國立交通大學

電信工程學系

博士論文

中文聽障語者的強健性辨認研究

Robust Distributed Recognition of Hearing-impaired Mandarin Speech over Wireless Networks

研究生：李承龍

指導教授：張文輝 博士

中華民國 九十八 年 二 月

中文聽障語者的強健性辨認研究

研究生：李承龍　　　　　　　　指導教授：張文輝 博士

摘 要

本篇論文旨在分散式語音辨認架構下，針對語者變異與傳輸錯誤的影響分別提供其強健性處理。語者變異所造成的效能失真，其影響源自於語音辨認器在模型訓練與實際測試兩個階段的語者不匹配。針對聽障中文語者的發聲，我們提出語音轉換機制使其能匹配辨認模型所蘊含的語音特性。此轉換系統的設計乃是基於中文語音的特性，考慮聲母-韻母組合的音節結構及聲調變化，分別針對頻譜與韻律兩層次的特徵參數進行轉換，而特徵參數的擷取則是依據正弦語音模型。頻譜轉換需考慮不同音類在聲學特性的明顯差異，並據以針對聲母及韻母所屬的次音節參數分別設計其最佳化轉換函數。此外，構音速度的調變亦針對不同類型的次音節，設計其線性或非線性的轉換機制。至於聲調的調變，則考慮中文四聲變化的結構，先藉由正交轉換分析基頻變化曲線的特徵參數，再利用向量對應機制估算最佳的基頻轉換曲線。系統模擬證實，語音轉換機制可有效改善聽障者語音的清晰度，進而有效提升其語音辨認的正確率。分散式辨認系統的另一研究重點是語音特徵參數於無線傳輸過程中，將遭遇叢發性通道錯誤而導致其辨認效能衰減。有鑑於此，我們設計一錯誤隱匿解碼機制，其關鍵在於有效整合訊源編碼輸出的殘餘冗息以及通道錯誤的相關特性。在辨認特徵參數的冗息分析中，編碼輸出的量化索引序列仍存在大量的相關特性，而行動通訊的叢發性錯誤則適於以馬可夫模型來模擬。我們結合這兩種訊息，再依據最大後驗機率準則設計一合併訊源通道解碼演算法。實驗結果證實訊源通道解碼器在無線傳輸環境能有效提升其錯誤隱匿效能。

Robust Distributed Recognition of Hearing-impaired Mandarin Speech over Wireless Networks

Student: Cheng-Lung Lee   Advisor: Dr. Wen-Whei Chang

Department of Communication Engineering, National Chiao Tung University
Hsinchu, Taiwan, Republic of China

ABSTRACT

This study focuses on the robustness of distributed speech recognition (DSR) systems against the inter-speaker and channel variabilities. In the first part, we develop joint source-channel decoding algorithms with increased robustness against channel errors in mobile DSR applications. An MAP symbol decoding algorithm which exploits the combined a priori information of source and channel is proposed. This is used in conjunction with a modified BCJR algorithm for decoding convolutional channel codes based on sectionalized code trellises. Performance is further enhanced by the use of the Gilbert channel model that more closely characterizes the statistical dependencies between channel bit errors. In the second part, we develop voice conversion approaches based on the feature transformation to perform speaker adaptation for hearing-impaired Mandarin speech. The basic strategy is the combined use of spectral and prosodic conversions to modify the hearing-impaired Mandarin speech. The analysis-synthesis system is based on a sinusoidal representation of the speech production mechanism. By taking advantage of the tone structure in Mandarin speech, pitch contours are orthogonally transformed and applied within the sinusoidal framework to perform pitch modification. Also proposed is a time-scale modification algorithm that finds accurate alignment between hearing-impaired and normal utterances. Using the alignments, spectral conversion is performed on subsyllabic acoustic units by a continuous probabilistic transform based on a Gaussian mixture model.

Acknowledgements

這博士班研讀的過程中，首先要感謝我的指導老師張文輝教授悉心的引導，更感染自老師對於教學及研究的熱誠與認真，故期許自己對於電信相關領域能有多一分的貢獻。此外，個人非常感謝交通大學所提供優良的學術環境，讓學生在科技與人文的素養均能有所成長。再者，特別感謝新竹教育大學江源泉教授對於研究所提供的協助與指導。在語音通訊實驗室裡的研究生活中，有學長蔡偉和、許亨仰與同學何宗仁、吳政麟、蔡若望、楊雅茹、李世耀、李維晟、葉志杰、蔡淑羚、吳俊鋒、傅泰魁、曹正宏、林宜德、許忠安、蔡知鑑、何依信等相互提攜與砥礪，由於大家的相伴，即使在最苦悶的日子裡依然能開心的向目標邁進，研究也方能如此順利。最後，感謝父母親以及家人在這段時間裡的支持跟鼓勵，若個人能對研究與社會能有絲毫的貢獻，願將一切的成就與榮耀與你們一同分享。

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Recent progress in automatic speech recognition (ASR) technology has enabled the development of more sophisticated spoken language interface applications. These applications make use of speech to replace or complete an interface for communicating with a machine, e.g. for accessing a service or controlling a functionality of an equipment. Moreover, combined with use of distributed client-server operation mode, ASR has become a common service for mobile communications and computing devices. The fantastic growth of the Internet has also created a demand for easy ways of accessing and retrieving all the available information and services. Through the use of distributed ASR capability, people can access information anytime and anywhere [1].

The block diagram of the speech recognition system is shown in Figure 1.1. The basic strategy begins with an extraction of parameter set from the speech signal. These parameters describe the speech by their variation over time and hence can be used to build up a pattern that characterizes the speech. In the training phase statistic models are estimated for every phoneme used in the target application, which are then concatenated to form the word-based models. In the recognition phase testing speech signals are analyzed to compute the acoustic parameters and the acoustical decoding block is used to search for the closest speech templates whose corresponding models are

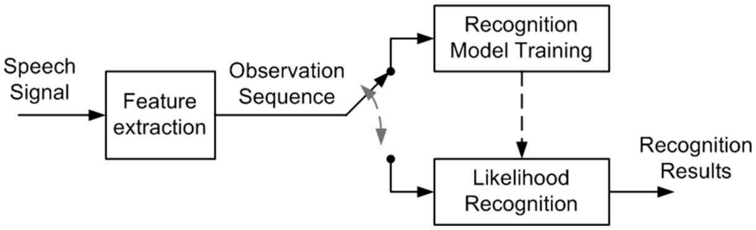


Figure 1.1: Structure of a speech recognition system based on HMM models.

the closest to the observed sequence of acoustic parameters. The most common way to characterize acoustic modelling is based on the Hidden Markov Models(HMMs).

Due to the diversity of the capabilities and characteristics of terminal devices and networks, it is expected that various client-server modalities of speech recognition exist in the mobile environment. The client-server framework of distributed speech recognition (DSR) is shown in Figure 1.2. Various kinds of devices such as personal computers, smart devices, wire and wireless telephones can act as speech-enabled client devices. Through the data channel, the characteristic features of speech are transmitted to the engine server for back-end recognition. Finally, the server recognizes the speech according to an application-specific servers and sends the result string or action back to the client.

In a DSR system, centralized servers can share the computational burden between users and enable the easy upgrade of new services without any additional cost for the user. However, transmitting acoustic data over communication networks changes the encoded information and consequently leads to severe degradation in the recognizer performance. In the case of packet-erasure channels, several packet loss compensation techniques such as interpolation [13] and error control coding [14] have been introduced for DSR. For wireless channels, joint source-channel decoding (JSCD) techniques

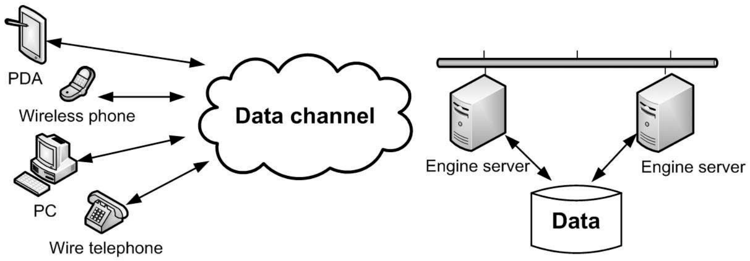


Figure 1.2: The principle framework of DSR.

[15,16,17] have been shown effective for error mitigation using the source residual redundancy and assisted with the bit reliability information provided by the soft-output channel decoder. However, the usefulness of these techniques may be restricted because they only exploited the bit-level source correlation on the basis of a memoryless AWGN channel assumption. In the thesis, we attempt to capitalize more fully on the a priori knowledge of source and channel and then develop a DSR system with increased robustness against channel errors. The first step toward realization is to use quantizer indexes rather than single index-bits as the bases for the JSCD, since the dependencies of quantizer indexes are much stronger than the correlations of the index-bits. The next knowledge source to be exploited is the channel error characteristics on which the decoder design is based [18]. For this investigation, we focused on the Markov chain model proposed by Gilbert [19]. This model can characterize a wide range of digital channels and has a recursive formula for computing the channel transition probabilities.

Speech is a dynamic acoustic signal with many sources of variation. As the production of phonemes involves different movements of the speech articulators in different environments, there is much freedom in the timing and degree of vocal tract movements. A more difficult challenge is that the speech recognition system is very sensitive to variations and mismatches between training and testing environments. Several speech variabilities can be due to the following [2].

- Inter-speaker variability: Physiological differences, articulatory habits and speaking styles are important sources of variation between speakers. For example, male-female differences account for some basic differences between speakers, since a shorter vocal tract length generally yields higher formant values.

- Intra-speaker variability: A speaker can change his/her voice quality, speaking rate, fundamental frequency or articulation patterns. Small changes in articulation patterns can result in big changes at the acoustic level. The environment changes also induce intra-speaker variability. Background noise or stress conditions yield an increase in the speakers' vocal effort and a modification of speech production.

- Environment variability: The environment in which speech is produced plays an important role and affects its production, perception, and acoustic representation. The elements of this variability include: room acoustics and reverberation, recording equipment, microphone placement, background noise, and transmission channel. Often environmental changes are difficult to simulate in the laboratory because of their large variation. This explains why there is a big difference between laboratory and filed performance. To combat the environment variability, researchers have developed algorithms for environment normalization, microphone independence, and noise robustness.

- Linguistic variability: Linguistic variations are generally associated with audible variations in terms of accents and dialects. Often the major differences which occur between dialects are apparent in their phonological, phonetic, and lexical composition. It is still not clear where the boundary is between speaker characteristics and the linguistic variations.

- Contextual variation: There exists different kinds of contextual relevant sources, such as coarticulation, local phonetic environment, and linguistic context including syntax, semantics, and pragmatics. Coarticulation is a language-dependent phenomenon, which involves changes in the articulation and acoustics of a phoneme due to its phonetic context. Recent research has shown that it is able to provide important cues and should be exploited in ASR.

The aim of this thesis is to enhance the robustness of DSR systems against the inter-speaker and channel variabilities. First, it is well known that the performance of ASR systems is sensitive to mismatches between training and testing conditions. Especially, when the acoustic characteristics of a new speaker are very different from those of the speakers in the training data, the recognition accuracy for the new speaker might be far below the average accuracy. Several different approaches to solve this problem have been proposed. They are roughly grouped into two categories, namely feature transformation methods [3,4] and model adaptation methods [5,6,7]. Feature transformation approaches attempt to transform the speaker's feature space to match the space of the training population. These approaches have the advantage of simplicity. In addition, if the number of free parameters is small, then transformation techniques adapt to the new user with only a small number of adaptation data. Among model-adaptation approaches, the maximum a posteriori (MAP) technique [5] and maximum likelihood linear regression (MLLR) technique [6] have been widely used. The MAP adaptation process is referred to as Bayesian adaptation, which involves the use of priori knowledge about the model parameter distribution. With a large amount of adaptation data, the MAP method can adapt model parameters to be converged to the corresponding speaker-dependent model parameters. On the other hand, the MLLR methods are popular due to their effectiveness and computational advantages. MLLR adaptation formulae take only a limited amount of adaptation data from a new speaker and update the HMM model parameters to maximize the likelihood of the adaptation data.

Unlike previous work, we investigate the use of voice conversion based on feature transformation to perform speaker adaptation for hearing-impaired Mandarin speaker. In Figure 1.3, this approach behaves as a preprocessing step at the speech recognizer in order to reduce the speaker variability. The goal of voice conversion is to improve the

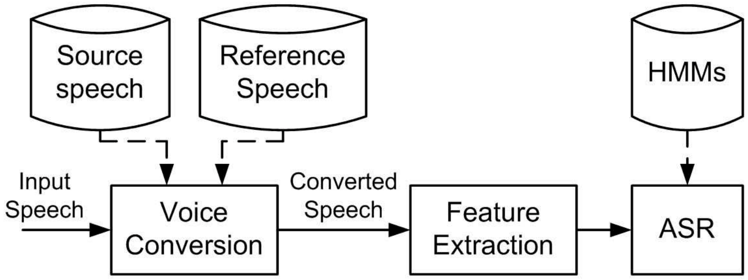


Figure 1.3: Speaker adaptation approach using voice conversion for HMM-based ASR.

intelligibility and the naturalness of hearing-impaired speech. By controlling speech individuality or adding individual cues to converted speech, it can be used to convert voice quality from one speaker to another. The technique of voice conversion has many applications, such as text-to-speech synthesis [8] and improving the quality of alaryngeal speech [9]. Most current systems [10,11] concentrate on the spectral envelope transformation, while the conversion of prosodic features is essentially obtained through a simple normalization of the average pitch. Such systems may lead to an unsatisfactory conversion quality for tonal languages, such as Chinese, which uses lexical tones to distinguish meanings of syllables that have the same phonetic compositions. In view of the important roles of prosody in Mandarin speech perception, further enhancement is expected by better modelling of pitch contour dynamics and by additionally incorporating prosodic transformation into the voice conversion system. The key to solving the problem of voice conversion lies in the detection and exploitation of characteristic features that distinguish the source speech from the reference speech [12]. To proceed with this, we found the phonological structure of Chinese language could be used to advantage in the search for the basic speech units. Also proposed is a subsyllable-based approach to voice conversion that takes into consideration both the prosodic and the spectral characteristics.

This thesis is organized as follows. In Chapter 2, we give an overview of distributed speech recognition system and examine alternative architectures for the implementation of client-server system. Chapter 3 investigates the error mitigation algorithms for DSR systems to increase the robustness against wireless channels. In Chapter 4, we present the combined use of spectral and prosodic conversion to enhance the quality of hearing-impaired Mandarin speech and therefore reduce inter-speaker variability for their use of commercial speech recognition systems. Finally, chapter 5 concludes this thesis and outlines some directions for future research.

Chapter 2

Distributed Recognition of Mandarin Speech

The primary objective of using speech recognition technique is to enable an easy access of the information. For multimedia communication over wireless network, the concept of distributed client-server system is considered by the mobile devices in lack of computational power. In this chapter we briefly examine alternative architectures for the design of distributed recognition systems and discuss their core techniques.

2.1 Distributed Speech Recognition System

Figure 2.1 illustrates the functional blocks for a distributed speech recognition system. Typically, the front-end processing performs a short-time Fourier analysis and extracts a sequence of feature vectors used for speech recognition. The source encoder removes the redundancy from the speech features to achieve lower data rate. In addition, the channel encoder adds controlled redundancy to overcome the adverse effects of noise and interference. At the receiver the channel decoder and source decoder are used to reconstruct the desired speech parameters. After that the speech recognizer

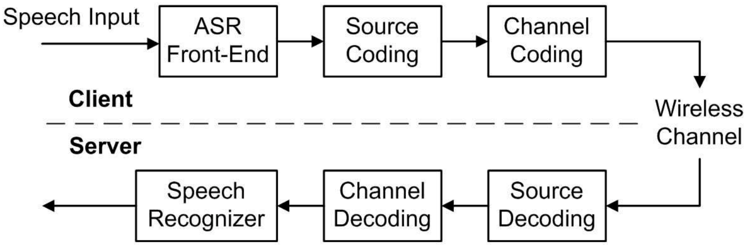


Figure 2.1: Block diagram of a DSR system.

based on the acoustic and language models is used to perform the features recognition and language understanding. The final goal of our system design is to provide high recognition accuracy over wireless channel, which keeping low bit rate and complexity for the client device. A detailed description of each functional block is given as follows.

- **Front-end signal processing** The way how to extract speech feature is an important task in the speech signal processing. Depending on the problem to be solved, the extracted features can be very simple such as zero-crossing rate, energy and pitch period, or more complex. For speech recognition applications, the power spectrum representing information about the source signal energy and vocal tract is generally used. The human ear resolves nonlinear frequency response across the audio spectrum. Empirical evidence suggests that designing a front-end to operate in a nonlinear manner improves recognition performance. The front-end processing performs a short-time Fourier analysis and extracts a sequence of acoustic vectors. Mel Frequency Cepstral Coefficients (MFCC) are typical feature vectors for ASR. The nonlinear Mel frequency scale, which is used by the MFCC representation, approximates the behavior of the human auditory response. The MFCC c_i is defined as the discrete cosine transform of the M filter outputs as follows

$$c_i = \sqrt{\frac{2}{N}} \sum_{j=1}^{M} m_j cos(\frac{\pi i}{N}(j - 0.5)), 1 \leq i \leq N \qquad (2.1)$$

where m_j is the jth *log* filterbank amplitude, M is the number of filterbank channels and N is the number of cepstral coefficients. Davis and Mermelstein [20] showed MFCC parameters are beneficial for speech recognition with increased robustness to noise and spectral estimate error.

- **Source coding** The task of the source coder is to compress the source signal so that the signal can be reconstructed with as little distortion as possible, under the constraint that the source coding rate cannot exceed the channel capacity. Various data compression techniques can be applied to remove redundancy from the original signal to achieve low bit rate for transmission and storage. Among them, vector quantization (VQ) has been widely used in many applications and allows optimum mapping a large set of input vectors into finite set of representative codevectors.

- **Channel coding** The function of the channel encoder is to introduce some artificial redundancy, which can be used at the receiver to combat the noise encountered in the data transmission. The encoding process generally involves taking k information bits and mapping each k-bit sequence into a unique n-bit sequence. The amount of redundancy introduced is measured by the ratio k/n, also called the code rate. The added redundancy serves to increase the reliability of the received data and aids the receiver in decoding the desired information sequence.

- **Communication Channel** The channel is a transmission medium which provides the connection between the transmitter and the receiver, and introduces distortion and noise to the transmitted signals. Transmission errors encountered in most real communication channels exhibit various degrees of statistical dependencies that are contingent on the transmission medium and on the particular modulation technique used. A typical example occurs in digital mobile radio channels, where speech parameters suffer severe degradation from error bursts due to the combined effects of fading and multipath propagation. In this thesis we focused on the Markov channel model. This model has several practical advantages over the Gaussian channel [4]-[6] and binary symmetric channels [9].

- **HMM-Based Speech Recognition** Hidden Markov models are often used to characterize the non-stationary stochastic process represented by the sequence of observation vectors. In HMM-based speech recognition, it is assumed that the sequence of observed vectors O corresponding to each word is generated by a Markov model. In the recognition phase, the acoustical decoding searches the state sequence for each vocabulary word and finds the most likely sequence of words W with the highest accumulated probability. This can be done by using Bayes' rule

  $$\widehat{W} = \arg\max_{W} P(W|O) = \arg\max_{W} \frac{P(W)P(O|W)}{P(O)}. \tag{2.2}$$

  The probability $P(W)$ of the word sequence W is obtained from the language model, whereas the acoustic model determines the probability $P(O|W)$. A Markov model is a finite state machine which changes state once every time unit and the state sequence S is not observed. The probability of O is obtained by summing the joint probability over all possible state sequences S, giving $P(O|W) = \sum_S P(O|S,W)P(S|W)$, where the probability $P(S|W)$ is governed by the state transition probability and the probability $P(O|S,W)$ is based on the observation distribution.

A client-server speech recognition system is implemented to provide speech-enabled applications over the Internet network. This system uses two major signal processing technologies, source coding and speech recognition, to provide efficient transmission and recognition performance. There are several alternative architectures for the client-server applications [21] over wireless communication, as shown in Figure 2.2.

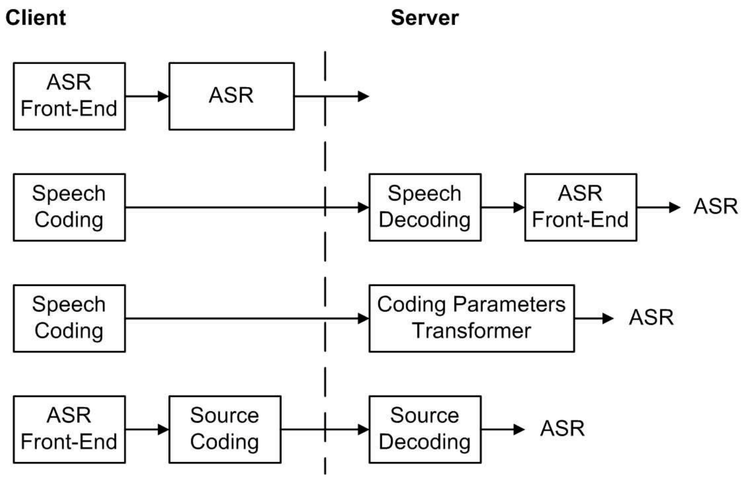


Figure 2.2: Block diagram of the different approaches for DSR.

- The first strategy(client-only processing) is to perform most of the speech recognition processing at the client side and then transmit results to the remote server. The recognition can obtain high quality speech parameters with less transmission channel error and more reliability. However, these client systems must be powerful enough to perform the recognition decoding with the heavy computation and memory resources. In addition, the recognition processing may not be inconvenient for upgrading applications.

- The second alternative(server-only processing) is to perform speech compression and coding at the client side, and transmit the user's voice parameters to the server for recognition processing. Alternatively, the recognition engine may utilize the synthesized speech as an input to ASR feature extraction or directly bitstream-based feature extraction without the synthesis process. This approach has the smallest computational and memory requirements on the client and allows a wide range of mobile communication systems to access the speech-enable applications. The disadvantage of this approach is that the recognition performance is degraded in low bit-rate connections.

- The third(client-server processing) is to perform only the front-end processing at the client and transmit characteristic features for speech recognition to the server. This approach only has a small part of computation for the front-end processing at the client, allows a wide range of mobile systems to operate various applications, and also enables the easy upgrade of technologies and services provided. To make speech recognition servers available from variety systems, front-end processing and compression need to be standardized and ensure compatibility between the client machines and the remote recognizer.

2.2 The ETSI-DSR Framework

The standard ETSI ES 202 212 [22] describes the speech processing, transmission, and quality aspects of a DSR system. The block diagram is shown in Figure 2.3. In the client side, the specification defines three major parts: the algorithm for front-end feature extraction, the processing to compress these features, and the formatting of these features. In the feature extraction part, noise reduction is performed on a frame-by-frame basis, and then mel-cepstral features are extracted. Noise reduction is performed based on Wiener filter theory in the frequency domain. After subdividing the input signal, the linear spectrum of each speech frame and the frequency-domain Wiener filter coefficients are calculated by using the estimates of speech spectrum and noise spectrum. The noise spectrum is estimated only within the silence frames, which are located by a voice activity detector. An adaptive noise reduction filter is used to subtract an additive noise from the input signal so as to improve the signal-to-noise ratio. For the cepstral analysis, noise-reduced speech signals are analyzed using a 25 ms Hamming window with 10 ms frame shift. The ceptsral coefficients are calculated from the mel-frequency warped Fourier transform representation of the log-magnitude spectrum. Mel-cepstral coefficients contain important cues in characterizing the speech sounds and are widely used in speech recognition applications.

The local front-end consists of a feature extraction algorithm and an encoding

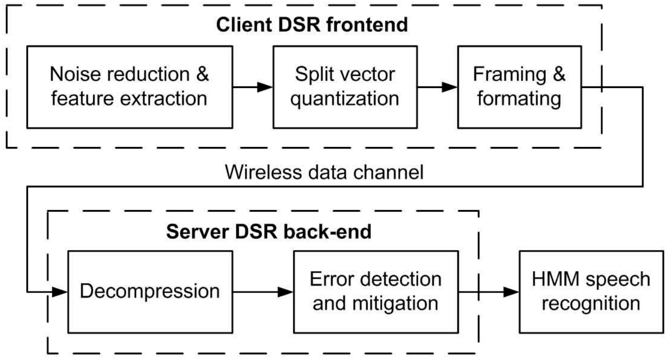


Figure 2.3: Block diagram of the ETSI-DSR system.

scheme for speech input to be transmitted to a remote recognizer. Each speech frame is represented by a 14-dimension feature vector containing log-energy $logE$ and 13 Mel-frequency cepstral coefficients (MFCCs) ranging from C_0 to C_{12}. These features are further compressed based on a split vector codebook where the set of 14 features is split into 7 subsets with two features in each. Each feature pair is quantized using its own codebook to obtain a lower transmission data rate. MFCCs C_1 to C_{10} are quantized with 6 bits each pair, (C_{11}, C_{12}) is quantized with 5 bits, and $(C_0, logE)$ is quantized with 8 bits. Experiments on small and large vocabularies indicated that the compression of mel-cepstrum parameters does not produce a significant degradation in recognition performance. After the split vector quantization, two quantized frames are grouped together and protected by a 4-bit cyclic redundancy check creating a 92-bit frame-pair packet. Twelve of these frame-pairs are combined and appended with overhead bits resulting in an 1152-bit multiframe packet representing 240 ms of speech. Multiframe packets are concatenated into a bit-stream for transmission via a data channel with an overall data rate of 4800 bits/s.

The remote back-end server performs three procedures step-by-step, including the error mitigation, decompression and recognition decoding. The error mitigation algorithm consists of two stages: detection of speech frames received with error and substitution of parameters when error are detected. The error detection includes a

CRC checking and a data consistency test. When an incorrect CRC is detected, the corresponding frame pair is classified as received with error. Besides, the consistency test is used to determine whether adjacent frames in a frame pair have a minimal continuity. When a frame is labelled as having errors, then the whole frame is replaced with the copy of the parameters from the nearest correct frame received. The front-end parameter are decompressed to reconstitute the DSR mel-cepstrum features. These are passed to the recognition decoder residing on the server. The reference recognizer is based on the HTK software package from Entropic. HTK is primarily designed for building HMM-based speech processing tools, in particular recognizers. HMM approach to speech recognition is a well-known statistical method used for characterizing the spectral properties of the speech.

In our work, the recognition of Mandarin digit strings is considered as the task without restricting the string length. A mandarin digit string database recorded by 50 male and 50 female speakers was used in our experiments. Each speaker pronounced 10 utterances and 1-9 digits in each utterance. The speech of 90 speakers (45 male and 45 female) was used as the training data, and the speech of other 10 as test data. The number of digits included in the training and test data were 6796 and 642, respectively. The digits were modelled as whole word Hidden Markov Models (HMMs) with 8 states per word and 64 mixtures for each state. In addition, a 3-state HMM was used to model pauses before and after the utterance and a one-state HMM was used to model pauses between digits. For recognition the 12 Mel-cepstrum coefficients and log-energy plus the corresponding delta and acceleration coefficients are considered.

2.3 Chinese Language Characteristics

Mandarin Chinese is a tonal language in which each syllable, with few exceptions, represents a morpheme [23]. A distinctive feature of the language is that all the characters are monosyllabic. Traditional descriptions of the Chinese syllable structure divide syllables into combinations of initials and finals rather than into individual phonemes. An initial is the consonant onset of a syllable, while a final comprises a vowel or diphthong but includes a possible medial or nasal ending. There are 22 initials and 38 finals in Mandarin and all the syllables with the initial-final combinations have 408 possible candidates. To convey different lexical meanings, each syllable can be pronounced with four basic tones; namely, the high-level tone (tone 1), the rising tone (tone 2), the falling-rising tone (tone 3), and the falling tone (tone 4). The tones are acoustically correlated with different fundamental frequency (F0) contours, and they use duration and intensity of the vowel nucleus to provide secondary information. It has been found that for Mandarin speech the vocal tract shape or parameters are essentially independent of the tones, and the tones can be separately recognized using the pitch contour information. Therefore, the tone-syllable structure is able to provide a concise and practical recognition unit and is helpful to design the Mandarin speech recognition system.

Vowels and consonants are different in the manner of their production. Most vowels are pronounced with the vocal folds vibrating, with each vowel being modified by the particular shape of the vocal tract. Depending on the manner of articulation, initial consonants can be further categorized into five phonetic classes including fricatives, affricates, stops, nasals, and glides. The major distinction in consonant type is between resonant and occlusive. Occlusive consonants depend on the obstruction degree of the airstream. Stops are produced with the mouth completely closed, and the airstream is completely stopped. In fricative consonants the mouth is not shut and the airstream is only directed through a narrow space. Affricates can be considered a combination of a stop and a fricative. In general, they start out with complete closure of the vocal tract, but then they are released in a fricative. Moreover, resonant consonants, like nasals and glides, are closer to vowels, and the vocal tract is not obstructed. Furthermore, vowels can be considered as more steady-state in nature with several hundred milliseconds in duration. By contrast, consonants are characterized by a more rapid changing articulation with the specific acoustic information in more fixed duration.

Chapter 3

Channel-Robust DSR over Wireless Networks

The increasing use of mobile communications has lead to DSR systems being developed. Transmitting DSR data over wireless environments can suffer from channel errors and consequently leads to degraded recognition performance. The ETSI Aurora DSR standard includes a basic error mitigation algorithm that has been shown effective for medium and good quality channels. In the case of packet-erasure channels, several packet loss compensation techniques such as interpolation and error control coding have been introduced for DSR. However, better mitigation algorithms have been derived from the joint source-channel decoding. In this chapter we attempt to capitalize more fully on the a priori knowledge of source and channel, and investigate the error mitigation algorithms for DSR systems with increased robustness against channel errors.

3.1 Joint Source-Channel Coding

This work is devoted to channel error mitigation for DSR over burst error channels. Figure 3.1 gives the block diagram of the transmission scheme for each DSR feature pair. Suppose at time t, the input vector $\mathbf{v}_t$ is quantized to obtain a codevector $\mathbf{c}_t \in \{\mathbf{c}^{(i)}, i = 0, 1, \ldots, 2^k - 1\}$ that, after bit mapping, is represented by a k-bit combination $u_t = (u_t(1), u_t(2), \ldots, u_t(k))$. Each bit combination u_t is assigned to a quantizer index $i \in \{0, 1, \ldots, 2^k - 1\}$ and we write for simplicity $u_t = u_t^i$ to denote that u_t represents the i-th quantizer index. Unlike source coding, the goal of DSR front-end for speech recognition is not to obtain a very low bit rate by removing the redundancy in the speech signal. Therefore, the VQ encoder exhibits considerable redundancy within the encoded index sequence, either in terms of a non-uniform distribution or in terms of correlation. If only the non-uniform distribution is considered and the indexes are assumed to be independent of each other, the redundancy is defined as the difference of between the index length k and the entropy given by

$$H(u_t) = -\sum_{u_t} P(u_t) \cdot \log_2 P(u_t). \tag{3.1}$$

If inter-frame correlation of indexes is considered by using a first-order Markov model with transition probabilities $P(u_t, u_{t-1})$, the redundancy is then defined as the index length k and the conditional entropy given by

$$H(u_t|u_{t-1}) = -\sum_{u_t} \sum_{u_{t-1}} P(u_t, u_{t-1}) \cdot \log_2 P(u_t|u_{t-1}). \tag{3.2}$$

Table 3.1 shows the index lengths and entropies for the seven feature pairs of the ETSI DSR frond-end. For each column in Table 3.1, the probabilities $P(u_t)$ and $P(u_t, u_{t-1})$ have to be estimated in advance from a training speech database. There is considerable residual redundancy left in the encoded index. From it we see that the DSR index sequence is better characterized by a first-order Markov process. For error protection individual index-bits are fed into a binary convolutional encoder consisting

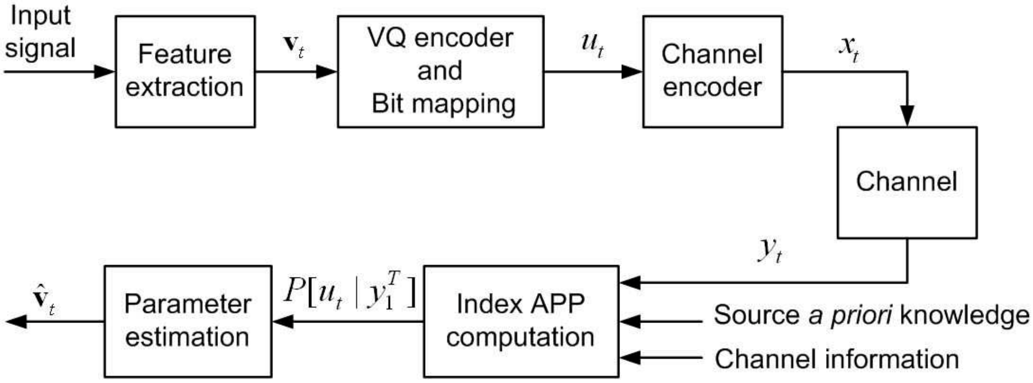
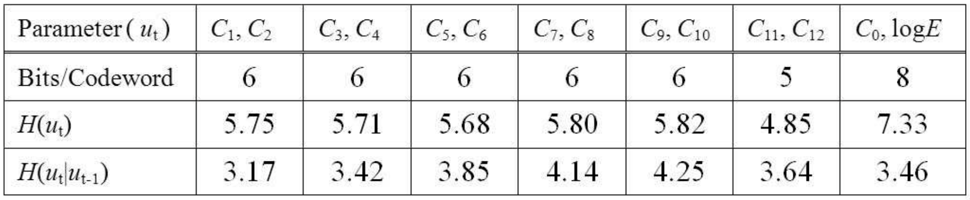


Figure 3.1: Transmission scheme for each DSR feature pair.

of M shift registers. The register shifts one bit at a time and its state is determined by the M most recent inputs. After channel encoding, the code-bit combination corresponding to the quantizer index u_t is denoted by $x_t = (x_t(1), x_t(2), \ldots, x_t(n))$ with the code rate $R = k/n$. One of the principal concerns in transmitting VQ data over noisy channels is that channel errors corrupt the bits that convey information about quantizer indexes. Assume that a channel's input x_t and output y_t differ by an error pattern e_t, so that the received bit combination is $y_t = (y_t(1), y_t(2), \ldots, y_t(n))$ in which $y_t(l) = x_t(l) \oplus e_t(l)$, $l = 1, 2, \ldots, n$, and $\oplus$ denotes the bitwise modulo-2 addition. At the receiver side, instead of using a conventional codebook-lookup decoder, the JSCD decoder will find the most probable transmitted quantizer index given the received sequence. The decoding process starts with the formation of *a posteriori* probability (APP) for each of possibly transmitted indices $u_t = i$, which is followed by choosing the index value $\hat{i}$ that corresponds to the maximum *a posteriori* (MAP) probability for that quantizer index. Once the MAP estimate of the quantizer index is determined, its corresponding codevector becomes the decoded output $\hat{\mathbf{v}}_t = \mathbf{c}^{(\hat{i})}$. The APP that a decoded index $u_t = i$ can be derived from the joint probability $P(u_t^i, s_t, y_1^T)$, where s_t is the channel encoder state at time t and $y_1^T = (y_1, y_2, \ldots, y_T)$ is the received sequence

Table 3.1: Entropies for DSR feature pairs.



Parameter (u_t)	C_1, C_2	C_3, C_4	C_5, C_6	C_7, C_8	C_9, C_{10}	C_{11}, C_{12}	C_0, log$E$
Bits/Codeword	6	6	6	6	6	5	8
$H(u_t)$	5.75	5.71	5.68	5.80	5.82	4.85	7.33
$H(u_t\|u_{t-1})$	3.17	3.42	3.85	4.14	4.25	3.64	3.46

from time $t = 1$ through some time T. We have chosen the length $T = 24$ in compliance with the ETSI bit-streaming format, where each multiframe message packages speech features from 24 frames. Proceeding in this way, the symbol APP can be obtained by summing the joint probability over all encoder states, as follows:

$$P(u_t = i|y_1^T) = \sum_{s_t} \frac{P(u_t^i, s_t, y_1^T)}{P(y_1^T)}, \qquad i = 0, 1, \ldots, 2^k - 1. \tag{3.3}$$

As measure of quality the parameter signal-to-noise ratio can be formulated as

$$SNR = 10 \log_{10} \frac{E\{\mathbf{v_t^2}\}}{E\{(\hat{\mathbf{v}}_t - \mathbf{v_t})^\mathbf{2}\}}. \tag{3.4}$$

3.2 Modified BCJR Algorithm

Depending upon the choice of the symbol APP calculator, a number of different MAP decoder implementations can be realized. For decoding convolutional codes, conventional BCJR algorithm [24] has been devised based on a bit-level code trellis. In a bit-level trellis diagram, there are two branches leaving each state and every branch represents a single index-bit. Proper sectionalization of a bit-level trellis may result in useful trellis structural properties [25] and allow us to devise MAP decoding algorithms which exploits bit-level as well as symbol-level source correlations. To advance with this, we propose a modified BCJR algorithm which parses the received code-bit

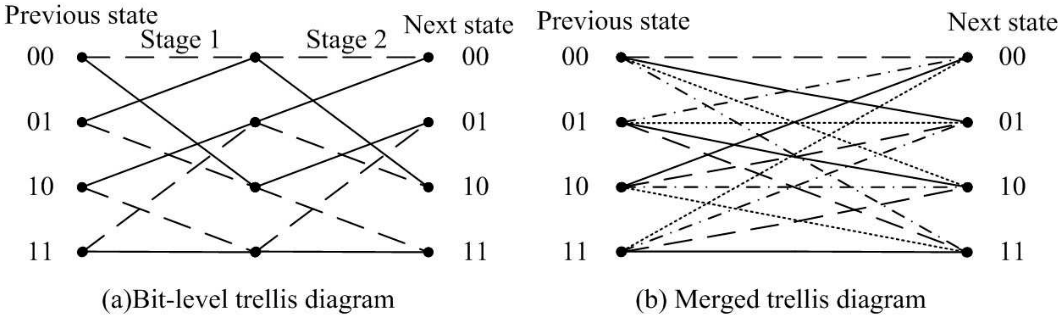


Figure 3.2: Trellis diagrams used for (a) the encoder and (b) the MAP decoder.

sequence into blocks of length n and computes the APP for each quantizer index on a symbol-by-symbol basis. Unlike conventional BCJR algorithm that decodes one bit at a time, our scheme proceeds with decoding the quantizer indexes in a frame as nonbinary symbols according to their index length k. By parsing the code-bit sequence into n-bit blocks, we are in essence merging k stages of the original bit-level code trellis into one. As an example, we illustrate in Figure 3.2 two stages of the bit-level trellis diagram of a rate 1/2 convolutional encoder with generator polynomial $(5, 7)_8$. The solid lines and dashed lines correspond to the input bits of 0 and 1, respectively. Figure 3.2 also shows the decoding trellis diagram when two stages of the original bit-level trellis are merged together. In general, there are 2^k branches leaving and entering each state in a k-stage merged trellis diagram. Having defined the decoding trellis diagram as such, there will be one symbol APP corresponding to each branch which represents a particular quantizer index $u_t = i$. For convenience, we say that the sectionalized trellis diagram forms a finite-state machine defined by its state transition function $S(u_t^i, s_t)$ and output function $X(u_t^i, s_t)$. Viewing from this perspective, the code-bit combination $x_t = X(u_t^i, s_t)$ is associated with the branch from state s_t to state $s_{t+1} = S(u_t^i, s_t)$ if the corresponding quantizer index at time t is $u_t = i$.

We next modified the BCJR algorithm based on sectionalized trellis to exploit the combined a priori information of source and channel. We begin our development of the modified BCJR algorithm by rewriting the joint probability in (3.3) as follows:

$$P(u_t^i, s_t, y_1^T) = \alpha_t^i(s_t)\beta_t^i(s_t), \tag{3.5}$$

where $\alpha_t^i(s_t) = P(u_t^i, s_t, y_1^t)$ and $\beta_t^i(s_t) = P(y_{t+1}^T|u_t^i, s_t, y_1^t)$. For the MAP symbol decoding algorithm, the forward and backward recursions are to compute the following metrics:

$$\begin{aligned}\alpha_t^i(s_t) &= \sum_{s_{t-1}}\sum_j P(u_t^i, s_t, u_{t-1}^j, s_{t-1}, y_t, y_1^{t-1}) \\ &= \sum_{s_{t-1}}\sum_j \alpha_{t-1}^j(s_{t-1})\gamma_{i,j}(y_t, s_t, s_{t-1}) \end{aligned}\tag{3.6}$$

$$\begin{aligned}\beta_t^i(s_t) &= \sum_{s_{t+1}}\sum_j P(u_{t+1}^j, s_{t+1}, y_{t+1}, y_{t+2}^T|u_t^i, s_t, y_1^t) \\ &= \sum_{s_{t+1}}\sum_j \beta_{t+1}^j(s_{t+1})\gamma_{j,i}(y_{t+1}, s_{t+1}, s_t) \end{aligned}\tag{3.7}$$

in which

$$\begin{aligned}\gamma_{i,j}(y_t, s_t, s_{t-1}) &= P(u_t^i, s_t, y_t|u_{t-1}^j, s_{t-1}, y_1^{t-1}) \\ &= P(s_t|u_{t-1}^j, s_{t-1}, y_1^{t-1})P(u_t^i|s_t, u_{t-1}^j, s_{t-1}, y_1^{t-1}) \\ &\cdot P(y_t|u_t^i, s_t, u_{t-1}^j, s_{t-1}, y_1^{t-1}). \end{aligned}\tag{3.8}$$

Having a proper representation of the branch metric $\gamma_{i,j}(y_t, s_t, s_{t-1})$ is the critical step in applying MAP symbol decoding to error mitigation and one that conditions all subsequent steps of the implementation. As a practical manner, several additional factors must be considered to take advantage of source correlation and channel memory. First, making use of the sectionalized structure of a decoding trellis, we write the first term in (3.8) as

$$P(s_t|u_{t-1}^j, s_{t-1}, y_1^{t-1}) = P(s_t|u_{t-1}^j, s_{t-1}) = \begin{cases} 1\ , & s_t = S(u_{t-1}^j, s_{t-1}) \\ 0, & otherwise. \end{cases} \tag{3.9}$$

The next knowledge source to be exploited is the residual redundancy remaining in the DSR features. Assuming that the quantizer index is modelled as a first-order Markov process with transition probabilities $P(u_t|u_{t-1})$, the second term in (3.8) is reduced to

$$P(u_t^i|s_t, u_{t-1}^j, s_{t-1}, y_1^{t-1}) = P(u_t = i|u_{t-1} = j). \tag{3.10}$$

In addition to source a priori knowledge, specific knowledge about the channel memory must be taken into consideration. There are many models describing the correlation of bit error sequences. If no channel memory information is considered, which means that the channel bit errors are assumed to be random, the third term in (3.8) is reduced to

$$P(y_t|u_t^i, s_t, u_{t-1}^j, s_{t-1}, y_1^{t-1}) = P(y_t|x_t = X(u_t^i, s_t)) = P(e_t) = \epsilon^l(1-\epsilon)^{n-l} \tag{3.11}$$

where ϵ is the channel bit error rate (BER) and l is the number of ones occurring in the error pattern e_t. When intraframe and interframe memory of the channel are considered, the third term in (3.8) becomes

$$\begin{aligned} P(y_t|u_t^i, s_t, u_{t-1}^j, s_{t-1}, y_1^{t-1}) &= P(y_t|x_t = X(u_t^i, s_t), x_{t-1} = X(u_{t-1}^j, s_{t-1}), y_{t-1}) \\ &= P(e_t|e_{t-1}). \end{aligned} \tag{3.12}$$

3.3 Probability Recursions for Gilbert channel

Designing a robust DSR system requires that parameterized probabilistic models be used to summarize some of the most relevant aspects of error statistics. It is apparent from previous work on channel modelling [26] that we are confronted with contrasting

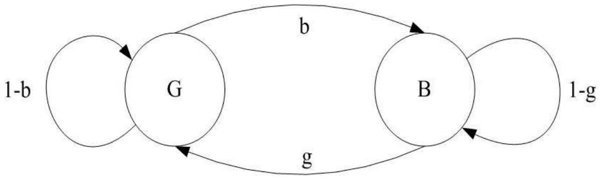


Figure 3.3: Gilbert channel model.

requirements in selecting a good model. A model should be representative enough to describe real channel behavior and yet it should not be analytically complicated. To permit theoretical analysis, we assumed that the encoded bits of DSR features were subjected to the sample error sequences typical of the Gilbert channel [27]. The Gilbert channel model consists of a Markov chain having an error-free state G and a bad state B, in which errors occur with the probability $(1 - h)$. The state transition probabilities are b and g for the G to B and B to G transitions, respectively. The model state-transition diagram is shown in Figure 3.3. The effective BER produced by the Gilbert channel is $\epsilon = (1 - h)b/(g + b)$. Notice that in the particular case of a Gilbert model with parameter values $b = 1$, $g = 0$, $h = 1 - \epsilon$, the channel model reduces to a memoryless binary symmetric channel with the BER ϵ.

The effectiveness of the MAP symbol decoding depends crucially on how well the error characteristics are incorporated into the calculation of channel transition probabilities $P(e_t|e_{t-1})$. Although using channel memory information was previously proposed for MAP symbol decoding [28], their emphasis were placed upon channels with no interframe memory. When only access to the intraframe memory is available, it was shown [27] that the channel transition probabilities of the Gilbert channel have closed-form expressions that can be represented in terms of model parameters $\{h, b, g\}$. Under such conditions, we can proceed the MAP symbol decoding in a manner similar to the work of [28]. Extensions of these results to channels with both intraframe and interframe memory has been found difficult. Recognizing this, we next develop a general treatment of probability recursions for the Gilbert channel. The main result is a recursive implementation of MAP symbol decoder being closer to the optimal for channels with memory. For notational convenience, channel bit error $e_t(l)$ will be denoted as r_m, in which the bit time m is related to the frame time t as $m = n(t-1) + l, l = 1, 2, \ldots, n$. Let $q_m \in \{G, B\}$ denote the Gilbert channel state at bit time m. The memory of the Gilbert channel is due to the Markov structure of the state transitions, which lead to a dependence of the current channel state q_m on previous state q_{m-1}.

To develop a recursive algorithm, it is more convenient to rewrite the channel transition probabilities as

$$P(e_t|e_{t-1}) = \prod_{m=n(t-1)+1}^{nt} P(r_m = 1|r_{m_0}^{m-1})^{r_m} P(r_m = 0|r_{m_0}^{m-1})^{1-r_m} \tag{3.13}$$

where $r_{m_0}^{m-1} = (r_{m_0}, r_{m_0+1}, \ldots, r_{m-1})$ represents the bit error sequence starting from bit $m_0 = n(t-2) + 1$. The following is devoted to a way of recursively computing of $P(r_m = 1|r_{m_0}^{m-1})$ from $P(r_{m-1} = 1|r_{m_0}^{m-2})$. The Gilbert channel has two properties, $P(q_m|q_{m-1}, r_{m_0}^{m-1}) = P(q_m|q_{m-1})$ and $P(r_m|q_m, r_{m_0}^{m-1}) = P(r_m|q_m)$, which facilitate the probability recursions. By successively applying Bayes rule and the Markovian property of the channel, we have

$$\begin{aligned} P(r_m = 1|r_{m_0}^{m-1}) &= P(r_m = 1|q_m = B, r_{m_0}^{m-1}) P(q_m = B|r_{m_0}^{m-1}) \\ &= (1-h) P(q_m = B|r_{m_0}^{m-1}) \end{aligned} \tag{3.14}$$

in which

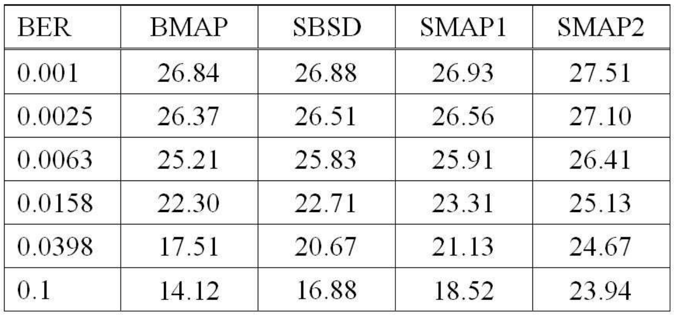
$$\begin{aligned} P(q_m = B|r_{m_0}^{m-1}) &= P(q_m = B|q_{m-1} = G, r_{m_0}^{m-1}) P(q_{m-1} = G|r_{m_0}^{m-1}) \\ &\quad + P(q_m = B|q_{m-1} = B, r_{m_0}^{m-1}) P(q_{m-1} = B|r_{m_0}^{m-1}) \\ &= b + (1-g-b) \frac{P(r_{m-1}|q_{m-1} = B)}{P(r_{m-1}|r_{m_0}^{m-2})} \frac{P(r_{m-1} = 1|r_{m_0}^{m-2})}{1-h}. \end{aligned} \tag{3.15}$$

3.4 Experimental Results

Computer simulations were conducted to evaluate three MAP-based error mitigation schemes for DSR over burst error channels. First a bit-level trellis MAP decoding scheme BMAP is considered that uses the standard BCJR algorithm to decode the index-bits. The decoders SMAP1 and SMPA2 exploit the symbol-level source redundancy by using a modified BCJR algorithm based on a sectionalized trellis structure. The SMAP1 is designed for a memoryless binary symmetric channel, whereas the SMAP2 exploits the channel memory though the Gilbert channel characterization. The channel transition probabilities to be used for the SMAP1 is $p(e_t)$ in (3.11), and $p(e_t|e_{t-1})$ in (3.12) for the SMAP2. For purpose of comparison, we also investigated an error mitigation scheme [15] which applied the concept of softbit speech decoding (SBSD) and achieved good recognition performance for AWGN and burst channels. A preliminary experiment was first performed to evaluate various decoders for reconstruction of the feature pair $(C_0, logE)$ encoded with the DSR front-end. A rate $R = 1/2$ convolutional code with memory order $M = 6$ and the octal generator $(46, 72)_8$ is chosen as the channel code. Table 3.2 presents the signal-to-noise ratio (SNR) obtained from transmission of the index-bits over Gilbert channel with BER ranging from 10^{-3} to 10^{-1}. The results of these experiments clearly demonstrate the improved performance achievable using the SMAP1 and SMAP2 in comparison to those of BMAP and SBSD. Furthermore, the improvement has a tendency to increase for noisy channels with higher BER. This indicates that the residual redundancy of quantizer indexes is better to be exploited at the symbol level to achieve more performance improvement. A comparison of the SMAP1 and SMAP2 also revealed the importance of matching the real error characteristics to the channel model on which the MAP symbol decoder design is based. The better performance of SMAP2 can be attributed to its ability to compute the symbol APP taking interframe and intraframe memory of the channel into consideration, as opposed to the memoryless channel assumption made in the SMAP1.

We further validate the proposed decoding algorithms for the case where error se-

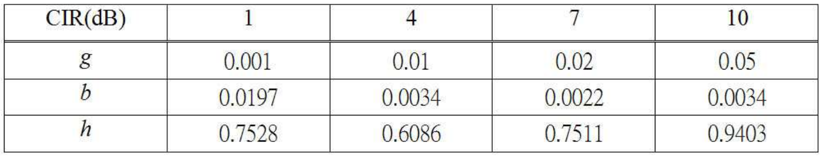
Table 3.2: SNR(dB) performance for various decoders on a Gilbert channel.



BER	BMAP	SBSD	SMAP1	SMAP2
0.001	26.84	26.88	26.93	27.51
0.0025	26.37	26.51	26.56	27.10
0.0063	25.21	25.83	25.91	26.41
0.0158	22.30	22.71	23.31	25.13
0.0398	17.51	20.67	21.13	24.67
0.1	14.12	16.88	18.52	23.94

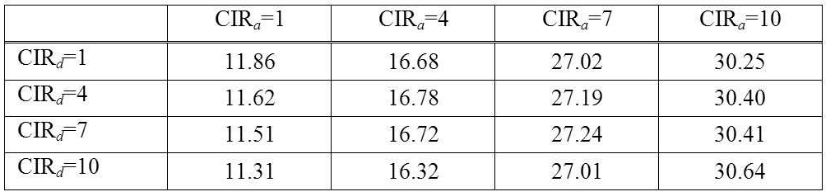
quences were generated using a complete GSM simulation. The simulator is based on the CoCentric GSM library [29] with TCH/F4.8 data and channel coding, interleaving, modulation, a channel model, and equalization. The channel model represents a typical case of a rural area with 6 propagation paths and a user speed of 50 km/h. Further, cochannel interference was simulated at various carrier-to-interference ratios (CIR). In using the SMAP1 and SMAP2 schemes, the channel transition probabilities have to be combined with a priori knowledge of Gilbert model parameters which can be estimated once in advance using the gradient iterative method [30]. For each of simulated error sequences, we first measured the error-gap distribution by computing the probability that at least l successive error-free bits will be encountered next on the condition that an error bit has just occurred. The optimal identification of Gilbert model parameters was then formulated as the least square approximation of the measured error-gap distribution by exponential curve fitting. Table 3.3 gives estimated Gilbert model parameters for the GSM TCH/F4.8 data channels operating at $CIR = 1, 4, 7, 10$ dB. The next step in the present investigation concerned the performance degradation that may result from using the SMAP2 scheme under channel mismatch conditions. In Table 3.4, CIR_d refers to the CIR value assumed in the design process, and CIR_a refers to the true CIR used for the evaluation. The best results are in the main diagonal of the table, where channel-matched Gilbert model parameters are used for the channel tran-

Table 3.3: Estimated Gilbert model parameters for GSM TCH/F4.8 data channels.



CIR(dB)	1	4	7	10
g	0.001	0.01	0.02	0.05
b	0.0197	0.0034	0.0022	0.0034
h	0.7528	0.6086	0.7511	0.9403

Table 3.4: SNR performance of the SMAP2 over the GSM data channel under channel mismatch conditions.



	CIR$_a$=1	CIR$_a$=4	CIR$_a$=7	CIR$_a$=10
CIR$_d$=1	11.86	16.68	27.02	30.25
CIR$_d$=4	11.62	16.78	27.19	30.40
CIR$_d$=7	11.51	16.72	27.24	30.41
CIR$_d$=10	11.31	16.32	27.01	30.64

sition probability computation of (3.13). The performance decreases in each column below the main diagonal when the CIR_d is increased. The investigation further showed that the SMAP2 is not very sensitive to a channel mismatch between the design and evaluation assumptions.

We next considered the speaker-independent recognition of Mandarin digit strings as the task without restricting the string length. A Mandarin digit string database recorded by 50 male and 50 female speakers was used in our experiments. Each speaker pronounced 10 utterances and 1-9 digits in each utterance. The speech of 90 speakers (45 male and 45 female) was used as the training data, and the other 10 as test data. The total numbers of digits included in the training and test data were 6796 and 642, respectively. The DSR results obtained by various error mitigation algorithms for the

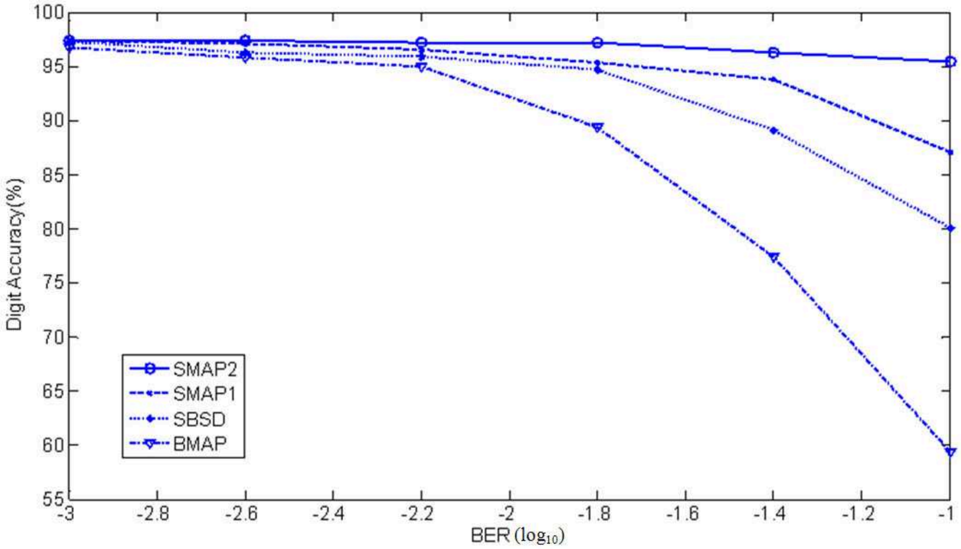


Figure 3.4: Recognition performances for DSR transmission over a Gilbert channel.

Gilbert channel are shown in Figure 3.4. It can be seen that employing the source a priori information, sectionalized trellis MAP decoding, and channel memory constantly improves the recognition accuracy. The SMAP2 scheme performs the best in all cases, showing the importance of combining the a priori knowledge of source and channel by means of a sectionalized code trellis and Gilbert channel characterization.

3.5 Summary

A JSCD scheme which exploits the combined source and channel statistics as an a priori information is proposed and applied to the channel error mitigation in DSR applications. We first investigate the residual redundancies existing in the DSR features and find ways to exploit these redundancies in the MAP symbol decoding process. Also proposed is a modified BCJR algorithm based on sectionalized code trellises which uses Gilbert channel characterization for better decoding in addition to source a priori knowledge. Experiments on Mandarin digit string recognition indicate that the proposed decoder achieved significant improvements in recognition accuracy for DSR over burst error channels.

Chapter 4

Speaker-Adaptive DSR for Hearing-Impaired Mandarin Speech

In this chapter we investigate the use of voice conversion based on feature transformation to perform speaker adaptation and to reduce the speaker variability for hearing-impaired speaker. Due to the lack of adequate auditory feedback, the hearing-impaired speakers produce speech with segmental and suprasegmental errors. This motivates our research into trying to devise a voice conversion system that modifies the speech of a hearing-impaired (source) speaker to be perceived as if it was uttered by a normal (target) speaker. The key to our proposed conversion lies in the detection and exploitation of characteristic features that distinguish the impaired speech from the normal speech at segmental and prosodic levels. Segmental features that contribute to speech individuality are encoded in the spectral envelop, whereas prosodic information can be found in pitch, energy, and duration variations that span across segments. Thus, we present that speech waveforms are modelled by the sinusoidal framework which decomposes speech signals into the product of excitation and system spectra and makes the reconstruction a best fit to the original speech [31]. Next, the conversion techniques were applied on the framework to enhance the hearing-impaired Mandarin speech.

4.1 Characteristics of Hearing-Impaired Mandarin Speech

Speech communication by profoundly hearing-impaired individuals suffers not only from the fact that they cannot hear other people's utterances, but also from the poor quality of their own productions. Due to the lack of adequate auditory feedback, the hearing-impaired speakers produce speech with segmental and suprasegmental errors [32]. It is common to hear their speech flawed by misarticulated phonemes, with varying degrees of severity associated with their hearing thresholds [33,34]. Their speech intelligibility is further affected by abnormal control over phoneme duration and pitch variations. Specifically, the duration of vowels, glides, and nasals were longer while the duration of fricatives, affricates, and plosives were shorter than in normal speech, and the pitch contour over individual syllables is either too varied or too monotonous. Their intonation also shows limited pitch variation, erratic pitch fluctuations, and inappropriate average F0 [35].

Recent perceptual work on Chinese deaf speech [36,37] has shown that speakers with greater than moderate degrees of losses ($\geq$ 50 dB HL bilaterally) were perceived with an average accuracy of 31% in phoneme production, and further, that the most errors in the consonants were affricates and fricatives. This finding may have more serious implications for Mandarin than for other languages as these two phonetic classes make up more than half of the consonants in Mandarin Chinese. Moreover, since most of them are palatal or produced without apparent visual cues, they are difficult to correct through speech training. In tone production, their accuracy only reached an average of 54%, with most errors involving confusions between tones 1 and 4, tones 1 and 2, and tones 2 and 3. The results also showed that tones produced by speakers with profound losses were only half as likely to be judged correct as those produced by speakers with less loss. Again, as tones are produced by phonatory, rather than articulatory control, they are almost impossible to correct through non-instrumental-based speech therapy.

The four basic Mandarin tones mentioned earlier have distinctive shapes of F0 contours, whose perception is correlated with the starting frequency, the initial fall and the timing when the turning point appears, as involved in tones 2 and 3 [38]. Our teenage data supported the general conclusion with a different measure. Specifically, instead of focusing on the interactions between the frequency and temporal aspects, we recorded the frequency differences between the highest and the lowest point found on the contours. The results showed a clear trend for the normal speakers with the difference increased when going from tone 1 to tone 4 (e.g., 19.6, 24.8, 53, 113.1 Hz), which was less orderly (e.g., 9.3, 16, 28, 66.4 Hz) for the impaired speakers. The most frequent perceptual mistakes made by our impaired speakers were substitutions of tone 3 with tone 2, which left only three perceptual categories 1, 2, and 4 in their tonal repertoire. Unstable tonal productions across recorded tokens were also common.

As stated earlier [35], the speech of the hearing-impaired speakers contains numerous timing errors, including a lower speaking rate, insertion of long pauses, and failure to modify segment duration as a function of phonetic environment. In Figure 4.1 the mean phoneme durations produced by the hearing impaired were plotted against those produced by the normal speakers. Data were collected from two normal speakers (one male and one female) and three hearing-impaired speakers (one male and two females), all aged 15. The phonemes tested were five fricatives, six affricates, and three vowels. It can be seen that the mean duration ratios of impaired-to-normal utterances were quite different for different phonemes and that vowels, as a group, stayed much in line with the normal production than the two consonant groups, with the mean ratios for vowels, fricatives, and affricates being 1.12, 0.4, and 0.34, respectively. All consonants, with the exception of /h/, of the hearing impaired were shorter, as indicated by their uniform appearances on the lower half of the graph. Our perceptual judgment showed that this shortening that could measure 10 to 1 (as seen in /sh/ and /shi/) was the result of substituting the two consonant classes with stops.

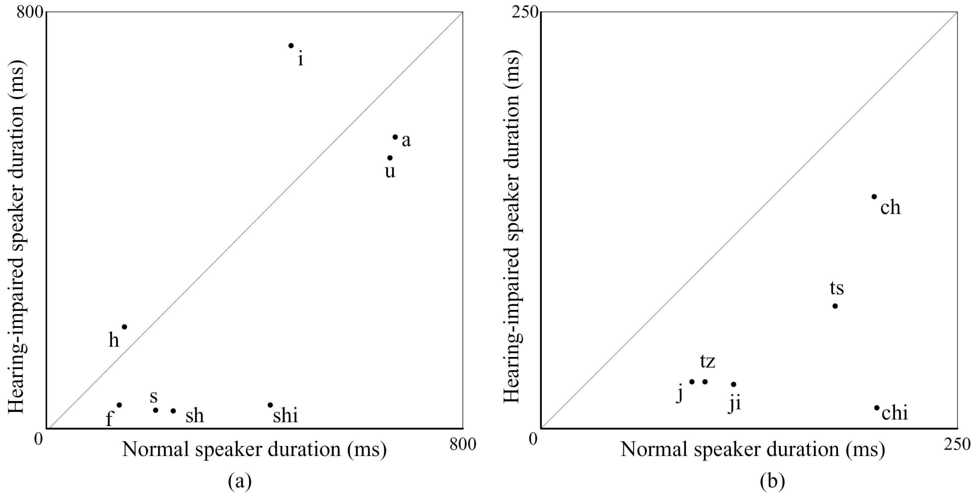


Figure 4.1: Phoneme duration statistics for (a) vowels and fricative consonants and (b) affricate consonants.

4.2 Sinusoidal Framework for Voice Conversion

The general approach to voice conversion consists of first analyzing the input speech to obtain characteristic features, then applying the desired transformations to these features, and synthesizing the corresponding signal. Essentially, the production of sound can be described as the output of passing a glottal excitation signal through a linear system representing the characteristics of the vocal tract. To track the nonstationary evolution of characteristic features, both the spectral and prosodic manipulations will be performed on a frame-by-frame basis. In this work, speech signals were sampled at 11 kHz and analyzed using a 46.4 ms Hamming windows with a 13.6 ms frame shift. Therefore, the analysis frame interval Q was fixed at 13.6 ms. For the speech on the mth frame, the vocal tract system function can be described in terms of its amplitude function $M(w;m)$ and phase function $\Phi(w;m)$. Usually the excitation signal is represented as a periodic train during voiced speech, and is represented as a noise-like signal during unvoiced speech. An alternative approach [39] is to represent the excitation signal by a sum of $K(m)$ sine waves, each of which is associated with the frequency

$w_k(m)$ and the phase $\Omega_k(m)$. Passing this excitation signal through the vocal tract system results in a sinusoidal representation of speech production. As noted elsewhere [40], this sinusoidal framework allows flexible manipulation of speech parameters such as pitch and speaking rate while maintaining high speech quality.

A block diagram of the proposed voice conversion system is shown in Figure 4.2. The system has five major components: speech analysis, spectral conversion, pitch modification, time-scale modification, and speech synthesis. The analysis begins by estimating from the Fourier transform of input speech the pitch period $P_0(m)$, the voicing probability $P_v(m)$, and the system amplitude function $M(w;m)$. The voicing probability will be used to control the harmonic spectrum cutoff frequency, $w_c(m) = \pi P_v(m)$, below which voiced speech was synthesized and above which unvoiced speech was synthesized. The second step in the analysis is to represent the system amplitude function $M(w;m)$ in terms of a set of cepstral coefficients $\{c_l(m)\}_{l=0}^{24}$. The main attraction of cepstral representation is that it exploits the minimum-phase model, where the log-magnitude and phase of the vocal tract system function can be uniquely related in terms of the Hilbert transform [41]. A more comprehensive discussion of the sine-wave speech model and the corresponding analysis-synthesis system can be found in [39].

The main part of the modification procedure involves the manipulation of functions which describe the amplitude and phase of the excitation and vocal tract system contributions to each sine-wave component. The effectiveness of voice conversion depends on a successful modification of prosodic features, especially of the time-scale and the pitch-scale. With reference to the sinusoidal framework, speech parameters included in the prosodic conversion are $P_0(m)$, $P_v(m)$, and the synthesis frame interval. The time-scale modification involves scaling the synthesis frame of original duration Q by a factor of $\rho(m)$, i.e., $Q'(m) = \rho(m)Q$. The pitch modification can be viewed as a transformation which, when applied to the pitch period $P_0(m)$, yields the new pitch period $P_0'(m)$, with an associated change in the F0 as $w_0'(m) = 2\pi/P_0'(m)$. It is worth noting also that the change in pitch period also corresponds to modification of the

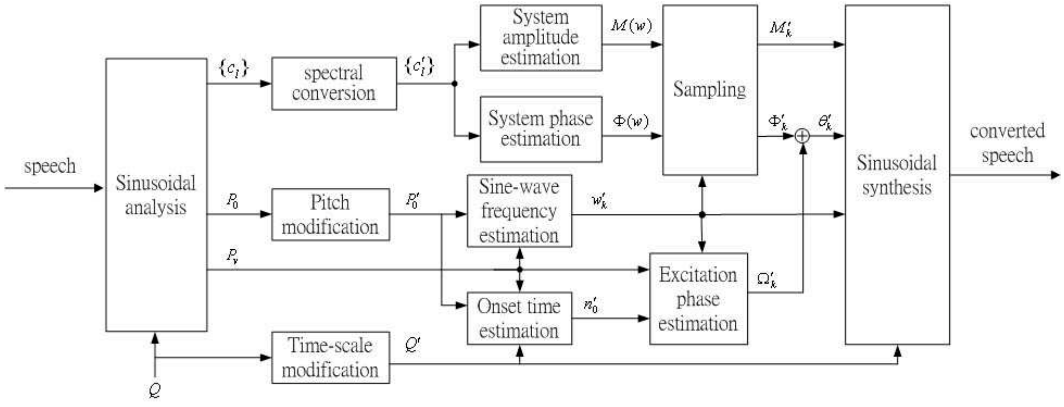


Figure 4.2: Block diagram of the voice conversion system.

sine-wave frequencies $w'_k(m)$ and the excitation phases $\Omega'_k(m)$ used in the reconstruction. Below the cutoff frequency the sine-wave frequencies are harmonically related as $w'_k(m) = kw'_0(m)$, whereas above the cutoff frequency $w'_k(m) = k^*w'_0(m) + w_u$, where k^* is the largest value of k for which $k^*w'_0 \leq w_c(m)$, and where w_u is the unvoiced F0 corresponding to 100 Hz. A two-step procedure is used in estimating the excitation phase $\Omega'_k(m)$ of the kth sine wave. The first step is to obtain the onset time $n'_0(m)$ relative to both the new pitch period $P'_0(m)$ and the new frame interval $Q'(m)$. This is done by accumulating a succession of pitch periods until a pitch pulse crosses the center of the mth frame. The location of this pulse is the onset time $n'_0(m)$ at which sine waves are in phase. The second step is to compute the excitation phase as follows:

$$\Omega'_k(m) = -n'_0(m)w'_k(m) + \epsilon'_k(m), \tag{4.1}$$

where the unvoiced phase component $\epsilon'_k(m)$ is zero for the case of $w'_k(m) \leq w_c(m)$ and is made random on $[-\pi, \pi]$ for the case of $w'_k(m) > w_c(m)$.

In addition to prosodic conversion, the technique of spectral conversion is also needed to modify the articulation-related parameters of speech. The problem with the spectral conversion lies with the corresponding modification of the vocal tract system function. Thus there is a need to estimate the amplitude function $M'(w;m)$ and the phase function $\Phi'(w;m)$ of the vocal tract system. If it is assumed that the vocal tract system function is minimum phase [41], the log-magnitude and phase functions form a Hilbert transform pair and hence can be estimated from a set of new cepstral coefficients $\{c'_l(m)\}_{l=0}^{24}$. The system amplitudes $M'_k(m)$ and phases $\Phi'_k(m)$ are then given by samples of their respective functions at the new frequencies $w'_k(m)$, i.e., $M'_k(m) = M'(w'_k;m)$ and $\Phi'_k(m) = \Phi'(w'_k;m)$. Finally, in the synthesizer the system amplitudes are linearly interpolated over two consecutive frames. Also, the excitation and system phases are summed and the resulting sine-wave phases, $\theta'_k(m) = \Omega'_k(m) + \Phi'_k(m)$, are interpolated using the cubic polynomial interpolator. The final synthetic speech waveform on the mth frame is given by

$$s(n) = \sum_{k=1}^{K(m)} M'_k(m) cos[nw'_k(m) + \theta'_k(m)],\ t_m \le n \le t_{m+1} - 1 \tag{4.2}$$

where $t_m = \sum_{i=1}^{m-1} Q'(i)$ denotes the starting time of the current synthesis frame.

4.3 Spectral Conversion

Mandarin is a syllable-timed language in which each syllable consists of an initial part and a final part. The primary difficulties in the recognition of Mandarin syllables are tied to the durational differences between the syllable-initial and syllable-final part. Specifically, the initial part of a syllable is short when compared with the final part, which usually causes distinctions among the initial consonants in different syllables to be swamped by the following irrelevant differences among the finals. This may help explain why early approaches that used whole-syllable models as the conversion units did not produce satisfactory results for Mandarin speech conversion. To circumvent this pitfall, we perform spectral conversion only after decomposing the Mandarin syllables into smaller sound units as in phonetic classes. The acoustic features included in the conversion are cepstral coefficients derived from the smoothed spectrum. The conversion system design involves two essential problems: 1) developing a parametric model representative of the distribution of cepstral coefficients, and 2) mapping the spectral envelopes of the source speaker onto those of the target. In the context of spectral transformation, Gaussian mixture models (GMMs) have been shown to provide superior performance to other approaches based on VQ or neural networks [42]. Our approach began with a training phase in which all cepstral vectors of the same phonetic class were collected and used to train the corresponding GMM associated with the phonetic class by a supervised learning procedure. We consider that the available data consists of two sets of time-aligned cepstral vectors $\mathbf{x}_t$ and $\mathbf{y}_t$, corresponding, respectively, to the spectral envelopes of the source and the target speakers. The GMM assumes that the probability distribution of the cepstral vectors $\mathbf{x}$ takes the following parametric form

$$p(\mathbf{x}) = \sum_{i=1}^{I} \alpha_i \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_i^x, \boldsymbol{\Sigma}_i^{xx}) \tag{4.3}$$

where α_i denotes a weight of class i, $I = 24$ denotes the total number of Gaussian mixtures, and $\mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_i^x, \boldsymbol{\Sigma}_i^{xx})$ denotes the normal distribution with mean vector $\boldsymbol{\mu}_i^x$ and covariance matrix $\boldsymbol{\Sigma}_i^{xx}$. It therefore follows the Bayes theorem that a given vector $\mathbf{x}$ is generated from the ith class of the GMM with the probability:

$$h_i(\mathbf{x}) = \frac{\alpha_i \mathcal{N}(\mathbf{x}, \boldsymbol{\mu}_i^x, \boldsymbol{\Sigma}_i^{xx})}{\sum_{j=1}^{I} \alpha_j \mathcal{N}(\mathbf{x}, \boldsymbol{\mu}_j^x, \boldsymbol{\Sigma}_j^{xx})}. \tag{4.4}$$

With this, cepstral vectors are converted from the source speaker to the target speaker by the conversion function that utilizes feature parameter correlation between the two speakers. The conversion function that minimizes the mean squared error between converted and target cepstral vectors was given by [42],

$$\mathcal{F}(\mathbf{x}_t) = \sum_{i=1}^{I} h_i(\mathbf{x}_t)[\boldsymbol{\mu}_i^y + \boldsymbol{\Sigma}_i^{yx}(\boldsymbol{\Sigma}_i^{xx})^{-1}(\mathbf{x}_t - \boldsymbol{\mu}_i^x)], \tag{4.5}$$

where for class i, $\boldsymbol{\mu}_i^y$ denotes the mean vector for the target cepstra, $\boldsymbol{\Sigma}_i^{xx}$ denotes covariance matrix for the source cepstra, and $\boldsymbol{\Sigma}_i^{yx}$ denotes the cross-covariance matrix.

Within the GMM framework, training the conversion function can be formulated as one of the optimal estimation of model parameters $\lambda = \{\alpha_i, \boldsymbol{\mu}_i^x, \boldsymbol{\mu}_i^y, \boldsymbol{\Sigma}_i^{xx}, \boldsymbol{\Sigma}_i^{yx}\}$. Our approach to parameter estimation is based on fitting a GMM to the probability distribution of the joint vector $\mathbf{z}_t = [\mathbf{x}_t, \mathbf{y}_t]^T$ for the source and target cepstra. Covariance matrix $\boldsymbol{\Sigma}_i^z$ and mean vector $\boldsymbol{\mu}_i^z$ of class i for joint vectors can be written as

$$\boldsymbol{\Sigma}_i^z = \begin{bmatrix} \boldsymbol{\Sigma}_i^{xx} & \boldsymbol{\Sigma}_i^{xy} \\ \boldsymbol{\Sigma}_i^{yx} & \boldsymbol{\Sigma}_i^{yy} \end{bmatrix}, \quad \boldsymbol{\mu}_i^z = \begin{bmatrix} \boldsymbol{\mu}_i^x \\ \boldsymbol{\mu}_i^y \end{bmatrix}. \tag{4.6}$$

The expectation-maximization (EM) algorithm [43] is applied here to estimate the model parameters which guarantees a monotonic increase in the likelihood. Starting with an initial model λ, the new model $\bar{\lambda}$ is estimated by maximizing the auxiliary function

$$Q(\lambda, \bar{\lambda}) = \sum_{t=1}^{T} \sum_{i=1}^{I} p(i|\mathbf{z}_t, \lambda) \cdot \log p(i, \mathbf{z}_t|\bar{\lambda}), \tag{4.7}$$

where

$$p(i, \mathbf{z}_t|\bar{\lambda}) = \bar{\alpha}_i \mathcal{N}(\mathbf{z}_t, \bar{\boldsymbol{\mu}}_i^z, \bar{\boldsymbol{\Sigma}}_i^z), \tag{4.8}$$

and

$$p(i|\mathbf{z}_t, \lambda) = \frac{\alpha_i \mathcal{N}(\mathbf{z}_t, \boldsymbol{\mu}_i^z, \boldsymbol{\Sigma}_i^z)}{\sum_{j=1}^{I} \alpha_j \mathcal{N}(\mathbf{z}_t, \boldsymbol{\mu}_j^z, \boldsymbol{\Sigma}_j^z)}. \tag{4.9}$$

On each EM iteration, the reestimation formulas derived for individual parameters of class i are of the form

$$\bar{\alpha}_i = \frac{1}{T} \sum_{t=1}^{T} p(i|\mathbf{z}_t, \lambda), \tag{4.10}$$

$$\bar{\boldsymbol{\mu}}_i^z = \frac{\sum_{t=1}^{T} p(i|\mathbf{z}_t, \lambda)(\mathbf{z}_t)}{\sum_{t=1}^{T} p(i|\mathbf{z}_t, \lambda)}, \tag{4.11}$$

$$\bar{\boldsymbol{\Sigma}}_i^z = \frac{\sum_{t=1}^{T} p(i|\mathbf{z}_t, \lambda)(\mathbf{z}_t - \boldsymbol{\mu}_i^z)(\mathbf{z}_t - \boldsymbol{\mu}_i^z)^T}{\sum_{t=1}^{T} p(i|\mathbf{z}_t, \lambda)}. \tag{4.12}$$

The new model $\bar{\lambda}$ then becomes λ for the next iteration and the reestimation process is repeated until the likelihood reaches a fixed value.

4.4 Prosodic transformation

Most current approaches to voice conversion make little or no use of pitch measures, despite evidence showing that intonational information is highly correlated to speech individuality. The main reason for this is the difficulty in finding an appropriate feature set that captures linguistically relevant intonational information. This problem is alleviated in Mandarin speech conversion task as its tonal system allows relatively non-overlapping characterizations of the corresponding F0 contour dynamics. Speech enhancement can therefore be realized by a proper analysis and control of the F0 contour dynamics. Since pitch is defined only for voiced speech, the pertinent tone-related portions of syllables are the vowel or diphthong nuclei from which distinctive pitch changes are perceived. Recognizing this, we need only to concatenate F0 values of the final subsyllable into a vector and represent it by a small linguistically motivated parameter set. Unlike the conventional frame-based VQ approaches [10], this segment-based approach makes it possible to convert not only the static characteristics but also the dynamic characteristics of F0 contours.

Choosing an appropriate representation of F0 contour is the first step in applying pitch modification to the voice conversion. By taking advantage of the simple tone structure of F0 contours in mandarin speech, the polynomial curve fitting technique is used to decompose the F0 contour into mutually orthogonal components in transform domain [44]. The F0 contour can therefore be represented by a smooth curve formed by orthogonal expansion using some low order transform coefficients. In describing the source speaker's F0 contour, F0 are measured only for the final subsyllable and are in the form of $\{w_0(m_x), B_x \leq m_x \leq T_x\}$. For notational convenience, the F0 contour of a segment with $I_x + 1$ frames is rewritten as $\{w_0(i_x), 0 \leq i_x \leq I_x\}$, where $i_x = m_x - B_x$ and $I_x = T_x - B_x$. Parameters for pitch modification are then extracted from the F0 contour segment by the orthogonal polynomial transform:

$$b_j^{(x)} = \frac{1}{I_x + 1} \sum_{i_x=0}^{I_x} w_0(i_x) \cdot \Psi_j(\frac{i_x}{I_x}),\ j = 0, 1, 2, 3. \tag{4.13}$$

Due to the smoothness of an F0 contour segment [44], the first four discrete Legendre polynomials are chosen as the basis functions $\Psi_j(\cdot)$ to represent it. Based on this orthogonal polynomial representation, the source F0 contour is characterized by a 4-dimensional feature vector, $\mathbf{b}^{(x)} = (b_0^{(x)}, b_1^{(x)}, b_2^{(x)}, b_3^{(x)})^T$, which will be quantized using vector quantization (VQ) technique. Similarly, $\mathbf{b}^{(y)} = (b_0^{(y)}, b_1^{(y)}, b_2^{(y)}, b_3^{(y)})^T$ is a feature vector representing the F0 contour of the target speaker.

Our conversion technique is based on the codebook mapping and consists of two steps: a learning step and a conversion-synthesis step. In the learning step, the source and target F0 codebooks were separately generated using an orthogonal polynomial representation of F0 contours in training utterances. Each of the two codebooks includes 16 codevectors and is designed using the well-known LBG algorithm [45]. Next, a histogram of correspondence between codebook elements of the two speakers is calculated. Using this histogram as a weighting function, the mapping codebook is defined as a linear combination of target F0 codevectors. In the conversion-synthesis step, the F0 contour of input speech was orthogonally transformed and vector-quantized using the source F0 codebook. Then, the pitch modification was carried out by decoding them using the mapping codebook. If the decoded codevector is $\hat{\mathbf{b}} = (\hat{b}_0, \hat{b}_1, \hat{b}_2, \hat{b}_3)^T$, the modified F0 for frame $m_x = i_x + B_x$ can be approximated as

$$w_0'(i_x + B_x) = \sum_{j=0}^{3} \hat{b}_j \cdot \Psi_j(\frac{i_x + B_x}{I_x}),\ 0 \leq i_x \leq I_x. \tag{4.14}$$

Hearing-impaired speech is generally characterized by a much lower speaking rate and by excessive shortening of consonants. For the converted speech to carry the naturalness of human speech, the duration of individual phonemes needs to match those found in the natural speech. This can be done by modifying the interval of each synthesis frame by a time-varying factor $\rho(m)$ in a way of $Q'(m) = \rho(m)Q$. The case $\rho(m) > 1$ corresponds to a time-scale expansion, while the case $\rho(m) < 1$ corresponds to a time-scale compression. The next step is to determine the time-scaling factor $\rho(m)$ based on spectral representations of the same syllable uttered by the source and target speakers. In describing the source speaker's spectral envelope, cepstral coefficients are measured frame by frame and are of the following form: $\mathbf{X} = \{\mathbf{x}(m_x), m_x = 1, 2, \ldots, T_x\}$, where T_x is the syllable duration in frames. Similarly, $\mathbf{Y} = \{\mathbf{y}(m_y), m_y = 1, 2, \ldots, T_y\}$ is the sequence of T_y cepstral vectors representing the target speaker's spectral envelope. Acoustic analysis of Mandarin hearing-impaired speech has indicated that unvoiced sound such as consonants may not be subjected to the same scaling as the vowels. Thus for time-scaling of speech, different approaches should be applied in the time-intervals where the frames corresponding to both speakers were marked as Mandarin initials or finals. The boundary between the initial and final parts of an isolated syllable is relatively easy to detect by a voiced/unvoiced decision based on the voicing probability P_v. Let B_x and B_y represent the starting frame for the final subsyllables in the source and target utterances, respectively. For constituent frames of the initial consonant, a linear time normalization was applied with a fixed factor $\rho = (B_y - 1)/(B_x - 1)$. With regards to the final subsyllables, two sets of paired cepstral vectors, $\{\mathbf{x}(m_x), B_x \leq m_x \leq T_x\}$ and $\{\mathbf{y}(m_y), B_y \leq m_y \leq T_y\}$, were time aligned using the procedure of dynamic time warping (DTW) [46]. Usually the problem of DTW is formulated as a path finding problem over a finite range of grid points (m_x, m_y). The basic strategy applied here is to interpret the slope of the DTW path as a time-scaling function, which indicates on a frame-by-frame basis how much to shorten or lengthen each frame of the source utterance in order to reproduce the same duration as in target utterance.

The DTW aims to align two utterances with a path through a matrix of similarity distances that minimizes the sum of the distances. We begin by defining a partial accumulated distance $D_A(m_x, m_y)$, representing the accumulated distance along the best path from the point (B_x, B_y) to the point (m_x, m_y). For an efficient implementation, a dynamic programming recursion is applied to compute $D_A(m_x, m_y)$ for all local paths that reach (m_x, m_y) in exactly one step from an intermediate point (m'_x, m'_y) using a set of local path constraints. Table 4.1 summarizes the local constraints and slope weights for three local paths, $\wp_1, \wp_2$, and $\wp_3$, chosen for the implementation. The local distance $d(m_x, m_y)$ between the time-aligned pairs of cepstral vectors is defined by a squared Euclidean distance. We summarize the dynamic programming implementation for finding the time-scaling factor at every frame of a final subsyllable as follows.

1) Initialization: Set $D_A(B_x, B_y) = d(B_x, B_y)$

2) Recursion: For $B_x + 1 \leq m_x \leq T_x$ and $B_y + 1 \leq m_y \leq T_y$, compute

$$D_A(m_x, m_y) = \min_{(m'_x, m'_y)} [D_A(m'_x, m'_y) + \varsigma((m'_x, m'_y), (m_x, m_y))] \tag{4.15}$$

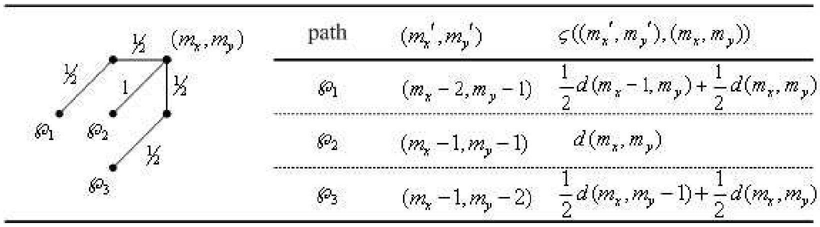
where the incremental distortion $\varsigma((m'_x, m'_y), (m_x, m_y))$ and the intermediate point (m'_x, m'_y) along three local paths $\wp_1, \wp_2$, and $\wp_3$ are given in Table 4.1.

3) Path backtracking: According to the optimal DTW path, we define the time-scaling factor $\rho(m)$=0.5, 1, or 2, for the case where the move from the point (m'_x, m'_y) to the point (m_x, m_y) is via the local path $\wp_1, \wp_2$, or $\wp_3$, respectively.

4.5 Experimental Results

Experiments were carried out to investigate the potential advantages of using the proposed conversion algorithms to enhance the hearing-impaired Mandarin speech. Our efforts began with the collection of a speech corpus that contained two sets of monosyllabic utterances, one for system learning and one for testing in our voice conversion

Table 4.1: Incremental distortions and slope weights for local paths.



½
(m_x, m_y)
½
1
½
$\wp_1$
$\wp_2$
½
$\wp_3$

path	(m_x', m_y')	$\varsigma((m_x', m_y'), (m_x, m_y))$
$\wp_1$	$(m_x - 2, m_y - 1)$	$\frac{1}{2} d(m_x - 1, m_y) + \frac{1}{2} d(m_x, m_y)$
$\wp_2$	$(m_x - 1, m_y - 1)$	$d(m_x, m_y)$
$\wp_3$	$(m_x - 1, m_y - 2)$	$\frac{1}{2} d(m_x, m_y - 1) + \frac{1}{2} d(m_x, m_y)$

experiment. The text material consisted of 76 isolated tonal CV syllables (19 base syllables × 4 tones), formed by pairing the three prominent vowels /a,i,u/ with 11 consonants, the five fricatives and the six affricates of Mandarin Chinese, but excluding combinations that were phonologically unacceptable. The choice of these two classes was based on the research findings showing these consonants appeared as the most frequently misarticulated sounds made by the hearing-impaired Mandarin speakers [47]. Speech samples were produced by two male adult speakers, one with normal hearing sensitivity and the other with congenital severe-to-profound (> 70 dB) hearing loss. The speech of the impaired speaker was largely intelligible in sentences but often caused misunderstanding if produced in syllable forms due to prosodic deviations and misarticulated initial consonants.

Figure 4.3 presents the results of our pitch modification method for transforming F0 contours. Panels 4.3(a) and 4.3(c) are the F0 contours for the source and the target syllable /ti/ spoken with four different tones, and panel 4.3(b) is the converted F0 contour using VQ and orthogonal polynomial representation. Comparison of F0 variations as a function of time found in panel 4.3(b) with 4.3(a) clearly shows the improvements on tones 2 and 3. Our next examination focused on how the converted F0 contours were perceived in relation to those of the source. For easy judgments of the tonal categories, only syllables with one consonant class (affricate) were used, with a total of 40 tonal syllables (10 for each tone). Four male and one female adult native

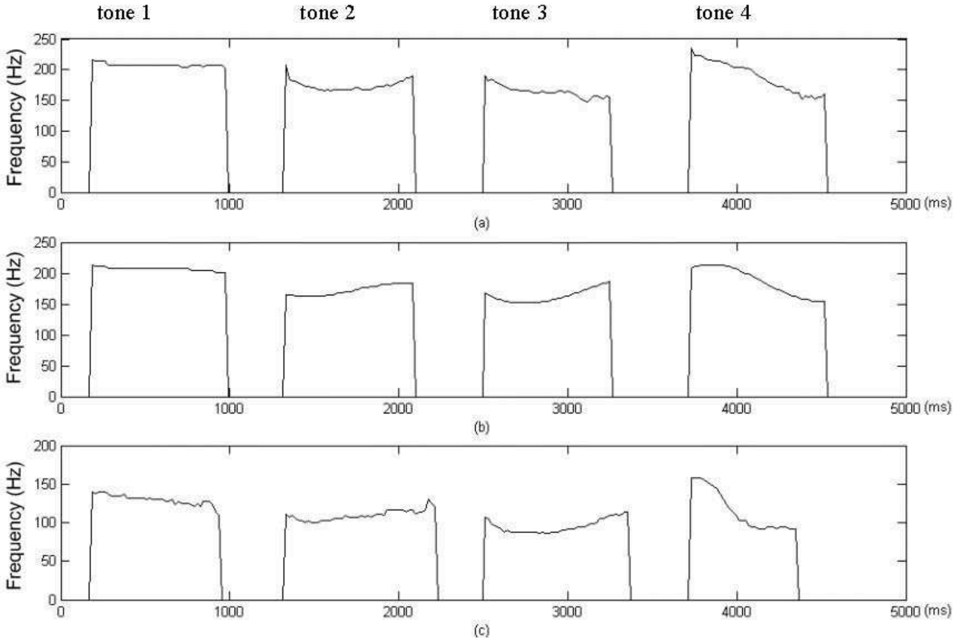


Figure 4.3: F0 contours for syllable /ti/ spoken with four different tones: (a) source speech, (b) converted speech, and (c) target speech.

Table 4.2: Confusion matrix showing tone recognition results for source syllables.

Stimulus Response	Tone 1	Tone 2	Tone 3	Tone 4
Tone 1	$\boxed{98}$	32	16	0
Tone 2	0	$\boxed{43}$	35	1
Tone 3	2	23	$\boxed{45}$	9
Tone 4	0	2	4	$\boxed{90}$

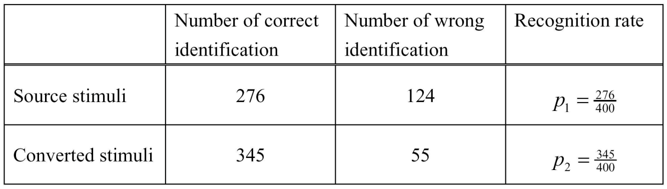
Table 4.3: Confusion matrix showing tone recognition results for converted syllables.

Stimulus Response	Tone 1	Tone 2	Tone 3	Tone 4
Tone 1	$\boxed{97}$	0	1	0
Tone 2	3	$\boxed{81}$	26	1
Tone 3	0	19	$\boxed{73}$	5
Tone 4	0	0	0	$\boxed{94}$

speakers of Mandarin Chinese, all with normal hearing status, served as the listeners. Tables 4.2 and 4.3 present the confusion matrices showing the tone recognition results for the source and the converted set, respectively. The results in each table were based on the listeners' judgments of 400 responses (40 tonal syllables× 5 listeners× 2 sessions). It is clear that the proposed system resulted in more intelligible stimuli with an average tone recognition score of 86.25%, compared with 69% for the source stimuli. The results further showed an improvement of 38% and 28% for syllables with tone 2 and tone 3, respectively.

To establish the statistical significance of these results, we calculated the P-value using a Z-test [48]. If we let p_1 and p_2 denote the recognition rates for the source and

Table 4.4: Raw data and tone recognition rates derived from Tables 2 and 3.



	Number of correct identification	Number of wrong identification	Recognition rate
Source stimuli	276	124	$p_1 = \frac{276}{400}$
Converted stimuli	345	55	$p_2 = \frac{345}{400}$

converted set, respectively, our objective was to test the null hypothesis $H_0 : p_1 \geq p_2$. Based on the statistics in Table 4.4, the Z-test yielded a small P-value ($P < 0.0002$); therefore, the null hypothesis was strongly rejected. Further evidence of improvement is seen on Figure 4.4, which shows our prosodic modification applied to continuous speech. A four-syllable utterance, containing tones 4-4-3-3, was used. According to the tone-sandhi rule, the first tone 3 should be produced with a tone 2 F0 pattern. The audio presentation, however, showed that the first tone 3 was produced more like tone 1 than the targeted tone 2. A comparison of the F0 contours for the source and the target utterances showed that the former exhibited fewer fine fluctuation details, even though the variation ranges were both within 100 Hz. Further, the first tone 4 was essentially carrying a tone 1 F0 pattern and the last tone 3 was produced with the rising part truncated. The improvement due to prosodic modification can be seen in the following areas. First, the missing falling part in the fist tone 4 and the dipping of the last tone 3 were fully restored. Second, the rising part of the first tone 3 segment was steeper in slope, making it more appropriate for the targeted tone 2. To hear audio examples of the voice conversion system, please visit the web site at http://a61.cm.nctu.edu.tw/demo.

Results of the spectral conversion were analyzed acoustically with software spectrograph to assess how closely the converted speech resembled the target speech in rendering acoustic cues for phoneme perception. The improvement for the fricatives is shown

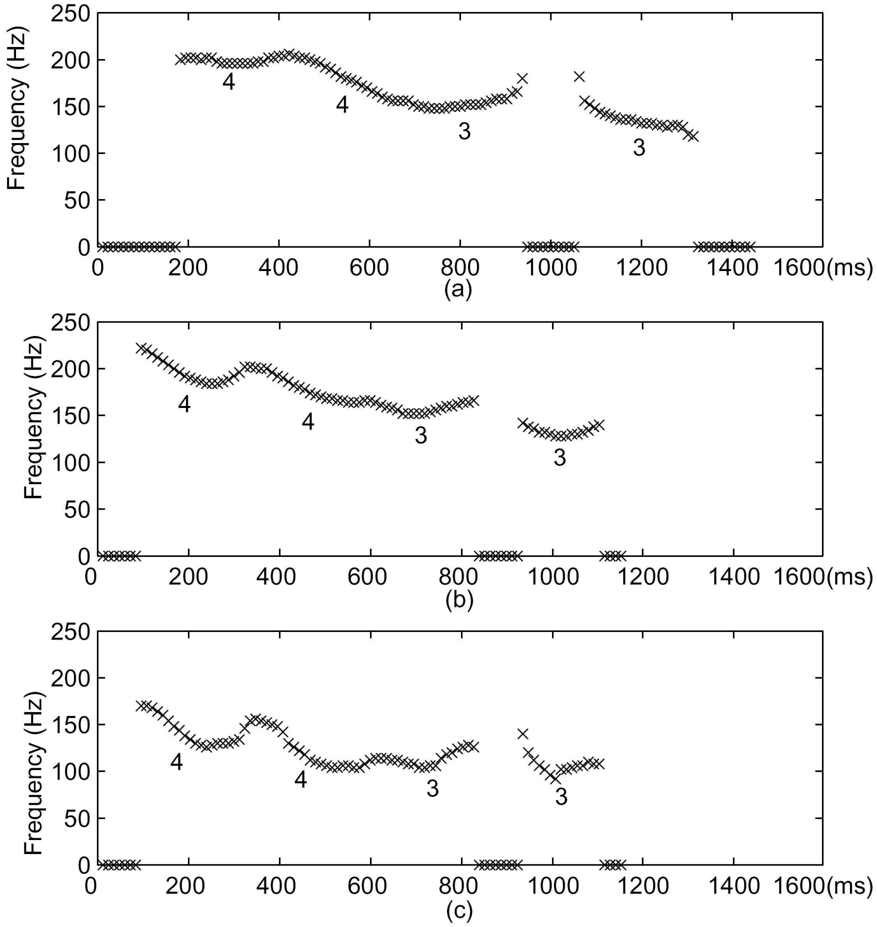


Figure 4.4: F0 contours for a four-syllable utterance /ying-4 yong-4 ruan-3 ti-3/(Application software): (a) source speech, (b) converted speech, and (c) target speech.

in three aspects: (1) lengthening of the consonant duration, (2) a less abrupt transition, or a gradual blending of the acoustic energy, near the consonant-vowel boundary, and (3) a redistribution of acoustic energy around appropriate frequency regions, such as an elevation to 3 kHz for the syllable /shu/ or to 4 kHz for the syllable /shii/. An example of such spectral differences for the syllable /shu/ is shown in Figure 4.5. Even closer spectrographic matches were obtained for the affricates, as shown in Figure 4.6 using /chii/ as an example. In normal production, affricates are stops followed by fricatives, which are individually represented on the spectrograph as a burst with its energy concentrated at higher frequencies to be blended immediately with those of the following fricative. The distorted affricate, however, was translated spectrographically into a stop that included a full voicing gap but not much of frication. Our analysis revealed that the conversion filled the gap, softened the burst, removed the low frequency energy and elevated the fricative portion to normal frequency ranges. When examined along with audio presentations, this modification also resulted in a change of the vowel percept from the erroneous, high front but lip-rounding, vowel /yu/ to the correct /i/, even though formant modification for the vowel was less apparent.

Two listening tests, preference and intelligibility, were conducted to determine whether the above spectrographic enhancement could also be realized perceptually. The five listeners for the previous tone recognition test were used. In the preference test, the listeners were asked to give their preference judgments over pairs of source vs. converted syllables. A two-alternative-forced-choice (2AFC) test paradigm was used, in which the presentation order of the two stimuli was randomized. For converted stimuli, two sets of converted syllables were used: (1) those with spectral conversion only and (2) those with combined spectral and time-scaled conversions. The results showed 62% of the 380 responses (2 stimulus sets × 19 base syllables × 5 listeners × 2 sessions) preferred spectrally modified syllables to source syllables, while 84% preferred those with combined modifications. To further validate the effect of the proposed approach, intelligibility measures were obtained for 19 base syllables before and after spectral conversion. The listeners were instructed to write down their responses using Mandarin phonetic symbols. Figure 4.7 shows comparison of the percent correct phoneme recognition scores for the source and the converted stimuli. Individual phonemes were arranged from left to right into three groups, fricative, affricate, and vowel. Recognition of vowels /a,u/ was near perfect even without the modification. In contrast, recognition for the affricates and the fricatives (with the exception of /h/) was either near or at 0%, a finding consistent with our earlier observation that these two consonant classes are frequently substituted with stops by the hearing-impaired speakers. The relatively good recognition for /h/, even for the source, could be explained by the fact that little oral modification of the glottal air source was required during articulation. With the converted stimuli, an improvement was seen in all three groups. An average increase of 47.25% was obtained for the fricatives, with /h/ counted out. The amount was further increased by 20% (=67.17%) for the affricates, with /ji, chi/ showing a total correction, making this group the phoneme class that benefited the most from our application. The vowels, despite their small improvement, were the only group showing a total correction for all its members.

Again, we considered a recognition experiment only based on the spectral conversion to perform supervised speaker adaptation for hearing-impaired speaker. The recognition system of Mandarin digit strings as the task without restricting the string length was described in Section 4.2. The source speaker and the target speaker are male. The recognition accuracy of the target speaker is 96.04% on the training tokens. After conversion, the recognition accuracy of the source speaker can be improved from 19.51% of the original speech to 36.02% of the converted speech.

4.6 Summary

This chapter presents a novel means of exploiting spectral and prosodic transformations in enhancing disordered speech. In spectral conversion, subsyllable-based GMMs were applied within the sinusoidal framework to modify the articulation-related parameters of speech. In prosodic conversion, we found the tone structure of F0 contour in Mandarin speech could be used to advantage in orthogonal polynomial representation of pitch contours. The results also suggest a new approach to time-scaling modification in which the initial part of a syllable is linearly normalized with a fixed factor, and then a DTW algorithm is used to control the time-varying scaling factor for the final part. Evaluations by objective tests and listening tests show that the proposed techniques can improve the intelligibility and naturalness of the hearing-impaired Mandarin speech.

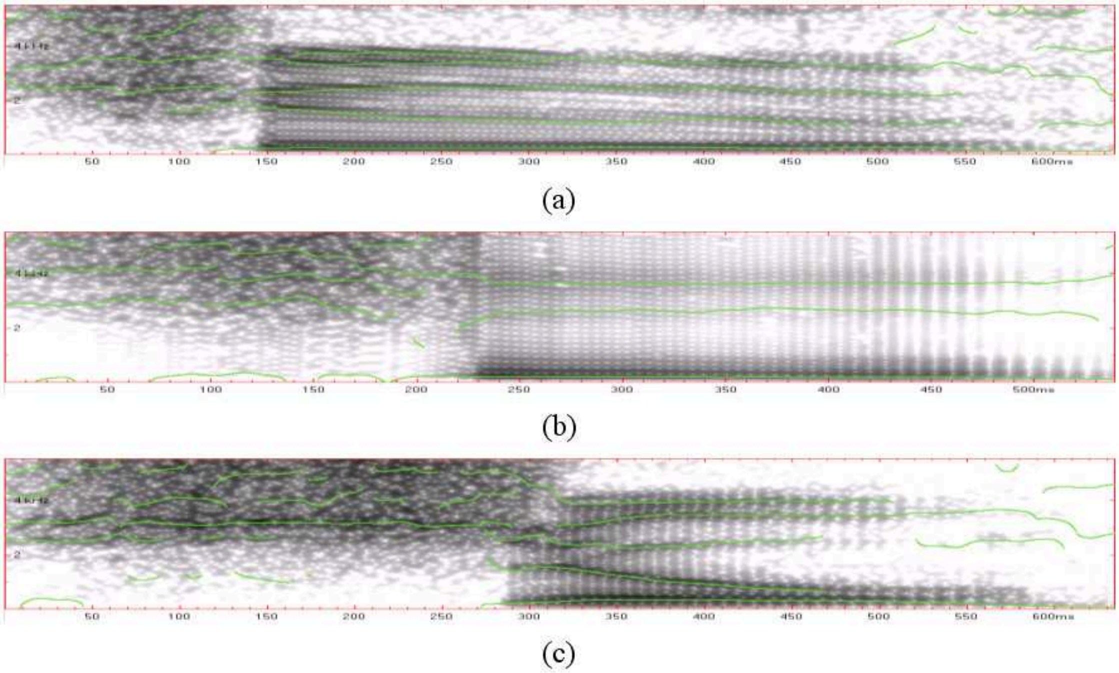
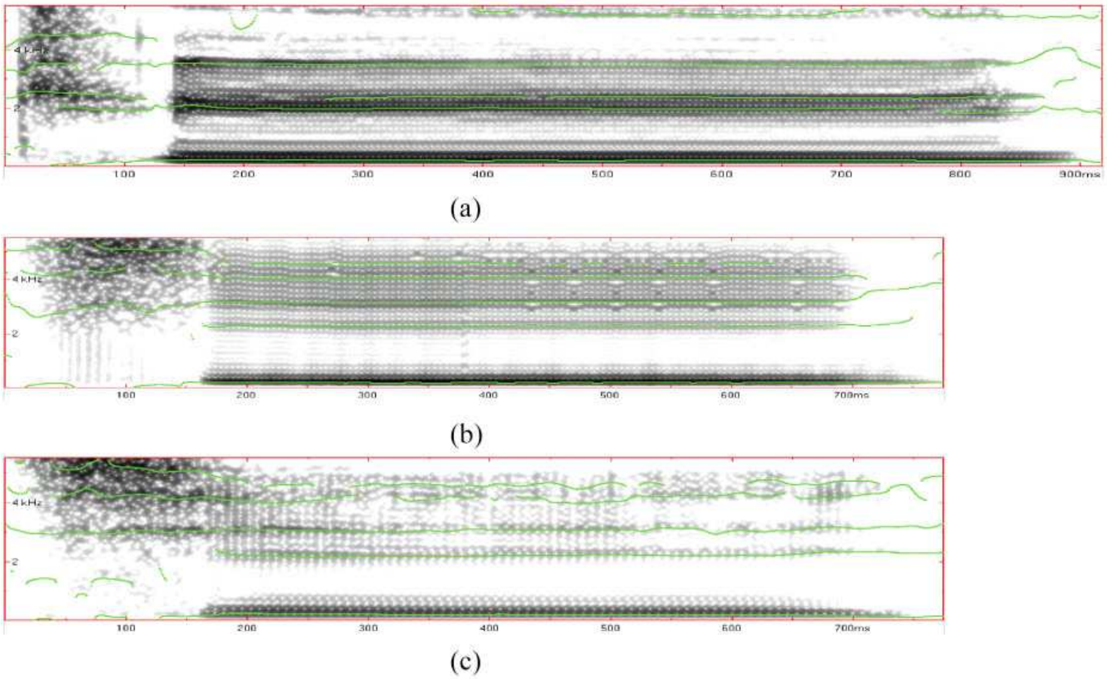


Figure 4.5: Spectrograms for syllable /shu/: (a) source speech, (b) converted speech, and (c) target speech.



(a)

(b)

(c)

Figure 4.6: Spectrograms for syllable /chii/: (a) source speech, (b) converted speech, and (c) target speech.

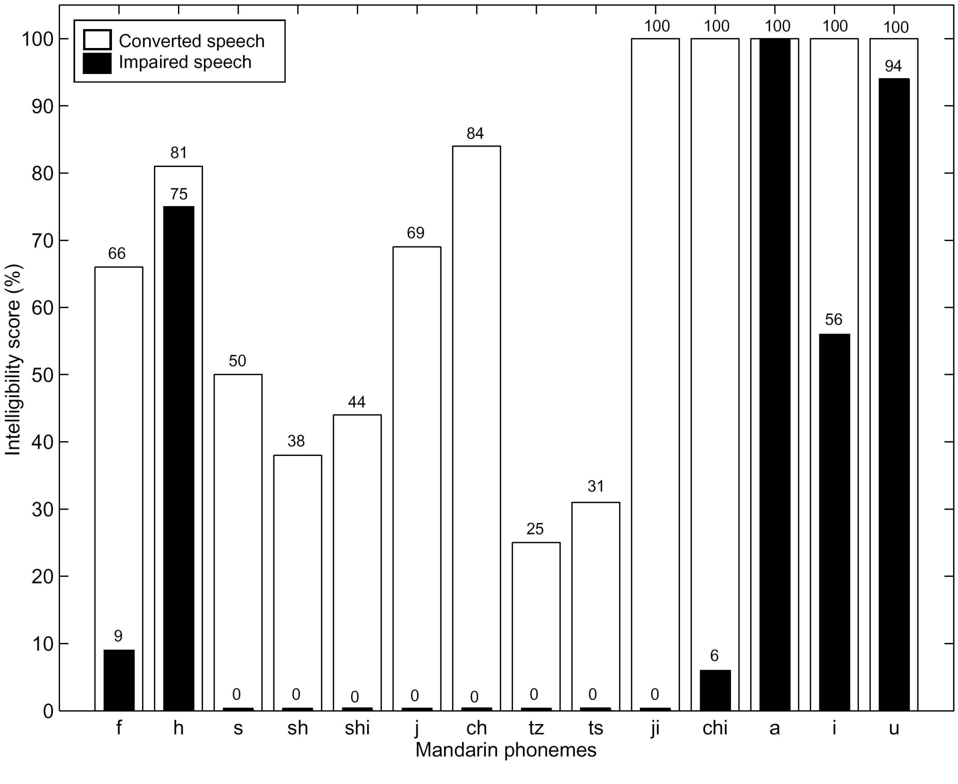


Figure 4.7: Percent correct phoneme recognition scores for source and converted speech.

Chapter 5

Conclusions and Future Work

5.1 Summary

Robust DSR systems provide substantial benefits for mobile applications, which necessitates ubiquitous access from communication networks with a guaranteed level of recognition performance. This study addressed two robustness issues of DSR systems. One is the inter-speaker variability, which is caused by the mismatches between training and testing conditions in the recognizer. The other issue is how to combat channel errors, which often leads to severe degradation in the recognition performance. The inter-speaker variability is a typical and well-known problem in ASR applications. Unlike the state-of-art approaches, we presented the voice conversion system based on feature transformation to perform speaker adaptation for hearing-impaired speaker. By taking advantage of the syllable phonetic structures of Mandarin, spectral and prosodic conversions were applied within the sinusoidal speech framework to modify the spectral envelop, pitch contour and rate of articulation. Simulation results indicated that the proposed system can achieve good adaptation performance in speech recognition applications and also improve intelligibility and naturalness of hearing-impaired Mandarin speech. In the second part of this study, we presented JSCD algorithms with increased robustness against channel errors in mobile DSR applications. Through the use of the modified-BCJR algorithm, the MAP symbol decoder which exploits the combined source and channel statistics as an a priori information is proposed. We first investigated the residual redundancies existing in the DSR features and exploited these redundancies in decoding process. Also proposed is a modified BCJR algorithm based on sectionalized code trellises which uses Gilbert channel characterization for better decoding in addition to source a priori knowledge. Simulation results indicated that the proposed decoder achieved significant improvements in recognition accuracy for DSR over wireless networks.

5.2 Future Work

Through the consideration of personalization and humanization, mobile communication devices can evolve to meet people's needs. DSR facilitates the creation of an exciting new set of applications and services combining voice and data. The proposed solution to robust DSR systems is only the beginning in the development of Human-Machine Interface service over wireless networks. Each of our proposed algorithms may be further examined to discover some possible contributions. This section briefly outlines some directions of future work.

The goal of voice conversion is to control speech individuality or add individual cues to speech processing algorithms. In this thesis, nice conversion technologies have be applied to convert voice quality from hearing-impaired speaker to normal speech. The key strategy is the detection and exploitation of characteristic features in spectral and prosodic levels. When voice personality can be more accurately characterized and exploited, more technologies can be integrated into voice-controlled services. For example, as the personal communication system becomes pervasive in mobile financial transactions and information retrieval services, the utility of speaker identification and authentication based on voice individuality increases. A speaker identification system based on Gaussian mixture models for characterizing spectral shapes can attain high identification accuracy [49]. Further, the GMM framework allows a direct integration with robust well-developed speech recognition systems. In addition, voice conversion can be applied to computer-assisted language learning. A learning system needs to provide the utility for detecting and correcting errors by mining the speech signal for information about learner's deviations from reference speakers' pronunciation. Recent research has found that a better solution for pronunciation learning should address not only the phone articulation but also the speech prosody.

In this thesis, a JSCD scheme which exploits the combined source and channel statistics as an a priori information is proposed for the channel error mitigation. The basic strategy is to exploit the large amount of residual redundancy existing in the DSR features. Similar analysis also indicated that substantial residual redundancy existed in the source-encoded speech parameters. Therefore, the proposed JSCD scheme can be applied to speech transmission systems in order to attain robust performance over wireless networks. The channel information considered in the decoding algorithm is the error statistics averaged over a training sequence. However, in the real-world communication the statistics of channel information also vary with time. Adaptively exploiting time-varying channel information is an important issue for the design of JSCD algorithms.

Bibliography

[1] D. Goddeau, W. Goldenthal, and C. Weikart, ”Deploying Speech Applications over the Web,” in *EUROSPEECH-1997*, pp. 685-688, 1997.

[2] J. C. Junqua and J. P. Haton, ”Robustness in Automatic Speech Recognition: Funamentals and Applications,” Kluwer Academic Pub, Boston, 1996.

[3] H. C. Choi and R.W. King, ”On the Use Spectral Transformation for Speaker Adaptation in HMM based Isolated-Word Speech Recognition,” *Speech Communication*, vol. 17 , pp. 13-143, 1995.

[4] C. H. Lee, ”On Stochastic Feature and Model Compensation Approaches to Robust Speech Recognition,” *Speech Communication*, vol. 25, pp. 29-47, 1998.

[5] C. H. Lee, C. H. Lin and B.H. Juang, ”A Study on Speaker Adaptation of the Parameters of Continuous Density Hidden Markov Models,” *IEEE Trans. Signal Processing*, vol. 39, pp. 806-814, May 1996.

[6] C. J. Leggetter and P. C. Woodland, ”Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density Hidden Markov Models,” *Computer Speech and Language*, vol. 9, pp. 171-185, Apr. 1995.

[7] A. Sankar and C. H. Lee, ”A Maximum-Likelihood Approach to Stochastic Matching for Robustspeech Recognition,” *IEEE Trans. Speech and Audio Processing*, vol. 4, pp.190-202, May 1996.

[8] A. Kain and M. W. Macon, "Spectral voice conversion for text-to-speech synthesis," in *Proceedings ICASSP'98*, pp. 285-288, 1998.

[9] N. Bi and Y. Qi, "Application of speech conversion to alaryngeal speech enhancement," *IEEE Trans. Speech Audio Processing*, vol. 5, pp. 97-105, 1997.

[10] M. Abe, S. Nakamura, K. Shikano and H. Kuwabara, "Voice conversion through vector quantization," in *Proceedings ICASSP'88*, pp. 655-658, 1988.

[11] Y. Stylianou, O. Cappe and E. Moulines, "Continuous probabilistic transform for voice conversion," *IEEE Trans. Speech Audio Processing*, vol. 6, pp. 131-142, 1998.

[12] R. N. Ohde and D. J. Sharf, "Phonetic Analysis of Normal and Abnormal Speech," Merrill, New York, 1992.

[13] A. Bernard and A. Alwan, "Low-bitrate distributed speech recognition for packet-based and wireless communication," *IEEE Trans. Speech and Audio Processing*, vol. 10, no. 8, pp. 570-579, 2002,

[14] C. Boulis, M. Ostendorf, E. Riskin and S. Otterson, "Graceful degradation of speech recognition performance over packet-erasure networks," *IEEE Trans. Speech and Audio Processing*, vol. 10, no. 8, pp. 580-590, 2002.

[15] A. M. Peinado, V. Sanchez, J. L. Perez-Cordoba and A. Torre, "HMM-based channel error mitigation and its application to distributed speech recognition," *Speech Communication*, vol. 41, pp. 549-561, 2003.

[16] H. U. Reinhold and I. Valentin, "Soft features for improved distributed speech recognition over wireless networks," in *Proc. Int. Conf. Spoken Language Processing*, pp. 2125-2128, Jeju Island, Korea, 2004.

[17] T. Fingscheidt and P. Vary, "Softbit speech decoding: a new approach to error concealment," *IEEE Trans. Speech and Audio Processing*, vol. 9, no. 3, pp. 240-251, 2001.

[18] L. N. Kanal and A. R. K. Sastry, "Models for channels with memory and their applications to error control," in *Proc. IEEE*, vol. 66, pp. 724-744, 1978.

[19] E. N. Gilbert, "Capacity of a burst-noise channel," *The Bell System Technical Journal*, vol. 39, pp. 1253-1265, 1960.

[20] S Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentence," *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. 28, pp. 357- 366, Aug. 1980.

[21] V. V. Digalakis, L. G. Neumeyer and M. Perakakis, "Quantization of cepstral parameters for speech recognition over the World Wide Web," *IEEE Journal on Selected Areas in Communications*, vol. 17, pp. 82-90, Jan. 1999.

[22] ETSI ES 202 212 v1.1.1. Digital speech recognition; extended advanced front-end feature extraction algorithm; compression algorithms; back-end speech reconstruction algorithm, Nov. 2003.

[23] L. S. Lee, "Voice dictation of Mandarin Chinese," *IEEE Signal Processing Magazine*, pp. 63-101, 1997.

[24] L. R. Bahl, J. Cocke, F. Jelinek and J. Raviv, "Optimal Decoding of Linear Codes for Minimizing Symbol Error Rate," *IEEE Trans. Inform. Theory*, vol. IT-20, pp. 284-287, 1974.

[25] S. Lin and D. J. Costello, "Error Control Coding," Prentice Hall, New Jersey, 2004.

[26] L. N. Kanal and A. R. K. Sastry, "Models for Channels with Memory and their Applications to Error Control," in *Proc. IEEE*, vol. 66, pp. 724-744, 1978.

[27] E. N. Gilbert, ”Capacity of a Burst-noise Channel,” *The Bell System Technical Journal*, vol. 39, pp. 1253-1265, 1960.

[28] W. Turin, ”MAP Symbol Decoding in Channels with Error Bursts,” *IEEE Trans. Inform. Theory*, vol. 47, no. 5, pp. 1832-1838, 2001.

[29] CoCentric System Studio-Referenec Design Kits, Mountain View, CA: Synopsys, Inc., 2003.

[30] J.Y. Chouinard, M. Lecours and G. Y. Delisle, ”Estimation of Gilbert's and Fritchman's Models Parameters Using the Gradient Method for Digital Mobile Radio Channels,” *IEEE Trans. Veh. Technol.*, vol. 37, no. 3, pp.158-166, Aug. 1988.

[31] R. J. McAulay and T. F. Quatieri, ”Speech analysis-synthesis based on a sinusoidal representation,” *IEEE Trans. Acoust. Speech Sig. Process.*, vol. 34, pp. 744-754, 1986.

[32] I. Hochberg, H. Levitt and M. J. Osberger, ”Speech of The Hearing Impaired: Research, Training, and Personnel Preparation,” University Park Press, Maryland 1983.

[33] R. Monsen, ”Toward Measuring How Well Hearing-impaired Children Speak,” *Journal of Speech and Hearing Research*. vol. 21, pp. 197-219, 1978.

[34] N. S. McGarr and K. S. Harris, Articulatory control in deaf speaker, in I. Hochberg, H. Levitt and M. J. Osberger (Eds.), *Speech of the Hearing Impaired*, University Park Press, Baltimore 1983.

[35] M. J. Osberger and H. Levitt, ”The Effect of Timing Errors on the Intelligibility of Deaf Children's Speech,” *J. Acoust. Soc. Amer.*, vol. 66 (5), pp. 1316-1324, 1979.

[36] B. L. Chang, ”The Perceptual Analysis of Speech Intelligibility of Students with Hearing Impairments,” *Bulletin of Special Education*, vol. 18, pp. 53-78, 2000.

[37] B. G. Lin and Y. C. Huang, ”An Analysis on The Hhearing Impaired Students' Chinese Language Abilities and its Error Patterns,” *Bulletin of Special Education*, vol. 15, pp. 109-129, 1997.

[38] X. S. Shen and M. Lin, ”A perceptual study of Mandarin tones 2 and 3,” *Language and Speech*, vol. 34(2), pp. 145-156, 1991.

[39] R. J. McAulay and T. F. Quatieri, ”Sinusoidal Coding: Speech Coding and Synthesis,” Elsevier, Amsterdam, 1995.

[40] T. F. Quatieri and R. J. McAulay, ”Shape Invariant Time-scale and Pitch Modification of Speech,” *IEEE Trans. Signal Processing*, vol. 40 (3), pp. 497-510, 1992.

[41] A. V. Oppenhein and R. W. Schafer, ”Discrete-time Signal Processing,” Prentice Hall, New Jersey, 1989.

[42] Y. Stylianou, O. Cappe and E. Moulines, ”Continuous probabilistic transform for voice conversion,” *IEEE Trans. Speech Audio Processing*, vol. 6, pp. 131-142, 1998.

[43] A. Dempster, N. Laird and D. Rubin, ”Maximum Likelihood from Incomplete Data via the EM Algorithm,” *J. Royal Stat. Soc.*, vol. 39, pp. 1-38, 1977.

[44] S. H. Chen and Y. R. Wang, ”Vector quantization of pitch information in Mandarin speech,” *IEEE Trans. Communications*, vol. 38, pp. 1317-1320, 1990.

[45] Y. Linde, A. Buzo and R. M. Gray, ”An Algorithm for Vector Quantizer Design,” *IEEE Trans. Communications*, vol. 28, pp. 84-95, 1980.

[46] L. Rabiner and B. H. Juang, ”Fundamentals of Speech Recognition,” Prentice Hall, New Jersey, 1993.

[47] P. C. Lee, "A Study on Acoustic Characteristic of Mandarin Affricates of Hearing-impaired Speech," *Bulletin of special Eduation and Rehabilitation*, vol. 7, pp. 79-112, 1999.

[48] R. A. Johnson and G. K. Bhattacharyya, "Statistics: Principles and Methods," John Wiley and Sons, New York, 1996.

[49] D. A. Reynolds and R. C. Rose, "Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models," *IEEE Trans. Speech and Audio Processing*, vol. 3, pp. 72-83, 1995.