Oxenham et al., 2004). For unresolved harmonics (e.g., high frequencies), the extraction of F0 may rely on phase locking to the temporal envelope of complex tones, or phase locking to the temporal fine structures (e.g., the local peaks near the temporal envelope peaks), or both (Schouton et al., 1962). Also, Moore and Moore (2003) suggested that for the harmonics that were partially resolved (e.g., the High harmonics in the current study), listeners may be able to extract partial harmonics to perceive pitch of the complex tone. When vowel stimuli were presented in multi-talker babble in the present study, the resolved low-frequency harmonics in the cochlea were degraded, making tone identification with individual and Low harmonics more difficult. The temporal envelope and temporal fine structure of vowel stimuli were corrupted by babble such that phase locking to these temporal features was reduced, thus resulting in greater difficulty to perceive tonal pitch for High harmonics. Moreover, the low-frequency distortion product (e.g., F0 component) produced by the cochlear nonlinearity (Pressnitzer and Patterson, 2001) may be masked by multi-talker babble, also leading to the lower performance in babble.

In addition to background noise and harmonic configuration, other factors such as vowel category and F0 also significantly affected tone identification. The amplitude of low harmonics is related to an interaction between harmonic frequencies and formant structures of speech sounds; that is, when a low harmonic is near to or coincide with the first formant (F1) of vowel signals, its amplitude may be affected by the resonatory structure of the vocal tract. For instance, the F1 frequency of vowel /i/ is usually low near 300 Hz, and the average F0 frequency of male and female speakers was 150 and 250 Hz in the present study; thus, the second harmonic of the male speaker and the first harmonic of the female speaker were near to F1, resulting in relatively higher amplitude than the amplitude of other low-frequency harmonics (see Table 4.9). On the other hand, for high F1 frequencies of vowels like /ɑ/, the H3 of the female speaker that was closer to F1 than H1 and H2 (e.g., 1125 Hz on average for the female speaker) may have a higher amplitude in the resolved harmonics compared with other low-frequency harmonics. However, it should be noted that higher amplitude of an individual low harmonic may not necessarily lead to higher tone identification, which was defined by harmonic SNRs, not just harmonic amplitude.

On the other hand, the durational cue was removed by duration equalization. Although tone duration plays a secondary cue in tone identification in quiet (Lin and Repp, 1989) and also tone identification with equalized durations was well above 90% in quiet, the role of tone duration may become more important in noise in which harmonic information is masked. In particular, duration equalization was conducted by shortening vowel durations, possibly resulting in the greatest effect on tone 3 that has the longest duration among the four lexical tones (Howie, 1976). Moreover, the duration equalization in this study also reduced the extent of pitch contour movement, possibly contributing to the challenge of tone identification in noise. More research is needed to reveal the effects of tone duration and the interaction between tone duration and harmonic structure on tone perception in noise.

The results of this study suggest that to improve Mandarin tone perception in noise for normal-hearing listeners, the key factor may be to increase local SNRs for low-frequency harmonics that are resolvable, especially in a listening scenario in which speech signals may be able to be enhanced prior to delivery via communication system to a receiver (e.g., telephone or public-address systems). For example, telephone speech typically has a bandwidth of 300 Hz and 3400 Hz, and therefore low-frequency harmonics below 300 Hz are severely reduced (Hu and Loizou, 2007). Thus, in applications of telephone and cellphone speech processing with severely degraded low-frequency harmonics, cue pre-enhancement can be applied to restore and amplify low-frequency harmonics of speech signals. Such speech enhancement in low-frequency harmonics embedded in unprocessed surrounding noise (i.e., enhancing the speech signal is implemented before the speech being mixed with the competing background signal) may improve tone recognition in noise. The benefit of improving the SNRs of low-frequency harmonics, however, may be dependent on vowel category (e.g., front or back vowels) and speaker gender (e.g., F0 range).



	Female			Male		
Amplitude	H1	H2	H3	H1	H2	H3
ɑ	58.7	54.2	53.6	54.7	56.4	53.4
ɤ	64.1	64.8	61.5	58.6	64.0	62.1
i	66.8	53.1	34.7	60.2	64.9	50.3

Table 4.9: Average amplitude (dB SPL) over the four tones for individual low-frequency harmonics (H1, H2, and H3) for the three vowels and for the female and male speaker.

4-3-5-SUMMARY

In summary, we found:

1- Resolved individual low harmonics, such as the first three harmonics, led to high accuracy in Mandarin tone identification in quiet, while their contribution to tone perception in noise may be limited by noise masking.

2- The local signal-to-noise ratios of individual harmonics were strongly related to tone identification scores in noise across vowel category, F0, and tone category.

3- When speech signals contain multiple harmonics, their tone identification in noise was significantly better than individual low harmonics, indicating that listeners had integrated tone information from different groups of harmonics.

Chapter Five – Methods

Multi-channel Voice Conversion algorithm

The aim of this research is to improve the intelligibility of speech for cochlear implant patients and help them to have good hearing performance even in noisy environments. In the positive side, as we discussed in motivation, voice conversion increasing the intelligibility of speech and in the result of various researches show good performance for regular listeners. However, all of the current techniques for voice conversion reducing the quality of speech and speech coherence. And as we know CI patients are more sensitive to noise than normal listeners. Therefore, none of those techniques can make good for CI users.

In this research, we proposing Multi-channel voice conversion algorithm to reduce the noise and increasing intelligibility. One of the advantages of this technique is we can reduce the limitations, related to acoustic waveform reconstruction in conversion stage by transporting the energy in frequency channels directly to electric outputs. We will adapt this technique to improve several most well-known techniques such the state of art Gaussian mixture model (GMM) and Artificial Neural Network (ANN).

Multi-channel procedures have also been used in the literature for "Improves speech intelligibility in noise for normal-hearing listeners" Gibak Kim et al. (2009) [86] and Noise reduction strategies in Cochlear Implants Kokkinakis et al. (2012) [62]. In next several pages, we describe noise reduction techniques used in this research. After that, we explain our feature extraction and preparation before modeling and at last we demonstrate the presented technique incorporating with each of three primary algorithms.

Noise reduction strategies in Cochlear Implants

It is not offbeat that intelligibility of speech reduces with interferes of background noise. And also the impact of this issue will increase by raising the level of background noise. In our daily events, we are able to understand speech even a high level of noise combine with the source of speech and it is because human speech is a very redundant signal. Therefore, even if portions of the speech signal are concealed by noise, less affected parts of the speech signal have enough information to understand the signal and protect intelligibility of speech.

Around five decades ago, Schroeder introduced the conventional speech enhancement strategies to improve speech corrupted by additive noise. After that, other researchers tried systematically to formulate the challenging problem of noise reduction and also compared different algorithms known at the time (e.g., see Boll, 1979 [63]; Ephraim & Malah, 1984[64]; Ephraim & VanTrees, 1995[65]; Lim, 1978[66]; Lim & Oppenheim, 1979[67]). Since that, scientists tried to establish techniques such as noise removal, noise suppression, noise reduction, and speech enhancement that improve the intelligibility of speech in the presence of background noise. With the growing usage of cochlear implants, it comes to researchers to using noise reduction solutions for CI users to enhance speech intelligibility when background noise is present.

There are two kinds of noise reduction algorithms fitting for sound processing approaches in CI systems: one is, depends on preprocessing the noisy acoustic signals on the front-end side located before the radio-frequency (RF) link, which transmits the audio stream to the inside implanted receiver. This strategy is similar to speech enhancement technique currently used in most modern communication devices. The other type of single-microphone noise reduction strategy is, based on applying some type of attenuation directly on the noisy electrical envelopes.

Figure 5.1 explains the difference between these two strategies.

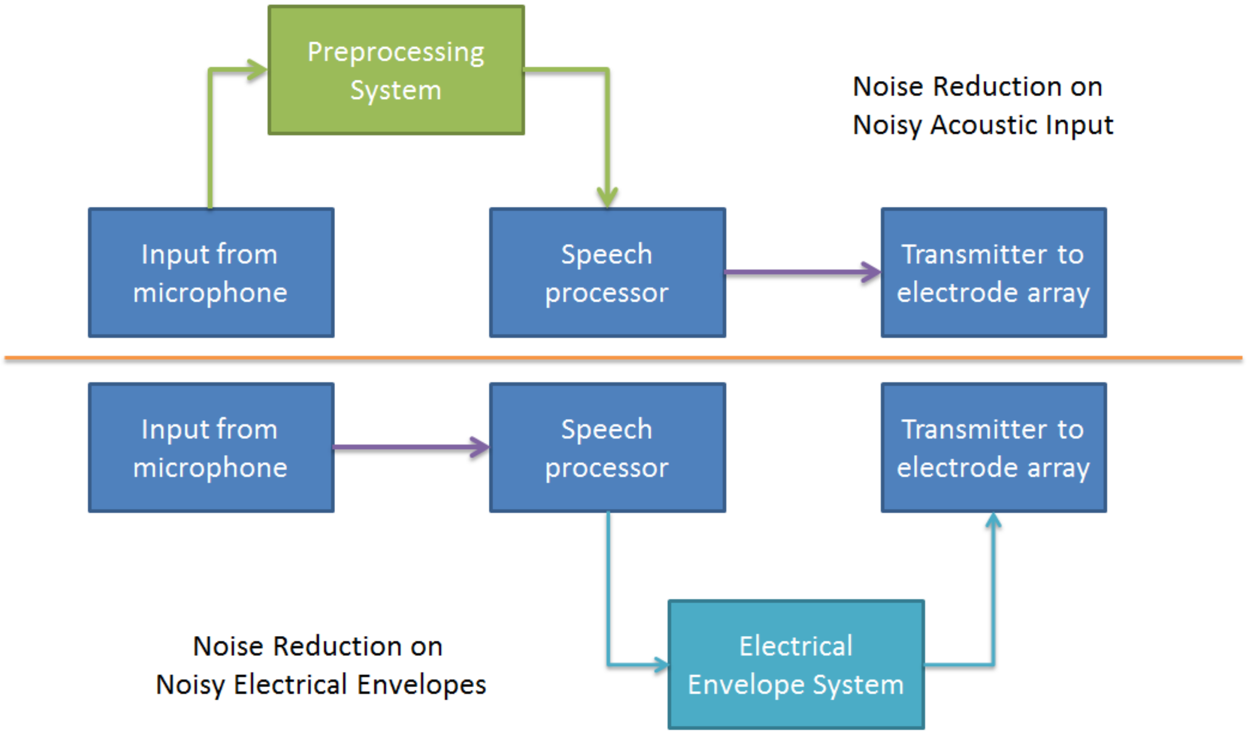

Preprocessing System
Noise Reduction on Noisy Acoustic Input
Input from microphone
Speech processor
Transmitter to electrode array
Input from microphone
Speech processor
Transmitter to electrode array
Noise Reduction on Noisy Electrical Envelopes
Electrical Envelope System

Figure 5.1 Top side shows position of preprocessing technique in signal processing of cochlear implants; Bottom side shows noise reduction on noisy Electrical envelopes.

Noise reduction on noisy acoustic inputs

The idea of using a digital single-channel noise suppression algorithm (INTEL) to process noisy speech inputs acoustically for CI devices provided by Hochberg et al. (1992)[68] and Weiss (1993)[69]. On that study, the enhanced acoustic stimuli generated with the INTEL technique and then presented to a group of normal-hearing persons and Nucleus 22 implant device users. Test performed on several input signal-to-noise ratios and results in both groups shows INTEL reduced noise steadily.

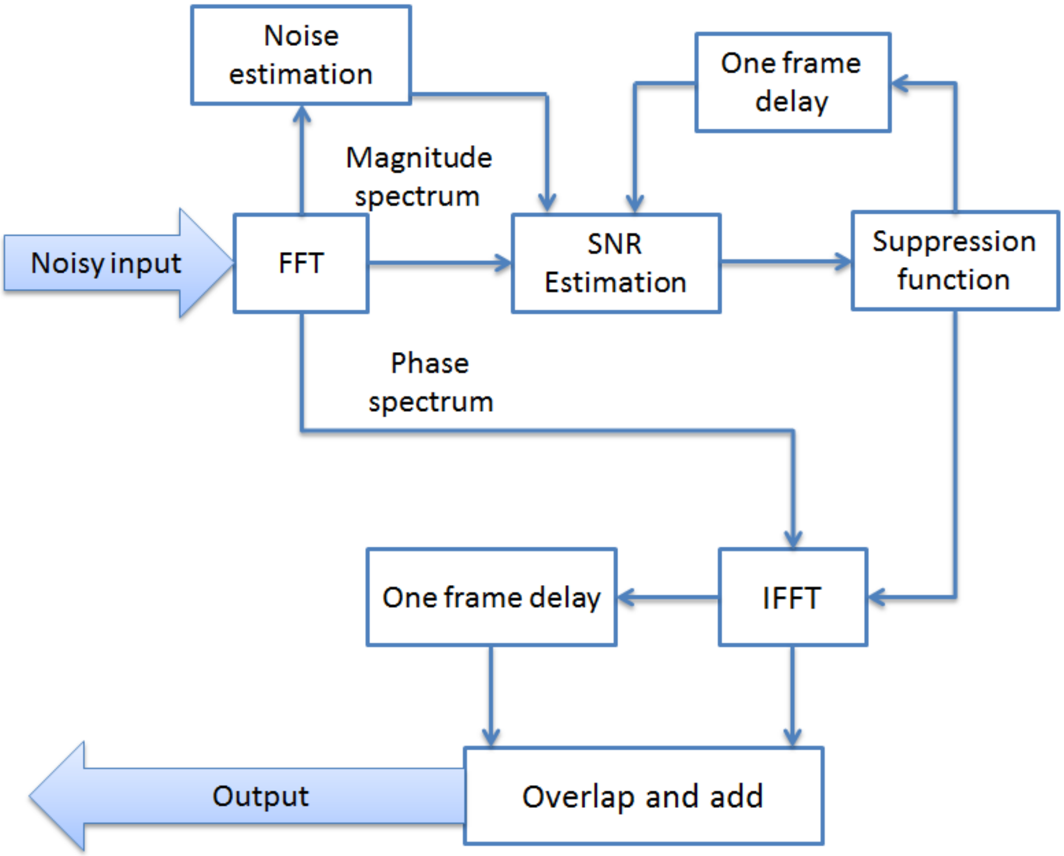


Figure 5.2 Block diagram of single-microphone noise reduction methods based on preprocessing the noisy acoustic input signals.

Also, this study shows this technique (INTEL) could improve the phoneme recognition performance of consonant-vowel-consonant (CVC) words corrupted with a speech-shaped random noise for CI patients remarkably. After that, in research performed by Yang and Fu (2005) [70], indicates notable improvements to sentences recognition in stationary speech-shaped noise situation at different SNRs for a group of seven CI users that used different CI devices. Figure 5.2 illustrates the noise reduction algorithms implemented in the frequency domain.

Also, Loizou et al. (2005)[71] studied the potential advantages of first preprocessing the noisy acoustic input with a custom subspace-based noise reduction algorithm. The subspace algorithm was first developed by Ephraim and Van Trees (1995) [65] for suppressing white input noise. However, several year after that extended to reducing colored noise (e.g.,speech-shaped noise) by Hu and Loizou (2002) [72].

The subspace algorithm is relying on the projection of the noisy speech vector (i.e., consisting of a segment of speech) onto two subspaces: the “noise” subspace and the “signal” subspace (Ephraim & VanTrees, 1995) [65]. Signal components stored in the noise subspace and the clean signal kept in the signal subspace. Consequently, an initial estimate of the clean signal can be determined by eliminating the components of the signal in the noise subspace and keeping only the elements of the signal in the signal subspace. This action can cause to suppressed noise to a large amount.

In following, mathematical equations and functions for the subspace-based noise reduction algorithm described. (Loizou et al. (2005) [71]) Estimated clean speech signal defined in equation below:

$$\hat{x} = Hy$$

Where $\hat{x}$ is estimated clean speech signal vector, y is noisy speech vector, and H is transformation matrix for the estimation of the clean speech signal. In brief, the aim is finding the transformation matrix H, which can estimate clean signal from the noisy vector. Next step is determining the error between the estimated signal vector $\hat{x}$ and the clean signal vector x. Equation below computes such error:

$$\varepsilon = \hat{x} - x = (H - 1) \times x + H \times n$$

Where n is the noise vector.

However, estimated transformation matrix will not be absolute, it will introduce some speech distortion, which is determined by the first error term $(H - 1) \times x$ also the amount of noise distortion introduced by the transformation matrix determined by the second error term $(H \times n)$.

After separating the speech and noise distortion, we can compute the optimal transformation matrix H which is falling below the threshold defined below (Hu and Loizou (2002) [72]):

$$H_{opt} = V^{-T}\, \Lambda(\Lambda\ + \mu I)^{-1} \Lambda^{-T}$$

Where μ is a parameter (typical values for $\mu = 1–20$), V is an eigenvector matrix and Λ is a diagonal eigenvalue matrix obtained from the noisy speech vector

In Loizou et al. (2005) [71] research, they tested this subspace based noise reduction algorithm using HINT sentences (Nilsson et al., 1994 [73]) with 5 dB SNR level speech-shaped noise. 14 Clarion CI users attend to speech intelligibility experiment, and results were compared against the users' standard sound processing strategy, either the continuous interleaved stimulation (CIS) strategy or the simultaneous analog stimulation (SAS) strategy. Results indicated that when using the subspace-based single-microphone noise reduction algorithm almost all subjects (CI users) significantly raised their speech perception scores.

Noise Reduction on Noisy Electrical Envelopes

The method of preprocessing noisy acoustic inputs with a proper noise reduction strategy has statistically significant results. Nevertheless, that noise reduction technique is not without drawbacks, such as:

1. Preprocessing algorithms often introduce unwanted acoustic distortion in the signal
2. Some algorithms (e.g., subspace algorithms) are computationally complex and fail to integrate well with existing CI strategies
3. It is not straightforward to always fine tune (or optimize) the operation of a particular algorithm to individual users.

Ideally, noise reduction algorithms should be easy to implement and be integrated into existing coding strategies. From a computational attitude, using directly apply attenuation to the electrical envelopes according to the intensity contrast between the speech signal and the noise signal is the easiest way to degrade this problem, and we can reduce the limitations, related to acoustic waveform reconstruction by transferring the energy in frequency channels directly to electric outputs.

An envelope-weighting is a strategy that can efficiently reduce noise by attenuating electrical envelopes while avoiding the intermediate acoustic waveform reconstruction stage. This method can directly effect on noisy electrical envelopes, therefore, can be easily used in existing procedures worked in commercially available implant devices (Hu et al., 2007 [74]). The stated technique is based on envelope-weighting of each spectral channel, and it is part of noise suppression by spectral modification algorithms. Figure 5.3 describes the envelope-weighting strategy block diagram.

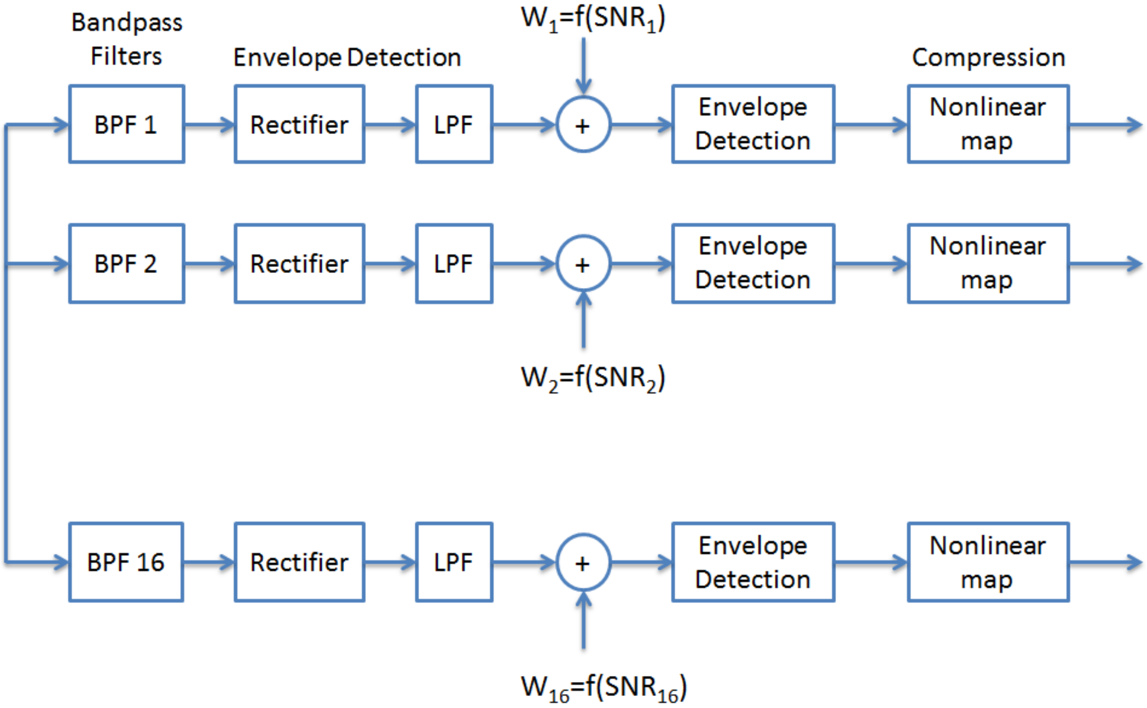


Figure 5.3 describes the envelope-weighting strategy block diagram.

In this algorithm, the enhanced signal envelopes are selected from noisy envelopes of each channel by applying a weight to them (taking values in the range 0-1). The weights have to be inversely proportional to the calculated SNR of each channel. In channels with high SNR, envelope amplitudes are multiplied by a weight close to one (i.e., left unchanged), on the other hand, in spectral channels with a low SNR level, envelope amplitudes are multiplied by a weight close to zero (i.e., heavily attenuated).

The main idea is that channels with low SNR are heavily masked by noise and therefore provide insufficient, if any, information about the speech signal. In essence, these low-SNR channels are heavily attenuated and keeping only the high-SNR channels, which are expected to provide more valuable information to the listener. It means we are using weighting functions that implement high attenuation in channels with low SNR and at the same time little or no attenuation in channels with high SNR levels (e.g., see Hu et al., 2007 [74]).

The proper decision is using the sigmoidal-shaped function such as:

$$g(i,l) = \exp(\frac{-b}{SNR(i,l)})$$

Where $g(i,l)$ indicates the weighting function ($0 < g(i,l) < 1$), b is equal to 2, and $SNR(i,l)$ indicates the estimated instantaneous SNR in the i^{th} channel and at stimulation cycle l.

Next, the enhanced temporal envelope can be consequently calculated by using equation below:

$$s(i,l) = g(i,l) \times \mathrm{y(i,l)}$$

Where $s(i,l)$ represents the enhanced signal and $\mathrm{y}(\mathrm{i},\mathrm{l})$ denotes the noisy envelope of the i^{th} channel at stimulation cycle ℓ.

As explained in Hu et al. (2007) [74], the stated sigmoidal-shaped function is capable of providing notable benefits to speech understanding of CI users in a presence of noise. The sigmoidal function maintains the envelope peaks and also expands the envelope valleys thereby increasing the effective envelope dynamic range within each channel.

The noise reduction technique explained, have shown promising results, however, it does not quite fix all issues CI users faced in noisy environments. The main reason is we can't estimate the SNR at each frequency band or each spectral channel. In 2008 Hu and Loizou [75] present another strategy to avoid this problem. In that approach, we assume that the real SNR values in each spectral channel are known a priori. Therefore, each frequency channel (or equivalent envelope) is selected only, if its corresponding SNR is larger than or equal to 0 db. In a similar way, a channel with SNR level which is smaller than 0 dB is dismissed. The principal opinion in this approach is spectral channels with a low level of SNR (e.g., SNR <0) carried masker-dominated envelopes and their effect on the speech signal is less or nothing. On the other hand channels with higher SNR levels (e.g., SNR ≥ 0 dB) carry target-dominated envelopes and can be preserved as they bear reliable information about the target input. Now, for implementing the SNR envelope-selection algorithm, we just need to multiply the noisy signal by a binary time-frequency (T-F) mask or equivalently a binary gain function to the electrical envelopes of the noisy signal.

This approach was tested on CI users, and the researchers showed that the SNR channel selection model can improve speech intelligibility of CI listeners in noisy environments (with extremely low input SNR levels, e.g., −10 dB) to become similar to listening in the quiet area. For this reason, this strategy is referred to as the optimal ACE (opACE; see Hu & Loizou, 2008) [75].

This envelop-selection strategy is one the best technique for suppressing noise. However, performing this strategy in the real world is not without challenge. Because computing the SNR level for each spectral channel from the mixture envelopes (corrupted speech) is a very challenging task. The research performed by Hu et al. [74] in 2007, confirmed that most conventional noise estimation algorithms perform poorly in estimating the SNR.

These algorithms are obtained by minimizing an appropriate optimization rule such as mean-square error, and they are assumed to work well in all noisy environments, which is definitely a very ambitious goal. Strictly speaking, common noise estimation algorithms are not optimized for a special listening situation, and therefore they cannot use the differences temporal and spectral characteristics of real-world.

In the study by Hu and Loizou [76] in 2010, they present an algorithm to select channels for stimulation based on estimated SNRs in each spectral channel. The authors proposed to take advantage of the distinctive temporal and spectral characteristics of different real-world noise reductions, which can be learned by using machine-learning techniques instead of relying on knowledge of the local SNR. The proposed noise reduction algorithm uses Gaussian mixture models (GMMs) to find out how to use the distinctive temporal and spectral characteristics of the different maskers in practice.

The proposed algorithm consists of two steps:

1. A training stage
2. A speech intelligibility enhancement stage.

In the training stage, by using the temporal envelopes of the speech signals (typically from a large corpus) and the noise signals, the SNRs are computed for each spectral channel. Then the binary status of the channels being defined speech dominated (with a binary gain of 1) or being defined masker dominated (with a binary gain of 0).

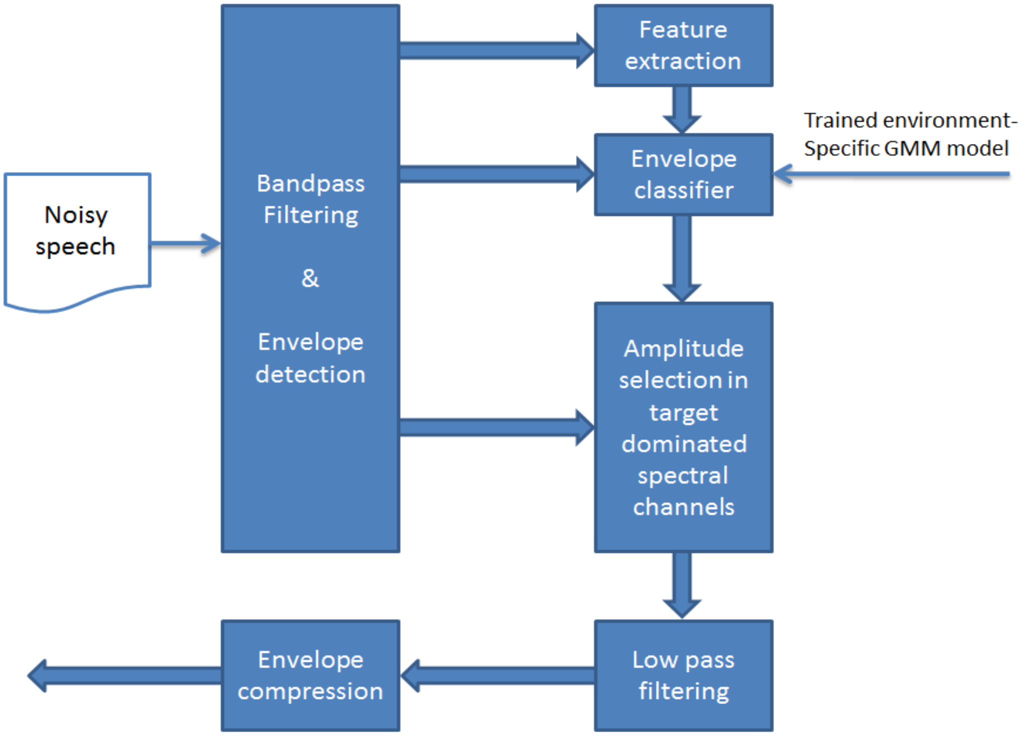

Feature extraction
Trained environment-Specific GMM model
Envelope classifier
Noisy speech
Bandpass Filtering & Envelope detection
Amplitude selection in target dominated spectral channels
Envelope compression
Low pass filtering

Figure 5.4 Block diagram of single-microphone noise reduction methods based on attenuating the noisy electrical envelopes through envelope-selection

In the next step, the features extracted from the noisy mixture temporal envelopes. Then for each spectral channel, the corresponding binary gains are used to train. Two GMMs representing two corresponding feature classes: target-dominated and masker-dominated. Note that in this stage features similar to amplitude modulation spectrograms (AMS) can be used (e.g., see Kollmeier & Koch, 1994 [77]; Tchorz & Kollmeier,2003 [78]).

In the enhancement stage, a Bayesian classifier used on the extracted features to classify the spectral channels into two classes: target-dominated and masker-dominated. Then in case spectral channel is target-dominated, the relevant electrode is chosen for stimulation. Figure 5.4 represents the block diagram of noise reduction methods based on envelope-selection as described in Hu and Loizou (2010)[76] .

Feature Extraction

The first step in any speech processing system is to extract features. In this step, we identify the components of the audio signal that are suitable for the classification of the Linguistic content and dumping all the other parts such as background noise. As we know human speed passing through vocal tract and based on its unique vocal tract components such as tongue, teeth, etc., filtered speech passed out of the mouth. The shape of output signal defines the speech. Therefore by determining the form of signal accurately we can get an accurate representation (feature) of the generated phonemes. The shape of vocal tract can estimate by the envelope of the short-time power spectrum.

The most popular features extraction method in speech processing systems is Mel-frequency cepstral coefficients (MFCCs), and its job is to represent this envelope accurately. This feature extraction method was first mentioned by Bridle and Brown in 1974[79], further developed by Mermelstein in 1976[80] and finally they were presented by Davis and Mermelstein in the 1980’s [81], and have been state-of-the-art ever since.

In continue, we describe two approaches to extracting features using MFCC. The first one is traditional approach. This method contains steps below, and Figure 5.5 shows block diagram of this method:

1. Frame the signal into short frames.
2. For each frame calculate the Fourier Transform.
3. Apply the Mel filterbank to the power spectra, sum the energy in each filter.
4. Take the logarithm of all filterbank energies.

5. Take the DCT of the log filterbank energies.
6. Keep DCT coefficients 2-13, discard the rest.

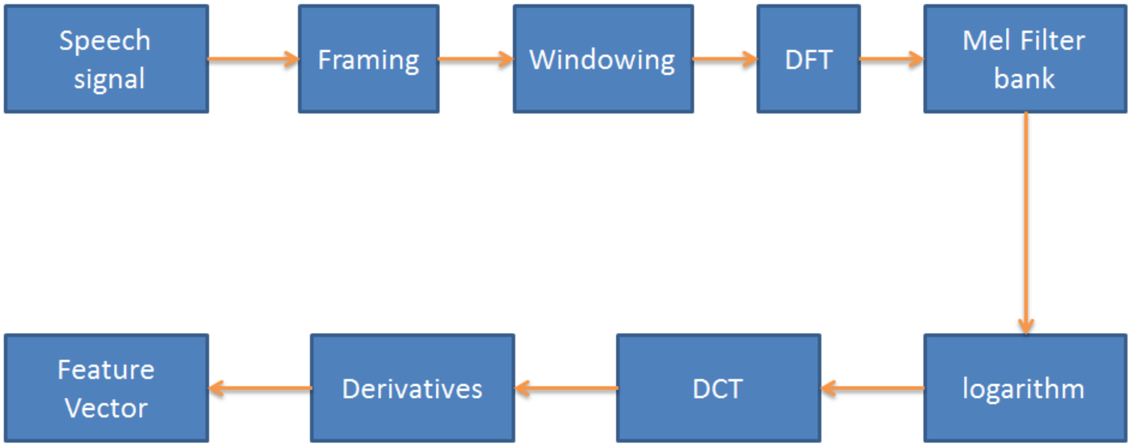


Figure 5.5 the block diagram of traditional MFCC features extraction.

Framing

Usually, speech signals are changing very fast. In order to work with that, we need to split it to short time scales. Therefore, the audio signal doesn't change enormously. That is the main reason we select our frame length between 20ms - 40ms. If we select less than 20ms, then the number of samples is not enough to get a decent spectral estimate. And if we chose longer than 40ms, then the signal changes too much during the frame. For example for 25ms frame and 16 kHz signal, frame length calculates as follow: 0.025*16000= 400. It means 400 samples per frame. And if we use 15ms overlap it means 240 samples is common between two frames and step frame is 160 samples. Figure 5.6 explain this example.

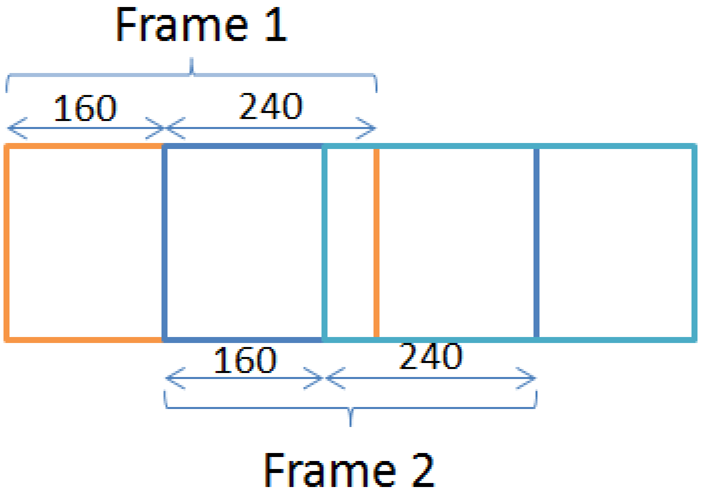


Figure 5.6 shows the frame with 400 samples length and frame step 160 samples

Windowing

The next step in the feature extraction process is, to windowing each frame, to reduce the signal interruption at the both sides of each frame (beginning and end). In this way, the signal at begging and end of each frame decreased to zero, therefore, minimize the spectral distortion. Hamming window is used as window shape in this method. Definition of hamming window given as:

$$Y(n) = X(n) \times W(n) \qquad 0 \leq n \leq N-1$$

Where: N is the number of samples in each frame, $Y(n)$ is Output signal, $X(n)$ is input signal and $W(n)$ is hamming window.

$$W(n) = 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right) \qquad 0 \leq n \leq N-1$$

Discrete Fourier Transform (DFT)

The next step is to compute the signal's Spectrum or Periodogram of each frame. To do that we need to determine the frequency domain representation of the input signal. Discrete Fourier Transform converts each frame of N samples from the time domain into the frequency domain. The DFT is defined as follow:

$$X_k = \left| \frac{1}{N} \sum_{n=0}^{N-1} x_n e^{\left[-i2\pi \frac{nk}{N}\right]} \right| \qquad \text{k=0, \ldots, N-1}$$

Where N is the number of samples, X_k is DFT output, and for each output X_k requires a sum of N terms.

Mel Filtering

Next step is the calculation of the Mel-frequency spectrum. As we know incoming audio signal vibrates in different spots of the human cochlea, and it depends on a frequency of the signal. Therefore, different nerves activate and inform the brain that certain frequencies are received. For this reason, the spectrum calculated in the last step need to filter with separate band-pass filters and the power of each frequency band need to compute. The Mel-scale is a non-linear scale that is accommodated to the non-linear pitch perception of the human auditory system. All attributes of the band-pass filters (such as number, the shape, and the center frequency) can be varied. The typical Mel filter bank setting is set of 20-40 (26 is standard) with triangular filters.

The first filter is shown the energy exists near 0 Hz frequency, and it is very narrow. By increasing the frequency of the filter, its bandwidth gets wider. We need to determine the value of energy in each spot. The Mel scale gives us the space between filter-banks and their bandwidth. The formula for converting from frequency to Mel scale is shown below:

$$M(f) = 2595 \log_{10}\left(1 + \frac{f}{700}\right) = 1127 \ln\left(1 + \frac{f}{700}\right)$$

And for calculating frequency from Mel value:

$$f = M^{-1}(m) = 700\left(10^{\frac{m}{2595}} - 1\right) = 700\left(e^{\frac{m}{1127}} - 1\right)$$

For example, assume we want to estimate ten filter-banks that start from 300Hz to 8000Hz. By using above equation, we can calculate Mel value for each of them same as below:

$$M(300) = 2595 \log_{10}\left(1 + \frac{300}{700}\right) = 401.97$$

$$M(8000) = 2595 \log_{10}\left(1 + \frac{8000}{700}\right) = 2840.02$$

Now we need 12 points and distribute them between frequencies calculated in the last step, so first, we need to add ten more points and after that convert all Mel numbers to the corresponding frequency. Table 5.1 shows the Mel value and frequency for these twelve points.

Frequency(Hz)	**300**	**517.33**	**781.90**	**1103.98**	**1496.05**
Mel value	401.97	623.61	845.25	1066.89	1288.53
Frequency(Hz)	1973.34	2554.35	3261.64	4122.66	5170.80
Mel value	1510.17	1731.81	1953.45	2175.09	2396.74
Frequency(Hz)	6446.74	8000			
Mel value	2618.38	2840.02			

Table 5.1 shows frequency and corresponding Mel values for 10 filter-banks.

Now with these points, we can design out filters. Point one is the start point for the first filter-bank, point number two is its peak and point number three is its end point. In same order points number two, three and four are the start, peak and end points for second filter-bank. Figure 5.7 shows these ten filters.

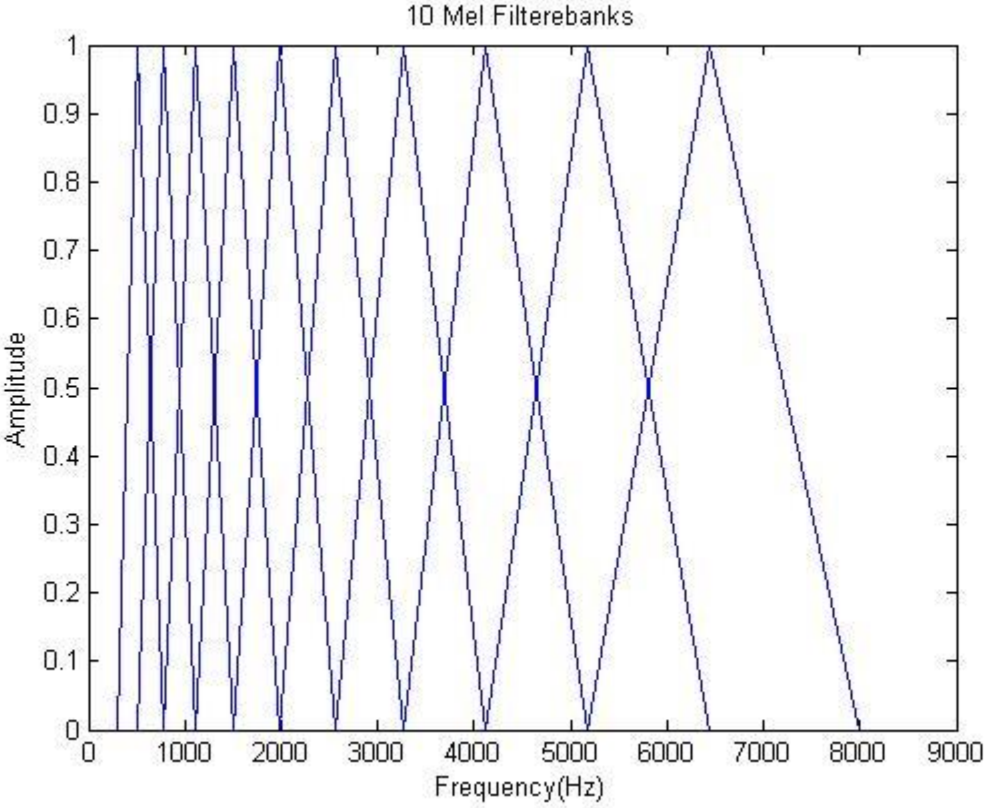


Figure 5.7 shows 10 Mel filter-banks start from 300Hz and ends at 8000Hz

Logarithmic scale

The next processing step is computing the logarithm of filter-bank energies. This process is similar to human hearing; we don't hear loudness on a linear scale. Experiments demonstrated that humans perceive loudness is on a logarithmic scale.

$$C_i = \log(e_i) \qquad\qquad 0 < i < N_f$$

Where e_i is energy of filter bank i calculated in last step and N_f is the number of Filter-banks.

Discrete Cosine Transform (DCT)

After calculating the logarithm of the energy in each filter-bank, we need to separate the speaker-dependent characters by computing the cepstral coefficients. The cepstrum can be defined as the spectrum of a spectrum. The cepstrum of a signal is calculated by

$$D_i = F^{-1}\{\log(F\{C_i\})\}$$

Where C_i is the input signal, F is the Fourier Transformation and D_i is cepstrum.

Because the logarithm of the signal was calculated in the previous step, we can drop estimate of the logarithm. Also instead of using the Fourier Transform, we can use the Discrete Cosine Transform (DCT). Another reason for using DCT is that because our filter-banks are all overlapping, the filter-bank energies are entirely correlated with each other. Therefore, DCT can de-correlate the energies of filter banks. However, only lower order DCT coefficients are kept for further processing, because higher DCT coefficients describe the fast changes in the filter-bank energies also contain speaker dependent. For example in the case of having 26 DCT coefficients, we are only interested to 12 of them.

The cepstral coefficients are computed by

$$C_k = \sum_{j=1}^{N_f} C_j \cos\left[\frac{k(2j-1)}{2N_f}\right] \quad k = 0, .., N_m < N_f$$

Where N_m is the number of chosen cepstral coefficients. Typical value for N_m is between twelve and twenty.

Derivatives

Also known as delta and acceleration coefficients. The MFCC feature vector describes only the power spectral envelope of a single frame, but it seems like to represent the dynamic nature of speech the first and second order derivatives of the cepstral coefficients needed too. The delta coefficients can calculate from the following formula:

$$d_t = \frac{\sum_{n=1}^{N} n(C_{t+n} - C_{t-n})}{2\sum_{n=1}^{N} n^2}$$

Where d_t is a derivative coefficient, from frame t and N can be one or two.

Also, some other sources defined derivatives same as below:

$$\Delta C_t = C_{t+1} - C_{t-1}$$

And

$$\Delta\Delta C_t = \Delta C_{t+1} - \Delta C_{t-1}$$

Where ΔC_t is a derivative coefficient, from frame t.

Alternative approach

Mel-frequency warping and the filter-bank can be performed smoothly in the frequency domain. However, there are a lot of disadvantage in this technique and different methods proposed to fix some of the issues. Some of the disadvantages listed below:

- The large dynamic range of the power spectrum.
- If the spectral resolution is not choosing properly, it can cause the lowest filters to contain very few spectral lines only, and the maximum of one of the filters may fall just in between two spectral lines.
- It is not clear how many filters are required and which filter shape is optimal.

S. Molau et al. (2001) [82] investigated an alternative method to estimate Mel frequency warped cepstral coefficients directly on the power spectrum and thereby avoided possible problems of the standard approach. Figure 5.8 compare traditional and integrated approach.

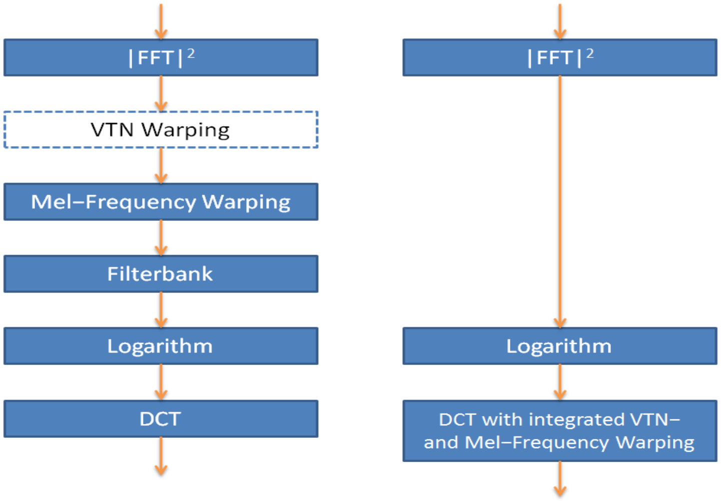


Figure 5.8: Comparison of the traditional MFCC (left) with the integrated approach (right) [82].

Multi-channel GMM Voice conversion algorithm

As we mentioned before Gaussian Mixture Model is the state of the art technique for speech processing. Now we are going to combine this technique with our filtering method to increase the quality and intelligibility of speech for Cochlear Implant users. We compare two different approaches both using GMM. One of them Uses MFCC features and the other uses HSM analyze [84]. In both systems, we used parallel recordings from two Male speakers (Source and Target) with 16 KHz sampling rate. All corpuses for training and transformation selected from IEEE database. All speech data from source and target go through Multi-Channel Process before they deliver for feature extraction. Figure 5.9 shows all steps for both methods in Training and transformation stage.

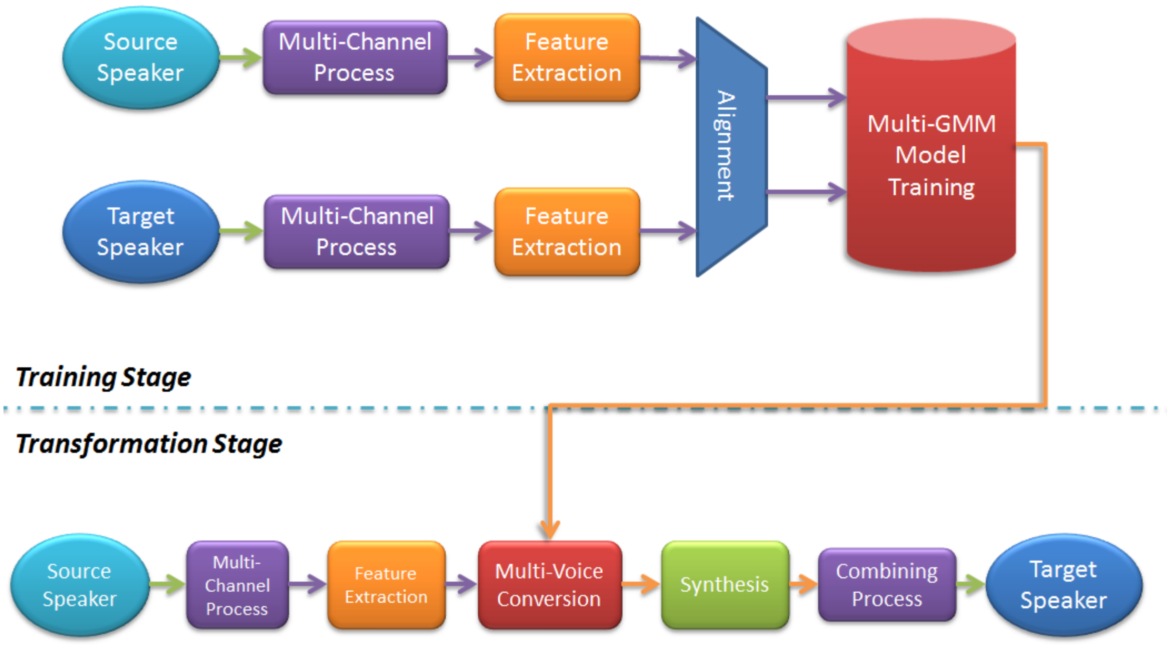


Figure 5.9 shows Training and transformation stage for Multi-GMM Voice conversion.

The multi-channel process includes phoneme segmentation, epoch detection, and multi-channels separation. The main difference between these two methods is in Feature extraction. In MFCC Feature extraction based system, feature vectors are produced. Also, F0 extract to create Mean and Standard Deviation. However in the other method features extracted using Harmonic/Stochastic Modeling (HSM) analyzes. In next step feature vectors are aligned for all channels and then deliver to GMM model training module. And the output of the last step in Training Stage is our Multi GMM conversion model. In Transformation Stage, Speech from source speaker goes through phoneme segmentation, epoch detection, and multi-channels separation processes and then delivered to feature extraction process depend on method. Extracted features present to Multi-Voice conversion module to convert to Target speaker by using Multi-GMM Model created in Training stage. Transformed data go through Synthesis process. This synthesis process is also different in these two methods. Finally, all channels merge to create speech of target speaker. Next, we describe each step of Training and Transformation stage.

Training Stage - Multi Channel process

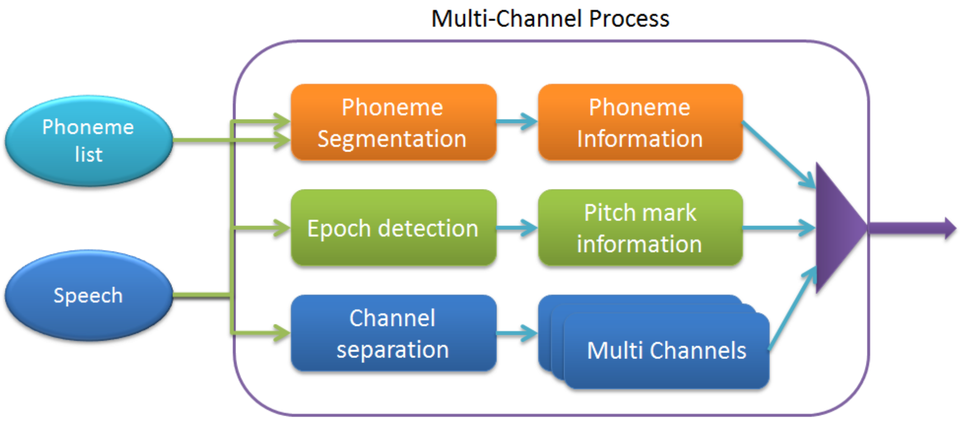


Figure 5.10 block diagram of submodules in Multi-Channel process

This Multi-Channel process is a combination of several parallel sub-processes such as Phoneme segmentation, Epoch detection, and Channel separation. Figure 5.10 shows submodules in this process.

Phoneme Segmentation

The aim of this section is creating a list of phonemes with respect to their location in speech sentences in the time domain. There is a lot of technique to do automatic phoneme segmentation. Main reason to use those techniques is Automatic Speech Recognition (ASR). However, there are different applications such as Voice conversion and speech enhancement that use this technique too. These methods usually rely on Signal processing techniques such as Vector Quantization (VQ), Cognitive techniques such as Artificial Neural Networks (ANN), and Statistical Techniques such as Hidden Markov Models (HMM).

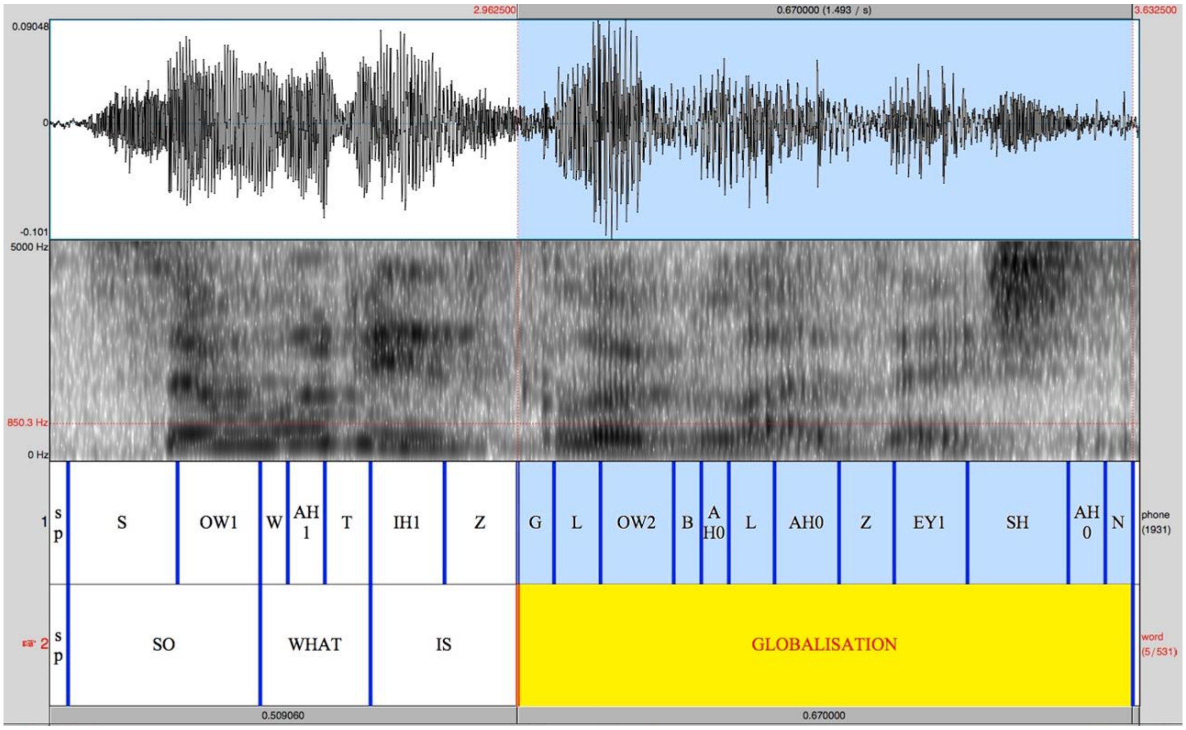


Figure 5.11 Phoneme Segmentation with Praat

In this research, we used the system based on statistic Hidden Markov Models (HMM), Also results verified with well-known software in speech analysis "Praat". The figure below shows an example of segmentation with Praat. Therefore, input for this sub-system is speech data with its corresponding text and result is phoneme information. These results contain the starting time, duration and stopping time for each phoneme. The table below demonstrates the result of phoneme segmentation for words “The birch”.

intervals	Start (Second)	End (Second)	Item
1	0.000000	0.130000	<p:>
2	0.130000	0.210000	D
3	0.210000	0.250000	@
4	0.250000	0.370000	b
5	0.370000	0.530000	3:
6	0.530000	0.700000	tS

Table 5.2 result of phoneme segmentation section

Epoch detection or Pitch marking

Pitch marking or epoch detection algorithms (EDA) attempt to recognize a single instant in each period that may serve as an “anchor” for future analysis. These positions are usually known as pitch marks or epochs or instants of glottal closure (IGC). Another definition of Epochs is a time instant of significant excitation of the vocal tract system during the production of speech. Accurate identification of epochs is not an easy task because of nonstationary nature of excitation source and human vocal tract system.

There is a lot of different technique for Epochs detection. However, most of them are based on the short-time Fourier transform (STFT). In this research, we used the method presented by K. Sri Rama Murty and B. Yegnanarayana(2008) [83] and compare its results with Praat. The figure below shows the epoch detection in Praat.

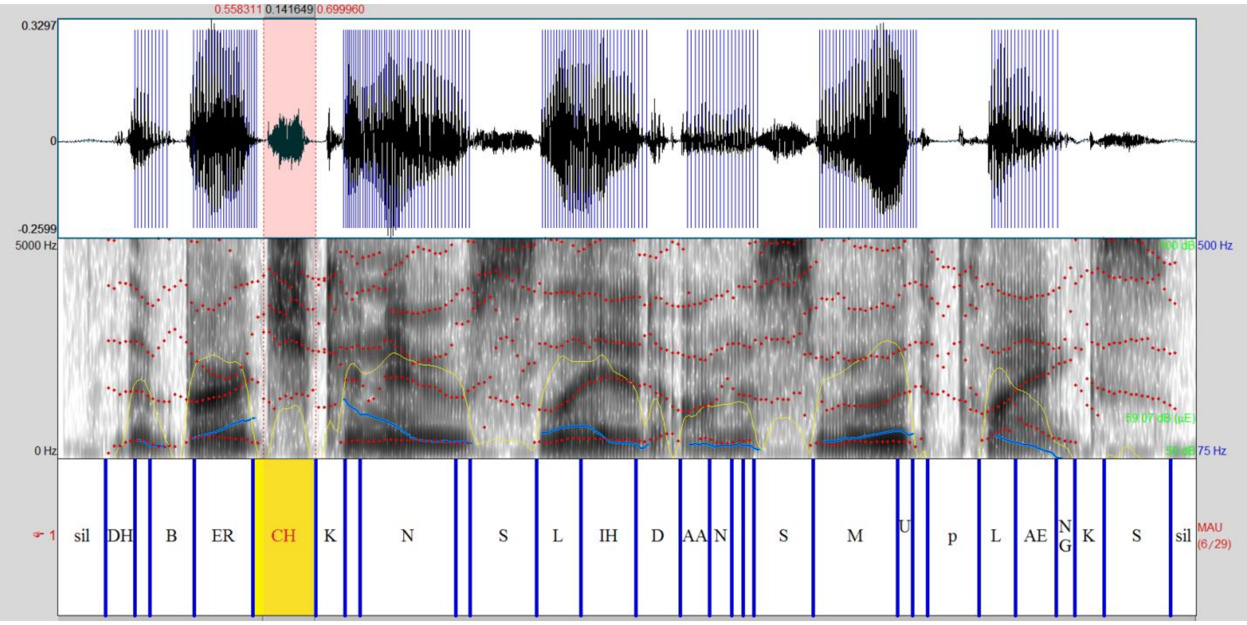


Figure 5.12 Epoch detection in Praat

Multi-Channel process

In this step, we using multiple Bandpass filter to divide speech signal to several different frequency channels. This method is based on Kokkinakis et al., 2007 [62] Single and Multiple Microphone Noise Reduction Strategies in Cochlear Implants. Essentially, it mimics the structure of Cochlear Implants. Frequency range is arbitrarily and selected between 350Hz to 5500Hz. Also, center frequencies distributed by log spacing. We eight different group of channels and compare their performance in the result section. The outcome of this step directly sends to Harmonic/Stochastic Modeling (HSM) analyzes [84] or MFCC feature extraction. The figure below shows the block diagram of this step.

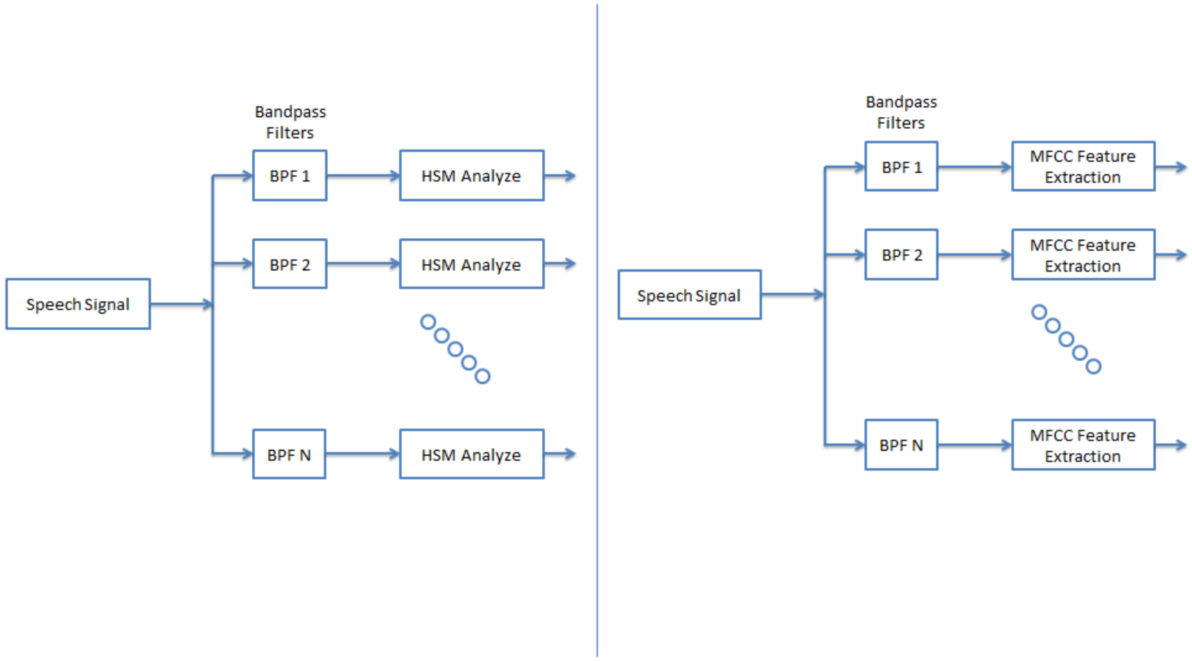


Figure 5.13 channeling Block Diagram

Feature Extraction

For this step, we have two different approaches, in one we are going to use MFCC, and we already explain the detail of this method before. However, the second one uses HSM analyzes [84], in continue we describe this process.

Harmonic/Stochastic Modeling (HSM)

Harmonic/Stochastic Modeling (HSM) presented by Y.Stylianou in 1996 [84]. Also, D. Erro and A. Moreno in 2007 [85] modified this model for voiced conversion system to analysis and reconstruction of the speech signal. In this technique, the speech signal is modeled by a harmonic component and a stochastic component. The harmonic component is a sum of sinusoids so for each frame amplitudes, frequencies and phases need to determine. The stochastic component is, modeled by a linear predictive coding (LPC) filter driven by white noise. The signal parameters are, measured at a fixed frame rate of fs/N frame/sec. Where fs is the sampling frequency and N is a time interval of 8 or 10ms. Pitch and voiced/unvoiced determination are, obtained at each frame. In voiced frames, the amplitudes and phases of the harmonics below 5 KHz used to describe the harmonic component.

Speech Alignment

In this section, we need to align all recorded parallel sentences for both speakers. In one of the methods (MFCC method), we used Dynamic Time Warping (DTW) to align the frames of the source and target speaker in the time domain. Then, we used ‘Dynamic Frequency Warping’ (DFW) to align each frame of speakers with the frequency axis.

On the other hand, We HMM-based forced recognition used to aligned parallel sentences in second method (HSM method) [85].

GMM model

In training section, we used the features of 120 sentences to create 8th order GMM model. Each channel trained independently and conversion model stored separately. Set of time-aligned LSF vectors of the source and target speakers used to determine the parameters of a mutual model of m Gaussian mixtures (α_i,μ_i,δ_i).[86] after training, the transformation function F(x) will generate using equation below:

$$F(x) = \sum_{i=1}^{m} p_i(x)\left[\mu_i^y + \delta_i^{yx}(\delta_i^{xx})^{-1}(x - \mu_i^x)\right]$$

$$p_i(x) = \frac{\alpha_i N(x,\mu_i^x,\delta_i^{xx})}{\sum_{j=1}^{m} a_j N\left(x,\mu_j^x,\delta_j^{xx}\right)} \quad \mu_i = \begin{bmatrix}\mu_i^x \\ \mu_i^y\end{bmatrix} \quad \delta_i = \begin{bmatrix}\delta_i^{xx} & \delta_i^{xy} \\ \delta_i^{yx} & \delta_i^{yy}\end{bmatrix}$$

Where $p_i(x)$ is the probability that a LSF vector x exists in the i^{th} Gaussian component of the GMM. After transformation, the signal reconstructed by adding and overlapping all frames then all channels combined together to create new speech.

Multi-channel ANN Voice conversion algorithm

In this section, we review the effect of multi-channel technique on common voice conversion using the artificial neural network (ANN) method. Figure 5.14 shows blog diagram of the multi-channel voice conversion using the artificial neural network.

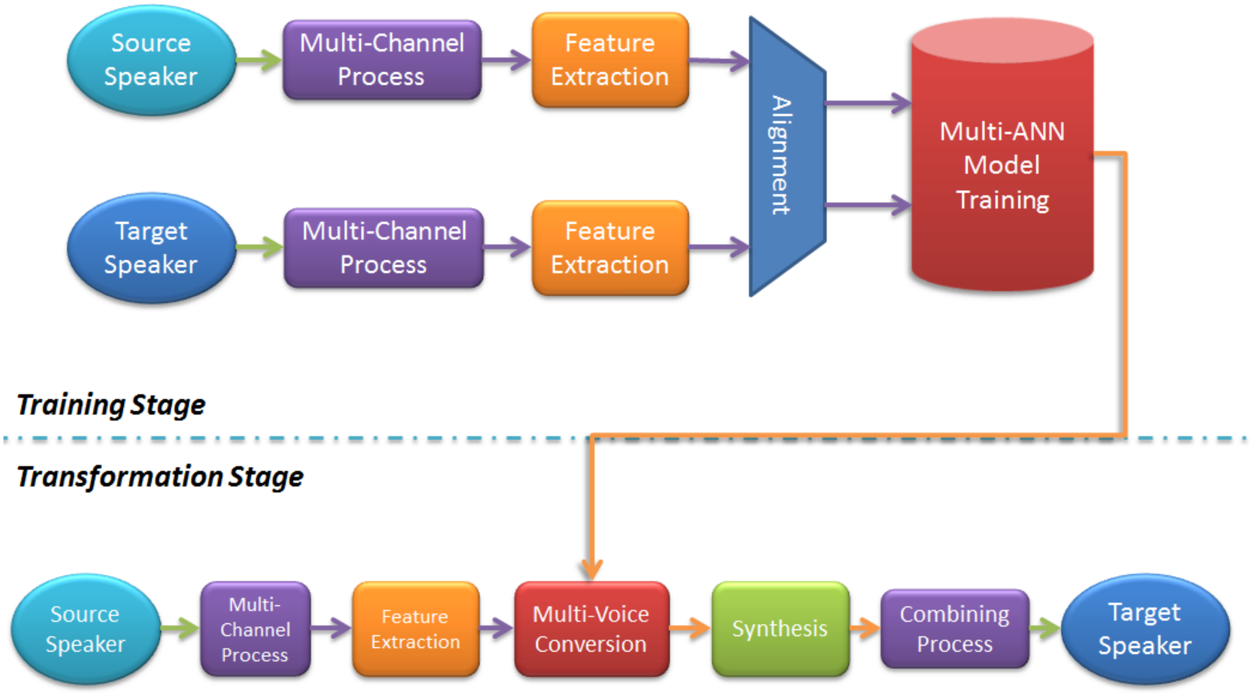


Figure 5.14: multi-channel VC with artificial neural network

In this method, we extract the features with a different number of Mel filter bands, and number 70 indicates the best performance. For this modeling, Levenberg-Marquardt training algorithm selected and performance plot is shown in figure 5.15. The best performance will happen when the validation error stops increasing.

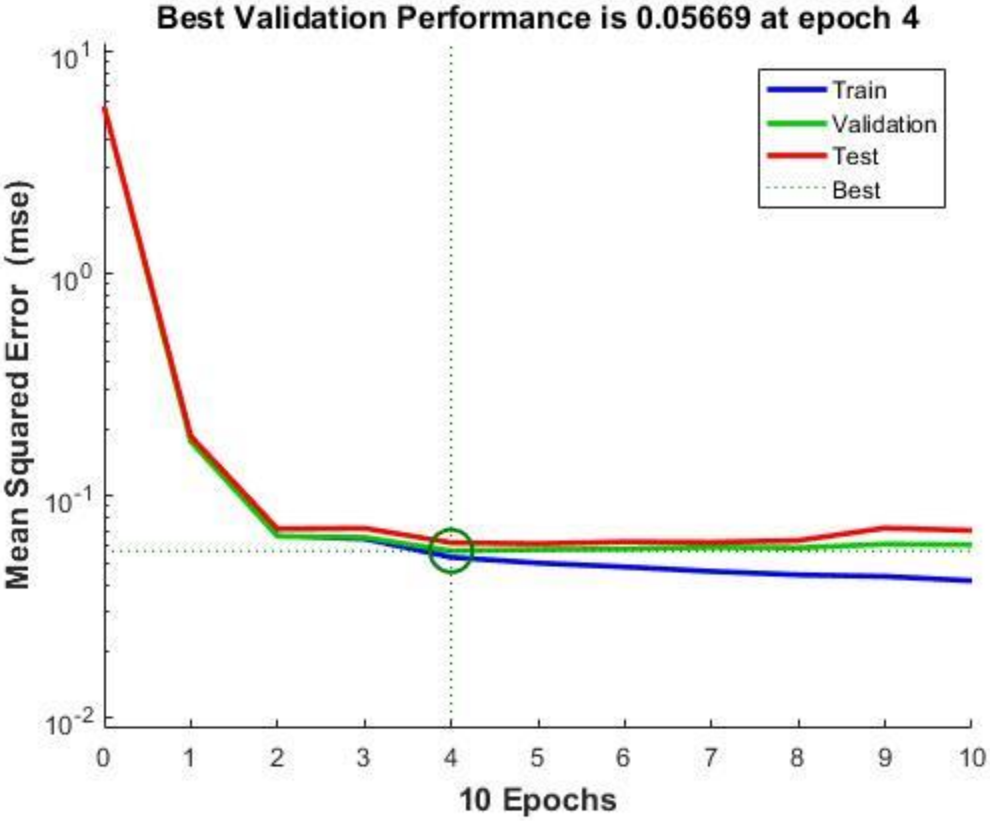


Figure 5.15: Best validation performance is 0.05669 at epoch 4, Blue line shows the decreasing error on training data, Green shows error on Validation data and red shows the error on test data.

Chapter Six - Results and Conclusion

In this section, we compare the result of multichannel voice conversion with traditional GMM and ANN methods. For evaluating the performance of presented techniques, we adopt mean opinion scores (MOS). MOS belongs to the family of test known as Perceptual Evaluation of Speech Quality (PESQ). PESQ contained methods for automated assessment of the speech quality as experienced by a user of a telephony system.

Figure 6.1 compares the quality of traditional voice conversions such as GMM and ANN with Multichannel enhanced voice conversion present in this study. For every method, 120 sentences of each speaker used to train the corresponding model and another 200 sentences used for voice conversion.

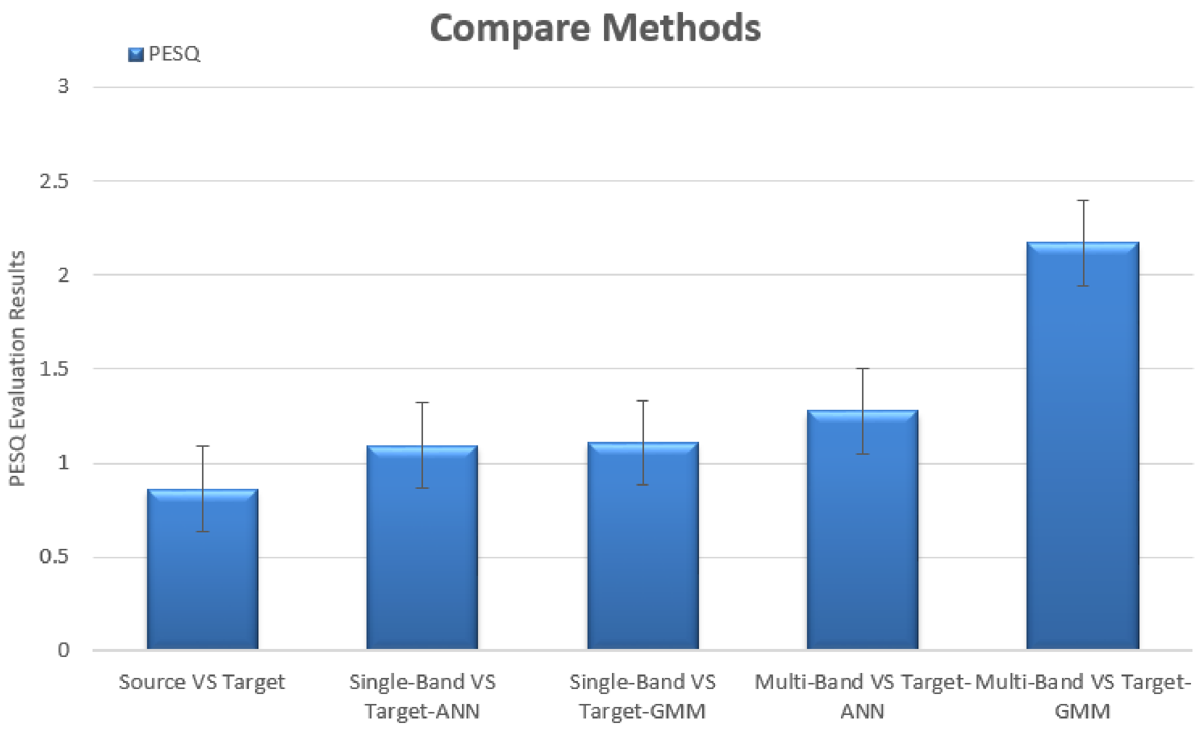


Figure 6.1: compares the quality of Single band and multiband with GMM and ANN methods.

In next figure (6.2) we examine an effect of a different number of channels to the quality of speech. There is some consideration to select a number of channels. The important one can be a trade-off between number of channels and process time. By adding more channel quality may increase as well as process load. Also by increasing number of filter bank we reduce the range of frequency in each channel and it can cause to losing the structure of speech. The result below shows selecting 16 channels increases the quality of speech in compare with 2,4,6,8,10,12,14 channels. In this study, 120 sentences used to train the conversion model and 200 sentences used for quality comparison.

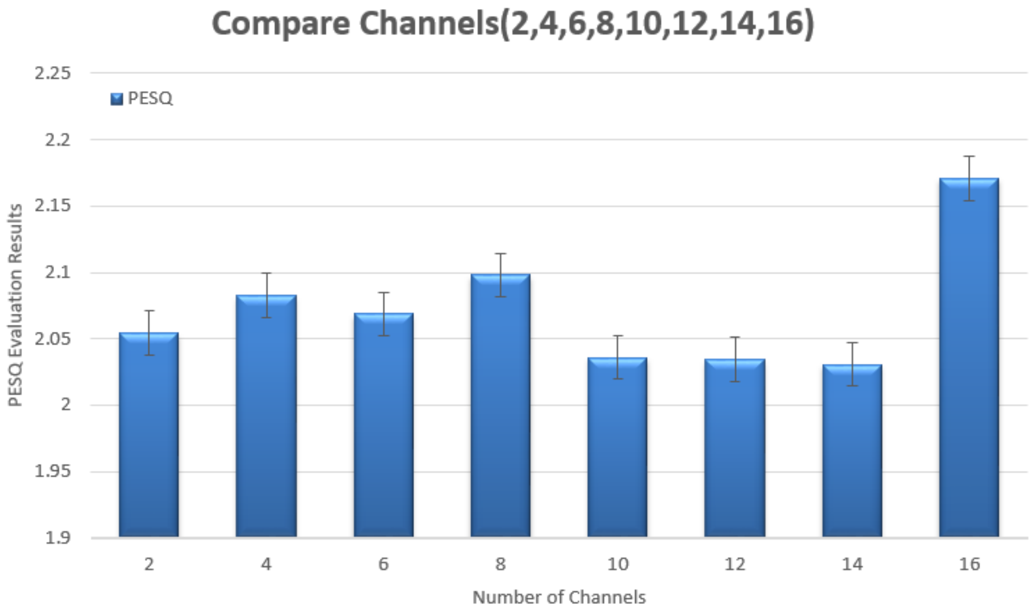


Figure 6.2: compares the quality of a various number of channels.

This technique can quickly adopt by cochlear implant devices because of their similarity in signal processing.

References

[1]- A. Boothroyd, "Profound deafness," in Cochlear Implants: Audiological Foundations (R. Tyler, ed.), pp. 1-34, Singular Publishing Group, Inc, 1993.

[2]- R. Shannon, F. Zeng, V. Kamath, J. Wygonski, and M. Ekelid, "Speech recognition with primarily temporal cues," Science, vol. 270, pp. 303-304, 1995.

[3]- M. Dorman, P. Loizou, and D. Rainey, "Speech intelligibility as a function of the number of channels of stimulation for signal processors using sine-wave and noise-band outputs," Journal of the Acoustical Society of America, vol. 102, pp. 2403-2411, 1997.

[4]- Qin, M. K., and Oxenham, A. J.(2003). “Effects of simulated cochlear-implant processing on speech reception in fluctuating maskers,” J. Acoust.Soc. Am.114, 446–454.

[5]- Fu, Q., and Nogaki, G.(2004). “Noise susceptibility of cochlear implant users: The role of spectral resolution and smearing,” J. Assoc. Res. Oto-laryngol.6, 19–27

[6]- Firszt, J., Holden, L., Skinner, M., Tobey, E., Peterson, A., Gaggl,W., . . . Wackym, P. A. (2004). Recognition of speech presented at soft to loud levels by adult cochlear implant recipients of three cochlear implant systems. Ear and Hearing, 25, 375-387.

[7]- Nilsson, M., Soli, S., and Sullivan, J.(1994). “Development of the hearing in noise test for the measurement of speech reception thresholds in quiet and in noise,” J. Acoust. Soc. Am.95, 1085–1099.

[8]- Susannah V. Levi, Stephen J. Winters, David B. Pisoni.(2011) “Effects of cross-language voice training on speech perception:Whose familiar voices are more intelligible?”, JASA Vol. 130, No. 6.

[9]- Magnuson, J. S., Yamada, R. A., and Nusbaum, H. C. (1995). “The effects of familiarity with a voice on speech perception,” Proceedings of the 1995 Spring Meeting of the Acoustical Society of Japan, pp. 391–392.

[10]- Nygaard, L. C., and Pisoni, D. B. (1998). “Talker-specific learning in speech perception,” Percept. Psychophys. 60(3), 355–376.

[11]- Sommers, M. S., Nygaard, L. C., and Pisoni, D. B. (1994). “Stimulus variability and spoken word recognition. I. Effects of variability in speaking rate and overall amplitude,” J. Acoust. Soc. Am. 96(3), 1314–1324.

[12]- Hazan, V., and Markham, D. (2004). “Acoustic-phonetic correlates of talker intelligibility for adults and children,” J. Acoust. Soc. Am. 116, 3108–3118.

[13]- Ladefoged P., and Broadbent D. E. (1957). “Information conveyed by vowels,” J. Acoust. Soc. Am. 29, 98–104. 10.1121/1.1908694

[14]- Ladefoged P. (1978). “Expectation affects identification by listening,” Lang. Speech 21(4), 373–374.

[15]- Johnson K. (1990). “The role of perceived speaker identity in F0 normalization of vowels,” J. Acoust. Soc. Am. 88 , 642–654.

[16]- Johnson K., Strand E. A., and D’Imperio M. (1999). “Auditory-visual integration of talker gender in vowel perception,” J. Phonetics 27, 359–384. 10.1006/jpho.1999.0100

[17]- Allen J. S., and Miller J. L. (2004). “Listener sensitivity to individual talker differences in voice-onset-time,” J. Acoust. Soc. Am. 115, 3171–3183. 10.1121/1.1701898

[18]- Eisner F., and McQueen J. M. (2005). “The specificity of perceptual learning in speech processing,” Percept. Psychophys. 67, 224–238. 10.3758/BF03206487

[19]- Kraljic T., and Samuel A. G. (2005). “Perceptual learning for speech: is there a return to normal?” Cogn. Psychol. 51, 141–178. 10.1016/j.cogpsych.2005.05.001

[20]- Kraljic T., and Samuel A. G. (2006). “Generalization in perceptual learning for speech,” Psychon. Bull. Rev. 13(2), 262–268. 10.3758/BF03193841

[21]- Kraljic T., and Samuel A. G. (2007). “Perceptual adjustments to multiple talkers,” J. Mem. Lang. 56, 1–15. 10.1016/j.jml.2006.07.010

[22]- Kraljic T., Brennan S. E., and Samuel A. G. (2008). “Accommodating variation: Dialects, idiolects, and speech processing,” Cognition 107, 54–81. 10.1016/j.cognition.2007.07.013

[23]- Green, T., Katiri, S., Faulkner, A., and Rosen, S. (2007). “Talker intelligibility differences in cochlear implant listeners,” J. Acoust. Soc. Am. 121, EL223–EL229.

[24]- Machado A. and Queiroz M. (2010). “Techniques for Crosslingual Voice Conversion” Multimedia (ISM), 2010 IEEE International Symposium

[25]- D. G. Childers, B. Yegnanarayana, and K. Wu, “Voiceconversion: Factors responsible for quality,” ICASSP,pp. 748–751, 1985.

[26]- K. Shikano, K. Lee, and R. Reddy, “Speaker adaptation through vector quantization,” in ICASSP, vol. 11, 1986.

[27]- H. Valbret, E. Moulines, and J. Tubach, “Voice transformation using PSOLA technique,” in 2nd ECSCT, 1992.

[28]- Y. Stylianou, “Continuous probabilistic transform forvoice conversion,” IEEE TSAP, no. 6, pp. 131–142,1998.

[29]- D. Rentzos, S. Vaseghi, E. Turajlic, Q. Yan, and C. Ho,“Transformation of speaker characteristics for voice conversion,” in IEEE WASRU, pp. 706–711, 2003.

[30]- H. Ye and S. Young, “Quality-enhanced voice morphingusing maximum likelihood transformations,” IEEE TASLP, vol. 14, no. 4, pp. 1301–1312, 2006.

[31]- K. Rao and B. Yegnanarayana, “Voice conversion by prosody and vocal tract modification,” in 9th ICIT, pp. 111–116, 2006.

[32]- M. Zhang, J. Tao, J. Nurminen, J. Tian, and X. Wang,“Phoneme cluster based state mapping for textindependent voice conversion,” in ICASSP, pp. 4281– 4284, 2009.

[33]- A. F. Machado and M. Queiroz. “Voice conversion: A critical survey”. In SMC, 2010.

[34]- Gidon Eshel : “The Yule Walker Equations for the AR Coefficients.” University of South Carolina.

[35]- D. Srinivas , E. V. Raghavendra , B. Yegnanarayana , A. W. Black and K. Prahallad, "Voice conversion using artificial neural networks", Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., pp. 3893-3896, 2009

[36]- Abe, M., Nakamura, S., Shikano, K. & Kuwabara, H. (1988). Voice conversion through vector quantization, Proc. of ICASSP, pp. 655–658.

[37]- J. Gandour, “The perception of tone,” in Tone: A Linguistic Survey, edited by V. Fromkin (Academic,New York, 1978), pp. 41–76.

[38]- Y. R. Chao, A Grammar of Spoken Chinese (University of California Press, Berkeley, CA, 1968).

[39]- D. H. Whale and Y. Xu, “Information for Mandarin tones in the amplitude contour and in brief segments,” Phonetica 49, 25–47 (1992).

[40]- Q.-J. Fu, F.-G. Zeng, R. V. Shannon, and S. D. Soli, “Importance of tonal envelope cues in Chinese speech recognition,” J. Acoust. Soc. Am. 104, 505–510 (1998).

[41]- Q.-J. Fu, and F.-G. Zeng, “Identification of temporal envelope cues in Chinese tone recognition,” Asian Pac. J. Speech, Lang., Hear. 5, 45–57 (2000).

[42]- X. Luo and Q.-J. Fu, “Enhancing Chinese tone recognition by manipulating amplitude envelope: Implications for cochlear implants,” J. Acoust. Soc. Am. 116, 3659–3667 (2004).

[43]- X. Luo and Q.-J. Fu, “Contribution of low-frequency acoustic information to Chinese speech recognition in cochlear implant simulations,” J. Acoust. Soc. Am. 120, 2260–2266 (2006).

[44]- R. V. Shannon, “Temporal modulation transfer functions in patients with cochlear implants,” J. Acoust. Soc. Am. 91, 2156–2164 (1992).

[45]- A. Oxenham, “Pitch perception and auditory steam segregation: implications for hearing loss and cochlear implants,” Trend Amp. 12, 316–331 (2008).

[46]- ANSI S3.6-2010: Specification for Audiometers (Acoustical Society of America, New York, 2010).

[47]- B. Glasberg and B. Moore, “Derivation of auditory filter shapes from notched-noise data,” Hear. Res. 47, 103–138 (1990).

[48]- K. J. Van Engen, “Similarity and familiarity: second language sentence recognition in first-and secondlanguage multi-talker babble,” Speech Commun. 52, 943–953 (2010).

[49]- G. Studebaker, “A ‘rationalized’ arcsine transform,” J. Speech, Lang. Hear. Res. 42, 56–64 (1985).

[50]- N. Zhou and L. Xu, “Lexical tone recognition with spectrally mismatched envelopes,” Hear. Res. 246,36–43 (2008).

[51]- Buss, E., Pillsbury, H. C., Buchman, C. A., & Pillsbury, C. H., Clark, M. S., Haynes, D. S., . . . Barco, A. L. Multicenter “U.S. bilateral MED-EL cochlear implantation study: Speech perception over the first year of use.” Ear and Hearing, 29, 20–32. (2008)

[52]- Litovsky, R., Johnstone, P. M., Godar, S., Agrawal, S., Parkinson, A., Peters, R., & Lake, J. “Bilateral cochlear implants in children: Localization acuity measured with minimum audible angle.” Ear and Hearing, 27, 43–59. (2006).

[53]- Caselli, M. C., Rinaldi, P., Varuzza, C., Giuliani, A., & Burdo, S. “Cochlear implant in the second year of life: Lexical and grammatical outcomes.” Journal of Speech, Language, and Hearing Research, 55, 382–394. (2012).

[54]- Brown, K. D., & Balkany, T. J. “Benefits of bilateral cochlear implantation: A review.” Current Opinion in Otolaryngology & Head and Neck Surgery, 15, 315–318. (2007).

[55]- Gantz, B. J.,Tyler,T. S., Rubinstein, J.T.,&Wolaver,A., Lowder,M., Abbas, P., . . . Preece, J. P. “Binaural cochlear implants placed during the same operation.” Otology & Neurotology, 23, 169–180. (2002).

[56]- Schleich, P., Nopp, P., & D’Haese, P. “Head shadow, squelch, and summation effects in bilateral users of the MED-EL COMBI 40/40+ cochlear implant.” Ear and Hearing, 25, 197–204. (2004).

[57]- Bronkhorst, A. W., & Plomp, R. “Binaural speech intelligibility in noise for hearing-impaired listeners.” The Journal of the Acoustical Society of America, 86, 1374–1383. (1989).

[58]- Litovsky, R., Parkinson, A., Arcaroli, J., & Sammeth, C. “Simultaneous bilateral cochlear implantation in adults: A multicenter clinical study.” Ear and Hearing, 27, 714–731. (2006).

[59]- Chen, F., Wong, L. L. N., Tahmina, Q., Azimi, B., & Hu, Y. “The effects of binaural spectral resolution mismatch on Mandarin speech perception in simulated electric hearing.” The Journal of the Acoustical Society of America, 132, EL142–EL148 (2012).

[60]- Fu, Q.-J., & Shannon, R. V. “Effect of acoustic dynamic range on phoneme recognition in quiet and noise by cochlear implant users.’ The Journal of the Acoustical Society of America, 106, L65–L70. (1999).

[61]- G. Kim, Y. Lu, Y. Hu and P. C. Loizou, (2009). “An algorithm that improves speech intelligibility in noise for normal-hearing listeners.” Journal of the Acoustical Society of America, 126,1486-1494.

[62]- K. Kokkinakis, B. Azimi, Y. Hu and D. R. Friedland, (2012) “Single and Multiple Microphone Noise Reduction Strategies in Cochlear Implants” SAGE 16(2) 102–116

[63]- Boll, S. F. (1979). “Suppression of acoustic noise in speech using spectral subtraction.” IEEE Transactions on Acoustics, Speech and Signal Processing, 2, 113-120.

[64]- Ephraim, Y., & Malah, D. (1984). “Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator.” IEEE Transactions on Acoustics, Speech and Signal Processing, 32, 1109-1121.

[65]- Ephraim, Y., & VanTrees, H. L. (1995). “A signal subspace approach for speech enhancement.” IEEE Transaction on Speech and Audio Processing, 3, 251-266.

[66]- Lim, J. S. (1978). “Evaluation of a correlation subtraction method for enhancing speech degraded by additive white noise.” IEEE Transactions on Acoustics, Speech and Signal Processing, 26,471-472.

[67]- Lim, J. S., & Oppenheim, A. V. (1979). “Enhancement and bandwidth compression of noisy speech.” Proceedings of the IEEE, 67, 1586-1604.

[68]- Hochberg, I., Boothroyd, A., Weiss, M., & Hellman, S. (1992). “Effects of noise and noise suppression on speech perception for cochlear implant users.” Ear and Hearing, 13, 263-271.

[69]- Weiss, M. (1993). “Effects of noise and noise reduction processing on the operation of the Nucleus 22 cochlear implant processor.” Journal of Rehabilitation Research and Development, 30,117-128.

[70]- Yang, L.-P., & Fu, Q.-J. (2005). “Spectral subtraction-based speech enhancement for cochlear implant patients in background noise.” Journal of the Acoustical Society of America, 117, 1001-1004.

[71]- Loizou, P. C., Lobo, A., & Hu, Y. (2005). “Subspace algorithms for noise reduction in cochlear implants.” Journal of the Acoustical Society of America, 118, 2791-2793.

[72]- Hu, Y., & Loizou, P. C. (2002). “A subspace approach for enhancing speech corrupted with colored noise.” IEEE Signal Processing Letters, 9, 204-206.

[73]- Nilsson, M., Soli, S. D., & Sullivan, J. (1994). “Development of hearing in noise test for the measurement of speech reception thresholds in quiet and in noise”, Journal of the Acoustical Society of America, 95, 1085-1099.

[74]- Hu, Y., Loizou, P. C., Li, N., & Kasturi, K. (2007). “Use of a sigmoidal- shaped function for noise attenuation in cochlear implants.” Journal of the Acoustical Society of America, 122, 128-134.

[75]- Hu, Y., & Loizou, P. C. (2008). “A new sound coding strategy for suppressing noise in cochlear implants.” Journal of the Acoustical Society of America, 124, 498-509.

[76]- Hu, Y., & Loizou, P. C. (2010). “Environment-specific noise suppression for improved speech intelligibility by cochlear implant users”. Journal of the Acoustical Society of America, 127,3689-3695.

[77]- Kollmeier, B., & Koch, R. (1994). “Speech enhancement based on physiological and psychoacoustical models of modulation perception and binaural interaction.” Journal of the Acoustical Society of America, 95, 1593-1602.

[78]- Tchorz, J., & Kollmeier, B. (2003). “SNR estimation based on amplitude modulation analysis with applications to noise suppression.” IEEE Transactions on Speech and Audio Processing, 11, 184-192.

[79]- J. S. Bridle and M. D. Brown (1974), “An Experimental Automatic Word-Recognition System”, JSRU Report No. 1003, Joint Speech Research Unit, Ruislip, England.

[80]- P. Mermelstein, “Distance Measures for Speech Recognition – Psychological and Instrumental”, Pattern Recognition and Artificial Intelligence, pp. 374–388, 1976.

[81]- Davis, S. Mermelstein, P. (1980) “Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences.” In IEEE Transactions on Acoustics, Speech, and Signal Processing, Vol. 28 No. 4, pp. 357-366

[82]- S. Molau, M. Pitz, R. Schluter and H. Ney, “Computing mel-frequency cepstral coefficients on the power spectrum”, Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 1, pp. 73-76

[83]- K. Sri Rama Murty and B. Yegnanarayana, “Epoch Extraction From Speech Signals”, IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, VOL. 16, NO. 8, NOVEMBER 2008

[84]- Y.Stylianou, “Harmonic plus Noise Models for Speech, combined with Statistical Methods, for Speech and Speaker Modification”, PhD thesis, École Nationale Supérieure des Télécommunications, 1996.

[85]- D. Erro and A. Moreno, “Weighted frequency warping for voice conversion”, Proc. Interspeech, pp. 1965-1968, 2007

[86]- D. Erro , A. Moreno and A. Bonafonte, "Voice conversion based on weighted frequency warping", IEEE Trans. Audio, Speech, Lang. Process., vol. 18, no. 5, pp. 922-931, 2010

CURRICULUM VITAE

Behnam Azimi

Education

PhD, Electronic Engineering, EE Dep., University of Wisconsin- Milwaukee 2016.
M.S, Electronic Engineering, EE Dep., University of Wisconsin- Milwaukee 2011
B.S, Electronic Engineering, EE Dep., Azad Univ., Central, Tehran, Iran 2002

Research Interest

- CubeSat
- Signal processing
- Low power electronics
- Control, Robotic & intelligent Systems and
- Bioelectronics
- Embedded System

Research Experiences

- Noise reduction for cochlear implants.
- Bilateral Training App for cochlear implants user.
- Sound Localization Training App for cochlear implants user on Android and iOS devices.
- Parallel processing using GPU to reduce processing time of training for noise reduction.
- Designing low price hearing aid.
- At Home Robots
- Middle size Robo Soccer

Professional Experience

- Pathway at NASA GSFC Jan 2015 - Present

- Summer Internship at NASA GSFC July 2014 - Aug 2014

- Teaching Assistant at University of Wisconsin-Milwaukee Sep 2013 - Dec 2014
  Gave weekly lectures to students enrolled in Introduction to Microprocessors lecture and LAB. In this course I teach microprocessor Structure, embedded design and assembly language.

- Project Assistant at University of Wisconsin-Milwaukee Sep 2010 - Aug 2013
  Designed prototype portable audiometers for clinical use and several APP such Diotic and Dichotic training for cochlear implant patients.

Computer Skills

- Matlab , C ,C++, Vb.net, C# , ASP.net , Xml service, JAVA, SQL server, PYTHON
- ITOS
- Labview
- Android Programming, Pocket Pc Programming with .net and iOS
- Micro controller programming (ARM,AVR, PIC, ST)
- FPGA Xilinx
- PLC S5, S7
- Protel 99Se, DXP, Dip Trace, PAD(Mentor Graphics)
- GitLab

Volunteer activities

- Graduate School Senator at University of Wisconsin Milwaukee 2014
- Team Leader of IAUT Robotic Team, 1999-2004.
- Team Leader of the Satrap **Robocup** Team, 2004-2008.
- Member of the Iranian National Youth Society, 2004-2005.
- Elected President of the Iranian National Scientific Student's Organization of Electrical Engineering, 2002-2004.
- Technical Committee member, Student's Conference of Electrical Engineering, 2001

Journal papers

C Liu, **B Azimi**, M Bhandary, Y Hu (2014), "Contribution of low-frequency harmonics to Mandarin Chinese tone identification in quiet and six-talker babble background" The Journal of the Acoustical Society of America 135 (1), 428-438

F Chen, LLN Wong, J Qiu, Y Liu, **B Azimi**, Y Hu (2013)," The Contribution of Matched Envelope Dynamic Range to the Binaural Benefits in Simulated Bilateral Electric Hearing" Journal of Speech, Language, and Hearing Research 56 (4), 1166-1174

C. Liu, **B. Azimi**, Q. Tahmina and Y. Hu (2012), "Effects of low harmonics on tone identification in natural and vocoded speech", The Journal of the Acoustical Society of America 132 (5), EL378-EL384

K Kokkinakis, **B Azimi**, Y Hu, DR Friedland (2012), "Single and multiple microphone noise reduction strategies in cochlear implants" Trends in amplification 16 (2), 102-116

F. Chen, L. Wong, Q. Tahmina, **B. Azimi** and Y. Hu (2012), "The effects of spectral resolution mismatch on Mandarin speech perception in simulated electric hearing", The Journal of the Acoustical Society of America, 132(2), EL142-EL148.

Conference papers

V. Thongpriwan, ***B. Azimi***, M. Bhandary, Y. Hu, “Transforming evidence into better prevention of adolescent dating violence: The BetterMe App”, The Midwest Nursing Research Society annual conference, 2012.
Q. Tahmina, ***B. Azimi***, M. Bhandary, Y. Hu, R. Utianski and J. Liss,“The effect of visual information on speech perception in noise by electroacoustic hearing”, the 164th Meeting of the Acoustical Society of America, October 25 -- 29, Kansas City, Missouri, 2012.
E. Schafer, ***B. Azimi***, and Y. Hu, “An android auditory training app for sequential bilateral cochlear implant users”, 2011 Conference on Implantable Auditory Prostheses, July 24 - 29, Pacific Grove, California, 2011.