Enhancement of Throat Microphone Recordings Using Gaussian Mixture Model Probabilistic Estimator

by

Mehmet Ali Tuğtekin Turan

A Thesis Submitted to the
Graduate School of Sciences and Engineering
in Partial Fulfillment of the Requirements for
the Degree of

Master of Science

in

Electrical and Electronics Engineering

Koç University

August, 2013

Koç University

Graduate School of Sciences and Engineering

This is to certify that I have examined this copy of a master's thesis by

Mehmet Ali Tuğtekin Turan

and have found that it is complete and satisfactory in all respects,
and that any and all revisions required by the final
examining committee have been made.

Committee Members:

_____

Assoc. Prof. Engin Erzin

_____

Prof. Murat Tekalp

_____

Prof. Levent Arslan

Date: _____

In human life, you will find players of religion until the knowledge and proficiency in religion will be cleansed from all superstitions, and will be purified and perfected by the enlightenment of real science.

Mustafa Kemal Atatürk

ABSTRACT

The throat microphone is a body-attached transducer that is worn against the neck. It captures the signals that are transmitted through the vocal folds, along with the buzz tone of the larynx. Due to its skin contact, it is more robust to the environmental noise compared to the acoustic microphone that picks up the vibrations through air pressure, and hence the all interventions. The throat speech is partly intelligible, but gives unnatural and croaky sound. This thesis tries to recover missing frequency bands of the throat speech and investigates envelope and excitation mapping problem with joint analysis of throat- and acoustic-microphone recordings. A new phone-dependent GMM-based spectral envelope mapping scheme, which performs the minimum mean square error (MMSE) estimation of the acoustic-microphone spectral envelope, has been proposed. In the source-filter decomposition framework, we observed that the spectral envelope difference of the excitation signals of throat- and acoustic-microphone recordings is an important source of the degradation in the throat-microphone voice quality. Thus, we also model spectral envelope difference of the excitation signals as a spectral tilt vector, and propose a new phone-dependent GMM-based spectral tilt mapping scheme to enhance throat excitation signal. Experimental evaluations are performed to compare the proposed mapping scheme using both objective and subjective evaluations. Objective evaluations are performed with the log-spectral distortion (LSD) and the wide-band perceptual evaluation of speech quality (PESQ) metrics. Subjective evaluations are performed with A/B pair comparison listening test. Both objective and subjective evaluations yield that the proposed phone-dependent mapping consistently improves performances over the state-of-the-art GMM estimators.

ÖZETÇE

Gırtlak mikrofonu, ses tellerindeki titreşimi gırtlaktan gelen sinyallerle beraber ileten ve kullanan kişinin boynuna taktığı insan bedeniyle temas eden bir mikrofon türüdür. Bu bağlantı sayesinde, titreşimleri havadan alan akustik mikrofonlara nazaran gürültü gibi çevresel etmenlere karşı daha gürbüz bir iletişim sağlar. Gırtlak mikrofonu ile kaydedilen sesler kısmen de olsa anlaşılmasına rağmen, doğal olmayan ve kulağı rahatsız edici bir yapıdadır. İşte bu çalışma gırtlak mikrofonlarındaki üretilemeyen frekans aralıklarını geri kazanabilmeyi amaçlarken aynı zamanda sesin kaynak ve süzgeç kısımlarını doğru tahmin edebilme sorununu, gırtlak ve akustik kayıtları müşterek bir şekilde çözümleyerek irdelemektedir. Bu bağlamda, ortalama kare hatasını en aza indirerek, ses birimlerine bağlı Gauss karışım modeli tabanlı bir kestirici sistemi öne sürülmüştür. Kaynak–süzgeç ayrıştırması çerçevesinde, gırtlak ve akustik süzgecinin görüngesel farklılıklarının, gırtlak mikrofonundan gelen ses kalitesini düşüren önemli bir etmen olduğunu gözlemledik. Bu sebepten ötürü, yukarıda bahsedilen farkı görüngesel eğim vektörü olarak modelleyip, gırtlak süzgecini iyileştirici bir sistemi ayrıca öne sürdük. Ortaya konulan sistemlerin katkılarını yorumlayabilmek için hem nesnel hem de öznel deneyler tasarladık. Nesnel deneyler, logaritmik görünge tahribatı ve ses kalitesinin algısal değerlendirilmesi kıstasları üzerinden incelendiler. Bununla birlikte, öznel değerlendirmeler ise A/B eş karşılaştırma deneyi şeklinde tatbik edildi. Hem nesnel hem de öznel deneyler gösterdi ki öne sürülen ses birimi tabanlı kestirimler, halihazırda bulunan Gauss karışım modeli tabanlı kestirimlere göre tutarlı bir şekilde iyileştirmeler sağlamaktadır.

TABLE OF CONTENTS

List of Tables viii

List of Figures ix

Nomenclature x

Chapter 1: Introduction 1

1.1 Non-Acoustic Sensors 1

1.2 Scope 4

1.3 Related Work 5

1.4 Contributions 9

1.5 Organization 10

Chapter 2: Enhancement System 11

2.1 System Overview 11

2.2 Source-Filter Separation 12

2.3 Gaussian Mixture Model (GMM) 17

2.4 GMM-Based Probabilistic Mapping 18

2.5 Enhancement Framework 19

2.6 Spectral Envelope Enhancement 20

2.7 Excitation Enhancement 21

Chapter 3: Experimental Evaluations 24

3.1 Observations on Throat-Microphone Speech Attributes 24

3.2 Objective Evaluations 26
3.2.1 Envelope Enhancements 28
3.2.2 Excitation Enhancements 30
3.3 Subjective Evaluations 33
3.3.1 Envelope Enhancements 34
3.3.2 Excitation Enhancements 35

Chapter 4: Conclusion **37**

Bibliography **38**

LIST OF TABLES

3.1 The average LSD scores of the data M1 between throat- and acoustic-microphone spectrums for different phonetic attributes with relative occurrence frequencies in the test database. 25

3.2 The Turkish METUbet phonetic alphabet with classification into 8 articulation attributes. 27

3.3 The average LSD and PESQ scores for different mapping schemes for enhancement of the throat-microphone recordings. 28

3.4 The average PESQ scores for different mapping schemes using acoustic residual. 29

3.5 The average PESQ scores for evaluation of the targeted excitation and filter enhancement strategies. 30

3.6 The average PESQ scores for different excitation and spectral envelope mapping schemes. 33

3.7 The average preference results of the subjective A/B pair comparison test for envelope mapping 34

3.8 The average preference results of the subjective A/B pair comparison test for excitation mapping 35

LIST OF FIGURES

1.1 Bone-Conducting Headset [1] 2
1.2 Non-Audible Murmur (NAM) Microphone [2] 3
1.3 Throat Microphone (TM) [3] 4

2.1 Block Diagram of Enhancement System 12
2.2 Source-Filter Model of Speech 13
2.3 Framework of the GMM 17

3.1 Sample spectrograms of (a) AM excitation, (b) enhanced TM excitation with the original spectral tilt, (c) enhanced TM excitation with the estimated spectral tilt, and (d) TM excitation. 32

NOMENCLATURE

TM	:	Throat-Microphone
AM	:	Acoustic-Microphone
GMM	:	Gaussian Mixture Model
$MMSE$	:	Minimum Mean Square Error
LSD	:	Log-Spectral Distortion
$PESQ$	:	Perceptual Evaluation of Speech Quality
NAM	:	Non-Audible Murmur
LSF	:	Line Spectrum Frequency
LPC	:	Linear Predictive Coding
HMM	:	Hidden Markov Model
VQ	:	Vector Quantization
EM	:	Expectation Maximization
SM	:	Soft Mapping
HM	:	Hard Mapping
POF	:	Probabilistic Optimum Filter
$PDSM$	:	Phone-Dependent Soft Mapping
$PDHM-G$	:	Phone-Dependent Hard Mapping with the GMM Classifier
$PDHM-M$	:	Phone-Dependent Hard Mapping with the HMM Phone Recognition
$PDHM-T$	:	Phone-Dependent Hard Mapping with the True Phone Class

Chapter 1

INTRODUCTION

1.1 Non-Acoustic Sensors

Multi-sensor configurations have recently been applied to speech enhancement problem which mainly aims to obtain high performance quality of speech. Environmental effects such as background noise or wind turbulence motivated researchers to use different mediums such that speech can be spread other than by means of air. Other mediums, such as human tissue, infrared ray, light wave, and laser also can be used to detect the non-air conducted speech or acoustical vibrations. A few type of non-acoustic sensors, i.e. sensors that don't not use the air, are developed due to this reason, however, their application are limited since detection materials are usually difficult to obtain [4].

The traditional acoustic microphones use air as a medium of sound conduction. Thus, it is ineffective in extreme conditions such as acoustic noise. On the other hand, piezoelectric transducers in non-acoustic sensors can pick up voice signals by absorbing the vibrations generated from human body. Thus, all of the non-acoustic sensors are insensitive to environmental conditions. However, they only capture low-frequency portion of sound and may distort speech signals due to their low energy static noise. This causes a reduced frequency bandwidth unfortunately. Another advantage of them is the ability to reveal certain speech attributes lost in the noisy acoustic signal; for example, low-energy consonant voice bars and nasality excitation. These sensors provide measurements of functions of the glottal excitation and, more generally, of the vocal tract articulator movements that are relatively resistant to acoustic disturbances [5].

One of the early studies about non-acoustic sensors have been originated from human auditory system. Since hearing is actualized both air and bone conduction pathways, a playback from recorded speech is perceived different to us. Likewise, bone conducting microphones, see Figure 1.1, can catch signals from the inner ear through the bones of the skull [1]. It has many advantages over air conduction owing to its robustness to noise. For example, it can be used for getting information whether the user is talking or not [6]. It is commonly used in military environments such as headphone-based communication interfaces. Bone conduction sensors provides an effective transmission without interfering with hearing protection devices. In quiet environments, the soldier could receive radio communications through bone conduction without obscuring the ears, thereby maintaining full awareness of the surrounding acoustic environment [7]. However, one of the main disadvantages of current bone conduction systems is that they are restricted to single channel operation. Due to this reason, it is typically used for enhancements as a supplementary speech source. In [8], combining the two channels from the air- and bone- conductive microphone, it is possible to remove background speech.

Figure 1.1: Bone-Conducting Headset [1]

.

Another favored non-acoustic sensor is Non-Audible Murmur (NAM) microphone that is attached behind the speaker's ear (see Figure 1.2). The specialty of NAM is

the ability of capturing very quietly uttered speech that cannot be heard by listeners through human tissue [9]. Since it captures inaudible speech produced without vibration of glottis, it is difficult to identify the differences between whisper and NAM speech. The principle behind NAM sensor is based on medical stethoscope used for monitoring internal sounds of human body. Similarly the NAM microphone is mainly used for privacy purposes while communicating with speech recognition engines. The NAM users don't pay so much effort owing to quiet utterance so this provides communication without hearing from others. On the other hand, it can be useful for diseased people who have physical difficulties in speech [10].

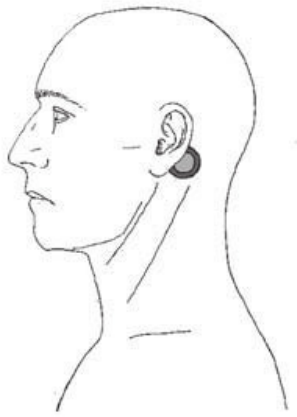

Figure 1.2: Non-Audible Murmur (NAM) Microphone [2]

Another specialized non-acoustic sensor is throat microphone (TM) that have been used in military applications and radio communication for several years (see Figure 1.3). It can capture speech signals in the form of vibrations and resonances of vocal cords through skin-attached piezoelectric sensors [11]. Since the signals are acquired from the throat, they have lower bandwidth speech signals compared to acoustic-microphone (AM) recordings. Like other non-acoustic sensors, the TM recordings are significantly more robust to environmental noise conditions, however they suffer from the perceived speech quality [12]. Since TM recordings are strongly robust and highly correlated with the acoustic speech signal, they are attractive candidates for robust speech recognition

applications under adverse noise conditions, such as airplane, motorcycle, military field, factory or street crowd environments. Likewise with the NAM, it can be used for patients who have lost their voices due to injury or illness, or patients who have temporary speech loss after a tracheotomy. One of the biggest problem is the quality degradation of TM speech caused by deficiency of oral cavity transmission such as lack of lip radiation. This problem can be handled with synchronous analysis of AM and TM speech. Moreover, the TM is useful in terms of its distinct formant-like structures that serve as acoustic cues that can be used to resolve the highly confusable voiced phones into classes based on the place of articulation [11]. Since the TM conveys more information about the AM speech characteristics among other non-acoustic sensors, we aim to improve its perceived quality by mapping its spectral features in this thesis.

Figure 1.3: Throat Microphone (TM) [3]

1.2 Scope

It is emphasized in previous section that the TM is lack of perceptual quality because of its muffled speech. In the literature, there are a few attempts which aim to enhance the quality of throat-only speech. The published works mostly deal with robust speech recognition by means of the TM in the extreme environments. In this research, we target to enhance the naturalness and the intelligibility of the TM speech by mapping not only its filter but also its excitation spectra closer to the one that belongs to

the AM speech via Gaussian mixture model (GMM) probabilistic estimators. This mapping is trained using simultaneously recorded acoustic- and throat-microphone speech and is formulated by context independent and dependent methods. Moreover, the outcome of this work may expand to a linear system that uses the direct filtering so that it can be speaker independent. Thus, all of the proposed schemes try to improve the understandability of throat-only speech.

1.3 Related Work

In one of the early studies, Viswanathan et al. presented a two sensor system involving an accelerometer mounted on the speaker's throat and a noise-canceling microphone is located close to the lips [13]. Close talking first- and second-order differential microphones are designed to be placed close to the lips where the sound field has a large spatial gradient and the frequency response of the microphone is flat. Second-order differential microphones using a single element piezoelectric transducer have been suggested for use in very noisy environment of aircraft communication systems to enhance a noisy signal for improved speech recognition.

A device that combines a close-talk and a bone-conductive microphone is proposed by the Microsoft research group for speech detection using a moving-window histogram [8]. They tried to handle non-stationary noises in both automatic speech recognition and audio enhancement. Note that the bone sensor is highly insensitive to ambient noise and provides robust speech activity detection. They showed that such devices can be used to determine whether the speaker is talking or not, and furthermore, the two channels can be combined to remove overlapping noise. It works by training a piecewise linear mapping from the bone signal to the close-talk signal. One drawback of this approach is that it requires training for each speaker. This problem can be solved with a technique called *Direct Filtering* that does not require any training. It is based on learning mappings in a maximum likelihood framework and investigated in [14]. Later, direct filtering is improved to deal with the environmental noise leakage into the bone sensor and with the teeth-clack problem [15].

The use of non-acoustic sensors in multi-sensory speech processing has been studied for speech enhancement, robust speech modeling and improved speech recognition [5, 12, 16, 17]. Multi-sensory speech processing for noisy speech enhancement and improved noise robust speech recognition are discussed in [16, 18, 19]. In these works, Subramanya et al. proposed an algorithm based on the SPLICE technique and a speech detector based on the energy in the bone channel. In another multi-sensory study, speech recorded from throat and acoustic channels is processed by parallel speech recognition systems and later a decision fusion yields robust speech recognition to background noise [20]. Due to the approximation of the sensor to the voice source, the signal is naturally less exposed to background noise. In [20], Dupont et al. proposed to use the information from both signals by combining the probability vectors provided by both models.

Graciarena et al. proposed estimation of clean acoustic speech features using the probabilistic optimum filter (POF) mapping with combined throat and acoustic microphone recordings [21]. The POF mapping is a piecewise linear transformation applied to noisy feature space to estimate the clean ones [22]. It maps the temporal sequence of noisy mel-cepstral features from the standard and the throat microphone. Thus, this mapping allows for an optimal combination of the noisy and the throat speech. In [12], Erzin developed a framework to define a temporal correlation model between simultaneously recorded throat- and acoustic-microphone speech. This framework aims to learn joint sub-phone patterns of throat and acoustic microphone recordings that define temporally correlated neighborhoods through a parallel branch hidden Markov model (HMM) structure. The resulting temporal correlation model is employed to estimate acoustic features, which are spectrally richer than throat features, from throat features through linear prediction analysis. The throat and the estimated acoustic microphone features are then used in a multi-modal speech recognition system.

Non-acoustic sensors can reveal speech attributes that are lost in the noisy acoustic signal such as, low-energy consonant voice bars, nasality, and glottal excitation. Quatieri et al. investigate methods of fusing non-acoustic low-frequency and pitch

content with acoustic-microphone content for low-rate coding of speech [5]. By fusing non-acoustic low-frequency and pitch content with acoustic-microphone content, they achieved significant intelligibility performance gains using the diagnostic rhyme test across a variety of environments.

Although throat-microphone recordings are robust to acoustic noise and reveal certain speech attributes, they often lack naturalness and intelligibility. There have been a few attempts in the literature that improve the perceived speech quality of non-acoustic sensor recordings. A neural network based mapping of the speech spectra from throat-microphone to acoustic-microphone recordings has been investigated in [11]. This neural network is used to capture the speaker-dependent functional relationship between the feature vectors, i.e. cepstral coefficients, of the speech signals. Moreover, speech spectra mapping techniques have been also studied extensively for the artificial bandwidth extension of telephone speech [23, 24]. This method aims to estimate wide-band speech (50 Hz - 7 kHz) from narrow-band signals (300 Hz - 3.4 kHz). Applying the source-filter model of speech, many existing algorithms estimate vocal tract filter parameters independently of the source signal. In another study [25], the transfer characteristics of bone-conducted and acoustic-microphone speech signals are modeled as dependent sources, and an equalizer, which is trained using simultaneously recorded acoustic and bone-conducted microphone speech, has been investigated to enhance bone-conducted speech.

Speech enhancement of non-acoustic sensor recordings also employs techniques used for voice conversion [26, 27] and artificial bandwidth extension [23, 24] to improve naturalness and intelligibility of the speech signal. One widely used framework for enhancement of the non-acoustic sensor recordings is the source-filter decomposition, which breaks down the problem into two, namely the enhancement of the excitation (source) and the spectral envelope (filter).

Enhancement of the spectral envelope has been both studied for the speech conversion and the artificial bandwidth extension problems. Stylianou et al. [26] presented one of the early works on continuous probabilistic mapping of the spectral

envelope for the voice conversion problem which is simply defined as modifying the speech signal of one speaker (source speaker) so that it sounds like be pronounced by a different speaker (target speaker). Their contribution includes the design of a new methodology for representing the relationship between two sets of spectral envelopes. Their proposed method is based on the use of a Gaussian mixture model of the source speaker spectral envelopes. The conversion itself is represented by a continuous parametric function which takes into account the probabilistic classification provided by the mixture model. Later Toda et al. [27] improved the continuous probabilistic mapping by incorporating not only static but also dynamic feature statistics for the estimation of a spectral parameter trajectory. Furthermore, they tried to deal with the over-smoothing effect by considering a global variance feature of the converted spectra.

Enhancement of the excitation has been studied on domain specific problems and not as widely as the enhancement of spectral envelope. Recently, conversion methodologies from NAM to acoustic and whispered speech have been developed to improve voice quality and intelligibility of NAM speech [28]. In [28], spectral and excitation features of acoustic speech are estimated from the spectral feature of NAM. Since NAM lacks fundamental frequency information, a mixed excitation signal is estimated based on the estimated fundamental frequency and aperiodicity information from NAM. The converted speech reported to suffer from unnatural prosody because of the difficulty of estimating the fundamental frequency of normal speech. In another study [25], the transfer characteristics of bone-conducted and acoustic-microphone speech signals are modeled as dependent sources, and an equalizer, which is trained using simultaneously recorded acoustic and bone-conducted microphone speech, has been investigated to enhance bone-conducted speech. Since the transfer function of the bone-conduction path is speaker and microphone dependent, the transfer function should be individualized for effective equalization. Then, Kondo et al. [25] propose a speaker-dependent short-term FFT based equalization with extensive smoothing. In the bandwidth extension framework, the extension of the excitation signal has

been performed by modulation, which attains spectral continuation and a matching harmonic structure of the baseband [23]. In other words, this method guarantees that the harmonics in the extended frequency band always match the harmonic structure of the baseband. Moreover, their pitch-adaptive modulation reacts quite sensitive to small errors of the estimate of the pitch frequency.

1.4 Contributions

In this thesis, we have the following contributions over state-of-art techniques that are investigated in [11, 23–25]:

- The main contribution of this work is the context-dependency of estimations, which is set at the phone level. We observe significant improvements when the true phone-context is available for the both envelope and excitation mappings. Based on this observation, we investigate some phone-dependent mapping schemes in the presence of predicted phone-context.

- We introduce a new phone-dependent GMM-based spectral envelope mapping scheme to enhance TM speech using joint analysis of TM and AM recordings. The proposed spectral mapping scheme performs the minimum mean square error (MMSE) estimation of the AM spectral envelope within the phone class neighborhoods. Objective and subjective experimental evaluations indicate that the phone-dependent spectral mapping yields perceivable improvements over the state-of-the-art context independent schemes.

- We also observed that the spectral envelope difference of the excitation signals of TM and AM recordings is an important source of the degradation for the TM voice quality. Thus, we model spectral envelope difference of the excitation signals as a spectral tilt vector, and propose a new context-dependent probabilistic spectral tilt mapping scheme based on MMSE estimation. We consider that

incorporating temporal dynamics of the spectral tilt to the probabilistic mapping expectedly attain further improvements for TM speech enhancement.

1.5 Organization

The organization of this thesis as follows: In chapter 2 we introduce the proposed throat-microphone speech enhancement system. By doing this, we first review the well-known source-filter model of human speech system and try to associate the proposed mappings to the statement of this theory. Chapter 3 discusses the experimental evaluations using both objective and subjective results. Finally, chapter 4 includes the conclusion and future direction of this study.

Chapter 2

ENHANCEMENT SYSTEM

2.1 System Overview

We start to learning part by source-filter separation of TM and AM speech. We use line spectral frequency (LSF) features to represent envelope spectra (filter) of the speech signal as training features, the prediction coefficients are firstly converted to the LSFs. We use Gaussian mixtures as probabilistic estimator model that is a parametric probability density function represented as a weighted sum of Gaussian component densities. The discourse about this model and our proposed modifications on it are discussed in Section 2.4.

For the envelope mapping, joint distribution of AM and TM spectral envelopes are modeled and we define tilt features $D(b)$ based on the spectral envelope difference of the TM and AM excitation signals for the enhancement of source (excitation). We also use cepstral feature vectors $C_T(n)$ to constitute the observable part of the excitation mapping. At the end, GMM-based training is applied to the time-aligned TM and AM features. Details about the calculations of these features are comprehensively examined in Section 2.5 – 2.7.

In the test stage, the throat-microphone test recordings are separated into source $R_T(z)$ and filter $W_T(z)$ through linear prediction analysis. The estimated acoustic filter $\hat{W}_A(z)$ is extracted from the throat filter $W_T(z)$ and the estimated acoustic source $\hat{R}_A(z)$ is computed using the throat cepstral feature vectors $C_T(n)$ and the throat filter $W_T(z)$ via different mapping schemes based on the minimum mean square error (MMSE) approach. Then, the enhanced throat-microphone recordings are synthesized using these estimated source and filter. The summary of whole system is depicted in Figure 2.1.

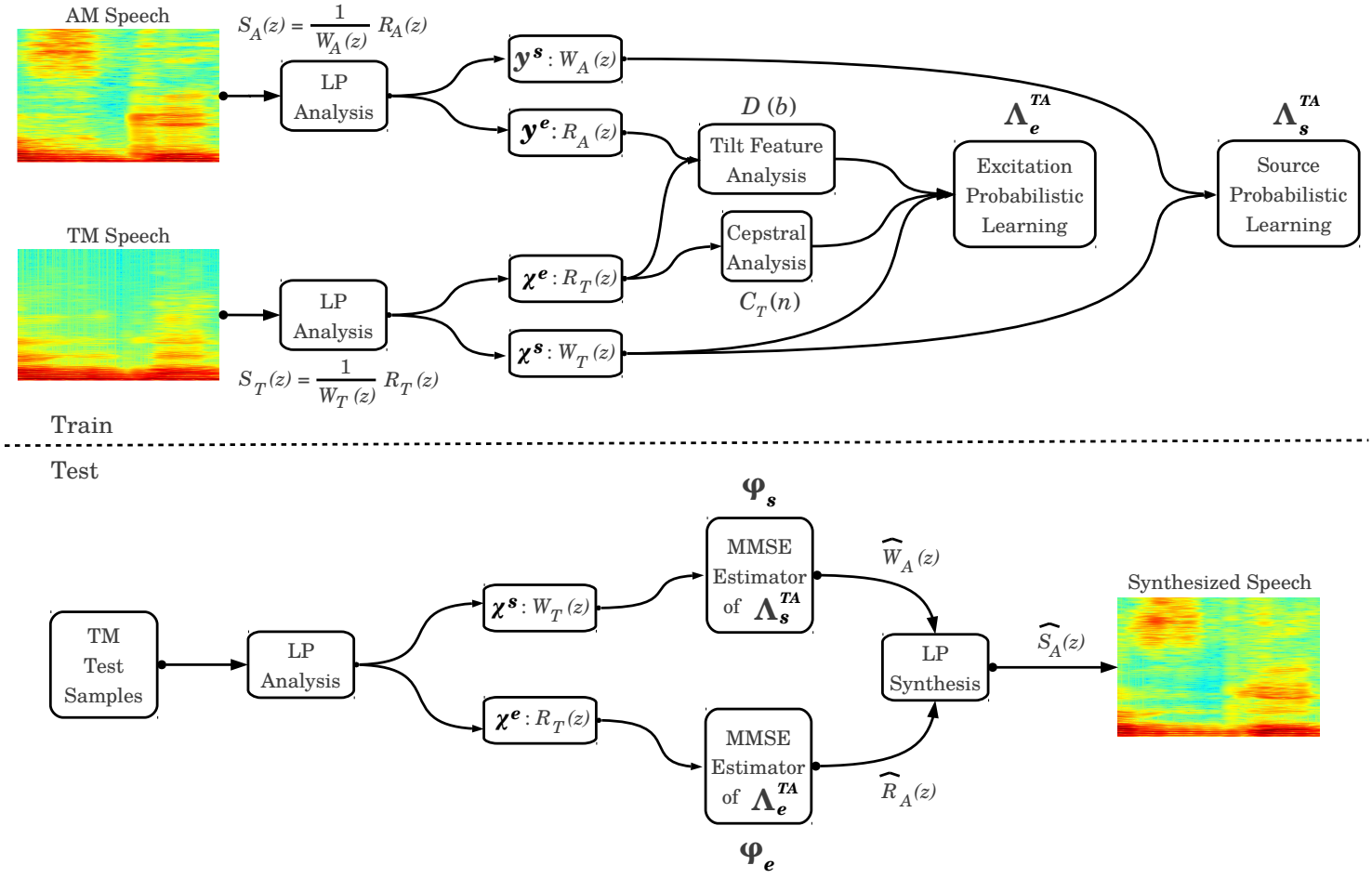

Figure 2.1: Block Diagram of Enhancement System

2.2 Source-Filter Separation

In 1960, Gunnar Fant, from Royal Institute of Technology (KTH), observed that the glottis and the vocal tract are totally distinct and both of them can be designed independently of each other. The *source-filter model* has been introduced based on the fact that speech is produced by an excitation signal generated in our throat, which is modified by resonances produced by different shapes of our vocal, nasal and pharyngeal tracts [29]. In other words, the speech signal can be decomposed as a source passed through a linear time-varying filter where the source represents the air flow at the vocal cords, and the filter represents the resonances of the vocal tract that changes over time (see Figure 2.2).

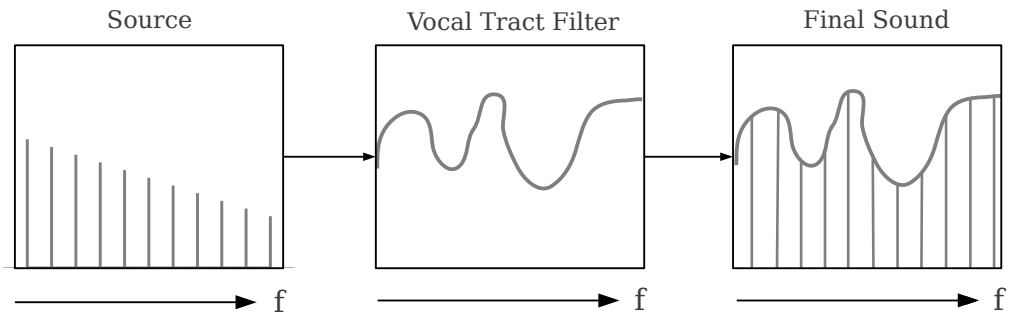

Figure 2.2: Source-Filter Model of Speech

Speech sounds are generally divided into two main groups, namely voiced and unvoiced. Voiced sounds such as vowels, liquids, glides and nasals are produced when the vocal tract is excited by air pressure due to opening and closing of the vocal cords. On the other hand, unvoiced sounds are produced by creating turbulent air flow, which acts as a random noise excitation of the vocal tract. According to source-filter model of speech, the excitation of a voiced sound is a quasi-periodic sequence of discrete glottal pulses whose fundamental frequency determines the perceived pitch of the voice whereas its unvoiced counterpart behaves like discrete-time noise signal with flat spectrum.

The question about the estimation of speech parameters are solved by means of linear predictive (LP) analysis. It simply states that the current features can be predicted using weighted sum of the past ones. By doing this, it analyses the speech by estimating its formants, resonances of the vocal tract, and then eliminates its effect from the speech signal. The process of removing the formants is called inverse filtering, and the remaining signal is called the residue [30]. Then, LP analysis realizes the reverse of this process to create a source signal, and uses the formants to construct a filter and run the source through this filter which results in actual speech. More clearly, this linear system is described by an all-pole filter of the form [31]:

$$H(z) = \frac{S(z)}{E(z)} = \frac{1}{1 - \sum_{k=1}^{p} a_k z^{-k}} \tag{2.1}$$

In this linear system, the speech samples $s[n]$ are related to the excitation $e[n]$ by the difference equation

$$s[n] = \sum_{k=1}^{p} a_k s[n-k] + e[n]. \tag{2.2}$$

A linear predictor with prediction coefficients, α_k, is defined as a system whose output is

$$\tilde{s}[n] = \sum_{k=1}^{p} \alpha_k s[n-k]\,, \tag{2.3}$$

and the prediction error, defined as the amount by which $\tilde{s}[n]$ fails to exactly predict sample $s[n]$ is

$$d[n] = s[n] - \tilde{s}[n] = s[n] - \sum_{k=1}^{p} \alpha_k s[n-k]. \tag{2.4}$$

It follows that the prediction error sequence is the output of an FIR linear system whose system function is

$$A(z) = 1 - \sum_{k=1}^{p} \alpha_k z^{-k} = \frac{D(z)}{S(z)} \tag{2.5}$$

It can be clearly seen that the speech signal obeys the source-filter model, and if $\alpha_k = a_k$ and $d[n] = e[n]$. Thus, the prediction error filter $A(z)$ becomes an inverse filter for the system $H(z)$ that is

$$H(z) = \frac{1}{A(z)} \tag{2.6}$$

After all these equations, a simple question arises about the determination of the predictor coefficients $\{\alpha_k\}$ from the speech [32]. The ubiquitous solution is to find a set that minimizes the mean square error of the speech waveform. There are some important motivations behind this approach, such as the predictor coefficients that comes from mean-squared minimization are identical to the coefficients of difference equation in (2.2). Moreover, if α_k is equal to a_k and $d[n]$ is equal to $e[n]$, $d[n]$ should be almost equal to impulse train except at isolated samples spaced by the current pitch period [32]. Eventually, one approach to computing the prediction coefficients is based

on the *covariance method* that is direct solution of well-known Yule-Walker equations. However, the most widely used method is called *autocorrelation method* because the covariance function, i.e. Yule-Walker equations, has no specific range [31]. In other words, we can use a analysis window in the autocorrelation method which provides zero prediction error outside the window interval [30]. Briefly, the latter method is based on the short-time autocorrelation function which is the inverse discrete Fourier transform of the magnitude-squared of the short-time Fourier transform of the windowed speech signal [32].

Since the important parameter of this linear model is the prediction coefficients $\{\alpha_k\}$, there is a considerable amount of equivalent representations for these coefficients. All of them have a distinct characteristic that is very important especially in speech coding because of the parameter quantization [29]. The first one comes from the roots of inverse filter in (2.5). It is clear that this representation is a polynomial, the coefficients can be interpreted as zeros of $A(z)$ that are poles of $H(z)$ by definition. Due to the this fact, the filter of LP model is also called as all-pole filter. However, the roots are quite sensitive to the quantization errors which makes the system unstable. Therefore, a much more robust option is proposed by Itakura in [33] called Line Spectrum Frequencies (LSF).

LSFs collectively describe the two resonance conditions arising from an interconnected tube model of the human vocal tract. This includes mouth shape and nasal cavity, and forms the basis of the underlying physical relevance of the linear prediction representation [34]. The two resonance conditions describe the modeled vocal tract as being either fully open or closed at the glottis respectively. The resonances of each condition give rise to odd and even line spectral frequencies respectively, and are provided into a set of LSFs which have monotonically increasing value [35]. In reality, however, the human glottis opens and closes rapidly during speech so it is neither fully closed nor open. Hence, actual resonances occur at frequencies located somewhere between the two extremes of each LSF condition. Nevertheless, this relationship between resonance and LSF position leads a significant physical basis to the representation.

In more detail, LSFs are a direct mathematical transformation of the set of LP parameters, and are used within many speech compression systems. It is very popular due to their excellent quantization characteristics and consequent efficiency of representation. To explain it in more detail, we can define two $(p + 1)$-th order polynomials related to $A(z)$, named $P(z)$ and $Q(z)$. These are referred to as antisymmetric (or inverse symmetric) and symmetric parts based on observation of their coefficients. The polynomials represent the interconnected tube model of the human vocal tract and correspond respectively to complete closure and opening at the source part of the interconnected tubes. In the original model, the source part is the glottis, and is neither fully open nor closed during the period of analysis, and thus the actual resonance conditions encoded in $A(z)$ are a linear combination of the two boundaries [35]. In fact this is simply stated as

$$A(z) = \frac{P(z) + Q(z)}{2}\,, \tag{2.7}$$

where $A(z)$ is the LP polynomial that is derived in (2.5).

The two polynomials are created from the LPC polynomial with an extra feedback term being positive to model energy reflection at a completely closed glottis, and negative to model energy reflection at a completely open glottis

$$P(z) = A(z) - z^{-(p+1)}A(z^{-1}) \tag{2.8}$$

$$Q(z) = A(z) + z^{-(p+1)}A(z^{-1}) \tag{2.9}$$

The roots of these two polynomials are the set of line spectral frequencies, ω_k that can be located on the unit circle in the z-plane if the original LPC filter was stable and alternate around the unit circle. Remember that any equivalent size set of roots that alternate in this way around and on the unit circle will represent a stable LPC filter. In practice, LSF are useful because of sensitivity (a quantization of one coefficient generally results in a spectral change only around that frequency) and efficiency (LSF results in low spectral distortion). At the end, as long as the LSFs are ordered, the

resulting LPC filter is stable.

2.3 Gaussian Mixture Model (GMM)

The Gaussian Mixture Model (GMM) is a classic parametric model used in many pattern recognition techniques to represent multivariate probability distribution. Any general distribution is approximated by sum of weighted Gaussian distributions. The overall framework is visualized in Figure 2.3.

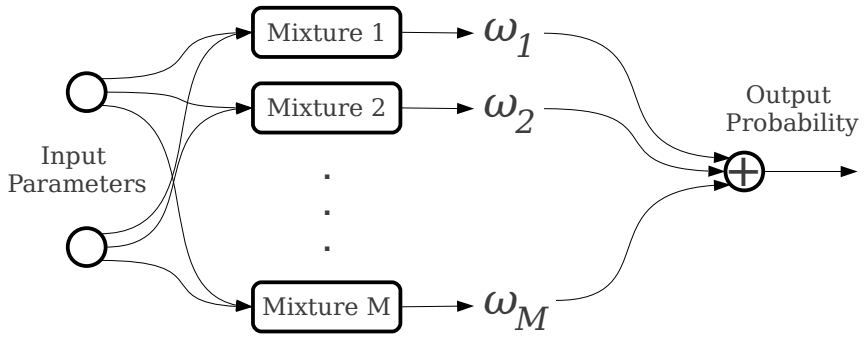

Figure 2.3: Framework of the GMM

In the estimation process, we use GMM-based density function to calculate output probability of a feature x using a weighted combination of multi-variate Gaussian densities. Briefly, the GMM is a weighted sum of D component densities and given by the equation

$$GMM_{\lambda}(x) = \sum_{i=1}^{D} \omega_i N_i(x), \tag{2.10}$$

where $N_i(x)$ is the multi-variate Gaussian distribution and defined as

$$N_i(x) = \frac{1}{\sqrt{2\pi^D |\sum|}} \exp\left(-\frac{1}{2}(x-\mu_i)^T \sum^{-1} (x-\mu_i) \right) \tag{2.11}$$

and ω_i is the mixture weight corresponding to the i-th mixture and satisfies

$$\sum_{i=1}^{D} \omega_i = 1 \text{ and } \omega_i \geq 0. \tag{2.12}$$

λ is the model and described by

$$\lambda = \left\{ \omega_i, \mu_i, \sum \right\}, \tag{2.13}$$

where μ_i is the mean of the i-th Gaussian mixture and $\sum$ is the diagonal covariance matrix.

In the GMM context, a speaker's speech is characterized by D acoustic classes representing broad phones in language. The probabilistic modeling of an acoustic class is important since there is variability in features coming from the same class due to variations in pronunciation and articulation. Thus, the mean vector μ_i represents the average features for the i-th acoustic class and the covariance matrix $\sum$ models the variability of features within the acoustic class. In this model, the covariance matrix is typically assumed to be diagonal because of computational concern. The GMM parameters are usually estimated by a standard iterative parameter estimation procedure, which is a special case of the *Expectation-Maximization* (EM) algorithm and the initialization is provided by the *Vector Quantization* (VQ) method.

2.4 GMM-Based Probabilistic Mapping

The Gaussian mixture model (GMM) estimator of [26,36] is a soft mapping (SM) from observable source $\mathcal{X}$ to hidden source $\mathcal{Y}$ with an optimal linear transformation in the minimum mean square error (MMSE) sense. This mapping can be formulated as the MMSE estimator from the observable source to the hidden source,

$$\hat{\boldsymbol{y}}_k^s = \sum_{l=1}^{L} p(\gamma_l | \boldsymbol{x}_k)[\mu_{y,l} + \boldsymbol{C}_{yx,l}(\boldsymbol{C}_{xx,l})^{-1}(\boldsymbol{x}_k - \mu_{x,l})], \tag{2.14}$$

where γ_l is the l-th Gaussian mixture and L represents the total number of Gaussian mixtures. The vectors $\mu_{x,l}$ and $\mu_{y,l}$ are respectively the centroids for the l-th Gaussian

for sources $\mathcal{X}$ and $\mathcal{Y}$, $\boldsymbol{C}_{xx,l}$ is the covariance matrix of source $\mathcal{X}$ in the l-th Gaussian, and $\boldsymbol{C}_{yx,l}$ is the cross-covariance matrix of sources $\mathcal{X}$ and $\mathcal{Y}$ for the l-th Gaussian mixture. The probability of the l-th Gaussian mixture given the observation $\boldsymbol{x}_k$ is defined as the normalized Gaussian pdf as,

$$p(\gamma_l|\boldsymbol{x}_k) = \frac{\mathcal{N}(\boldsymbol{x}_k; \mu_{x,l}, \boldsymbol{C}_{xx,l})}{\sum_{m=1}^{L} \mathcal{N}(\boldsymbol{x}_k; \mu_{x,m}, \boldsymbol{C}_{xx,m})}. \tag{2.15}$$

The GMM estimator can also be formulated as a hard mapping (HM) from the observable source $\mathcal{X}$ to the hidden source $\mathcal{Y}$ as,

$$\hat{\boldsymbol{y}}_k^h = p(\gamma_{l^*}|\boldsymbol{x}_k)[\mu_{y,l^*} + \boldsymbol{C}_{yx,l^*}(\boldsymbol{C}_{xx,l^*})^{-1}(\boldsymbol{x}_k - \mu_{x,l^*})], \tag{2.16}$$

where γ_{l^*} represents the most likely mixture component, that is,

$$l^* = \arg\max_l p(\gamma_l|\boldsymbol{x}_k). \tag{2.17}$$

2.5 Enhancement Framework

Let us consider having two simultaneously recorded TM and AM speech, which are represented as $s_T[n]$ and $s_A[n]$, respectively. Source-filter decomposition through the linear prediction filter model of speech can be defined as,

$$S_{TT}(z) = \frac{1}{W_T(z)} R_T(z) \tag{2.18}$$

$$S_{AA}(z) = \frac{1}{W_A(z)} R_A(z), \tag{2.19}$$

where $W_T(z)$ and $W_A(z)$ are the inverse linear prediction filters, and $R_T(z)$ and $R_A(z)$ are the source excitation spectra for the TM and AM speech, respectively. Then we can define the TM speech enhancement problem as finding two mappings, the first one from TM spectra to AM spectra, and the second one from TM excitation to AM

excitation,

$$\widehat{W}_A(z) = \varphi_S(W_T(z)|\Lambda_S^{TA}), \tag{2.20}$$

$$\widehat{R}_A(z) = \varphi_E(R_T(z)|\Lambda_E^{TA}), \tag{2.21}$$

where Λ_S^{TA} and Λ_E^{TA} are general correlation models of TM and AM spectral envelopes and excitation. These joint correlation models can be extracted using a simultaneously recorded training database. Replacing the TM speech spectra and excitation with the estimates,

$$\widehat{S}_{AA}(z) = \frac{1}{\widehat{W}_A(z)}\widehat{R}_A(z), \tag{2.22}$$

is expected to enhance the perceived quality of the TM speech. Similarly, we can replace only TM spectra or excitation and also none of them as (2.18) to see detailed the effect of each mapping individually.

$$\widehat{S}_{AT}(z) = \frac{1}{\widehat{W}_A(z)}R_T(z) \tag{2.23}$$

$$\widehat{S}_{TA}(z) = \frac{1}{W_T(z)}\widehat{R}_A(z) \tag{2.24}$$

2.6 Spectral Envelope Enhancement

In this study, the line spectral frequency (LSF) feature vector representation of the linear prediction filter is used to model spectral envelope. The TM and AM spectral representations are extracted as 16th order linear prediction filters over 10 ms time frames. We define the elements of this representation at time frame k as column vectors $\boldsymbol{x}_k^s$ and $\boldsymbol{y}_k^s$, respectively representing the TM spectral envelope as an observable source $\mathcal{X}^s$ and AM spectral envelope as a hidden source $\mathcal{Y}^s$. Throat-microphone recordings reveal certain speech attributes, and deliver varying perceptual quality for different sound vocalizations, such as nasals, stops, fricatives. Hence an acoustic phone dependent relationship between throat- and acoustic-microphone speech can be formulated to value the attributes of the throat-microphone speech. In order to

explore such a relationship between throat- and acoustic-microphone speech, we first define a phone-dependent soft mapping (PDSM),

$$\hat{\boldsymbol{y}}_k^{s|c} = \frac{1}{N}\sum_{n=1}^{N}\sum_{l=1}^{L_n} p(\gamma_l^{c_n}|\boldsymbol{x}_k^h)[\mu_{y,l}^{c_n} + \boldsymbol{C}_{yx,l}^{s|c_n}(\boldsymbol{C}_{xx,l}^{s|c_n})^{-1}(\boldsymbol{x}_k^s - \mu_{x,l}^{s|c_n})], \tag{2.25}$$

where N is the number of context elements and each phone c_n has a separate GMM, which is defined by phone-dependent mean vectors and covariance matrices. The phones are set as the context c_n.

Furthermore, a phone-dependent hard mapping (PDHM) can be defined as,

$$\hat{\boldsymbol{y}}_k^{h|c} = \sum_{l=1}^{L^*} p(\gamma_l^{c^*}|\boldsymbol{x}_k^h)[\mu_{y,l}^{h|c^*} + \boldsymbol{C}_{yx,l}^{h|c^*}(\boldsymbol{C}_{xx,l}^{h|c^*})^{-1}(\boldsymbol{x}_k^h - \mu_{x,l}^{h|c^*})], \tag{2.26}$$

where c^* is the given phone, and L^* is the total number of Gaussian mixtures for the phone class c^*. In this study we consider three different sources for the given context. The true context, c^T, is defined as the true phonetic class of the phone, which is considered as the most informative upper bound for the phone-dependent model. The likely context from the GMM, c^G, is defined as the most likely phonetic class, which can be extracted as,

$$c^G = \arg\max_{c_n} \mathcal{N}(\boldsymbol{x}_k; \mu_{x,l}^{c_n}, \boldsymbol{C}_{xx,l}^{c_n}). \tag{2.27}$$

Finally, the likely context from an HMM-based phoneme recognizer, c^M, is defined as the most likely phonetic class, which is decoded by an HMM-based phoneme recognition over the observable source $\mathcal{X}$.

2.7 Excitation Enhancement

In the source-filter decomposition framework we observed that the TM and AM recordings exhibit significant differences at the excitation signal spectra, which appears to be an important source of the degradation in the TM voice quality. Hence, we

model spectral envelope difference of the excitation signals as a spectral tilt vector, and we propose a new phone-dependent GMM-based spectral tilt mapping scheme to enhance TM excitation.

Let us first define a triangular filter-bank, which will help us to compute the average spectrum around a sequence of center frequencies,

$$w_b(n) = \begin{cases} 0 & n < f_{b-1} \text{ or } n > f_{b+1} \\ \frac{n-f_{b-1}}{f_b-f_{b-1}} & f_{b-1} \leq n \leq f_b \\ \frac{f_{b+1}-n}{f_{b+1}-f_b} & f_b < n \leq f_{b+1} \end{cases} \tag{2.28}$$

where f_b is the b-th center frequency index and w_b is the b-th triangular filter. We take number of bands as $B = 8$ and let $b = 0, 1, \ldots, B+1$, where $f_0 = 0$ and $f_{B+1} = N$ are taken as boundary frequency indexes. Then the average spectrum energy of the AM excitation signal is computed for frequency band b as,

$$E_A(b) = \log\{\sum_{n=1}^{N-1} w_b(n)|R_A(n)|^2\} \text{ for } b = 1, \ldots, B, \tag{2.29}$$

where R_A is the $2N$-point DFT of the excitation signal, and B is the total number of frequency bands. Similarly the average spectrum energies of the TM excitation can be computed and represented as E_T.

We can now define the spectral tilt vector between the TM and AM excitation signals as,

$$D(b) = E_A(b) - E_T(b) \text{ for } b = 1, \ldots, B. \tag{2.30}$$

The spectral tilt vector is considered as the representation of the hidden source $\mathcal{Y}^e$ for the probabilistic excitation mapping. We define the spectral tilt vector at time frame k as column vector $\boldsymbol{y}_k^e$ representing the hidden source $\mathcal{Y}^e$. The observable source of the spectral envelope mapping, which is the 16th order LSF feature vector $\boldsymbol{x}_k^s$ of TM speech can be considered as a valuable observation also for the probabilistic excitation mapping. However we also consider excitation spectrum of the TM speech to be

valuable. Hence we compute a cepstral feature vector representing the TM excitation spectrum as,

$$c_T(n) = \sum_{b=1}^{B} E_T(b) \cos(\pi n(b - 1/2)/B), \tag{2.31}$$

for $n = 1, 2, \ldots, B-1$. We form the observable source vector of the excitation mapping as $\boldsymbol{x}_k^e = [\boldsymbol{x}_k'^s \; c_T']'$, where c_T representing the cepstral column feature vector at frame k. Then, a phone-dependent mapping for the excitation enhancement of TM recordings is defined similarly as in (2.26) to estimate $\hat{\boldsymbol{y}}_k^{e|c}$, or equivalently the spectral tilt vector $\widehat{D}$. Finally, the enhanced excitation spectrum can be estimated by tilting the TM excitation spectrum as following,

$$\widehat{R}_A^{\widehat{D}}(n) = \sum_{b=0}^{B+1} w_b(n) 10^{\widehat{D}_b} R_T(n) \;\; n = 1, \ldots, N-1, \tag{2.32}$$

where boundary spectral tilt values are taken as $\widehat{D}_0 = \widehat{D}_1$ and $\widehat{D}_{B+1} = \widehat{D}_B$.

Note that in processing of the excitation signals, a 2048-point DFT is used over 20 ms hamming windowed excitation signals with a frame shift of 10 ms. The enhanced excitation signal is reconstructed from the $\widehat{R}_A^{\widehat{D}}$ spectrum with inverse DFT and overlap-and-add schemes.

Chapter 3

EXPERIMENTAL EVALUATIONS

We perform experiments on a synchronous TM and AM database which consists of two male speakers namely, M1 and M2. The latter one is recorded with a new IASUS-GP3 headset. Also, the AM data comes from Sony electret condenser tie-pin microphone. Each speaker utters 799 sentences that are recorded simultaneously at 16-kHz sampling rate. At the training stage, codebooks are established via varying number of Gaussian mixtures model using one-fold cross validation. In other words, we use 720 sentences as training data and the rest of the recordings as test data in our speaker dependent mapping schemes.

Experimental evaluations are divided into two sub-groups. As it discussed in Chapter 2 that speech can be separated into two independent parts, namely source and filter. Thus, firstly, different mapping schemes for the enhancement of filter are applied, then, excitation improvement is carried out with a branch of experiments to emphasize their individual effect more detailed. For envelope enhancement, we use a database from one male speaker only namely M1 and for excitation enhancements, both male records (M1 and M2) are used for comprehensive analysis.

3.1 Observations on Throat-Microphone Speech Attributes

It can be helpful to analyze the phones according to how they are articulated in oral cavity. The articulation of different phones come with its distinct character in terms of resonance shaping. Although they differ in realization across individual speakers, the tongue shape and positioning in the oral cavity do not change significantly. Since, the throat-microphone captures a reliable low-frequency energy, it can represent baseband spectrum, such as nasals and voice bars, sufficiently well. In articulatory phonetics,

manner of articulation is very important parameter for classification of phones. It describes the degree of narrowing in the oral cavity and certain acoustic or perceptual characteristics. For example, the phone /n/ and /k/ have same manner of articulation because they are articulated by the rapid release of a complete oral closure. Likewise, /s/ and /z/ have same manner of articulation and are articulated by forming a constriction that causes a turbulence in the flowing air so they produce a hissing sound.

In Table 3.1 we collect the average LSD scores between the acoustic and throat spectral envelopes, respectively $W_A(\omega)$ and $W_T(\omega)$, for the main phonetic attributes. The two minimum LSD scores occur for the nasals and stops, and the fricatives yield the maximum LSD score. Note that, nasals realized over closure of nasal cavity such as */m/* have smallest distortion, and fricatives realized over the friction of narrow-stream turbulent air such as */s/* have largest distortion due to its high-frequency energy. Clearly, the mapping of the throat-microphone speech spectra to the acoustic-microphone speech spectra is harder for the fricatives than for the nasals. That is one of the main reasons that we investigate a context-dependent mapping for the enhancement of throat-microphone speech.

Table 3.1: The average LSD scores of the data M1 between throat- and acoustic-microphone spectrums for different phonetic attributes with relative occurrence frequencies in the test database.

Attribute	**Freq**	**LSD (dB)**
Nasals	9.27	5.58
Stops	16.94	6.27
Liquids	9.59	7.05
Back Vowels	16.18	7.22
Front Vowels	13.93	7.65
Glide	2.36	7.81
Affiricate	2.72	9.54
Fricatives	11.10	11.81

3.2 Objective Evaluations

Evaluations of the TM speech enhancement are performed with two distinct objective metrics, the logarithmic spectral distortion (LSD) and the perceptual evaluation of wide-band speech quality (PESQ) metrics. The logarithmic spectral distortion (LSD) is a widely used metric for spectral envelope quality assessment. The LSD metric assesses the quality of the estimated spectral envelope with respect to the original wide-band counterpart, and is defined as

$$d_{LSD} = \sqrt{\frac{1}{2\pi}\int_{-\pi}^{\pi}\left(10\log\frac{|W_A(\omega)|^2}{|\hat{W}_A(\omega)|^2}\right)^2 d\omega} \tag{3.1}$$

where $W_A(\omega)$ and $\hat{W}_A(\omega)$ represent the original and estimated acoustic spectral envelopes, respectively. The ITU-T Standard PESQ [37] is employed as the second objective metric to evaluate the perceptual quality of the enhanced throat-microphone speech signal, which is constructed using the estimated spectral envelope and the excitation signal of the throat-microphone speech. The PESQ algorithm predicts subjective opinion scores of a degraded speech sample from 4.5 to 0.5 (higher score indicates better quality).

In phonetics and linguistics, a phone is defined as a unit of speech sound that is the smallest identifiable part found in a stream of speech. Since it is pronounced in a defined way, we can regard it as a context data. From the Table 3.2, there are 37 different phones despite of the fact that 29 letters are available in Turkish language. The recordings are phonetically transcribed using the Turkish phonetic dictionary METUbet and the phone level alignment is performed using forced-alignment and visual inspection. In [38], it is developed a new letter-to-phone conversion rule set that is based on the phonetic symbol set of Turkish language. These rules are formed by observing the phonetic transcriptions of the letters in the dictionary and determining the phonetic conditions in which they appear [38]. The choice of symbol formatting in METUbet is similar to that used within ARPAbet for American English. The METUbet phonetic alphabet is given in Table 3.2 where phones are categorized into 8

different manner of articulation.

Table 3.2: The Turkish METUbet phonetic alphabet with classification into 8 articulation attributes.

Back Vowels		**Stops**		**Fricatives**	
AA	a**n**ı	B	**b**al	H	**h**asta
A	l**a**f	D	**d**e**d**e	J	mü**j**de
I	**ı**s**ı**	GG	kar**g**a	F	**f**asıl
O	s**o**ru	G	**g**enç	S	**s**e**s**
U	k**u**lak	KK	a**k**ıl	SH	a**ş**ı
Front Vowels		K	**k**edi	VV	**v**ar
E	**e**lma	P	i**p**	V	ta**v**uk
EE	der**e**	T	ü**t**ü	Z	a**z**ık
IY	s**i**m**i**t	**Liquids**		ZH	yo**z**
OE	**ö**rtü	LL	ku**l**	**Affiricates**	
UE	**ü**mit	L	**l**ey**l**ek	C	**c**am
Nasals		RR	I**r**mak	CH	se**ç**im
M	da**m**	RH	bi**r**	**Glide**	
NN	a**n**i	R	**r**af	Y	**y**at
N	sü**n**gü				

As Salor et al. suggest in [38], the Turkish phone GH, soft g, has not been used for transcription and recognition, since it is used for lengthening of the previous vowel sound.

In machine learning, taking the contextual information into account undoubtedly improves the performance of any system. Since articulation of speech can be approximated as a stationary process, knowing the piece of a speech that surrounds a particular word can contribute its full analysis. Therefore, the proposed enhancement schemes take phone information into consideration as probabilistic mappings.

3.2.1 Envelope Enhancements

Table 3.3 presents the average LSD scores between the estimated filter $\hat{W}_A(z)$ and the original acoustic filter $W_A(z)$, and the average PESQ scores between the enhanced throat-microphone recordings and the original acoustic-microphone recordings. Note that for increasing the PESQ, the LSD scores decrease in a consistent manner.

Table 3.3: The average LSD and PESQ scores for different mapping schemes for enhancement of the throat-microphone recordings.

	LSD (dB)	**PESQ (MOS-LQO)**
PDHM-G	3.92	1.27
HM	3.80	1.29
SM	3.66	1.34
PDSM	3.65	1.36
PDHM-M	3.48	1.38
PDHM-T	3.18	1.43

The number of mixture components for the phone-dependent hard and soft mapping schemes are set as $L_n = 16$ for all phones. Similarly the number of mixture components for the GMM based hard and soft mapping schemes is set as $L = 256$. The worst performing scheme is observed as the phone-dependent hard mapping when phone recognition is performed with the GMM classifier (PDHM-G). The hard and soft mapping schemes, respectively HM and SM, achieve some performance improvement over the PDHM-G mapping. The best LSD and PESQ scores are attained with the phone-dependent hard mapping when the true phone class is known (PDHM-T). Also, the phone-dependent soft mapping (PDHM) performs close to the soft mapping (SM) scheme.

The phone-dependent hard mapping with the HMM-based phone recognition (PDHM-M) attains a performance improvement and performs closest to the PDHM-T

mapping scheme. The HMM-based phone recognition for the PDHM-M mapping is performed with 3-state and 256-mixture density phone level HMM recognizer, which is trained over the throat-microphone recordings of the 11 male speakers of the TAM database in [12]. Note that the test recordings in this study have been excluded from the training set of the phone level HMM recognizer. The HMM-based phone recognizer attains 62.22% correct phone recognition over the test database. We observe significant performance improvement when the true phone class is known to the phone-dependent hard mapping (PDHM-T) scheme. Furthermore the phone-dependent hard mapping with a reliable phone recognition, in this case the PDHM-M mapping, attains the best blind estimation for the spectral envelope to enhance the throat-microphone recordings.

Table 3.4: The average PESQ scores for different mapping schemes using acoustic residual.

	PESQ (MOS-LQO)
PDHM-G	1.66
HM	1.75
SM	1.97
PDSM	2.02
PDHM-M	2.16
PDHM-T	2.53

The throat-microphone recordings have a lower bandwidth at low-frequency bands compared to the reference acoustic-microphone recordings. Since the perceived intelligibility is poor for the throat-microphone recordings, the average PESQ scores stay at low values. In order to isolate the degradation, which is introduced by the throat source $R_T(z)$, we consider the case with the acoustic source $R_A(z)$ and throat filter $W_T(z)$ as a degraded speech signal. In this case we synthesized an enhanced speech

signal using the estimated filter $\hat{W}_A(z)$ and the acoustic source $R_A(z)$. Table 3.4 presents the average PESQ scores for this investigation. Note that the PESQ results are higher compared to Table 3.3. Furthermore the phone-dependent hard mapping PDHM-M scheme has the highest PESQ improvement.

3.2.2 Excitation Enhancements

In this part, we consider only phone-dependent mapping schemes with two possible sources of the phone. First source is obtained by forced-alignment, and the second one is picked up from an HMM-based phone recognition system over the observable TM source. The phone recognition performances for the M1 and M2 speakers are obtained respectively as 62.22% and 61.07%.

We first investigate possible best case scenarios for the enhancement of TM recordings when AM speech data are available. Table 3.5 presents average PESQ scores for the four scenarios, where reference condition is always the AM recordings and a form of the TM recordings are synthesized with the given excitation and filter models. The first row presents the average PESQ scores between TM and AM recordings for both speakers M1 and M2. The second row presents the average PESQ scores of the source-filter synthesis when the TM filter is replaced by the AM filter. We observe similar PESQ score improvements for both speakers, and these can be considered as best case improvements for the enhancement of the spectral envelope. The last two rows of Table 3.5 presents average PESQ scores when AM filter is used and TM excitation is tilted with the original spectral tilt vector with linearly spaced frequency bands (D_{lin}) and with mel-scaled frequency bands (D_{mel}). Note that when the TM excitation is tilted with the original spectral tilt vector we observe high PESQ score improvements. However, we do not discover any improvement with the use of mel-scaled frequency bands in the computation of the spectral tilt vectors. Hence, we keep using the linearly spaced frequency bands in the remaining parts of the experimental evaluations.

Table 3.5: The average PESQ scores for evaluation of the targeted excitation and filter enhancement strategies.

		PESQ (MOS-LQO)	
Exc.	**Filter**	**M1**	**M2**
R_T	W_T	1.22	1.42
R_T	W_A	1.70	1.74
$\widehat{R}_A^{D_{lin}}$	W_A	2.23	2.72
$\widehat{R}_A^{D_{mel}}$	W_A	2.26	2.63

Figure 3.1 presents spectrograms of the excitation signals to emphasize the effect of spectral tilt. Clearly, the third from the top (c) is the TM excitation, which is enhanced by the estimated spectral tilt vector using the proposed probabilistic mapping. Note that enhancement of the TM excitation significantly compensates spectral energy distribution with respect to the AM excitation.

Table 3.6 presents PESQ scores for different excitation and spectral envelope mapping schemes that we use for the enhancement of TM recordings. First two columns define excitation and filter mapping schemes. The PESQ scores are presented with respect to the source of phone information, either force alignment or HMM-based phone recognition, and speaker information M1 and M2. The first observation is on the source of the phone information. Given the PESQ scores with the reliable force aligned phone information, the PESQ scores with the HMM-based phone recognition information do not degrade significantly. For example, the first row of the results, where only the spectral envelope is mapped, PESQ score drop is from 1.59 to 1.58 for speaker M2 with the use of phone recognition information. The second observation is on the attained PESQ score improvements with the proposed excitation and spectral envelope mapping schemes. The PESQ scores 1.22 and 1.42 of the TM recordings as reported in Table 3.5 respectively for speakers M1 and M2 are improved to 1.46 and 1.58 with the spectral envelope mapping, and to 1.61 and 1.86 with the excitation and spectral envelope

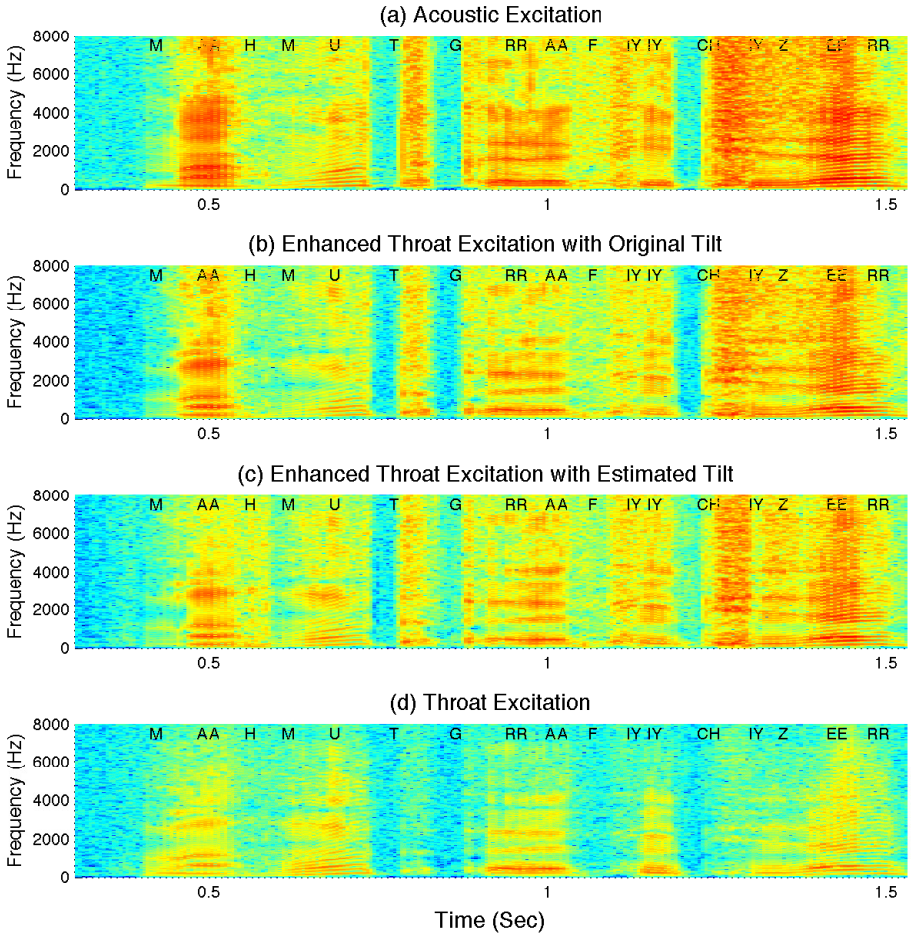

Figure 3.1: Sample spectrograms of (a) AM excitation, (b) enhanced TM excitation with the original spectral tilt, (c) enhanced TM excitation with the estimated spectral tilt, and (d) TM excitation.

mappings when phone information is extracted from the HMM-based phone recognizer. Here surprising observation is the contribution of the excitation enhancement, which brings a higher improvement than the spectral envelope enhancement. Finally the third observation is on the sole contribution of the excitation mapping, where last row of Table 3.6 presents PESQ scores with the proposed excitation mapping when TM filter is used. We do not observe PESQ score improvements for the sole use of excitation mapping, hence we can say that the proposed excitation mapping is

significant when used with the spectral envelope mapping.

Table 3.6: The average PESQ scores for different excitation and spectral envelope mapping schemes.

		PESQ (MOS-LQO)			
		Force-Align.		**Phone Recog.**	
Exc.	**Filter**	**M1**	**M2**	**M1**	**M2**
R_T	$\widehat{W}_A$	1.53	1.59	1.46	1.58
$\widehat{R}_A^{\widehat{D}}$	$\widehat{W}_A$	1.65	1.90	1.61	1.86
$\widehat{R}_A^{\widehat{D}}$	W_T	1.34	1.43	1.28	1.40

From the objective evaluations of both envelope and excitation enhancements, we can clearly notice that phone-dependent mapping is the best enhancement model in terms of intelligibility and spectral distortion. This is because of the accuracy in context information that suits the nature of speech enhancement well. In other words, if we know the precede and past of a particular speech very well, we can built more qualified model for the TM enhancement.

3.3 Subjective Evaluations

Since the reported PESQ scores stay at low values, a subjective evaluation of the proposed throat-microphone speech enhancement techniques is necessary to check whether the objective score improvements are subjectively perceivable. We performed a subjective A/B comparison test to evaluate the proposed enhancement techniques. During the test, the subjects are asked to indicate their preference for each given A/B test pair of sentences on a scale of (-2; -1; 0; 1; 2), where the scale corresponds to *strongly prefer A*, *prefer A*, *no preference*, *prefer B*, and *strongly prefer B*, respectively. The subjective tests are divided into two parts likewise in objective evaluations.

3.3.1 Envelope Enhancements

The subjective A/B test of envelope mapping schemes includes 21 listeners, who compared 20 sentence pairs randomly chosen from our test database to evaluate 5 conditions. The acoustic-microphone speech condition is compared to all conditions with 1 pair. The throat-microphone speech condition is compared to all three enhancement schemes with 2 pairs. The GMM-based soft mapping scheme is compared to the phone-dependent hard mapping schemes PDHM-T and PDHM-M with 3 pairs. Finally, the PDHM-T scheme is compared to the PDHM-M scheme with 3 pairs.

Table 3.7 presents the average subjective preference results. The rows and the columns of these tables correspond to A and B conditions of the A/B pairs, respectively. Also, the average preference scores that tend to favor B are given in bold to ease visual inspection. Speech samples from the subjective A/B comparison test are available for online demonstration [39].

Table 3.7: The average preference results of the subjective A/B pair comparison test for envelope mapping

		B			
	A	**1**	**2**	**3**	**4**
1	Acoustic	**0.05**			
2	Throat	**1.93**			
3	SM	**1.93**	-0.57		
4	PDHM-T	**1.83**	-1.12	-0.49	
5	PDHM-M	**1.83**	-0.83	-0.27	**0.54**

All the three enhancement schemes yield a perceivable difference compared to the throat-microphone speech. Among the three enhancement schemes, the PDHM-T, which uses the true phone class, has the highest perceivable improvement. The proposed PDHM-M scheme has the second best apprehensible improvement, which is inline with the objective evaluations.

3.3.2 Excitation Enhancements

The subjective A/B test includes 33 listeners, who compared 29 sentence pairs randomly chosen from our test database of speaker M2 to evaluate 6 conditions. The AM and TM speech conditions are compared to all conditions with 1 pair. The second row of the Table 3.5 (R_T, W_A) defines a target condition for the best possible spectral envelope mapping, which is evaluated as the third condition.

Table 3.8: The average preference results of the subjective A/B pair comparison test for excitation mapping

		B				
	A	**1**	**2**	**3**	**4**	**5**
1	Throat	0.00				
2	Acoustic	-1.97	**0.06**			
3	R_T, W_A	-1.21	**1.70**			
4	$\widehat{R}_A^D$, W_A	-1.85	**1.00**	-1.40		
5	R_T, $\widehat{W}_A$	-0.79	**1.82**	**0.34**	**1.72**	
6	$\widehat{R}_A^{\widehat{D}}$, $\widehat{W}_A$	-1.41	**1.49**	-0.57	**0.69**	-1.12

The fourth condition is defined as the target for the best possible excitation and spectral envelope mapping ($\widehat{R}_A^D$, W_A), which is the third row of the Table 3.5. The fifth and sixth conditions are set as the two proposed enhancement schemes, the sole spectral envelope mapping (R_T, $\widehat{W}_A$) and the excitation and spectral envelope mappings ($\widehat{R}_A^{\widehat{D}}$, $\widehat{W}_A$), which are respectively the first two rows of the Table 3.6. Conditions 3-6 are compared to each other with 3 pairs.

Table 3.8 presents the average subjective preference results. The rows and the columns of Table 3.8 correspond to A and B conditions of the A/B pairs, respectively. Also, the average preference scores that tend to favor B are given in bold to ease visual inspection. Speech samples from the subjective A/B comparison test are available for online demonstration [40].

The proposed excitation and spectral envelope mapping scheme ($\widehat{R}_A^{\widehat{D}}$, $\widehat{W}_A$), which is the condition 6, has perceivable improvements compared to all conditions except the AM recordings and the best target mapping ($\widehat{R}_A^D$, W_A). Furthermore it is significantly preferred over the sole spectral envelope mapping (R_T, $\widehat{W}_A$) with a preference score 1.82, and it is the second closest condition to the AM recordings after the best target mapping with a preference score 1.49.

Chapter 4

CONCLUSION

In this thesis, we introduce a new phone-dependent GMM-based spectral envelope mapping scheme to enhance throat-microphone speech using joint analysis of throat- and acoustic-microphone recordings. The proposed spectral mapping scheme performs the minimum mean square error (MMSE) estimation of the acoustic-microphone spectral envelope within the phone class neighborhoods. Objective and subjective experimental evaluations indicate that the phone-dependent spectral mapping yields perceivable improvements over the state-of-the-art context independent mapping schemes. Overall, the proposed phone-dependent spectral mapping, PDHM-M, introduces a significant intelligibility improvement over the throat-microphone speech. However, there is still a big room to further improve the perceive quality by modeling the source excitation signal of the throat-microphone recordings.

In the source-filter model of vocal tract, we observed that the spectral envelope difference of the excitation signals of TM and AM speech is an important source of the degradation in the throat-microphone voice quality. Thus, we model spectral envelope difference of the excitation signals as a spectral tilt vector using the same joint structure of TM and AM. Again, objective and subjective experimental evaluations indicate that the correction of the TM excitation spectrum has a strong potential to improve intelligibility for the TM speech. Although the proposed excitation mapping achieves a significant improvement (1.86 PESQ score for speaker M2) within the strong potential of correcting TM excitation with a spectral tilt (2.72 PESQ score of the best target mapping for speaker M2), there is still some unattempted improvement. We have to consider that incorporating temporal dynamics of the spectral tilt to the probabilistic mapping may attain further improvements for TM speech enhancement.

BIBLIOGRAPHY

[1] R. Lindeman, H. Noma, and P. de Barros, "Hear-through and mic-through augmented reality: Using bone conduction to display spatialized audio," in *Mixed and Augmented Reality, 2007. ISMAR 2007. 6th IEEE and ACM International Symposium on*, 2007, pp. 173–176.

[2] T. Toda, "Statistical approaches to enhancement of body-conducted speech detected with non-audible murmur microphone," in *Complex Medical Engineering (CME), 2012 ICME International Conference on.* IEEE, 2012, pp. 623–628.

[3] Iasus gp3 throat microphone system. [Online]. Available: http://iasus-concepts.com/productimages/GP3_1.jpg

[4] S. Li, J. Wang, J. Xijing, and T. Liu, "Nonacoustic Sensor Speech Enhancement Based on Wavelet Packet Entropy," in *Computer Science and Information Engineering, 2009 WRI World Congress on*, vol. 6, 2009, pp. 447–450.

[5] T. Quatieri, K. Brady, D. Messing, J. Campbell, W. Campbell, M. Brandstein, C. Weinstein, J. Tardelli, and P. Gatewood, "Exploiting Nonacoustic Sensors for Speech Encoding," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 2, pp. 533–544, Mar. 2006.

[6] R. Cutler and L. Davis, "Look who's talking: speaker detection using video and audio correlation," in *Multimedia and Expo, 2000. ICME 2000. 2000 IEEE International Conference on*, vol. 3, 2000, pp. 1589–1592.

[7] J. A. MacDonald, P. P. Henry, and T. R. Letowski, "Spatial audio through a bone conduction interface," *International Journal of Audiology*, vol. 45, no. 10, pp. 595–599, 2006.

[8] Y. Zheng, Z. Liu, Z. Zhang, M. Sinclair, J. Droppo, L. Deng, A. Acero, and X. Huang, "Air- and bone-conductive integrated microphones for robust speech detection and enhancement," in *IEEE Workshop on Automatic Speech Recognition and Understanding*, 2003, pp. 249–254.

[9] P. Heracleous, Y. Nakajima, H. Saruwatari, and K. Shikano, "A tissue-conductive acoustic sensor applied in speech recognition for privacy," in *Proceedings of the 2005 joint conference on Smart objects and ambient intelligence: innovative context-aware services: usages and technologies*. ACM, 2005, pp. 93–97.

[10] P. Heracleous, Y. Nakajima, A. Lee, H. Saruwatari, and K. Shikano, "Accurate hidden markov models for non-audible murmur (nam) recognition based on iterative supervised adaptation," in *Automatic Speech Recognition and Understanding, 2003. ASRU '03. 2003 IEEE Workshop on*, 2003, pp. 73–76.

[11] A. Shahina and B. Yegnanarayana, "Mapping Speech Spectra from Throat Microphone to Close-Speaking Microphone: A Neural Network Approach," *EURASIP Journal on Advances in Signal Processing*, vol. 2007, 2007.

[12] E. Erzin, "Improving Throat Microphone Speech Recognition by Joint Analysis of Throat and Acoustic Microphone Recordings," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 7, pp. 1316–1324, Sep. 2009.

[13] S. Roucos, V. Viswanathan, C. Henry, and R. Schwartz, "Word recognition using multisensor speech input in high ambient noise," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 11, 1986, pp. 737–740.

[14] Z. Liu, Z. Zhang, A. Acero, J. Droppo, and X. Huang, "Direct Filtering for Air- and Bone-Conductive Microphones," in *IEEE 6th Workshop on Multimedia Signal Processing*, 2004, pp. 363–366.

[15] Z. Liu, A. Subramanya, Z. Zhang, J. Droppo, A. Acero, and O. M. Way, "Leakage Model and Teeth Clack Removal for Air- and Bone-Conductive Integrated Microphones," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2005.

[16] A. Subramanya, Z. Zhang, Z. Liu, and A. Acero, "Speech modeling with magnitude-normalized complex spectra and its application to multisensory speech enhancement," in *IEEE International Conference on Multimedia and Expo*, 2006, pp. 1157–1160.

[17] S.-C. Jou, T. Schultz, and A. Waibel, "Whispery Speech Recognition using Adapted Articulatory Features," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. I, 2005, pp. 1009–1012.

[18] A. Subramanya, Z. Zhang, Z. Liu, J. Droppo, A. Acero, and O. M. Way, "A Graphical Model for Multi-Sensory Speech Processing in Air-and-Bone Conductive Microphones," in *Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2005.

[19] A. Subramanya, Li Deng, Zicheng Liu, and Zhengyou Zhang, "Multi-Sensory Speech Processing: Incorporating Automatically Extracted Hidden Dynamic Information," in *IEEE International Conference on Multimedia and Expo*, 2005, pp. 1074–1077.

[20] S. Dupont, C. Ris, and D. Bachelart, "Combined use of close-talk and throat microphones for improved speech recognition under non-stationary background noise." in *Proc. of Robust 2004 (Workshop on Robustness Issues in Conversational Interaction)*, 2004.

[21] M. Graciarena, H. Franco, K. Sonmez, and H. Bratt, "Combining standard and throat microphones for robust speech recognition," *IEEE Signal Processing Letters*, vol. 10, no. 3, pp. 72–74, 2003.

[22] L. Neumeyer and M. Weinrraub, "Probabilistic optimum filtering for robust speech recognition," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1994, pp. 417–420.

[23] P. Jax and P. Vary, "On artificial bandwidth extension of telephone speech," *Signal Processing*, vol. 83, no. 8, pp. 1707–1719, 2003.

[24] C. Yagli, M. Tugtekin Turan, and E. Erzin, "Artificial bandwidth extension of spectral envelope along a Viterbi path," *Speech Communication*, vol. 55, no. 1, pp. 111–118, Jan. 2013.

[25] K. Kondo, T. Fujita, and K. Nakagawa, "On Equalization of Bone Conducted Speech for Improved Speech Quality," in *IEEE International Symposium on Signal Processing and Information Technology*, Aug. 2006, pp. 426–431.

[26] Y. Stylianou, O. Cappe, and E. Moulines, "Continuous probabilistic transform for voice conversion," *Speech and Audio Processing, IEEE Transactions on*, vol. 6, no. 2, pp. 131–142, 1998.

[27] T. Toda, A. W. Black, and K. Tokuda, "Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, no. 8, pp. 2222–2235, 2007.

[28] T. Toda, M. Nakagiri, and K. Shikano, "Statistical Voice Conversion Techniques for Body-Conducted Unvoiced Speech Enhancement," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 9, pp. 2505–2517, Nov. 2012.

[29] T. Dutoit and F. Marqués, *Applied Signal Processing: A Matlab-Based Proof of Concept*. Springer-Verlag, 2009.

[30] S. Furui, *Digital Speech Processing: Synthesis, and Recognition*. New York and Basel: Marcel Dekker, Inc., 1989.

[31] X. Huang, A. Acero, and H.-W. Hon, *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development.* Prentice Hall PTR, 2001.

[32] L. Rabiner and R. Schafer, "Introduction to digital speech processing," *Foundations and trends in signal processing*, vol. 1, no. 12, pp. 1–194, 2007.

[33] F. Itakura, "Line spectrum representation of linear predictive coefficients of speech signals," *Journal of Acoustical Society of America*, vol. 1, no. 57, 1975.

[34] S. V. Vaseghi, *Advanced digital signal processing and noise reduction.* John Wiley & Sons, 2006.

[35] I. McLoughlin, *Applied speech and audio processing.* New York, NY, USA: Cambridge University Press, 2009.

[36] Y. Agiomyrgiannakis and Y. Stylianou, "Conditional Vector Quantization for Speech Coding," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 2, pp. 377–386, Feb. 2007.

[37] ITU, "Wideband extension to Recommendation P.862 for the assessment of wideband telephone networks and speech codecs," ITU-T, Tech. Rep., 2005.

[38] O. Salor, B. Pellom, T. Ciloglu, K. Hacioglu, and M. Demirekler, "On developing new text and audio corpora and speech recognition tools for the Turkish language," in *International Conference on Spoken Language Processing (ICSLP02)*, 2002, pp. 349–352.

[39] M. Turan and E. Erzin, "Speech samples of icassp2013 for the throat-microphone speech enhancement schemes," Nov. 2012. [Online]. Available: http://home.ku.edu.tr/∼eerzin/t2a-icassp13

[40] ——, "Speech samples of interspeech2013 for the throat-microphone speech enhancement schemes," May 2013. [Online]. Available: http://home.ku.edu.tr/∼eerzin/interspeech2013