Multisensory speech enhancement in noisy environments using bone-conducted and air-conducted microphones

Mingzi Li

Multisensory speech enhancement in noisy environments using bone-conducted and air-conducted microphones

Final Paper

As Partial Fulfillment of the Requirements for
the Degree Master of Science in Electrical Engineering

Mingzi Li

Submitted to the Senate of the Technion—Israel Institute of Technology

Kislev 5774 Haifa November 2013

Acknowledgement

The Final Paper Was Done Under The Supervision of Professor Israel Cohen in the Department of Electrical Engineering.

The Generous Financial Help Of The Technion Is Gratefully Acknowledged. This research was supported by the Israel Science Foundation (grant no. 1130/11).

I would like to express my sincere gratitude and appreciation to my advisor Professor Israel Cohen, for his valuable guidance and constant encouragement throughout all stages of this research. I would also like to express my thanks to Saman Mousazadeh for his valuable advises. Finally, I would like to thank my parents and friends who have supported me in the life.

Contents

List of Figures

List of Tables

Abstract

In this thesis, we address the speech enhancement problem by exploiting bone-conducted and air-conducted microphones. A bone-conducted microphone, being less sensitive to surrounding noise, may be utilized to complement a regular air-conducted microphone in noisy environments. However, since the high frequency components of bone-conducted microphone signals are attenuated significantly due to transmission loss, the quality of speech signals acquired by a bone-conducted microphone is relatively low. On the other hand, although an air-conducted microphone picks up the full-band speech, it is less immune to environmental noise. Therefore, we wish to enhance the speech signal by combining both microphones and producing high quality speech signal with low background noise.

Existing multisensory speech enhancement methods may be classified into two main categories, according to the role of the bone-conducted microphone: In one category, the bone-conducted microphone is used as a supplementary sensor, whereas in the other category the bone-conducted microphone is used as a dominant sensor. Implementation in the first category relies on the accuracy of a voice activity detection or pitch extraction facilitated by the bone-conducted microphone. When the bone-conducted microphone is exploited as the main acquisition sensor, algorithms are related to either equalization, analysis-and-synthesize, probabilistic approaches. Generally, equalization approaches reconstruct the clean speech through an FIR filter of the pre-enhanced air and bone conducted speech spectra ratio. Similar to equalization approaches, analysis-and-synthesize methods require a speech enhancement procedure priorly, whereas the reconstruction filter is the ratio of the linear prediction model of both speech signals. With Gaussian noise hypothesis in air and bone channels, probabilistic approaches can be conducted either in a maximum likelihood sense or a minimum mean square error sense. According to various

assumptions of speech and noise models, more complicated algorithms have been proposed as well, such as non-linear information fusion, model-based fusion, and bandwidth extension.

In this thesis, clean speech is restored through a family of functions named geometric harmonics, i.e., eigenfunction extensions of a Gaussian kernel. Geometric harmonics can describe the geometry of high dimensional data and extend these descriptions to new data points, as well as the function defined on the data. In our case, the high dimensional data is defined by concatenation of air-conducted and bone-conducted speech in the short time Fourier transform (STFT) domain. A nonlinear mapping to the STFT of clean speech defined on the new concatenation of speech signals can be obtained by a linear combination of geometric harmonics.

Application of geometric harmonics requires a careful setting of the correct extension scale and condition number. As a result, a multi-scale Laplacian pyramid extension is utilized to avoid scale tuning. Based on the kernel regression scheme, Laplacian pyramid extension approximates the residual of the previous representation via a series of Gaussian kernels.

Experiments are conducted on simulated air-conducted and bone-conducted speech in interfering speaker and Gaussian noise environments. Geometric methods provide a consistent reconstruction of speech spectrograms in a variety of noise levels and categories. Log spectral distance results obtained using the proposed methods are compared to an existing probabilistic approach. We show that the Laplacian pyramid method outperforms the other methods.

Notation

$A(k,l)$	Spectral speech amplitude
$a^{AC}(i)$	ith LP coefficient of AC speech
$a^{BC}(i)$	ith LP coefficient of BC speech
$B(l,k)$	BC speech in STFT domain
$D(f)$	BC speech distortion spectral
$e^{AC}(n)$	Excitation from AC speech
$e^{BC}(n)$	Excitation from BC speech
$G(f)$	Multichannel wiener filter
$G^c(f)$	Constrained multichannel wiener filter
G_t	Leakage of background noise into the bone sensor
G_{H_1}	Filter for speech presence
G_{H_0}	Filter for speech absence
g_{x_t}	Scales $\tilde{X}_t$ to match the clean speech X_t
$\hat{H}^{INV}$	Linear prediction filter
$H(l,k)$	Filter of bone-conducted speech in in STFT domain
$H_d(f)$	BC speech transfer function
$H_0(k,l)$	Hypotheses for speech absence
$H_1(k,l)$	Hypotheses for speech presence
H_t	Optimal linear mapping from X_t to B_t
$\hat{h}^{INV}$	Equalization filter in Discrete Fourier transform domain
$h(t)$	Filter of bone-conducted speech in time domain
M_t	Index of the mixture of distributions
N_f	Complex Fourier transform at frequency f of the noise signal
n_f	Log spectrum at frequency f of the noise signal

$n(t)$	Noise in time domain
P_d	Distortion speech power spectrum
P_n	Noise speech power spectrum
P_s	Clean speech power spectrum
$p(k,l)$	Conditional signal presence probability
$\hat{q}(k,l)$	A priori probability for speech absence
S_t	State (speech/non-speech) of the frame
s_x	Speech state
U_t	Sensor noise in the air channel
V_t	Background noise
W_l	Gaussian kernel
W_t	Sensor noise in the bone channel
$W_{1,2,3}$	Weights for information fusion filter
$X(l,k)$	Clean speech in STFT domain
X_f	Complex Fourier transform at frequency f of the clean signal
X_t^{AC}	Discrete Fourier transform of AC speech
X_t^{BC}	Discrete Fourier transform of BC speech
$\tilde{X}_t$	Magnitude-normalized complex spectra
x	Narrowband MFCC feature vector
x_f	Log spectrum at frequency f of the clean signal
x_{VTR}	VTRs sequence
$x(t)$	Clean speech in time domain
$Y(l,k)$	Noisy speech in STFT domain
Y_f	Complex Fourier transform at frequency f of the noisy signal
$Y_1(l,k)$	AC speech in STFT domain
$Y_2(l,k)$	BC speech in STFT domain
$Y_3(l,k)$	Synthesized speech in STFT domain
y	Noisy narrowband MFCC feature vector
y_f	Log spectrum at frequency f of the noisy signal
$y(t)$	Noisy speech in time domain
z	Wideband MFCC feature vector

φ_l	Eigenvectors of Gaussian kernel
λ_l	Eigenvalues of Gaussian kernel
δ	Condition number
ρ	Signal coherence
$\hat{\xi}(k,l)$	A priori SNR
$\zeta(k,l)$	Recursive average of the a priori SNR

Abbreviations

AC	Air Conducted
AMDF	Average Magnitude Difference Function
ANN	Artificial Neural Networks
ANC	Adaptive noise Cancellation
ASMDF	Average Squared Mean Difference Function
BC	Bone Conducted
BWE	Bandwidth Extension
CMS	Cepstral Mean Subtraction
DC	Direct Current
EGG	Electroglottograph
EM	Expectation-Maximization
GEMS	Electromagnetic Motion Sensor
GMM	Gaussian Mixture Models
HMM	Hidden Markov Models
HPF	High Pass Filter
KMs	Kernel Methods
LP	Linear Prediction
LPF	Low Pass Filter
LSD	Log-Spectral Distortion
LSF	Line Spectral Frequencies
MCRA	Minima Controlled Recursive Averaging
ML	Maximum Likelihood
MFCC	Mel Frequency Cepstral Coeficients
MMSE	Minimum Mean Square Error

MSE	Mean Squared Error
OM-LSA	Optimally-Modified Log Spectral Amplitude
PCA	Principal Components Analysis
P-mic	Physiological Microphone
RASTA	Relative Spectral Processing
SNR	Signal to Noise Ratio
STFT	Short Time Fourier Transform
SVM	Support Vector Machine
VoIP	Voice over IP

Chapter 1

Introduction

1.1 Background

1.1.1 Speech Enhancement

When speech signals are acquired by microphones, various sources of noise may be accidentally encompassed as well, including car noise, construction noise and other unexpected environmental noises. Simply as conducting a conversation on cellphones, the voice contaminated by the surronding environment on one side will degrade the user experience on the other side, let alone the noise induced by communication. Improving the quality of degraded speech, as a result, is strongly needed for comfortable communication. Speech enhancement, apart from mobile phones, has a wide range of applications including Voice over IP (VoIP), teleconferencing systems, speech recognition, and hearing aids.

There are generally two main categories of speech enhancement methods with respective to the category of the noise [34]. For additive noise, spectral subtraction, Wiener filtering, model-based processing and auditory masking are commonly used classical approaches. Spectral subtraction optimal in the maximum likelihood (ML) sense and Wiener filtering optimal in the minimum mean square error (MMSE) sense can be viewed as filtering operation where high SNR regions of the measured spectrum are attenuated less than low SNR regions [13]. Model-based processing is based on an all-pole vocal tract transfer function. However, for auditory masking method, the key point is calculation of the masking threshold. In the case of convolutional distortion, cepstral mean subtraction

(CMS) and relative spectral processing (RASTA) [18] are exploited for noise reduction. CMS method tries to find nonlinear transformations of time trajectories, while RASTA is a cepstral lifter that removes low and high modulation frequencies.

The speech enhancement related problems can also be grouped by the system modeling. Attempts have been made to model state changes over time in the frequency domain. State changes of speech and noise signals can be modeled as autoregressive process for short-term speech [28] and hidden Markov process for long-term speech [11]. While these models have been widely used in coding and recognition, they are not sufficiently refined for speech enhancement [12].

1.1.2 Multi-sensory Speech Enhancement

In general, while multi-channel methods tend to outperform single-channel algorithms, their performance in non-stationary noise still needs to improve. The reason is that all the sensors are of a similar type and capture signal and noise with analogous properties simultaneously. One solution to this problem is multisensory processing where different types of sensors containing complimentary information are utilized. However, in multi-sensory processing, simple modification of multi-channel techniques will not yield desired results. As a result, new schemes have been developed for the purpose of multiple sensors.

In the speech recognition scenario, an audio-visual speech processing has received a lot of attention. It exploits a regular microphone along with a camera that captures images of the speakers mouth/face region [33]. The visual stream is used to disambiguate between phones that are easily confused when using only the audio stream. In addition, as the video stream is immune to noise existing in the audio stream, it is robust to the speech recognition engine.

Speech enhancement researchers adopt this sensor fusion idea by combining air microphone with other sorts of microphones as well, including throat microphones [16], ear plug [48], or stethoscope device [17]. In [16], Graciarena et al. trained a mapping from the concatenated features of both microphone signals in a noisy environment to clean speech signal. The down side of their approach is that the mapping is environment dependent, which can lead to generalization problems in new environments. In [48], Strand et al. designed an ear plug to capture the vibrations in the ear canal, and used the

signals for speech recognition with MLLR adaptation. In [17], Heracleous et al. used a stethoscope device to capture the bone vibrations of the head for non-audible murmur recognition. Although the above-mentioned methods have demonstrated improvement for speech recognition, the non-acoustic signal is merely exploited for adapting the recognizer, rather than providing an enhanced speech. A thorough review of bone-conducted microphone involved speech enhancement methods, including the improvement of speech waveform, will be shown in the next chapter.

1.1.3 Non-acoustive Sensors

In this section we present an overview of several non-acoustic sensors exploited in the last 40 years, i.e., the electromagnetic motion sensor (GEMS), physiological microphone (P-mic), bone-conduction microphone, and electroglottograph (EGG) [35].

The GEMS is a microwave radar sensor originally developed at Lawrence Livermore National Laboratory. It measures tissue movement during voiced speech, i.e., speech involving vocal chord vibrations [5, 20, 21]. It usually straps or tapes an antenna on the throat at the laryngeal notch, but other facial locations are also feasible. This sensor emits a 2-GHz electromagnetic signal that penetrates the skin and reflects off the speech production anatomy, such as the tracheal wall, the vocal folds, or the vocal tract wall. It is relatively immune to degradation from external acoustic noise sources, since signals collected from a GEMS device merely depend on the tissue movement in the speech production anatomy.

The P-mic is composed of a gel-filled chamber and a piezo-electric sensor behind the chamber [41]. Vibrations permeating the liquid-filled chamber are measured by the piezo-electric sensor that provides an output signal in response to applied forces generated by movement, converting vibrations traveling through the liquid-filled chamber into electrical signals. The liquid-filled chamber is designed to have poor coupling between ambient background noise and the fluid-filled pad, thus attenuating vibrations of unwanted ambient background noise.

In a bone-conducted microphone, voice content is transmitted by way of bone vibrations [55]. As with the P-mic, bone-conduction microphones comprise a piezo-electric-like material. The bone-conducted microphone located in the top of the helmet picking up

bone vibrations at the top of the skull. At this skull location, the bone-conducted microphone is found to provide strong voicing, as well as significant vocal tract content in a mid-frequency range.

The EGG sensor measures vocal fold contact area by providing an electrical potential (of about 1 V rms and 23 MHz) across the throat at the level of the larynx [40]. With a pair of gold-plated electrodes, the sensor measures the change of impedance over time. When the vocal folds are closed, the impedance is decreased; when they are open, the impedance is increased. Thus, the opening and closing of the vocal folds, present in voiced speech, are measured by the EGG.

1.1.4 Bone-conducted Microphone

Among all above-mentioned non-acoustic sensors, bone-conducted (BC) microphone has a wide applications for its simplicity and low cost. As a result, we will address in this research the speech enhancement problem involving joint utilization of bone-conducted and air conducted microphones. One design of a bone-conductive sensor is illustrated in Figure 1.1 [56]. In this sensor, a soft elastomer bridge is adhered to the diaphragm of a normal air-conducted (AC) microphone. This soft bridge conducts vibrations from skin contact of the user directly to the microphone. The skin vibrates in accordance with the voices of the speakers, and the microphone converts the skin vibration into analog electrical signal.

BC speech is acquired by the vibration of skull, thus unlike AC speech, it is robust to the background noise. High frequency part of BC speech, however, is distorted due to channel loss. Figure 1.2 shows conducting path of AC and BC microphones.

The BC speech largely reflects the bone structure and facial tissue through which the signal propagates to reach the bone sensor, thus it is speaker dependent. In addition, different placements of the bone sensor may introduce further differences in the signals, which leads to placement dependence.

Another strong and non-stationary effect is articulation dependence. Different speech phones entail different distributions of source energy and different patterns of coupling. These physical differences result in pronounced differences in the log spectra between the two sensors. Figure 1.3 illustrates the smoothed log spectrum of examples of the three

Figure 1.1: Air- and bone-conductive integrated microphone headset.(after [56])

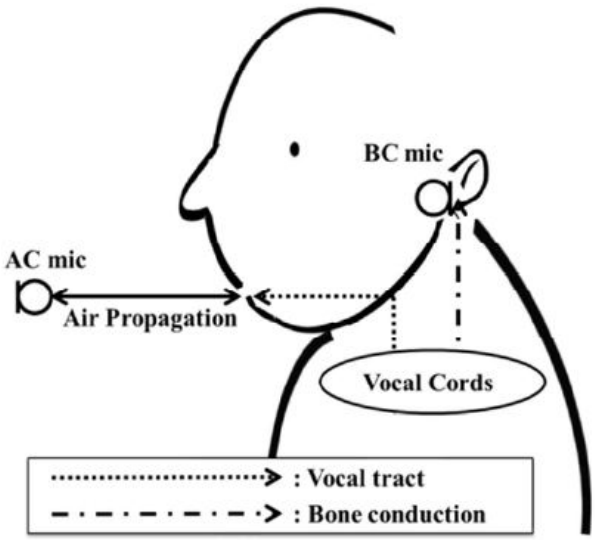


Figure 1.2: Conducting path of AC and BC microphones.(after [25])

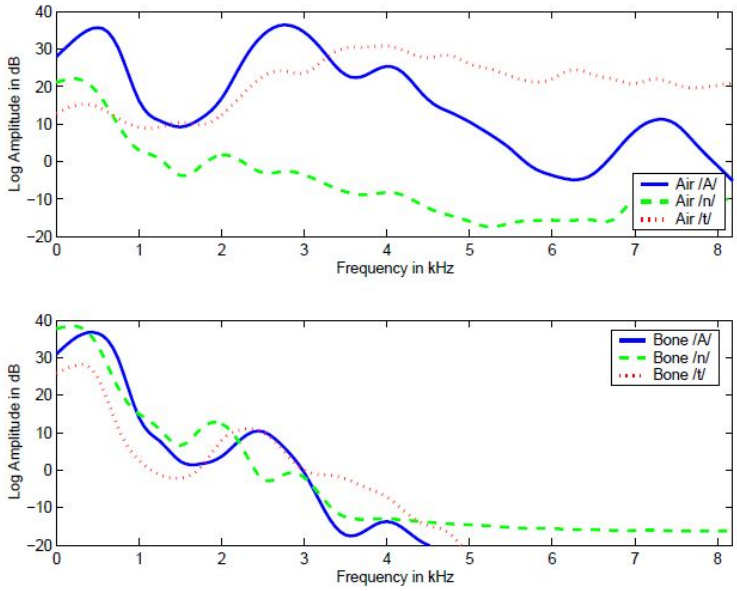


Figure 1.3: Smoothed log spectra received at the air (top) and bone (bottom) sensors for three phones /A/ (as in cake), /n/ and /t/, showing the differences in the relative patterns of the spectra in each sensor. In particular, notice how the /n/ spectrum is dramatically attenuated in the air sensor relative to the other phones, but not in the bone sensor. (after [19])

phones /A/, /n/, and /t/, taken from a single sentence for the air and bone sensors respectively [19].

One effect of articulation derives from the closing or stricture of the oral tract during speech. When the mouth is open as in a vowel sound such as /A/, the acoustic energy is well coupled to both the air and bone sensors. However, when the mouth is restricted, as in a nasal stop such as /n/, the acoustic coupling to the air sensor is greatly diminished. In this case, the coupling to the bone sensor may actually be increased with less energy escaping through the mouth.

Concerning the effect of articulation in response to the location and manner of the acoustic energy generation, voiced speech sounds originate in the throat with the vocal cords are well transmitted to the nearby bone sensors. In contrast, fricatives, such as /t/, generated at the place of articulation, as turbulent air passing through an aperture in the mouth, are thus typically much more readily transmitted to the air sensor than to the bone sensor.

A third articulation effect is the artifact of the frequency response for bone sensors. As the bone sensor is at the noise floor for frequencies above 2kHz, differences in the log

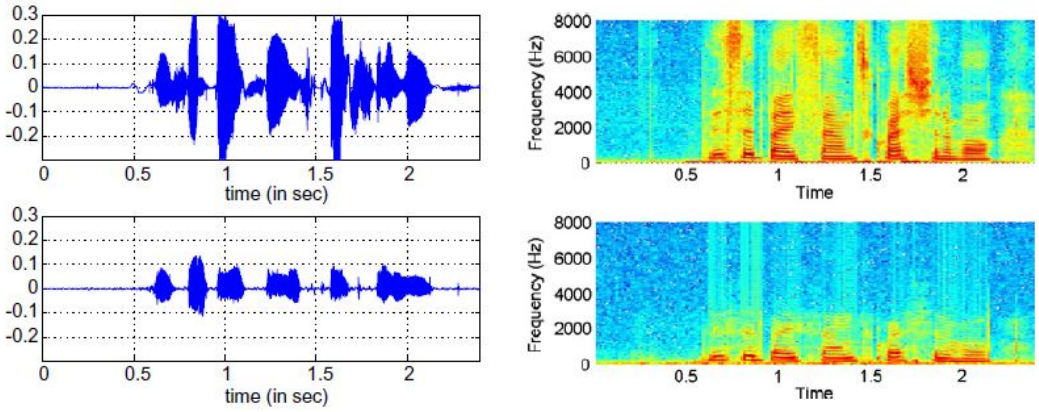


Figure 1.4: Waveforms and spectrograms of the signals captured by the ABC microphone. The first row shows the signal captured by the air microphone and the second row shows the signal captured by the bone microphone. (after [49])

spectrum between air and bone sensors will simply reflect the energy in the air sensor for those frequencies. This energy varies greatly for dinstinct phonemes, with fricatives such as /t/ having a great deal of high frequency energy, in contrast to nasal stops such as /n/.

In general, a bone sensor is relatively noise robust, and only captures frequency components up to 2 KHz [49]. Figure 1.4 shows signal and spectrogram of recorded speech in an office environment captured by both AC and BC microphone. In essence, in conjunction with the additional sensor, we have one stream undistorted but noise corrupted speech (air-conducted) and another stream distorted but fairly uncorrupted speech (bone-conducted). The challenge here is to intelligently fuse the signals from both sensors to obtain an enhanced speech.

1.2 Motivation and Goals

Due to the stability against the external noise, bone-conducted speech seems feasible to be used instead of noisy air-conducted speech in an extremely noisy environment. However, the quality of bone-conducted speech is degraded and restoring bone-conducted speech is still a challenging topic in speech processing field. Existing BC microphone-involved speech enhancement algorithms are classified into two categories in the role it plays, i.e., BC microphone employed as supplementary sensor or as dominant sensor. In the thesis, we will take into account the second case.

We first present the probabilistic approach assuming a mixture Gaussian model based on statistics of magnitude-normalized complex spectra for speech signal representation and dynamic Bayesian network for time update, as it allows a relatively enhanced speech reconstruction compared to other methods. Then we propose to reconstruct clean speech under the knowledge of samples of air-conducted and bone-conducted speech using the geometric method. A concatenation of both AC and BC speech can be regarded as a nonlinear function of the clean speech. We intend to find the nonlinear mapping using function extension methods, including geometric harmonics and Laplacian pyramids method which first describes the geometry of high dimensional data in terms of integral operator eigenfunctions and then extends these descriptions to new data points.

Finally, we will demonstrate reconstruction results with simulated AC and BC speech and compare them to those obtained by the probabilistic method.

1.3 Structure

We first review existing methods in Chapter 2, where most important and common bone-conducted microphone techniques are presented. We then review in more details the probabilistic approach with a different voice activity detection scheme in Chapter 3. In Chapter 4, we present geometric methods using BC microphone as a dominant sensor. Finally, in Chapter 5 we conclude with a summary and discussion on future research directions.

Chapter 2

Review of existing methods

2.1 Introduction

The BC microphone-involved speech enhancement algorithms are classified into two categories in the role it plays. In the early stage, BC speech was only used as a supplementary sensor. It allowed a more robust distinguishment between non-speech segment and speech segment [57] and a more accurate detection of voice activity in low SNR environment [58]. It has also been applied for glottal source information detection where the source-filter model needs accurate pitch information [38]. Enhancing low frequency components is another direction for BC speech enhancement [35, 37].

Then BC and AC speech are combined to generate even better results. A prototype hardware device was developed as a turning point [57]. From then on, a typical on-line or off-line training stage is required to calibrate characteristic differences between AC and BC microphones [44, 49]. This kind of AC-BC combined microphone for speech enhancement are characterized into three major groups, i.e., equalization, analysis-and-synthesize and probabilistic approach [45]. Based on various model assumptions, recently there have been a great number of more complicated algorithms involved with BC microphone [9, 19, 31, 42, 44].

In this chapter, we first review methods when BC microphone works as a supplementary sensor, where useful information is extracted from the BC microphone, including more accurate voice activity detection and pitch detection result. Then we briefly review the three main groups of methods and their variations where BC speech plays a dominant

role. Finally, we review several methods belonging to the second category, which may have significant impact on future study.

2.2 Model

At a preliminary step, we assume there is no leakage noise in the BC speech. As a result, BC speech can be regarded simply as the filtered clean speech. AC speech is assumed to be contaminated by an additive noise. The model is represented in time domain, where $x(t)$ is the clean speech signal that we wish to estimate given noisy AC speech $y(t)$ and BC speech $b(t)$.

$$y(t) = x(t) + n(t) \tag{2.1}$$

$$b(t) = h(t) * x(t) \tag{2.2}$$

where $n(t)$ is the corrupting noise and $h(t)$ is the convolving filter.

2.3 BC Microphone as a Supplementary Sensor

2.3.1 BC Speech for Voice Activity Detection

High performance of speech enhancement in non-stationary noisy environment is achieved by combining the following two techniques: the speech detection based on the bone-conductive microphone signal and non-stationary noise suppression by adaptive spectral subtraction of air-conductive microphone signal [58]. Since the bone speech signals capture relatively less noise, we are able to distinguish non-speech segments from speech segments even in low SNR environments. As a result, when two people are talking at the same time, bone microphone can detect when the main speaker talks.

Figure 2.1 shows the voice detection of bone microphone. The top figure illustrates the speech signal captured by the bone sensor when two people are talking at the same time. The middle figure shows the signal captured by the regular microphone and bottom figure presents the detection result. It can be seen from the figures that bone conducted microphone yields a more accurate voice activity detection result.

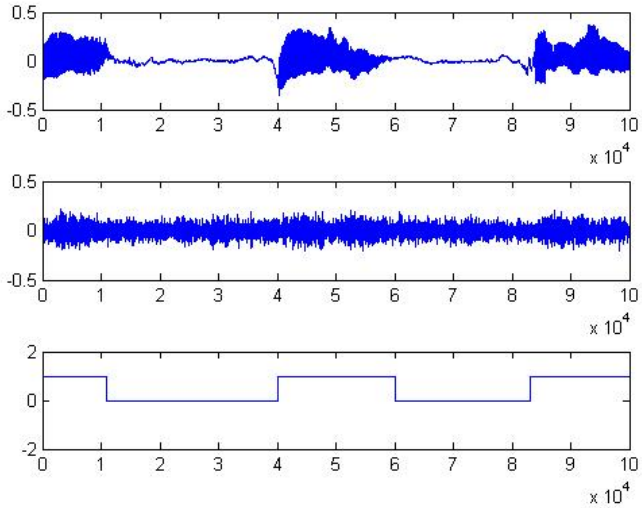


Figure 2.1: Speech detection using bone sensors: The top figure illustrates the speech signal captured by the bone sensor when two people are talking at the same time. The middle figure shows the signal captured by the regular microphone and bottom figure presents the detection result. (after [58])

Then the spectral power density of noise is estimated based on the speech presence detection estimated by BC microphone, which is much more accurate than the detection result of air microphone in non-stationary noise.

2.3.2 BC Speech for Pitch Detection

Being robust to ambient noise, bone conducted speech preserves a regular harmonic structure in the lower spectral region. In the paper [38], bone conducted speech are demonstrated to detect accurate pitch information in both noisy and noiseless case. Figure 2.2 illustrates the pitch tracking extracted by air and bone microphone in the noiseless case. Left column presents pitch tracking of four speeches using air conducted speech, middle column shows the spectrogram of air conducted speech, and right column shows the pitch tracking using bone conducted speech. Figure 2.3 shows the pitch tracking extracted by air and bone microphone in the noisy case. Left column shows pitch contours contaminated by white noise for air conducted speech (up), bone conducted speech (down); right column shows pitch contours contaminated by babble noise for air conducted speech (up), bone conducted speech (down).

We can see from the figures that in noiseless case, bone conducted microphone can

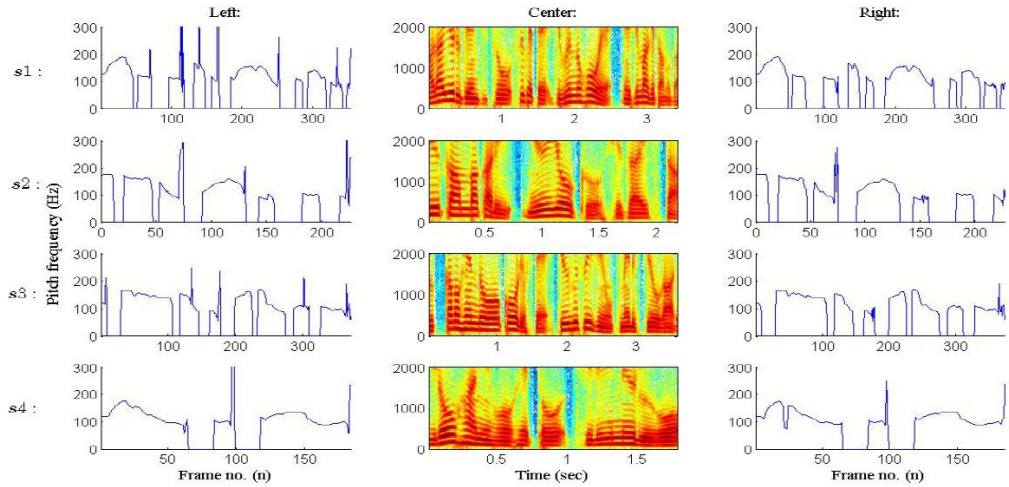
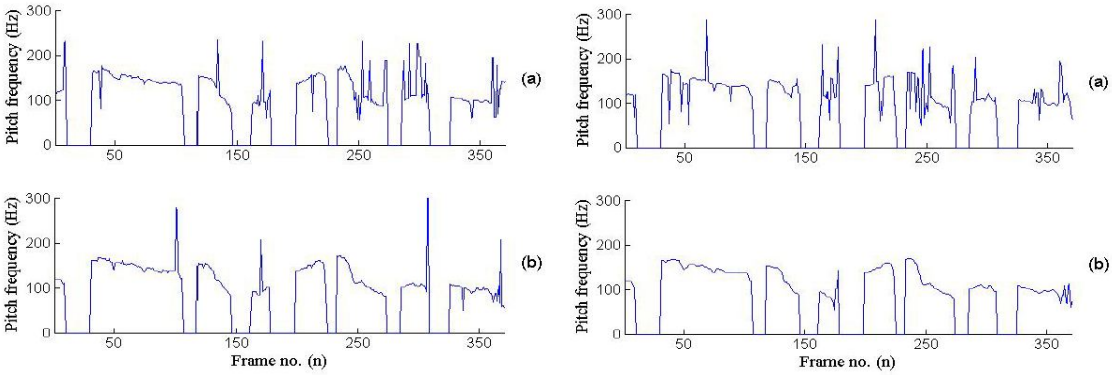


Figure 2.2: Left: Pitch tracking of air-conducted speech in noiseless condition. Center: Speech spectrogram. Right: Pitch tracking of bone-conducted speech in noiseless condition. The experiments have been conducted on four speeches. (after [38])



(a) white noise (b) babble noise

Figure 2.3: Pitch contours estimated from speech when corrupted by noise. a) using white noise, b) using babble noise. Up: Pitch contours estimated from air-conducted speech. Down: Pitch contours estimated from bone-conducted speech. (after [38])

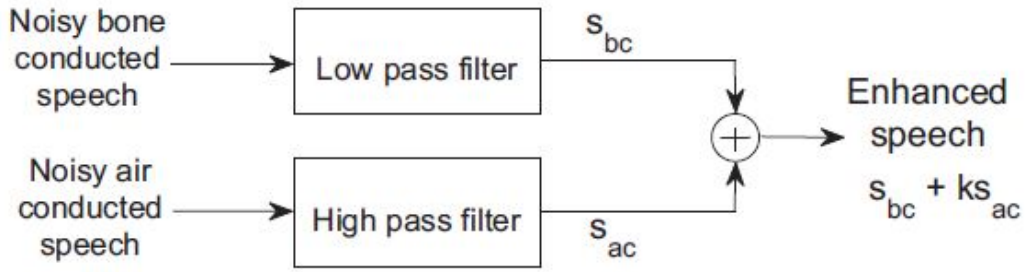


Figure 2.4: Block diagram when BC Speech is used for low frequency enhancement. (after [37])

provide an accurate pitch tracking, i.e., with result analogous to the air conducted microphone. However, in the noisy case, the pitch contour extracted by the bone conducted microphone contains less noise than the pitch contour extracted by the air conducted microphone.

2.3.3 BC Speech for Low Frequency Enhancement

In practice, speech is often corrupted by car and multi-talker babble noise that affect mostly the low-frequency region of speech signal. When AC speech is corrupted by low-frequency band noise, it can be readily enhanced by replacing the lower frequency region of noisy AC speech by the same region from BC speech [37]. Figure 2.4 shows the block diagram of the proposed method. Though the method appears to be very simple, it is effective irrespective of the nature of the noise, i.e., stationary (e.g. car) or non-stationary (e.g. babble).

2.4 BC Microphone as a Dominant Sensor

2.4.1 Equalization

In [44], Kondo et al. designed a linear phase Finite Impulse Response (FIR) filter to improve the quality of the bone-conducted speech. The filtering technique was utilized combining both pre-processing and post-processing speech enhancement method. After conducting a traditional speech enhancement method, a linear impulse response was designed in the training stage with inverse discrete Fourier transform (IDFT) for the long

term AC and BC spectra ratio. The traditional speech enhancement methods for single channel speech enhancement include spectral subtraction or Wiener filter, where the pre-processed AC and BC speech are denoted as X_t^{AC} and X_t^{BC}, and the equalization filter in the time domain is computed by:

$$\hat{h}^{INV} = IDFT\left[\left|X_t^{AC}\right| / \left|X_t^{BC}\right|\right] \tag{2.3}$$

Three types of AC and BC speech such as vowels, short and long sentences were utilized to train the filter. Then the filter was utilized for new speech captured by the bone-conducted microphone. Although the quality of the reconstructed speech has been improved, there was no consistency with respect to speakers and phrases. Besides, the output performance varies in response with filter length.

In [25], Liu et al. proposed a short-term DFT ratio filter. The equalization filter was estimated frame by frame and then averaged as:

$$\hat{h}^{INV} = E\left\{\left|X_t^{AC}\right| / \left|X_t^{BC}\right|\right\} \tag{2.4}$$

where $E\{\}$ is the expectation operator to average over the frames. Similar techniques were applied to learn the equalization filter. To remove fluctuations across adjacent frequency bins, smoothing process was applied, whereas the quality of the enhanced speech depended on both the filter length and the smoothing length.

A recursive least mean square (LMS) filter was proposed in [43] where no training stage was needed. The weighting vector $\hat{h}^{INV}(n)$ was recursively updated to minimize the mean square error of the difference between AC and BC speech $E\{e^2(n)\}$:

$$\hat{h}^{INV}(n) = \hat{h}^{INV}(n-1) + 2\mu e(n-1)\, b(n-1) \tag{2.5}$$

where $e(n)$ is the difference between AC and BC speech, μ is step size, $b(n)$ is BC speech.

Although the equalization method seems to be intuitive and simple, it suffers from a common problem that they are speaker-dependent thus cause problems for new speakers. Another defect is that it needs additional speech enhancement method where unexpected noise will be involved. Besides, their performance largely depends on the filter length, as a result, equalization method requires a careful setting of the filter length.

2.4.2 Analysis and Synthesize

Another group of method taking into account this problem was proposed by modeling denoised AC speech and BC speech as linear prediction (LP) filter [52]. The LP-based approach provided an estimation of the filter $\hat{H}^{INV}$ with LP coefficients of AC and BC speech or with only BC speech, corresponding to different assumption of models. At the early stage, the reconstruction filter was proposed as:

$$\hat{H}^{INV} = \frac{Z\left[e^{BC}(n)\right]\sum_{i=0}^{Q} a^{BC}(i)\, z^{-i}}{Z\left[e^{AC}(n)\right]\sum_{i=0}^{P} a^{AC}(i)\, z^{-i}} \tag{2.6}$$

where $a^{AC}(i)$ and $a^{BC}(i)$ represent the i-th LP coefficient of AC and BC speech with P and Q orders respectively, and $e^{AC}(n)$ and $e^{BC}(n)$ represent excitation of AC and BC speech.

An improvement was to predict line spectra frequency (LSF) via a recurrent neural network without using AC speech [51], as the LSF coefficients are more robust to quantization noise. Another method without AC speech was by modifying LP coefficients of BC speech [39]. The first and third poles of BC speech have been adjusted to have standard bandwidths as well as eliminate an irregular broadening of spectral peaks. This time the excitation from BC speech was unmodified, thus canceled out, where the reconstruction was as:

$$\hat{H}^{INV} = \frac{\sum_{i=0}^{Q} a^{BC}(i)\, z^{-i}}{\sum_{i=0}^{Q} \hat{a}^{BC}(i)\, z^{-i}} \tag{2.7}$$

Although the LP based filter improves the quality of BC speech, it is difficult for practical application as it needs preprocessing step for noise cancellation. In addition, it requires a careful framework for modeling both AC and BC speech in the linear prediction manner.

2.4.3 Probabilistic Approach

The probabilistic approach is based on the assumption of Gaussian noise in both AC and BC channels. In [30], Liu et al. utilized the maximum likelihood estimation to minimize the cost function R:

$$R = \sum_{n=l}^{N} \left\{ \begin{array}{l} \left|Y(l,k) - X(l,k)\right| / 2\sigma_V^2 \\ + \left|B(l,k) - H(l,k)\, X(l,k)\right| / 2\sigma_W^2 \end{array} \right\} \tag{2.8}$$

where l and k represent frame and frequency bin index. The spectra of background noise $V(l,k)$, and BC sensor noise $W(l,k)$ are distributed as $N(0,\sigma_V^2)$ and $N(0,\sigma_V^2)$. By taking a partial derivative of R with respective to $H(l,k)$ and $X(l,k)$, the estimated transfer function and reconstructed clean speech can be derived.

Liu et al. later refined the model by adding leakage noise induced by the BC microphone, which is also assumed to be Gaussian [29]. The enhanced speech is expressed as the weighted sum of AC speech and the leakage noise reomved BC speech. Although this method doesn't need training process, it lacks of speech model.

In [50], Subramanya et al. proposed a graphic speech model depending on a subset of variables [4]. The problem is solved by breaking the joint distribution into a product of factors, followed by the posteriors inferred according to the variables factorized on the subsets. The speech is assumed to have two states represented by $T(l)$. The MMSE estimator of the clean speech is

$$\hat{X}(l,k) = \sum_{t=0}^{1} \left\{ \begin{array}{l} P\left(T(l)=t\,|Y(l,k),B(l,k)\right) \\ \cdot E\left\{X(l,k)\,|Y(l,k),B(l,k),T(l)=t\right\} \end{array} \right\} \tag{2.9}$$

where t is 0 (non-speech) or 1 (speech).

Later a normalized complex spectral based on a mixture of Gaussian for speech model was proposed as an extended work in the probabilistic scenario [49]. Subramanya et al. adopted the dynamic Bayesian network to factorize joint distribution and infer posteriors, in conjunction with constraints to smooth the distortions between frames. As this method has induced a relatively pleasing result, we will further discuss it in chapter 3, and then compare it to our method in Chapter 4.

2.5 Other methods

2.5.1 Generalized Multichannel Wiener filter

In [31], McCree et al. extended previous work on fixed waveform fusion to an optimal dynamic waveform fusion algorithm that minimizes both additive noise and signal distortion in the estimated speech signal. McCree et al. proposed a minimum mean square error (MMSE) waveform matching criterion for a generalized multichannel Wiener filter,

simultaneously performing waveform fusion, noise suppression, and cross-channel noise cancellation.

In the frequency domain, the bone conducted speech is assumed to involve an additive distortion as

$$Y(f) = [1 + H_d(f)]\, S(f) \tag{2.10}$$

where $H_d(f)$ represents the unknown component of the overall transfer function. Note that if this transfer function noise is independent of the signal, it can be viewed as:

$$Y(f) = S(f) + D(f) \tag{2.11}$$

The signal coherence between clean speech and distortion is given to derive the distortion power spectrum:

$$\rho = \frac{\|s\|^2 + \langle S, D\rangle}{\sqrt{\|s\|^2 \left(\|s\|^2 + \|D\|^2 + \langle S, D\rangle\right)}} \tag{2.12}$$

Assuming the signal and distortion to be uncorrelated, we have the power spectrum of the distortion as:

$$P_d = P_s \left(\frac{1}{\rho^2} - 1\right) \tag{2.13}$$

where P_d is the power spectrum of distortion and P_s is the power spectrum of clean speech.

When we have two channels, the correlation is expressed as:

$$R_{sy} = \begin{bmatrix} P_s \\ P_s \end{bmatrix} \tag{2.14}$$

$$R_{yy} = \begin{bmatrix} P_s + P_n & P_s \\ P_s & P_s + P_d \end{bmatrix} \tag{2.15}$$

where P_n is the noise spectrum. And the multichannel wiener filter is derived as the ratio of cross correlation and autocorrelation:

$$G(f) = R_{yy}^{-1}(f) R_{sy}(f) \tag{2.16}$$

To eliminate the musical noise, a Lagrangian technique with a constraint on the sum of the coefficients is utilized:

$$E = \left\|S - \hat{S}\right\| + 2\lambda \|S\|^2 \left|1 - \sum_i G_i^c\right| \tag{2.17}$$

This leads to the constrained wiener filter:

$$G^c(f) = \frac{G(f)}{G_{sum}(f)} \tag{2.18}$$

They have developed a three-stage approach for robust signal power estimation. In the first step, the clean speech power is estimated by subtraction of noise power:

$$P_s^{(1)} = P_y - P_n \tag{2.19}$$

Then in the second step, speech power P_s is given by a gain-constrained waveform fusion scheme:

$$P_s^{(2)}(f) = (G^c)^T(f)(R_{yy} - R_{nn})G^c(f) \tag{2.20}$$

Finally, time smoothing is incorporated using the decision-directed a priori estimator of Ephraim and Malah [13], where the current frame signal power is smoothed with the previous filter output as:

$$P_s^{(3)}(f) = (1-\alpha)P_s^{(2)}(f) + \alpha\left|\hat{S}_{prev}\right|^2 \tag{2.21}$$

A Diagnostic Rhyme Test (DRT) has demonstrated a significant performance improvement for the fusion system as compared to the equalized resident microphone. Non-acoustic signals can be optimally exploiting by incorporating a multiplicative noise model in frequency domain into a multichannel Wiener filtering approach. It has been shown that the dynamic waveform fusion algorithm provides significant intelligibility and quality improvement for low-rate coding in difficult acoustic environments.

2.5.2 SPLICE

In [57], Zheng et al. presented a hardware device combining a regular microphone with a bone-conducted microphone. With appearance similiar to regular headset, the device can be plugged into any machine with a USB port. It demonstrates a robust speech detection result, eliminating more than 90% of background speech. Furthermore, by combining both channels, they significantly remove background speech even when the background speaker speaks at the same time as the speaker wears the headset.

A moving-window histogram approach is utilized for speech detection, by first computing the histogram of the audio data prior to this frame with a time efficient update

scheme. The second step is to find the valley after the first peak as the energy separator d . Given any frame, let e denote its energy and set $r = e/d$. The speech confidence for this frame is derived as:

$$p = \begin{cases} 0 & : \quad r < 1 \\ \frac{r-1}{\alpha-1} & : \quad 1 \leq r \leq \alpha \\ 1 & : \quad r > \alpha \end{cases} \tag{2.22}$$

where α is a user specified parameter which defines the grace transition period between the two states. The speech confidence p is used to modulate the audio signals captured by the regular microphone, removing the noise between speeches. The modulation factor f is set to be $f = h(p)$, where $h(x)$ is a Hermite polynomial whose derivatives at 0 and 1 are both zeros. In a certain frame, sample of the regular microphone signal x is modulated as $\tilde{x} = f * x$.

The scheme includes two steps, first predicting x from b, and then combining both b and y to obtain the final estimate of x. A SPLICE-like technique is exploited in order to predict x from b. SPLICE is a frame-based noise removal algorithm for cepstral enhancement in the presence of additive noise [40]. Instead of learning the mapping from the corrupted speech y to clean speech x as in the original SPLICE algorithm, here it learns the mapping from bone sensor signals b to clean speech x.

A piecewise linear representation is applied to model the mapping from b to x in the cepstral domain.

$$\begin{aligned} p\left(x, b\right) &= \sum_{s} p\left(x \left| b\right., s\right) p\left(b \left| s\right.\right) p\left(s\right) \\ p\left(x \left| b\right., s\right) &= N\left(x; b + r_s, \Gamma_s\right) \\ p\left(b \left| s\right.\right) &= N\left(b; \mu_b, \Gamma_b\right) \end{aligned} \tag{2.23}$$

where the bone sensor model $p(b)$, consisting of 128 diagonal Gaussian mixture components, is trained by standard EM techniques. The parameters r_s of the conditional pdf $p(x|b, s)$ is derived according to the maximum likelihood criterion.

$$r_s = \frac{\sum_n p\left(s \left| b_n\right.\right)\left(x_n - b_n\right)}{\sum_n p\left(s \left| b_n\right.\right)} \tag{2.24}$$

$$\Gamma_s = \frac{\sum_n p\left(s \left| b_n\right.\right)\left(x_n - b_n - r_s\right)\left(x_n - b_n - r_s\right)^T}{\sum_n p\left(s \left| b_n\right.\right)} \tag{2.25}$$

$$p\left(s \left| b_n\right.\right) = \frac{p\left(b_n \left| s\right.\right) p\left(s\right)}{\sum_n p\left(b_n \left| s\right.\right) p\left(s\right)} \tag{2.26}$$

With regard to additive noise, the air-conducted speech in the cepstral domain are commonly expressed as $y = x + C \ln\left(1 + e^{C^{-1}(n-x)}\right)$, where C is the DCT matrix and n is the noisy cepstral [22]. However, since the noise estimation in the cepstral domain involves highly nonlinear process, it is more feasible to handle both signals in the power spectral feature domain. Based on two assumptions that the noise level in the bone conducted speech is negligible and that the additive noise is uncorrelated with the speech signal, the problem can then be formulated as follows:

$$\begin{aligned} S_y(\omega) &= S_x(\omega) + S_n(\omega) \\ S_x(\omega) &= f\left(S_y(\omega), S_b(\omega)\right) \\ S_x(\omega) &= H(\omega) S_y(\omega) \end{aligned} \tag{2.27}$$

where S_y, S_x, S_b and S_n are the power spectrum for noisy speech, clean speech, bone speech, noise respectively, and $f(z)$ is a nonlinear transformation. The ultimate goal is to find the optimal Wiener filter H. Given two observations S_y and S_b under Gaussian approximation, the MMSE estimator is (omit ω for convenience):

$$\hat{S}_x = \left(\sum\nolimits_n^{-1} + \sum\nolimits_{x|b}^{-1}\right)^{-1} \left[\sum\nolimits_n^{-1}(S_y - \mu_n) + \sum\nolimits_{x|b}^{-1} \hat{S}_{x|b}\right] \tag{2.28}$$

where

$$\begin{aligned} \hat{S}_{x|b} &= e^{C^{-1}\hat{x}} \\ \hat{x} &= b + \sum_s p(s|b) r_s \\ \textstyle\sum_{x|b} &= Var\left(S_{x|b}\right) \\ \mu_n &= E[S_n], \quad \textstyle\sum_n = Var(S_n) \end{aligned} \tag{2.29}$$

μ_n and $\sum_n$ are estimated during non-speech section detected by the speech detection algorithm. Given S_y and $\hat{S}_x$, the estimated Wiener filter $\hat{H}$ can be calculated as:

$$\hat{H} = \frac{\hat{S}_x}{S_y} \tag{2.30}$$

To measure the quality of the reconstruction result, mean opinion score (MOS) evaluations are conduted, demonstrating a 0.6 improvement of enhanced speech for 0dB noise and 0.3 for 10dB noise.

2.5.3 Nonlinear Information Fusion

Nonlinear information fusion focuses on exploration of dynamic properties relatively invariant between the bone-conductive sensors signal and the clean speech signal [9]. The

property employed here is the unobserved (hidden) vocal tract resonances (VTR), with experimental results demonstrating strong correlations between VTR extracted from both sensors under noisy environments, implying the VTR extracted from the bone sensor under noisy conditions indeed reflect clean speech dynamic properties. The VTR sequence, obtained by adaptive Kalman filtering [8], is represented by resonance frequencies and bandwidths corresponding to the lowest P poles in the all-pole speech model as:

$$x_{VTR} = (f, b)' = (f_1, f_2, ..., f_P, b_1, b_2, ..., b_P)' \tag{2.31}$$

The state-space fomulation of speech dynamic model is constructed as:

$$x_{VTR}(t+1) = \Phi x_{VTR}(t) + [I - \Phi]\, u + w(t) \tag{2.32}$$

where Φ is the system matrix, and u is the averaged VTR target vector, constraining the phone-independent mean values of the VTR. The observation equation of the speech dynamic model is

$$o(t) = C[x_{VTR}(t)] + \mu + v(t) \tag{2.33}$$

where $o(t)$ is the observation sequence from the bone sensor in the form of LPC cepstra. The nonlinear function $C[x_{VTR}(t)]$ has the following explicit form:

$$C(i) = \sum_{p=1}^{P} \frac{2}{i} e^{-\pi i \frac{b_p}{f_s}} \cos\left(2\pi i \frac{b_p}{f_s}\right), i = 1, ..., m \tag{2.34}$$

where f_s is the sampling frequency, i is the order of the cepstrum up to m, and p is the pole order of the VTR up to P. Considering the modeling error due to the missing zeros and additional poles beyond P, the residual vector μ is trained to complement the zero-mean noise.

Then the linear cepstral sequence is generated with the extracted VTR sequence, followed by an inverse cepstrum transform to the synthesized speech $Y_3(l,k)$. The final clean speech is reconstructed by fusion of three input data streams

$$\begin{aligned} Y_1(l,k) &= X(l,k) + N[0, \textstyle\sum_1] \\ Y_2(l,k) &= H(k)X(l,k) + N[0, \textstyle\sum_2] \\ Y_3(l,k) &= G(k)X(l,k) + N[0, \textstyle\sum_3] \end{aligned} \tag{2.35}$$

where $Y_1(l,k)$, $Y_2(l,k)$ represents AC and BC speech, $H(k)$ represents the bone microphone channel distortion, and $G(k)$ represents the overall channel distortion from clean

speech to bone-sensor distorted speech and then to the synthesized speech. The normally-distributed random vector $N\left[0, \sum\right]$ represents the additive interfering noise spectrum.

The estimation can be regarded as a weighted sum of three streams by maximum likelihood fusion rule:

$$\hat{X}(l,k) = W_1 Y_1(l,k) + W_2 \left[\bar{H}^{-1} Y_2(l,k)\right] + W_3 \left[\bar{G}^{-1} Y_3(l,k)\right] \tag{2.36}$$

where

$$\begin{aligned} W_1 &= \frac{\sum_1^{-1}}{\sum_1^{-1} + \sum_2^{-1} \left|\bar{H}(k)\right|^2 + \sum_3^{-1} \left|\bar{G}(k)\right|^2} \\ W_2 &= \frac{\sum_2^{-1} \left|\bar{H}(k)\right|^2}{\sum_1^{-1} + \sum_2^{-1} \left|\bar{H}(k)\right|^2 + \sum_3^{-1} \left|\bar{G}(k)\right|^2} \\ W_3 &= \frac{\sum_3^{-1} \left|\bar{G}(k)\right|^2}{\sum_1^{-1} + \sum_2^{-1} \left|\bar{H}(k)\right|^2 + \sum_3^{-1} \left|\bar{G}(k)\right|^2} \end{aligned} \tag{2.37}$$

and $\bar{H}$ is the estimated channel distortion function for the bone sensor [30]. $\bar{G}$ is estimated in a similar way, with synthesized speech waveform based on the extracted hidden dynamics instead of on the bone-sensor data. Note that the three weights sum to one, i.e., $W_1 + W_2 + W_3 = 1$, each corresponding to the AC speech signal, inverse-filtered bone sensor signal and inverse-filtered synthetic signal respectively.

Experiments are conducted to improve accuracy of speech recognition. The baseline accuracy is 72.21% for two-stream fusion, and 55.00% for the one-stream noisy-speech input. With a careful setting of variance scaling factor, a sizable accuracy improvement is achieved, as the nonlinear VTR extraction process creates complementary signal properties in the new stream. However, the synthesized spectra induces significant distortions, as a result, better synthesized techniques need to be developed.

2.5.4 Model-based Fusion

In [19], a probabilistic framework with integrated bone and air microphone is formulated for robust speech recognition. A simple and tractable model to accomplish it is a Gaussian mixture model on the high-resolution log spectra of each sensor, with the frequency components conditionally independent given the state. In [26], a high resolution model extracting the harmonic structure of speech is used to extrapolate harmonics from the bone sensor to the air sensor, since the bone signal can preserve pitch information.

Denote the log amplitude of a short time Fourier transform of the clean air signal and bone signal for frequency f as x_f^a and x_f^b respectively. Then the model is defined in as:

$$p\left(x^a, x^a, s_x\right) = p\left(s_x\right) \prod_f p\left(x_f^a \left| s_x\right.\right) p\left(x_f^b \left| s_x\right.\right) \tag{2.38}$$

where s_x represent the discrete speech state.

A mixture of Gaussian model is built for the observation probabilities. Let $N\left(x; \mu, \sigma\right)$ denote the univariate normal distribution defined on x with mean μ and variance σ. Then formulate the model as follows:

$$\begin{aligned} p\left(s_x\right) &= \pi_{s_x} \\ p\left(x_f^a \left| s_x\right.\right) &= N\left(x_f^a; \mu_{s_x}^a, \sigma_{s_x}^a\right) \\ p\left(x_f^b \left| s_x\right.\right) &= N\left(x_f^b; \mu_{s_x}^b, \sigma_{s_x}^b\right) \end{aligned} \tag{2.39}$$

The parameters can be estimated using the standard expectation maximization (EM) algorithm [7]. A similar model is posited for the noise:

$$\begin{aligned} p\left(s_n\right) &= \pi_{s_n} \\ p\left(n_f^a \left| s_n\right.\right) &= N\left(n_f^a; \mu_{s_n}^a, \sigma_{s_n}^a\right) \\ p\left(n_f^b \left| s_n\right.\right) &= N\left(n_f^b; \mu_{s_n}^b, \sigma_{s_n}^b\right) \end{aligned} \tag{2.40}$$

The model for a given frame of noisy speech in the frequency domain is

$$Y_f = X_f + N_f \tag{2.41}$$

where X_f, N_f, and Y_f denote the complex Fourier transform at frequency f of the clean signal, the noise, and the noisy sensor signal respectively. We denote $x_f \triangleq \ln \left|X_f\right|^2$, and likewise for y_f and n_f, thus the relation in high resolution log power spectrum domain is

$$y_f = x_f + \ln\left(1 + \exp\left(n_f - x_f\right)\right) + \ln\left(1 + \frac{2\left|X_f\right|\left|N_f\right|\cos\theta}{\left|X_f\right|^2 + \left|N_f\right|^2}\right) \tag{2.42}$$

where θ is the angle X between N. The above equation can be assumed obeying a Gaussian distribution as [27]

$$p\left(y_f \left| x_f, n_f\right.\right) = N\left(y_f; x_f + \ln\left(1 + \exp\left(n_f - x_f\right)\right), \Psi\right) \tag{2.43}$$

where Ψ is the variance. Define the concatenation of the AC and BC speech as

$$x \triangleq \begin{bmatrix} x^a \\ x^b \end{bmatrix}, n \triangleq \begin{bmatrix} n^a \\ n^b \end{bmatrix}, z = \begin{bmatrix} x \\ n \end{bmatrix} \tag{2.44}$$

To obtain the expected value of the clean speech given the noisy speech captured by both air and bone sensors, first estimate the posterior $p\left(x^a \left| y^a, y^b\right.\right)$. According to the Bayesian rule, the posterior is computed by a mixture of individual posteriors:

$$p\left(x^a \left| y^a, y^b\right.\right) = \sum_{s^x, s^n} p\left(s^x, s^n \left| y^a, y^b\right.\right) p\left(x^a \left| y^a, y^b, s^x, s^n\right.\right) \tag{2.45}$$

The individual mixture components can be decomposed due to conditional independence given the state. As a result, with $p\left(x^a \left| y^a, y^b, s^x, s^n\right.\right) = p\left(x^a \left| y^a, s^x, s^n\right.\right)$, they are derived as:

$$p\left(x^a \left| y^a, s^x, s^n\right.\right) = \frac{1}{z} p\left(x^a \left| s^x\right.\right) \int p\left(y^a \left| x^a, n^a\right.\right) p\left(n^a \left| s^n\right.\right) dn^a \tag{2.46}$$

where z is a normalizing constant. This posterior is non-Gaussian and analytically intractable. To solve this problem an iterative Laplace method is utilized to approximate the posteriors in a framework known as Algonquin [27].

By linearizing the function $g\left(z\right) \triangleq x + \ln\left(1 + \exp\left(n - x\right)\right)$ using a first order Taylor series expansion at the point z_0, the likelihood can be written as:

$$p_l\left(y \left| x, n\right.\right) = p_l\left(y \left| z\right.\right) = N\left(y; g\left(z_0\right) + G\left(z_0\right)\left(z - z_0\right), \Psi\right) \tag{2.47}$$

where

$$G\left(z_0\right) = \left[diag\left(\frac{\partial g\left(z\right)}{\partial x}\right), diag\left(\frac{\partial g\left(z\right)}{\partial n}\right)\right]_{z_0} \tag{2.48}$$

is a matrix of the derivatives of $g(z)$, evaluated at z_0.

The Gaussian approximation to the posterior for a particular speech and noise combination can be written as $p(x, n|\, y, s^x, s^n) \sim \mathrm{N}(\eta_s, \Phi_s)$ where [27]

$$\eta_s = \Phi_s \left[\sum_s^{-1} \mu_s + G(z_0)^T \Psi^{-1}\left(y - g\left(z_0\right) - G\left(z_0\right) z_0\right)\right] \tag{2.49}$$

$$\Phi_s = \left[\sum_s^{-1} + G(z_0)^T \Psi^{-1} G\left(z_0\right)\right]^{-1} \tag{2.50}$$

and the posterior mixture probability $p(y|\, s^x, s^n)$ can be shown to be

$$\gamma_s = \left|\sum_s\right|^{-1/2} \Psi^{-1/2} |\Phi_s|^{-1/2} \exp\left[-\frac{1}{2}\left(\begin{array}{l} \mu_s^T \sum_s^{-1} \mu_s \\ +(y - g\left(z_0\right) - G\left(z_0\right) z_0)^T \\ \Psi^{-1}\left(y - g\left(z_0\right) - G\left(z_0\right) z_0\right) - \eta_s^T \Phi_s^{-1} \eta_s \end{array}\right)\right]^{-1} \tag{2.51}$$

The final estimation is expressed as:

$$\hat{x}^a = E\left(x^a \left| y^a, y^b\right.\right) = \frac{\sum_s \gamma_s \eta_s^a}{\sum_s \gamma_s} \tag{2.52}$$

The enhanced results are tested using a commercial speech recognition system with baseline percent accuracy in the air microphone 57.2% for the noisy speech, and 92.66% for the clean speech. Best result of 79% accuracy is obtained when using integrated microphones with additional scheme, i.e., noise adaptation and speech state posterior determined solely on the bone sensor observation.

Whereas the high-resolution model can focus on high signal-to-noise ratio peaks in the spectrum, hundreds of states is required to represent all combinations of pitches and formant information. The model posits conditional independence of the bone and air signals given state. Thus improvements could include a model with a direct state-dependent correlation between the air and bone sensor. Another direction of improvement could be online noise adaptation as well as adaptation to varying channel characteristics, which may allow for the development of a speaker independent system.

2.5.5 Bandwidth Extension Method

Bone conducted speech reconstruction can be regarded as bandwidth extension (BWE) of telephony speech as they share the same bandlimit property. In [42], Seltzer et al. presented a new bandwidth extension algorithm for converting narrowband telephone speech into wideband speech using a transformation in the mel cepstral domain. Unlike previous approaches, the proposed method is designed specifically for bandwidth extension of narrowband speech that has been corrupted by environmental noise. Performing BWE in noisy enviroments is problematic for several reasons. Firstly, most BWE algorithms utilize LPC-derived features, such as LPC-cepstra or LSF coefficients, to represent both the narrowband and the extended frequency spectral envelopes in the codebooks or mixture models. However, because additive noise introduces zeros in the speech spectrum, noise-corrupted speech is no longer well-represented by an all-pole model. Secondly, if the noisy signal is pre-processed using a speech enhancement algorithm prior to BWE, the enhanced speech may not be optimal for the subsequent BWE processing.

By exploiting previous research in mel cepstrum feature enhancement, a unified probabilistic framework is created under which the feature denoising and bandwidth extension processes are tightly integrated using a single shared statistical model, enabling both denoising the observed narrowband speech and robustly extending its bandwidth in a jointly optimal manner.

The feature extraction process for generating MFCCs can be summarized as:

$$z = C \log \left(W |Z|^2 \right) \tag{2.53}$$

where $|Z|^2$ is the power spectrum of a frame of speech in a Short-Time Fourier Transform, W is the matrix of weighting coefficients of the mel filterbank and C is a DCT matrix.

Define x to be a narrowband MFCC feature vector and z to be the corresponding wideband MFCC feature vector. The goal is to estimate z from x, i.e., to compute $E\left[z\,|x\right]$. Under the assumtion that the observed narrowband feature vectors are generated by a Gaussian mixture model (GMM), the probability distribution $p\left(x\right)$ can be written as :

$$p\left(x\right) = \sum_{s=1}^{S} p\left(x\,|s\right) p\left(s\right) = \sum_{s=1}^{S} N\left(x; \mu_s, \sum_s\right) p\left(s\right) \tag{2.54}$$

where $N\left(x; \mu_s, \sum_s\right)$ is a Gaussian distribution with mean μ_s and covariance $\sum_s$, $p\left(s\right)$ is the prior probability for state s, and S is the total number of Gaussians in the mixture. This model is trained from narrowband training data using conventional EM.

The transformation from narrowband to wideband feature vectors is modeled as a piecewise-linear transformation. More specifically, it is assumed that

$$z = A_s x + b_s + e \tag{2.55}$$

where $p(e) = N(e; 0, I)$, and the transformation parameters $\{A_s, b_s\}$ are dependent on the Gaussian mixture component s. As a result, the conditional probability of z is:

$$p\left(z\,|x, s\right) = N\left(z; A_s x + b_s, I\right) = N\left(z; A'_s x', I\right) \tag{2.56}$$

where $A'_s = [A_s\ b_s]\ \ and\, x' = [x\ 1]^T$.

The transformation parameters $\{A_1{}', ..., A_s{}'\}$ are learned using a corpus of stereo training data in which each narrowband feature vector x_t has a corresponding wideband feature

vector z_t. For each state s, the ML estimate of A'_s is

$$A'_s = \left(\sum_{t=1}^{T} p\left(s \left| x_t\right.\right) z_t {x'_t}^T \right) \left(\sum_{t=1}^{T} p\left(s \left| x_t\right.\right) x'_t {x'_t}^T \right)^{-1} \tag{2.57}$$

Then $E\left[z \left| x\right.\right]$ can be computed as

$$\begin{aligned} E\left[z \left| x\right.\right] &= \sum_{s=1}^{S} \int z p\left(z, s \left| x\right.\right) dz \\ &= \sum_{s=1}^{S} p\left(s \left| x\right.\right) \int z p\left(z, s \left| x\right.\right) \\ &= \sum_{s=1}^{S} p\left(s \left| x\right.\right) A'_s x' \end{aligned} \tag{2.58}$$

where the state posterior probability $p\left(s \left| x\right.\right)$ is derived from the GMM as

$$p\left(s \left| x\right.\right) = \frac{p\left(x \left| s\right.\right) p\left(s\right)}{\sum_{s'=1}^{S} p\left(x \left| s'\right.\right) p\left(s'\right)} \tag{2.59}$$

When the speech and noise are uncorrelated, the relationship in MFCC domain is [32]:

$$\begin{aligned} y &= x + C \log\left(1 + C^{\dagger} \exp\left(n - x\right)\right) \\ &= x + f\left(x, n\right) \end{aligned} \tag{2.60}$$

where x, y and n are the MFCC vectors for the clean speech, noise-corrupted speech, and the noise, respectively. Thus, the noisy cepstrum y is decribed by an additive bias f which is a non-linear function of the speech and noise cepstra.

This relationship has been widely used in the speech recognition community as the basis of a family of feature enhancement (FE) algorithms. It could be solved under an EM framework utilizing a prior speech model, e.g. a GMM or HMM, and an iterative Vector Taylor Series (VTS) approximation algorithm. At each iteration, a VTS expansion is used to linearize f, and then compute the ML estimate of the posterior distributions of the hidden variables x and n. Expectations of these distributions becomes the new VTS expansion point for the next iteration, and it is repeated until convergence. At this point, a MMSE estimate of the clean cepstral vector is generated as

$$E\left[x \left| y\right.\right] = \sum_{s=1}^{S} p\left(s \left| y\right.\right) \int x p\left(x \left| y, s\right.\right) dx \tag{2.61}$$

where $p(s|y)$ is the state posterior distribution, and $p(x|y,s)$ represents the state-conditional posterior distribution of the clean cepstra.

Using Bayes rule, it can be written as

$$\begin{aligned} E\left[z\left|y\right.\right] &= \sum_{s=1}^{S} p\left(s\left|y\right.\right) \int z\left(p\left(z,x\left|y,s\right.\right)dx\right)dz \\ &= \sum_{s=1}^{S} p\left(s\left|y\right.\right) \int zp\left(z\left|y,s\right.\right)dz \end{aligned} \tag{2.62}$$

where $p\left(s\left|y\right.\right)$ can be obtained directly from the output of the feature enhancement algorithm. Notice that rather than relying on a point estimation of the narrowband clean cepstral vector, x is marginalized over all values, enabling the solution much more robust to estimation errors.

Then it can be shown that the final expression for the expected value is

$$E\left[z\left|y\right.\right] = \sum_{s=1}^{S} p\left(s\left|y\right.\right) A'_s v'_s \tag{2.63}$$

given $p\left(x\left|y,s\right.\right) = N\left(x;v_s,\Phi_s\right)$, with the same GMM for both feature enhancement and bandwidth extension.

To evaluate the performance of the algorithm, the mean RMS log spectral distortion (RMS-LSD) of the smooth power spectral envelopes between the original wideband speech S_Z and the bandwidth-extended speech $\hat{S}_Z$ over the extended frequencies is computed. Three different methods of BWE on noisy speech are compared, including BWE performed directly on the noisy speech, first enhanced with conventional Wiener filtering followed by BWE and proposed algorithm. It has been shown that enhancing the speech prior to performing BWE results in minimal improvement over BWE processing on the noisy speech directly. On the other hand, the proposed FE-BWE algorithm results in significantly less distortion in both the observed narrowband spectral envelope and the extended high frequency envelope.

The bandwidth extension algorithm utilizes a GMM trained on narrowband speech and a state-conditional affine transformation in the MFCC domain to transform the narrowband spectral envelope into a wideband spectral envelope. The proposed bandwidth extension (FE-BWE) algorithm significantly outperforms a more conventional approach. However, a more robust BWE algorithm must also handle the effects of additive noise in the excitation signal.

2.6 Summary

In this chapter, we reviewed existing methods from two perspectives, first when BC microphone works as a supplementary sensor, second when BC microphone functions as a dominant sensor. In the first category, BC speech is employed for voice activity detection, pitch tracking and enhancing low frequency part of the reconstructed speech. In the second category, there are mainly three approaches: equalization, analysis-and-synthesize and probabilistic approach. Among them, equalization and analysis-and-synthesize methods need preprocessing, thus are not suitable in practical situation. Whereas probabilistic approach has steadily been improved from the initial version and provides a favorable result with dynamic Bayesian network assumption. After demonstrating three main groups of approaches, we also present other methods which need to be considered as well, i.e., generalized multichannel Wiener filter, SPLICE, nonlinear information fusion, model-based fusion and bandwidth extension method. These methods are derived based on various speech and noise model assumptions, whereas suffering from computational complexity.

Chapter 3

Probabilistic Approach

The probabilistic approach is based on the method proposed in [49]. In [49], Subramanya et al. present an integrated sensor, in particular the bone-conducted sensor with standard air-conducted microphone to deal with highly non-stationary noise. The speech signal is modeled on the magnitude-normalized spectra, enabling alleviating intra- and inter-speaker variation with a small number of Gaussian components. The algorithms are developed taking advantage of additional sensor streams to reliably estimate the clean signal. In this chapter, we describe the method and show a different speech presence detection procedure [46].

3.1 Model

An air-and-bone conductive (ABC) microphone is developed by making use of a bone-conducted microphone in addition to the regular air microphone. Figure 3.1 demonstrates two prototypes of the ABC microphone. In the case of the first prototype, the bone sensor resides on user's temple and the air channel is a close-talking microphone. In the case of the second prototype, the bone sensor is positioned behind the ear, while the air channel is a medium-sized boom of 45 mm. Either case, the bone-conducted microphone captures the vibrations of the speakers' bones and skin during the articulation.

In Bayesian statistics, we exploit the prior information of hidden variables for inference. For example, to estimate the hidden variable z given some observation o, we can utilize the Bayes rule $p(z|o) = kp(o|z)p(z)$, with constant k independent of z. In the above

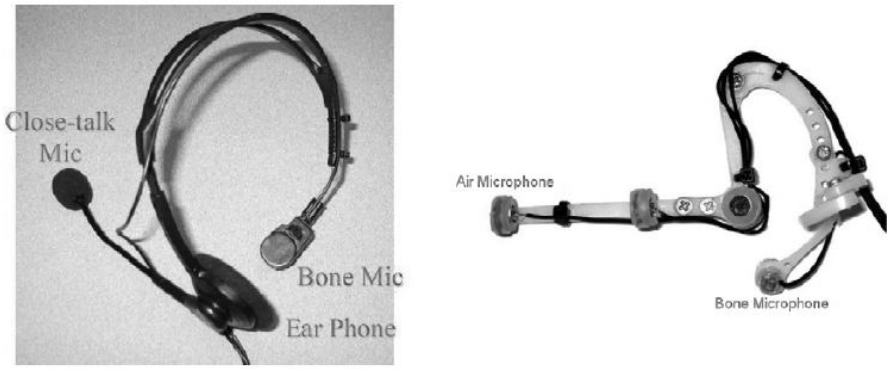


Figure 3.1: Two prototypes of air-and-bone conductive (ABC) microphone headsets.

equation, $p(z)$ is the prior information about z, and represents the knowledge about z that is known even before o is observed. In the case of speech enhancement, a speech model provides a prior on clean speech that is hidden given noisy speech. However, modeling human speech is an extremely complex problem owing to its large variability.

A proper domain selection ensuring overall user experience improvement and computation saving includes two constrains, 1) to reproduce both the magnitude and phase of the reconstructed signal; 2) to rule out the time domain. The Mel-cepstral and real-cepstral domain, though attractive from the perceptual point of view, does not take into account the phase information. It is also not entirely straightforward modeling phase in the log-spectral domain, let alone the non-linearity it brings. As a result, the normalized complex spectral domain is chosen for signal processing. For each frame, the speech signal is normalized with their energy, i.e.,

$$\tilde{X}_t = \frac{X_t}{\|X_t\|} \tag{3.1}$$

Althogh the normalizing procedure causes a reduction of the number of Gaussian mixture, it reduces the variance of the captured speech, thus a gain term g_{x_t} requires to compensate the model imperfection.

3.2 Training

Speech segments extracted from a collection of clean speech are used as the training data. The resulting speech frames are energy normalized and denoted by $\left\{\tilde{X}_t\right\}_{t=1}^{T}$. The

normalized speech frames are modeled into a mixture of Gaussian. The model is learned by k-means algorithm with the following distance metric:

$$\begin{aligned} d\left(\tilde{X}_i, \tilde{X}_j\right) &= \left\| \left[d\left(\tilde{X}_i^1, \tilde{X}_j^1\right), ..., d\left(\tilde{X}_i^f, \tilde{X}_j^f\right) ..., d\left(\tilde{X}_i^N, \tilde{X}_j^N\right) \right]^T \right\|, \ 1 \leq i, j \leq T \\ d\left(\tilde{X}_i^f, \tilde{X}_j^f\right) &= \log\left|\tilde{X}_i^f\right| - \log\left|\tilde{X}_j^f\right|, \ 1 \leq f \leq N \end{aligned} \tag{3.2}$$

to yield M clusters. The mean and variance of these Gaussians are set to the mean and variance of the clusters obtained above. The fraction of each Gaussian in the mixtures is represented by

$$\alpha_i = N(i)/T, \qquad 0 \leq i \leq M-1 \tag{3.3}$$

where $N(i)$ is the number of elements in the ith cluster, thus $\sum_{i=1}^{M} \alpha_i = 1$.

3.3 Network Description

A dynamic Bayesian network (DBN) represents a family of probability distributions defined in terms of a directed graph. The joint probability distribution over the variables, represented by the nodes in the graph, is obtained by taking products over functions on connected subsets of nodes. DBNs provide general algorithms for computing marginal and conditional probabilities of interest [3, 23, 59] by exploiting graph-theoretic representations. A characteristic of DBNs distinguishing them from other classes of graphical models with some edges of graph point in the direction of increasing time. DBNs have been applied to many tasks in the past including, speech recognition [59], vision applications such as tracking [1], natural language processing (NLP) applications such as parsing, tagging [24].

The AC speech $y(k)$ and BC speech $b(k)$ are transformed to the short-time Fourier transform (STFT) domain as Y_t and B_t respectively, where t is the frame index. Let X_t be a frame of the clean speech signal that we wish to estimate given Y_t and B_t. The model is given as:

$$\begin{aligned} Y_t &= X_t + V_t + U_t \\ B_t &= H_t X_t + G_t V_t + W_t \end{aligned} \tag{3.4}$$

where the air microphone Y_t is a sum of the clean speech signal X_t, the corrupting noise V_t and the sensor noise in the air microphone U_t. The signal captured by the bone

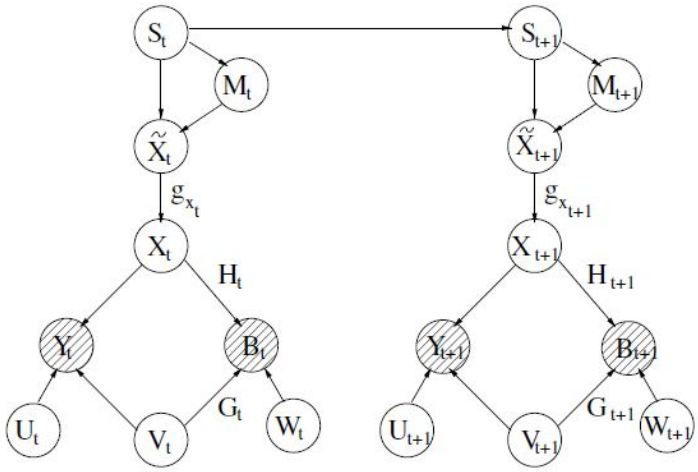


Figure 3.2: A dynamic Bayesian network (DBN) of the speech enhancement framework. (after [49])

microphone B_t is a sum of the transformed version of the clean speech signal $H_t X_t$, the much attenuated and distorted background noise that leaks into the bone sensor $G_t V_t$, and the sensor noise in the bone microphone W_t.

Figure 3.2 illustrates two frames of a DBN used to model the enhancement process in the complex spectral domain. The shaded nodes are observed variables. In this model,

• S_t is a discrete random variable representing the state (speech/non-speech) of the frame,

• M_t is a discrete random variable acting as an index of the mixture distributions modeling speech/non-speech,

• $\tilde{X}_t$ represents magnitude-normalized speech,

• g_{x_t} scales $\tilde{X}_t$ to match the clean speech X_t,

• V_t is the background noise,

• U_t and W_t represent the sensor noise in the air and bone channels respectively,

• H_t the optimal linear mapping from X_t to B_t,

• G_t models the leakage of background noise into the bone sensor.

Assumptions:

(1) The frequency components of $\tilde{X}_t$, X_t , Y_t, V_t, H_t, G_t, B_t, are independent.

(2) The variables S_t and M_t are scalars and are considered global over the frame at time t, i.e., the value of these variables are the same for all the frequency components for a given frame.

(3) Background noise is modeled using a zero mean Gaussian, i.e., $p(V_t) \sim N(0, \sigma_v^2)$.

(4) Sensor noise in the air channel is modeled with $p\left(U_t\right) \sim N\left(0, \sigma_u^2\right)$.

(5) Sensor noise in the bone channel is modeled with $p\left(W_t\right) \sim N\left(0, \sigma_w^2\right)$.

(6) $p\left(\tilde{X}_t \,|X_t\right) \sim \delta\left(X_t, g_{x_t}\tilde{X}_t\right)$, where δ is the Kronecker delta function with parameter g_{x_t}.

Following the graph model, the joint distribution over all the variables are factorized as follows:

$$\begin{aligned} &p(Y_t, B_t, X_t, \tilde{X}_t, V_t, S_t, M_t, U_t, W_t) \\ &= p(Y_t\,|X_t, V_t, U_t)p(B_t\,|X_t, V_t, W_t)p(X_t|\tilde{X}_t)p(\tilde{X}_t\,|M_t,\, S_t) \\ &\times p(M_t\,|S_t)p(S_t)p(V_t)p(U_t)p(W_t) \end{aligned} \tag{3.5}$$

3.4 Transfer Function and Leakage Factor

Consider the silent regions of an utterance, where $X_t^f = 0, \forall f$, we have

$$\begin{aligned} Y_t^f &= U_t^f + V_t^f, \;\; 1 \le f \le N \\ B_t^f &= W_t^f + G_t^f V_t^f, \;\; 1 \le f \le N \end{aligned} \tag{3.6}$$

where $t \in N_v$ and N_v is the set of non-speech frames in the region $[0, t]$.

The ML estimation of G_t is given by minimizing cost function R, represented by:

$$R = \sum\nolimits_{t \in N_v} \left(\frac{1}{2\sigma_u^2}|Y_t - V_t|^2 + \frac{1}{2\sigma_w^2}|B_t - GV_t|^2 \right) \tag{3.7}$$

The partial derivatives of R should be set to zero to get the optimum point, where we have the derivative of R with respective to V_t as:

$$\frac{\partial R}{\partial V_t} = \sum\nolimits_{t \in N_v} \left(-\frac{1}{\sigma_u^2}(Y_t - V_t)^* + \frac{-G}{\sigma_w^2}(B_t - GV_t)^* \right) \tag{3.8}$$

Setting it to zero, we get

$$V_t = \frac{\sigma_w^2 Y_t + \sigma_u^2 G^* B_t}{\sigma_w^2 + \sigma_u^2 |G|^2} \tag{3.9}$$

Substituting V_t back to R, we have

$$R = \sum\nolimits_{t \in N_v} \left(\frac{|B_t - GY_t|^2}{\sigma_w^2 + \sigma_u^2 |G|^2} \right) \tag{3.10}$$

Now set the derivative of R with respective to G as 0, namely, $\frac{\partial R}{\partial G} = 0$, we have

$$a^* \sigma_u^2 (G^*)^2 - bG^* - a\sigma_w^2 = 0 \tag{3.11}$$

where

$$\begin{aligned} a &= \textstyle\sum_{t\in N_v} {B_t}^* Y_t \\ b &= \textstyle\sum_{t\in N_v} (\sigma_u^2 |B_t|^2 - \sigma_w^2 |Y_t|^2) \end{aligned} \tag{3.12}$$

and the result of leakage factor is given

$$G_t = \frac{\sum_{t\in N_v} (\sigma_u^2 |B_t|^2 - \sigma_w^2 |Y_t|^2) + \sqrt{\left(\sum_{t\in N_v} (\sigma_u^2 |B_t|^2 - \sigma_w^2 |Y_t|^2)\right)^2 + 4\sigma_u^2\sigma_w^2 \left|\sum_{t\in N_v} {B_t}^* Y_t\right|}}{2\sigma_u^2 \sum_{t\in N_v} {B_t}^* Y_t} \tag{3.13}$$

Simiarly, by exploiting the set of speech frames in the region $[0, t]$ denoted by N_s, we have the transfer function H_t as:

$$H_t = G_t + \frac{\sum_{t\in N_s} (\sigma_u^2 |B_t'|^2 - \sigma_w^2 |Y_t|^2) + \sqrt{\left(\sum_{t\in N_s} (\sigma_u^2 |B_t'|^2 - \sigma_w^2 |Y_t|^2)\right)^2 + 4\sigma_u^2\sigma_w^2 \left|\sum_{t\in N_s} (B_t')^* Y_t\right|}}{2\sigma_u^2 \sum_{t\in N_s} (B_t')^* Y_t} \tag{3.14}$$

where $B_t' = B_t - G_t Y_t$. It can be intuitively understood as the scheme to remove the leakage noise in the bone sensor.

3.5 Probabilities

Recall the joint probability distribution in the dynamic Bayesian network is:

$$\begin{aligned} &p(Y_t, B_t, X_t, \tilde{X}_t, V_t, S_t, M_t, U_t, W_t) \\ &= p(Y_t|X_t, V_t, U_t) p(B_t\,|X_t, V_t, W_t) p(X_t|\tilde{X}_t) p(\tilde{X}_t\,|M_t, S_t) \\ &\times p(M_t\,|S_t) p(S_t) p(V_t) p(U_t) p(W_t) \end{aligned} \tag{3.15}$$

where $p(X_t|\tilde{X}_t)$, $p(V_t)$, $p(U_t)$, $p(W_t)$ have been defined previously. Furthermore, from the observation of both sensors, we have that $p(Y_t\,|X_t, V_t, U_t) \sim N(Y_t, X_t + V_t, \sigma_u^2)$ and $p(B_t\,|X_t, V_t, W_t) \sim N(B_t, H_t X_t + G_t V_t, \sigma_w^2)$.

Speech is modeled using a mixture of Gaussians (MG),

$$p(\tilde{X}_t\,|S_t) = \sum_m p(M_t = m\,|S_t) p(\tilde{X}_t\,|S_t, M_t) \tag{3.16}$$

with $p(\tilde{X}_t\,|S_t = s, M_t = m) \sim N\left(\mu_{s,m}, \sigma_{s,m}^2\right)$. $S_t = 0$ indicates the silence (non-speech) state and $S_t = 1$ indicates the speech state. Silence is represented using a single Gaussian, and thus $p(M_t = 0\,|S_t = 0) = 1$ with $p(\tilde{X}_t\,|S_t = 0) \sim N\left(\tilde{X}_t; 0, \sigma_{sil}^2\right)$. In speech segments, a mixture of Gaussian with $M = 4$ and $p(M_t = i\,|S_t = 1) = \alpha_i$ is fomulated.

As X_t and $\tilde{X}_t$ are related by a delta distribution, given g_{x_t}, estimating either one of these variables is equivalent. Thus, integrating out X_t from the joint distribution we have

$$\begin{aligned}
&\textstyle\int_{X_t} p(Y_t, B_t, X_t, \tilde{X}_t, V_t, S_t, M_t, U_t, W_t) dX_t \\
&= p(Y_t, B_t, \tilde{X}_t, V_t, S_t, M_t, U_t, W_t) \\
&= p(Y_t | g_{x_t} \tilde{X}_t, V_t, U_t) p(B_t | g_{x_t} \tilde{X}_t, V_t, W_t) p(\tilde{X}_t \,| M_t, S_t) \\
&\times p(M_t \,| S_t) p(S_t) p(V_t) p(U_t) p(W_t)
\end{aligned} \tag{3.17}$$

where

$$\begin{aligned}
&p(Y_t | g_{x_t} \tilde{X}_t, V_t, U_t) \sim N\left(g_{x_t} X_t + V_t, \sigma_u^2\right) \\
&p(B_t \,| g_{x_t} X_t, V_t, W_t\,) \sim N\left(g_{x_t} H_t \tilde{X}_t + G_t V_t, \sigma_w^2\right)
\end{aligned} \tag{3.18}$$

Now consider the posteriori probability of normalized clean speech spectra given two noisy observations in terms of the Bayes rule,

$$\begin{aligned}
&p(\tilde{X}_t \,| Y_t, B_t\,) = \sum_{s,m} p(\tilde{X}_t, S_t = s, M_t = m \,| Y_t, B_t\,) \\
&= \sum_{s,m} p(\tilde{X}_t \,| Y_t, B_t, S_t = s, M_t = m\,) \\
&\times p(M_t = m \,| Y_t, B_t, S_t = s)\; p(S_t = s \,| Y_t, B_t)
\end{aligned} \tag{3.19}$$

Let us first consider evaluating $p(\tilde{X}_t \,| Y_t, B_t, S_t = s, M_t = m\,)$. Using the definition of conditional probability, we have

$$p(\tilde{X}_t \,| Y_t, B_t, S_t = s, M_t = m\,) = \frac{p(\tilde{X}_t, Y_t, B_t, S_t = s, M_t = m)}{p(Y_t, B_t, S_t = s, M_t = m)} \tag{3.20}$$

And the numerator in the above equation can be obtained by integration,

$$p(\tilde{X}_t, Y_t, B_t, S_t = s, M_t = m) = \int_{V_t} \int_{U_t} \int_{W_t} p(Y_t, B_t, \tilde{X}_t, V_t, S_t, M_t, U_t, W_t) dU_t dW_t dV_t \tag{3.21}$$

Integrating out V_t, U_t and W_t, the joint probablity can be expressed as the production of several Guassians,

$$\begin{aligned}
p(\tilde{X}_t, Y_t, B_t, S_t = s, M_t = m) \sim\; & N\left(\tilde{X}_t; \alpha_{s,m}, \beta_{s,m}\right) N\left(B_t; \gamma_{s,m}, \eta_{s,m}\right) \\
& \times N\left(Y_t; g_{x_t} \mu_{s,m}, \sigma_1^2\right) p(M_t \,| S_t)\; p(S_t)
\end{aligned} \tag{3.22}$$

where the parameters of Guassians are

$$\begin{aligned}
\alpha_{s,m} &\triangleq \frac{\sigma_{s,m}^2\left(\sigma_1^2\left(\sigma_{uv}^2\mu_{s,m}+g_{x_t}Y_t\right)+g_{x_t}H_m^*\left(B_t\sigma_{uv}^2-G\sigma_{uv}^2Y_t\right)\right)}{\sigma_1^2\sigma_2^2+g_{x_t}^2\sigma_{s,m}^2\sigma_{uv}^2|H_m|^2} \\
\beta_{s,m} &\triangleq \frac{\sigma_1^2\sigma_{s,m}^2\sigma_{uv}^2}{\sigma_1^2\sigma_2^2+g_{x_t}\sigma_{s,m}^2\sigma_{uv}^2|H_m|^2} \\
\gamma_{s,m} &\triangleq g_{x_t}H_m\frac{\sigma_{uv}^2\mu_{s,m}+g_{x_t}\sigma_{s,m}^2Y_t}{\sigma_2^2}+\frac{G\sigma_v^2Y_t}{\sigma_{uv}^2} \\
\eta_{s,m} &\triangleq \sigma_1^2+g_{x_t}|H_m|^2\frac{\sigma_{s,m}^2\sigma_{uv}^2}{\sigma_2^2} \\
\sigma_{uv}^2 &\triangleq \sigma_u^2+\sigma_v^2 \\
\sigma_1^2 &\triangleq \sigma_w^2+\frac{|G|^2\sigma_u^2\sigma_v^2}{\sigma_{uv}^2} \\
\sigma_2^2 &\triangleq \sigma_{uv}^2+g_{x_t}^2\sigma_{s,m}^2 \\
H_m &= H-G\frac{\sigma_v^2}{\sigma_{uv}^2}
\end{aligned} \tag{3.23}$$

Plug it back into equation 3.20, we get

$$\begin{aligned}
p(\tilde{X}_t\,|Y_t,B_t,S_t=s,M_t=m) &= \frac{p(\tilde{X}_t,Y_t,B_t,S_t=s,M_t=m)}{p(Y_t,B_t,S_t=s,M_t=m)} \\
&= \frac{p(\tilde{X}_t,Y_t,B_t,S_t=s,M_t=m)}{\int_{\tilde{X}_t}p(Y_t,B_t,S_t=s,M_t=m)d\tilde{X}_t} \\
&\sim \frac{N\left(\tilde{X}_t;\alpha_{s,m},\beta_{s,m}\right)N(B_t;\gamma_{s,m},\eta_{s,m})N\left(Y_t;g_{x_t}\mu_{s,m},\sigma_1^2\right)p(M_t|S_t)\,p(S_t)}{\int_{\tilde{X}_t}N\left(\tilde{X}_t;\alpha_{s,m},\beta_{s,m}\right)N(B_t;\gamma_{s,m},\eta_{s,m})N\left(Y_t;g_{x_t}\mu_{s,m},\sigma_1^2\right)p(M_t|S_t)\,p(S_t)d\tilde{X}_t} \\
&\sim \frac{N\left(\tilde{X}_t;\alpha_{s,m},\beta_{s,m}\right)N(B_t;\gamma_{s,m},\eta_{s,m})N\left(Y_t;g_{x_t}\mu_{s,m},\sigma_1^2\right)p(M_t|S_t)\,p(S_t)}{N(B_t;\gamma_{s,m},\eta_{s,m})N\left(Y_t;g_{x_t}\mu_{s,m},\sigma_1^2\right)p(M_t|S_t)\,p(S_t)} \\
&\sim N\left(\tilde{X}_t;\alpha_{s,m},\beta_{s,m}\right)
\end{aligned} \tag{3.24}$$

In order to evaluate the posterior $p(\tilde{X}_t\,|Y_t,B_t)$ we still need to compute $p(M_t = m\,|Y_t,B_t,S_t=s)$ and $p(S_t=s\,|Y_t,B_t)$. First consider

$$p(M_t=m\,|Y_t,B_t,S_t=s) = \frac{p\left(M_t=m,Y_t,B_t,S_t=s\right)}{p\left(Y_t,B_t,S_t=s\right)} \propto p\left(Y_t,B_t,S_t=s,M_t=m\right) \tag{3.25}$$

Recall from the equation 3.22,

$$\begin{aligned}
&p(Y_t,B_t,S_t=s,M_t=m) = \textstyle\int_{\tilde{X}_t}p(\tilde{X}_t,Y_t,B_t,S_t=s,M_t=m)d\tilde{X}_t \\
&\sim N\left(B_t;\gamma_{s,m},\eta_{s,m}\right)N\left(Y_t;g_{x_t}\mu_{s,m},\sigma_1^2\right)p(M_t\,|S_t)\;p(S_t)
\end{aligned} \tag{3.26}$$

As a result,

$$p(M_t=m\,|Y_t,B_t,S_t=s) \sim N\left(B_t;\gamma_{s,m},\eta_{s,m}\right)N\left(Y_t;g_{x_t}\mu_{s,m},\sigma_1^2\right)p(M_t\,|S_t)\;p(S_t) \tag{3.27}$$

But, recall that M_t is global over a given frame. Thus let

$$k\left(m,f\right) \triangleq N\left(B_t^f;\gamma_{s.m}^f,\eta_{s.m}^f\right)N\left(Y_t^f;g_{x_t}\mu_{s.m}^f,\sigma_1^{f2}\right)p(M_t\,|S_t)\;p(S_t) \tag{3.28}$$

We can compute the value of the posterior using

$$p(M_t = m \,|Y_t, B_t, S_t = s) = \frac{\Pi_f k\,(m, f)}{\sum_m \Pi_f k\,(m, f)} \tag{3.29}$$

The posterior of S_t may be obtained in a similar manner by observing that

$$\begin{aligned} &p(S_t = s\,|Y_t, B_t) \propto \textstyle\sum_m p(Y_t, B_t, S_t = s, M_t = m) \\ &\sim \textstyle\sum_m N\left(B_t^f; \gamma_{s.m}^f, \eta_{s.m}^f\right) N\left(Y_t^f; g_{x_t}\mu_{s.m}^f, \sigma_1^{f2}\right) \\ &\times p(M_t\,|S_t)\; p(S_t) \end{aligned} \tag{3.30}$$

Now, we have derived all the components to our original problem, i.e., computing $p(\tilde{X}_t\,|Y_t, B_t)$. In practice, though, we are more interested in computing $E(\tilde{X}_t\,|Y_t, B_t)$. Thus, taking an expectation w.r.t $p(\tilde{X}_t\,|Y_t, B_t)$ we get

$$\begin{aligned} &\hat{X}_t = E(\tilde{X}_t\,|Y_t, B_t) = p(S_t = 0\,|Y_t, B_t) E(\tilde{X}_t\,|Y_t, B_t, S_t = 0, M_t = 0) \\ &+p(S_t = 0\,|Y_t, B_t) \sum_m p(M_t = m\,|Y_t, B_t\,, S_t = 1) \\ &E(\tilde{X}_t\,|Y_t, B_t, S_t = 1, M_t = m) \end{aligned} \tag{3.31}$$

Note that $E(\tilde{X}_t\,|Y_t, B_t, S_t = s, M_t = m) = \alpha_{s,m}$ and the estimator is essentially a MMSE estimator.

3.6 Estimating the gain g_{x_t}

Define $q\,(f)$ as the joint likelihood over some of the variables in the model,

$$q\,(f) = p\left(\tilde{X}_t^f, Y_t^f, B_t^f, S_t, M_t\right) \tag{3.32}$$

The joint log likelihood over the entire frame is given by

$$F = \log \prod_{all\ \ f} q\,(f) = \sum_{all\ \ f} \log q\,(f) \tag{3.33}$$

where the above equation follows as a result of the frequency independence assumption.

In order to maximize F, a numerical scheme is utilized, i.e, the EM algorithm. The E-step essentially consists of estimating the most-likely value of $\tilde{X}_t$ given the current estimate of g_{x_t} , i.e., $\hat{X}_t = E(\tilde{X}_t\,|Y_t, B_t\,, g_{x_t})$. The M-step involves maximizing the objective function F w.r.t. g_{x_t}. Taking the derivative of F w.r.t g_{x_t} and solving for g_{x_t} yields

$$g_{x_t} = \frac{\sum_{all\,f}\left[(Y_t^*\tilde{X}_t + Y_t\tilde{X}_t^*)\sigma_w^2 + C\sigma_v^2\right]}{\sum_{all\,f}\left[\left|\tilde{X}_t\right|^2\sigma_w^2 + |H - G|^2\left|\tilde{X}_t\right|^2\sigma_v^2\right]} \tag{3.34}$$

where

$$C = (B_t - GY_t)^* (H - G) \tilde{X}_t + (B_t - GY_t) (H - G)^* \tilde{X}_t^* \tag{3.35}$$

It should be noted here that we do not estimate g_{x_t} for the Gaussian that models silence, and in that case, it is set to 1.

3.7 Dynamics of S_t

The enhancement process starts off with both $S_t = 0$ and $S_t = 1$ being equally likely, i.e., $p(S_t = 0) = p(S_t = 1) = 0.5$. In order to enforce smoothness of the state we use the following state dynamics:

$$p(S_t = s \,|S_{t-1} = s) = \frac{0.5 + p(S_{t-1} = s \,|Y_{t-1}, B_{t-1})}{2} \tag{3.36}$$

and $p(S_t = s \,|S_{t-1} = 1 - s) = 1 - p(S_t = s \,|S_{t-1} = s)$. This way, if the previous frame is a speech, i.e., $p(S_{t-1} = s \,|Y_{t-1}, B_{t-1}) > 0.5$, the prior for the current frame being speech, i.e., $p(S_t = 1 \,|S_{t-1} = 1)$ is larger than 0.5. The same is also true for: $p(S_t = 0 \,|S_{t-1} = 0) > 0.5$. This introduces some bias towards the state of the previous frame, making frame to frame transition smoother.

3.8 Speech Detection

Here we use a more robust statistical model-based voice activity detector (VAD) method [46] instead of the speech detection method used therein. Under Gaussian random variables assumptions, we have the probability density functions conditioned on speech absent H_0 and speech present H_1 given by

$$p(x \,|H_0) = \prod_{k=0}^{L-1} \frac{1}{\pi \lambda_N(k)} \exp\left\{ -\frac{|X_k|^2}{\lambda_N(k)} \right\} \tag{3.37}$$

$$p(x \,|H_1) = \prod_{k=0}^{L-1} \frac{1}{\pi [\lambda_N(k) + \lambda_S(k)]} \exp\left\{ -\frac{|X_k|^2}{\lambda_N(k) + \lambda_S(k)} \right\} \tag{3.38}$$

where $\lambda_N(k)$and $\lambda_S(k)$ denote the variances of N_k and S_k, respectively. The likelihood ratio for the kth frequency band is

$$\Lambda_k \triangleq \frac{p(X_k \,|H_1)}{p(X_k \,|H_0)} = \frac{1}{1 + \xi_k} \exp\left\{ \frac{\gamma_k \xi_k}{1 + \xi_k} \right\} \tag{3.39}$$

where $\xi_k \triangleq \lambda_S(k)/\lambda_N(k)$ and $\gamma_k \triangleq |X_k|^2/\lambda_N(k)$, representing the a priori and a posteriori signal-to-noise ratios (SNRs), respectively [14]. The decision rule is established from the geometric mean of the likelihood ratios for the individual frequency bands, which is given by

$$\log\Lambda = \frac{1}{L}\sum_{k=0}^{L-1}\log\Lambda_k \underset{H_0}{\overset{H_1}{\gtrless}} \eta \tag{3.40}$$

We assume that $\lambda_N(k)$ are already known through the noise statistic estimation procedure. However, the unknown parameters ξ_k needs to be estimated. The ML estimator for ξ_k can easily be derived as follows:

$$\xi_k^{ML} = \gamma_k - 1 \tag{3.41}$$

Applying it into the LRT yields the Itakura Saito distortion (ISD) based decision rule [47]:

$$\log\hat{\Lambda}^{ML} = \frac{1}{L}\sum_{k=0}^{L-1}\log\{\gamma_k - \log\gamma_k - 1\} \underset{H_0}{\overset{H_1}{\gtrless}} \eta \tag{3.42}$$

Note that the left-hand side of equation can not be smaller than zero, which is the well-known property of ISD and implies that the likelihood ratio is biased to H_1.

In order to reduce this bias, the decision-directed (DD) a priori SNR estimation method is applied:

$$\hat{\xi}_k(n) = \alpha\frac{\hat{A}_k^2(n-1)}{\lambda_N(k,n-1)} + (1-\alpha)P[\gamma_k(n)-1] \tag{3.43}$$

where n is the frame index, $P[x] = x$ if $x > 0$, and $P[x] = 0$ otherwise. $\hat{A}_k(n-1)$ are the estimated signal amplitude of the previous frame derived from a minimum mean square error (MMSE) estimator [14]. The DD method provides a smoother estimation of the a priori SNR than the ML method [6], and consequently reduces the fluctuation of the estimated likelihood ratios during noise-only periods.

In practice, the initial decision is often modified to prevent clipping of weak speech tails by considering the previous decision results. Hang-over algorithms are thus adopted to delay the transition from H_1 to H_0, based on the idea that there is a strong correlation in the consecutive occurrences of speech frames. To express this property explicitly, the sequence of states is modeled as a first-order Markov process. Since the Markov

process assumes that the current state only depends on the previous state, the correlative characteristic of speech occurrence can be represented by $p\left(q_n = H_1 \left| q_{n-1} = H_1\right.\right)$ with the following constraint:

$$p\left(q_n = H_1 \left| q_{n-1} = H_1\right.\right) > p\left(q_n = H_1\right) \tag{3.44}$$

where q_n denotes the state of the nth frame.

By assuming that the Markov process is time invariant, we can use the notation, $a_{ij} \triangleq p\left(q_n = H_j \left| q_{n-1} = H_i\right.\right)$. We further assume the stationarity of the process, namely, $p\left(q_n = H_i\right) = p\left(H_i\right)$, where $p\left(H_0\right)$ and $p\left(H_1\right)$ are the steady-state probabilities obtained from $a_{01}p\left(H_0\right) = a_{10}p\left(H_1\right)$ and $p\left(H_0\right) + p\left(H_1\right) = 1$. Thus, the overall process can be characterized by only two parameters, a_{01} and a_{10}. In this Markovian frame state model, the current state depends on the previous observations as well as the current one, which is reflected on the decision rule in the following way:

$$\begin{aligned} \mathcal{L}\left(n\right) &\triangleq \tfrac{p(\mathcal{X}_n|q_n=H_1)}{p(\mathcal{X}_n|q_n=H_0)} \\ &= \tfrac{p(H_0)}{p(H_1)} \tfrac{p(q_n=H_1|\mathcal{X}_n)}{p(q_n=H_0|\mathcal{X}_n)} \underset{H_0}{\overset{H_1}{\gtrless}} \eta \end{aligned} \tag{3.45}$$

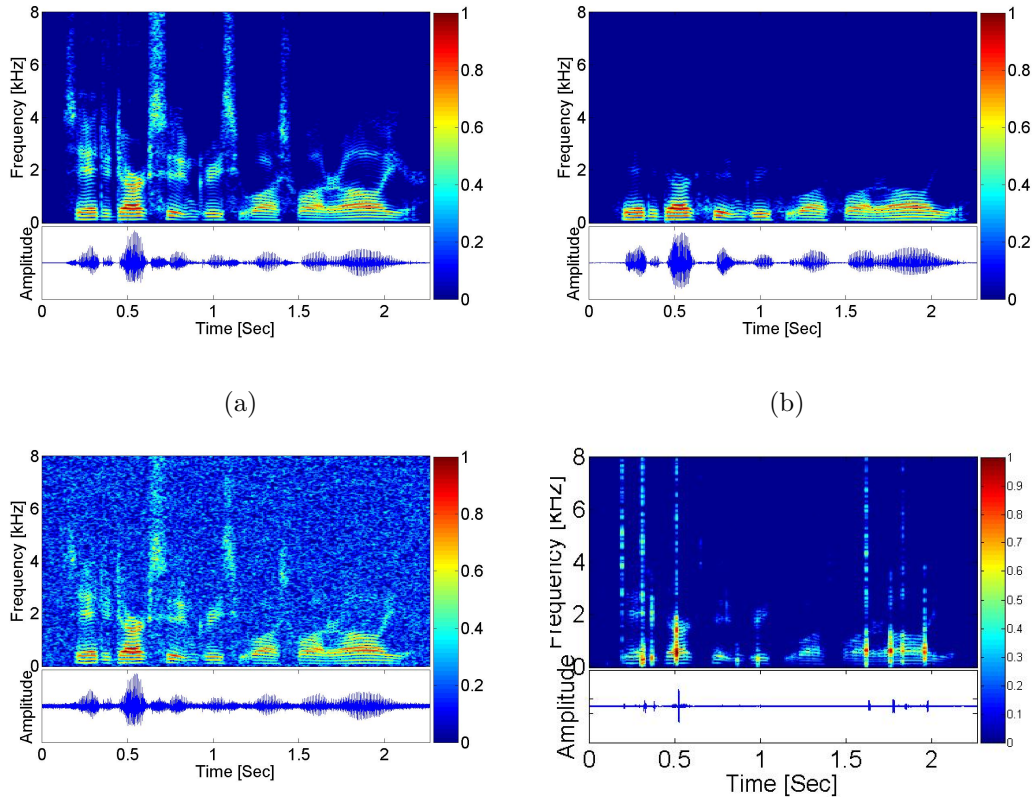
where $\mathcal{X}_n = \left\{X\left(n\right), ..., X\left(1\right)\right\}$ represents the set of observations up to the current frame n. For the efficient computation of the a posteriori probability ratio, $\Gamma\left(n\right) \triangleq p\left(q_n = H_1 \left|\mathcal{X}_n\right.\right) / p\left(q_n = H_0 \left|\mathcal{X}_n\right.\right)$,we define the forward variable as $\alpha_n\left(i\right) \triangleq p\left(q_n = H_i, \mathcal{X}_n\right)$. By using the forward procedure, we can solve for $\alpha_n\left(i\right)$ as follows:

$$\alpha_n\left(i\right) = \begin{cases} p\left(H_i\right) p\left(X\left(1\right) \left| q_1 = H_i\right.\right) & if\ n = 1 \\ \left(\alpha_{n-1}\left(0\right) a_{0j} + \alpha_{n-1}\left(1\right) a_{1j}\right) p\left(X\left(n\right) \left| q_1 = H_i\right.\right) & if\ n \geq 2 \end{cases} \tag{3.46}$$

Based on the above formulations, a recursive formula for $\Gamma\left(n\right)$ is obtained as

$$\Gamma\left(n\right) = \frac{\alpha_n(1)}{\alpha_n(0)} = \frac{a_{01} + a_{11}\Gamma\left(n-1\right)}{a_{00} + a_{10}\Gamma\left(n-1\right)}\Lambda\left(n\right) \tag{3.47}$$

where $\Lambda\left(n\right)$ denotes the likelihood ratio at nth frame. Consequently, the final decision statistic is obtained by $\mathcal{L}\left(n\right) = \left[p\left(H_0\right) / p\left(H_1\right)\right] \Gamma\left(n\right)$.



(a) (b) (c) (d)

Figure 3.3: Speech spectrograms and waveforms: (a) Clean signal; (b) bone-conducted signal (LSD=2.1317); (c) noisy air-conducted signal in Gaussian noise environment (SNR=10, LSD=2.4526); (d) Speech enhanced using the PA method (LSD=2.1592).

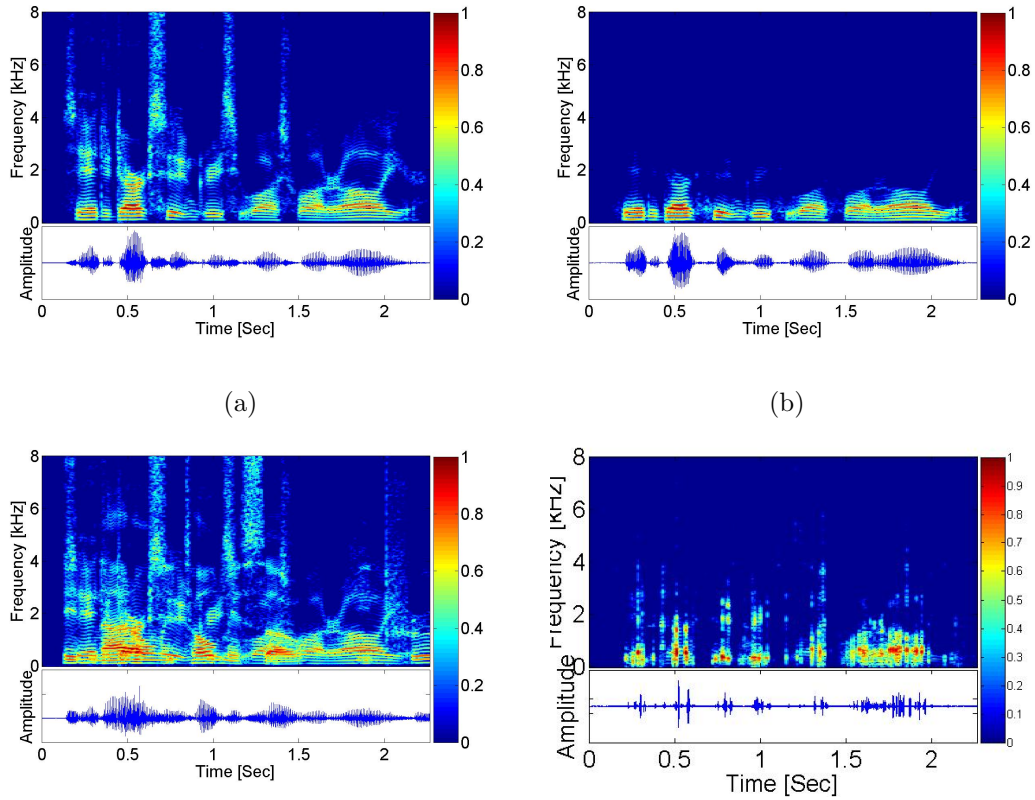
3.9 Experimental Results

For zero-mean Gaussian noise, an unbiased estimation of the variances is given by

$$\begin{aligned} \sigma_v^2 &= \tfrac{1}{|N_V|-1} \textstyle\sum_{t \in N_v} |Y_t|^2 \\ \sigma_w^2 &= \tfrac{1}{|N_V|-1} \textstyle\sum_{t \in N_v} |B_t - GY_t|^2 \end{aligned} \tag{3.48}$$

where N_V is the set of non-speech (or noise corrupted) frames. Recall that σ_v^2 is the variance of the corrupting noise, and σ_w^2 is the variance of the sensor noise in the bone microphone. Also we set $\sigma_u^2 = 10^{-4}\sigma_w^2$, as a consequence of less sensor noise of air-condcuted speech than the bone-conducted speech.

A collection of speech from the TIMIT database is used as the training data. In the STFT transform, the window is a Hanning window with length N = 256 and 50 percent overlap. After building the mixture of Gaussians of training data, we simulate the bone-



(a) (b) (c) (d)

Figure 3.4: Speech spectrograms and waveforms: (a) Clean signal; (b) bone-conducted signal (LSD=2.1317); (c) noisy air-conducted signal in an interfering speaker environment (SNR=-1.7821, LSD=1.2804); (d) Speech enhanced using the PA method (LSD=2.2672).

conducted speech by low pass filtering the clean speech. Noisy speech is created by contaminating Gaussian noise and interfering speaker. Figure 3.3 shows the spectrogram of clean, BC, noisy and reconstructed speech when SNR = 10. Figure 3.4 shows the spectrogram of clean, BC, noisy and reconstructed speech with interfering speaker. We can see from the spectrogram of the reconstructed speech that although it contains a relatively full frequecy band, it is still distorted. The result may be induced by the small number of mixure of Gaussians for representing the speech model. We will compare the result with our methods in the next chapter.

3.10 Summary

We have described the probabilistic approach for multi-sensory speech enhancement. It utilizes a mixture-of-Gaussians speech model built from magnitude-normalized complex spectra for speech enhancement. A dynamic Bayesian network representing a family of probability distributions defined in terms of a directed graph is used to estimate the posterior probability of the clean speech given both air and bone conducted measurements. The joint probability distribution over the variables, represented by the nodes in the graph, is obtained by taking products over functions on connected subsets of nodes. The final minimum mean square error estimator is computed by marginalizing the probabilities. We utilized a more robust speech detection method instead of the original method proposed in [49]. We showed the spectrogram of the reconstructed speech with a relatively full frequency band for the reconstructed speech with low SNR Gaussian noise and interfering speaker. Furthermore, it is computationally inexpensive and can be run in real time. We also compared this method to our method proposed in chapter 4.

Chapter 4

Geometric Extension Methods

4.1 Introduction

In this section, we model the problem as learning the nonlinear mapping from the concatenation of AC and BC speech to clean speech in the short time Fourier transform (STFT) domain. Here two function extension methods are used, i.e., Geometric harmonics [14] and Laplacian pyramids [36]. Geometric harmonics method is based on the Nystrom extension using Gaussian kernel for extending empirical functions f defined on a set X to a larger set $\bar{X}$. Laplacian pyramid extension is based on a kernel regression scheme where a series of Gaussian kernel with a decreasing scale are used to approximate the residual of previous representation. Both methods consist of two phases: an off-line phase (training phase) and an online phase (testing phase).

4.2 Problem Formulation

The training set is defined as $S = \{Y(k,l)\}_{k=1\ldots n, l=1\ldots m}$, where n is the total number of frequency bins, m is the total number of frames. We regard them as m data points in R^n. A function $f : S \rightarrow R^n$ is known on the points of the training set. Let $\bar{S}$ be a set in R^n such that $\bar{S} \subseteq R^n \backslash S$. The set $\bar{S}$ is denoted as the test set. The task is to extend the function f to $\bar{S}$ point of the test set . We concatenate noisy BC and AC speech together $Y_{AB}(k,l) = [Y_{AC}(k,l) \;\; Y_{BC}(k,l)]$ as the training set, with a nonlinear function f mapping it to the clean speech $X(k,l)$. After we learn the mapping given training noisy AB speech

and clean speech, we extend the function to new noisy AB speech and reconstruct the clean speech in the test data.

4.3 Prediction

Prediction is a part of statistical inference. Indeed, one description of statistics is that it provides a means of transfering knowledge about a sample of a population to the whole population, and to other related populations. Here we regard the denoising as a prediction process, where the reconstructed speech is predicted by the knowlege of the function mapping noisy concatenation speech to the clean speech.

In many applications it is possible to estimate the models that generate the observations. By expressing models as transfer functions or in terms of state-space parameters, the smoothed, filtered and predicted data can be estimated [10]. For linear underlying generating models, a minimum-variance Kalman filter and a minimum-variance smoother may be used to recover data of interest from noisy measurements. The afore-mentioned techniques rely on one-step-ahead predictors which minimize the variance of the prediction error. When the generating models are nonlinear then step-wise linearization may be applied to extended Kalman Filter and smoother recursions. However, in nonlinear cases, optimum minimum-variance performance guarantees no longer apply.

Statistical techniques used for prediction include regression analysis and time series analysis. Ordinary least squares, logistic regression, autoregressive moving average models, and vector autoregression models are their various sub-categories. In the process of regression analysis for prediction, data are collected on the variable to be predicted, called the dependent variable, and on one or more variables whose values are hypothesized to influence it, called independent variables. A functional form, often linear, is hypothesized for the postulated causal relationship, and the parameters of the function are estimated from the data chosen so as to optimize to fit the function. Then the independent variable values deemed relevant to unobserved values of the dependent variable are used in the parametrized function to generate predictions for the dependent variable.

4.4 Kernel methods

In computer science, kernel methods (KMs) are a class of algorithms for pattern analysis, whose best known element is the support vector machine (SVM). The general task of pattern analysis is to find and study general types of relations, for example clusters, rankings, principal components, correlations, classifications in general types of data, such as sequences, text documents, sets of points, vectors, images, etc. In the research, the function that maps the concatenation of AC and BC speech to clean speech is found by the kernel methods.

KMs approach the problem by mapping the data into a high dimensional feature space, where each coordinate corresponds to one feature of the data items, transforming the data into a set of points in a Euclidean space. In that space, a variety of methods can be used to find relations in the data. KMs owe their name to the use of kernel functions, which enable them to operate in the feature space without ever computing the coordinates of the data in that space, but rather by simply computing the inner products between the images of all pairs of data in the feature space. This operation is often computationally cheaper than the explicit computation of the coordinates. This approach is called the kernel trick. Kernel functions have been introduced for sequence data, graphs, text, images, as well as vectors.

Algorithms capable of operating with kernels include Support vector machine (SVM), Gaussian processes, Fisher's linear discriminant analysis (LDA), principal components analysis (PCA), canonical correlation analysis, ridge regression, spectral clustering, linear adaptive filters and many others. Because of the particular culture of the research community that has been developing this approach since the mid-1990s, most kernel algorithms are based on convex optimization or eigenproblems, which are computationally efficient and statistically well-founded.

4.5 Geometric Harmonics

Geometric Harmonics is derived from the Nyström extension, which has been widely used in partial differential solvers, and recently employed in machine learning and in spectral

graph theory as a way to subsample large data sets [2, 15, 54]. It is also used for the extension of functions from a training set to accommodate the arrival of new samples. Their work is also highly related to the Kriging technique widely employed in geostatistics [53]. The extension process involves the construction of a specific family of functions which are termed as geometric harmonics. These functions constitute a generalization of the prolate spheroidal wave functions of Slepian in the sense that they are optimally concentrated on X.

First assume the kernel $k : \bar{X} \times \bar{X} \to R$ satisfying

• k is symmetric $k(\bar{x}, \bar{y}) = k(\bar{y}, \bar{x})$ for all $\bar{x}$ and $\bar{y}$ in $\bar{X}$.

• k is positive semi-definite, i.e., for any $m \geq 1$ and any choice of real numbers $\alpha_1, ..., \alpha_m$ and of points $\bar{x}_1, ..., \bar{x}_m$ in $\bar{X}$, we have

$$\sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j k(\bar{x}_i, \bar{x}_j) \geq 0 \tag{4.1}$$

This property is not necessary to define geometric harmonics and extensions. However, it allows us to interpret the geometric harmonics as maximizing some concentration measure over X.

• k is bounded on $\bar{X} \times \bar{X}$ by a number $M > 0$. Although this assumption can be weakened, it is very convenient for the simplicity of the exposition.

When $\lambda_j \neq 0$, the eigenfunction ψ_j can be extended to $\bar{x} \in \bar{X}$ by

$$\psi_j(\bar{x}) = \frac{1}{\lambda_j} \int_X k(\bar{x}, y) \psi_j(y) \, d\mu(y) \tag{4.2}$$

This extension is called geometric harmonic.

The eigenvectors and eigenvalues of a Gaussian kernel on the training set S with scale ε are computed by

$$\lambda_l \varphi_l(x_i) = \sum_{x_j \in \Gamma} e^{-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2\varepsilon}} \varphi_l(x_j), \; x_i \in S \tag{4.3}$$

If $\lambda_l \neq 0$, the eigenvectors can be extended to any $y \in R^n$ by

$$\varphi_l(y) = \frac{1}{\lambda_l} \sum_{x_j \in \Gamma} e^{-\frac{\|y - \mathbf{x}_j\|^2}{2\varepsilon}} \varphi_l(x_j), \; y \in R^n \tag{4.4}$$

The extended distance of φ_l from the training set is proportional to ε. Let f be a function on the training set S . The eigenfunctions φ_l are the outcome of the spectral decomposition of a symmetric positive matrix, thus, they form a basis in R^n. Consequently,

Table 4.1: GH algorithm

Algorithm
Input: $X \in R^{n\times m}, f(x) \in R^{n/2}, \varepsilon, l, y$
Output: Extended function $f(y)$
1. Build a Gaussian kernel $k = e^{-\frac{\left\|x_i - x_j\right\|^2}{2\varepsilon}}$.
2. Compute the set of eigenvalues $\varphi_l(x)$ for this kernel
$\lambda_l \varphi_l(x_i) = \sum_{x_j \in S} e^{-\frac{\left\|\mathbf{x}_i - \mathbf{x}_j\right\|^2}{2\varepsilon}} \varphi_l(x_j),\ x_i \in S.$
3. Extend the eigenvalues $\varphi_l(y)$ via geometric harmonics
$\varphi_l(y) = \frac{1}{\lambda_l} \sum_{x_j \in S} e^{-\frac{\left\|y - \mathbf{x}_j\right\|^2}{2\varepsilon}} \varphi_l(x_j),\ y \in \bar{S}.$
4. Extend function via linear combination of the basis
$f(y) = \sum_{\lambda_l \geq \delta\lambda_0} \langle \varphi_l, f \rangle \tilde{\varphi}_l(y).$

any function f can be written as a linear combination of this basis:

$$f(x_i) = \sum_l \langle \varphi_l, f \rangle \varphi_l(x_i),\ x_i \in \Gamma \tag{4.5}$$

Using the Nystrom extension, f can be defined for any point in R^n by

$$f(y) = \sum_l \langle \varphi_l, f \rangle \varphi_l(y),\ y \in R^n \tag{4.6}$$

To avoid ill-conditioned eigenvalues, we cut off the sum keeping the eigenvalues (and the corresponding eigenfunctions) satisfying $\lambda_l \geq \delta\lambda_0$

$$\tilde{f}(y) = \sum_{\lambda_l \geq \delta\lambda_0} \langle \varphi_l, f \rangle \tilde{\varphi}_l(y),\ y \in R^n \tag{4.7}$$

The GH algorithm is described in table 4.1. Geometric hamornics method needs careful setting of the kernel scale ε and condition number l. Besides, the extended function is the projection of the function rather than the original function, thus may not represent the data ponits well. Last but not the least, the extension range has relation to the complexity of the function, namely, a more complex function will be more difficult to extend in a larger range. As a result, Laplacian pyramid method is utilized to circumvent these problems.

4.6 Laplacian Pyramid

The Laplacian pyramid is a method for multi-scale representation of data, with Gaussian kernels of decreasing widths. At each scale, the residual of the previous result is convolved with a Gaussian kernel.

A Gaussian kernel is defined on S as

$$W_0 = w_0\left(x_i, x_j\right) = e^{\frac{-\left\|x_i - x_j\right\|^2}{\sigma_0}} \tag{4.8}$$

Normalizing W_0 by its row sum, we have

$$K_0 = k_0\left(x_i, x_j\right) = q_0^{-1}\left(x_i\right) w_0\left(x_i, x_j\right) \tag{4.9}$$

$$\text{where } \; q_0\left(x_i\right) = \sum_j w_0\left(x_i, x_j\right) \tag{4.10}$$

yields a smoothing operator K_0.

Kernel regression is a popular non-parametric method rely on the data itself to dictate the structure of the model where the regression can be done by interpolation of surrounding data. The zero-th order weighted least square estimator, i.e., the Nadaraya-Watson estimator is

$$\hat{f}(x_k) = \sum_{i=1}^{n} k_0(x_i, x_k) f(x_i) \tag{4.11}$$

According to the Perron Frobenius thereom, the matrix K_0 is positive definite and row stochastic. The spectral radius of K_0 is one, with the highest eigenvalue 1 and its corresponding eigenvector $1/\sqrt{n}$. In addition, if the matrix K_0 is iterated infinity many times, it converges to the product of left and right eigenvectors, with eigenvalues smaller than 1 converges to 0, i.e., $\lim_{k\to\infty} K_0^k = v_1 u_1^T = 1_n u_1^T$. The matrix can also be viewed as probability transition matrix for markov chain or graphic models of data in machine learning senario.

Kernel regression can be applied for diverse problems ranging from denoising to upscaling and interpolation. Here two parameters can be varied to imrpove the regression result, including the variance of the kernel and the number of iteration. Iterating the kernel operator on the previous output will lead to a more improved result. By applying the kernel multiple times, the Duffision kernel is obtained:

$$\hat{f}_{l+1}(x_k) = \hat{f}_l(x_k) + \left(K_0 - I\right) \hat{f}_l(x_k) \tag{4.12}$$

The term diffusion comes from moving $\hat{f}_l(x_k)$ to the left hand side of the equation where the left side is a differential operator and right side is a Laplacian operator, which essentially is a discrete heat equation.

$$\hat{f}_{l+1}(x_k) - \hat{f}_l(x_k) = (K_0 - I)\,\hat{f}_l(x_k) \tag{4.13}$$

This time we multiply the kernel with the residual instead of the original function, as the residual involves more details,

$$\hat{f}_{l+1}(x_k) = \hat{f}_l(x_k) + K_0\left(f(x_k) - \hat{f}_l(x_k)\right) \tag{4.14}$$

Laplacian pyramid algorithm is used for constructing a coarse representation of a function f at a given scale l. The difference between f and the coarse approximation is the new input for the next iteration. The algorithm is iterated by approximating the difference using a Gaussian kernel of a finer scale. The approximated function at a given scale can be extended to new data points. The LP algorithm is described in table 4.2.

From the above description, we can see that the LP estimation is similar to the residual method but with a changing kernel function in each iteration.

$$\hat{f}_{l+1}(x_k) = \hat{f}_l(x_k) + K_l\left(f(x_k) - \hat{f}_l(x_k)\right) \tag{4.15}$$

Now we conduct the statistical analysis of Laplacian estimation, for additive noise, the mesurement is

$$y(x_k) = f(x_k) + n_k \tag{4.16}$$

The bias is bounded by the original bias of the algorithm,

$$\begin{aligned} &E\left[s_l(x_k)\right] = E\left[\sum_{i=1}^{n} k_l(x_i, x_k)\left(f(x_i) + n_i - \sum_{i=1}^{l-1} s_i(x_i)\right)\right] = s_l(x_k) \\ &Bias = E\left[\hat{f}(x_k)\right] - f(x_k) = E\left[\sum s_l(x_k)\right] - f(x_k) = \sum s_l(x_k) - f(x_k) = e_s \end{aligned} \tag{4.17}$$

The mean square error is bounded by the weighted sum of noise variance and square

Table 4.2: LP algorithm

Algorithm
Input: $X \in R^{n\times m}, f(x) \in R^{n/2}, \sigma_0, y$
Output: Extended function $f(y)$
Initialize: Build a normalized Gaussian kernel K_0.
Approximate the function $s_0(y) = \sum_{i=1}^{n} k_0(x_i, y) f(x_i)$
Repeat
1. Build the Gaussian kernel W_l with decreasing scale
$W_l = w_l(x_i, x_j) = e^{-\|x_i - x_j\|^2/\left(\frac{\sigma_0}{2^l}\right)}$.
2. Normalize the kernel by the row sum
$K_l = k_l(x_i, x_j) = q_l^{-1}(x_i)\, w_l(x_i, x_j)$
where $q_l(x_i) = \sum_j w_l(x_i, x_j)$.
2. Compute the residual $d_l = f - \sum_{i=0}^{l-1} s_i$.
3. Approximate the function $s_l(y) = \sum_{i=1}^{n} k_l(x_i, y) d_l(x_i)$.
End: Extend function via summation $f(y) = \sum_{k\le l} s_k(y)$.

of the system bias,

$$
\begin{aligned}
&E\left[\hat{f}(x_k) - f(x_k)\right]^2 \\
&= E\left[\sum_{i=1}^{n} k_0(x_i, x_k)\left(f(x_i) + n_i\right) + \ldots + \sum_{i=1}^{n} k_l(x_i, x_k)\left(f(x_i) + n_i - \sum_{i=1}^{l-1} s_i(x_i)\right) - \sum_{i=1}^{l-1} s_i(x_k) - e_s\right]^2 \\
&= E\left[\sum_{i=1}^{n} k_0(x_i, x_k) n_i + \ldots + \sum_{i=1}^{n} k_l(x_i, x_k) n_i - e_s\right]^2 \\
&= E\left[\sum_{i=1}^{n} k_0^2(x_i, x_k) n_i^2 + \ldots + \sum_{i=1}^{n} k_l^2(x_i, x_k) n_i^2 + e_s^2\right] \\
&= E\left[\sum_{i=1}^{n}\sum_{l=1}^{L} k_l^2(x_i, x_k) n_i^2 + e_s^2\right] = -\sum_{i=1}^{n} \frac{Ei(-2^l \ln(k_l^2(x_i, x_k)))}{\ln 2}\sigma_i^2 + e_s^2
\end{aligned}
\tag{4.18}
$$

Until now, the kernel we have utilized gives the nearby samples higher weight than samples far away. However, it fails to take into account distance between values of function, on the other hand, it should emphasize the contribution of values close to it. As a result, a data adaptive kernel, known as bilateral filter, combines the spatial distance

Table 4.3: LSD results for Gaussian noise with different SNR levels and interfering speech, obtained by using four different speech enhancement methods: GH, LP, OM-LSA and PA.

LSD	GH	LP	OM-LSA	PA
SNR =0	1.5726	1.1003	2.0901	1.9613
SNR = 10	1.5564	1.1028	1.4979	2.1592
SNR = 20	1.5755	1.1021	1.1085	2.2531
Interfering speech	1.5660	1.1768	1.3604	2.2672

together with the value distance can be utilized:

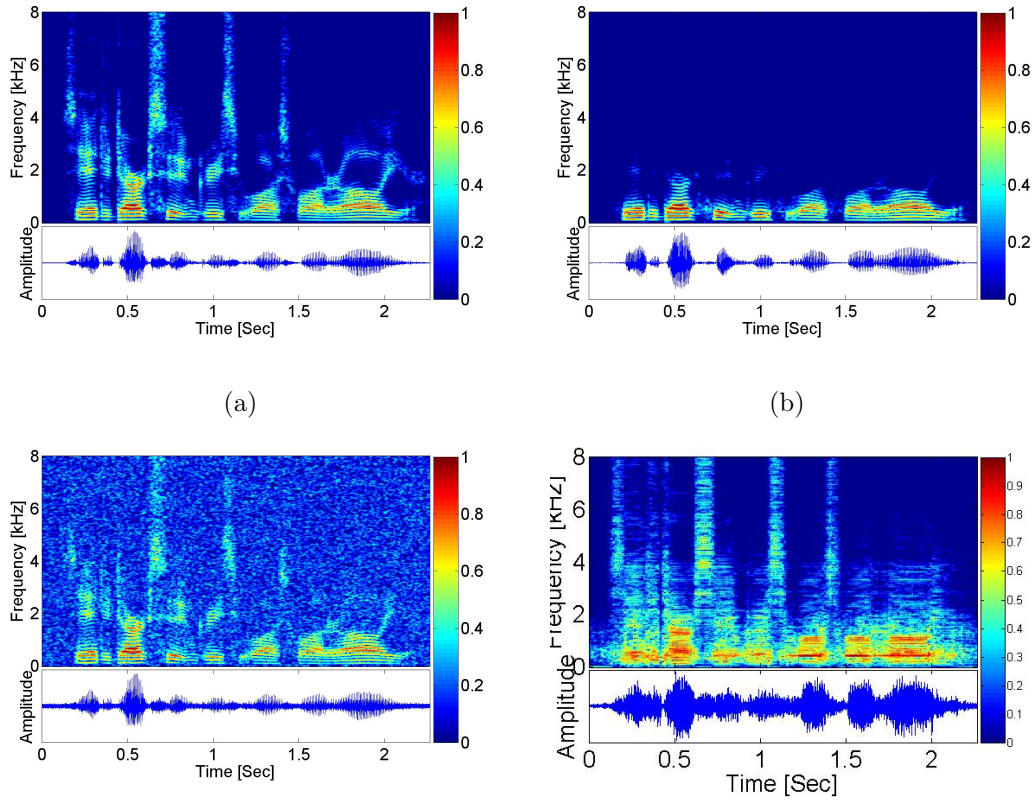
$$K_b = k\left(x_i, x_j\right) k\left(f\left(x_i\right), f\left(x_j\right)\right) \tag{4.19}$$

Here, the reconstructed result with classical kernel can be used as the function value.

4.7 Experimental Results

In this section, we present simulation results which demonstrate the performances of geometric harmonics and Laplacian pyramid compared to an existing probabilistic approach. Ten utterances of speech signals are taken from the TIMIT database. The sampling frequency is $f_s = 16$ kHz. The STFT window is a Hanning window of length N = 256 and the overlap between two successive STFT frames is 50 percent. The BC speech signals are obtained by low pass filtering the AC speech signals, where the pass–band cutoff frequency is 300 Hz and stop–band cutoff frequency is 3 kHz. Noisy AC speech signals are generated by adding Gaussian noise and interfering speaker.

The spectrograms of AC speech, BC speech, noisy clean speech and reconstructed speech via the geometric harmonics are demonstrated in Figures 4.1 and 4.2. Figure 4.1 illustrates the result for Gaussian noise and Figure 4.2 for an interfering speaker. The spectrograms of AC speech, BC speech, noisy clean speech and reconstructed speech via the Laplacian pyramid are demonstrated in Figures 4.3 and 4.4. Figure 4.3 illustrates the result for Gaussian noise and Figure 4.4 for an interfering speaker. The figures demonstrate that LP method facilitates enhancement of speech signals not only in stationary noise environments such as white Gaussian noise, but also in nonstationary noise envi-



(a) (b) (c) (d)

Figure 4.1: Speech spectrograms and waveforms: (a) Clean signal; (b) bone-conducted signal (LSD=2.1317); (c) noisy air-conducted signal in Gaussian noise environment (SNR=10, LSD=2.4526); (d) Speech enhanced using the GH method (LSD=1.5564).

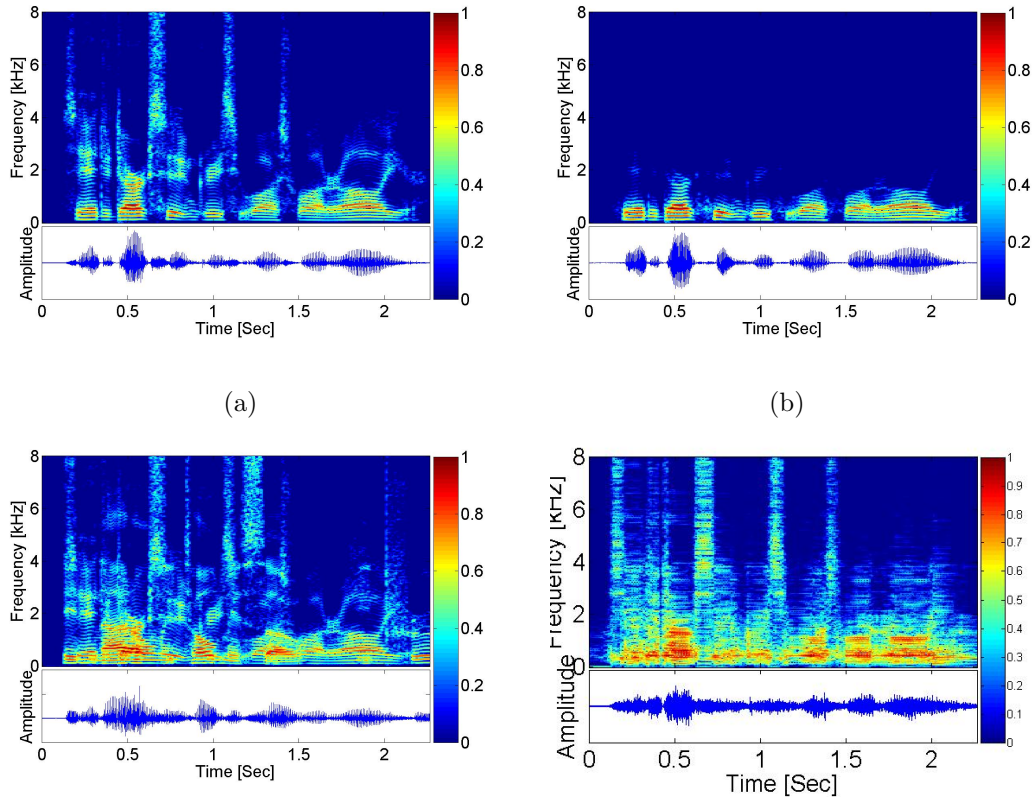
ronments such as an interfering speaker.

We then use the log spectral distortion (LSD) to evaluate the quality of the reconstructed speech. LSD is defined as the L_2 norm of the difference between the STFT transforms of $x(n)$ and $\hat{x}(n)$, namely $X(l,k)$ and $\hat{X}(l,k)$, in the lth frame:

$$LSD(l) = \left(\frac{1}{K} \sum_{k=0}^{K-1} \left| L\left\{ \hat{X}(l,k) \right\} - L\left\{ X(l,k) \right\} \right|^2 \right)^{\frac{1}{2}} dB \tag{4.20}$$

where $L\{X(l,k)\} = \max\{20\log_{10}(|X(l,k)|), \delta\}$ is the log spectrum confined to about 50 dB dynamic range ($\delta = \max_{l,k}\{20\log_{10}(|X(l,k)|)\} - 50$). The mean LSD is obtained by averaging over all speech frames.

Table 4.3 presents the LSD results for Gaussian noise with different SNR levels and interfering speech, obtained by using four different speech enhancement methods: GH,



(a) (b) (c) (d)

Figure 4.2: Speech spectrograms and waveforms: (a) Clean signal; (b) bone-conducted signal (LSD=2.1317); (c) noisy air-conducted signal in an interfering speaker environment (SNR=-1.7821, LSD=1.2804); (d) Speech enhanced using the GH method (LSD=1.5660).

LP, OM-LSA and PA. The table demonstrates that the LP method consistently provides the lowest distortion for all tested SNR levels and noise types. The noise level has little influence on the geometric extension methods, which may result from the fact that the nonlinear mapping learned via geometric methods implicitly involves the noise model. However, the OM-LSA methods are strongly influenced by the noise level. The GH method provides a relatively less accurate result, which may suggest that the geometric harmonics based on Gaussian kernel can not represent the concatenation data well. The downside of the LP method is the computation time, as the iteration will increase the computation complexity.

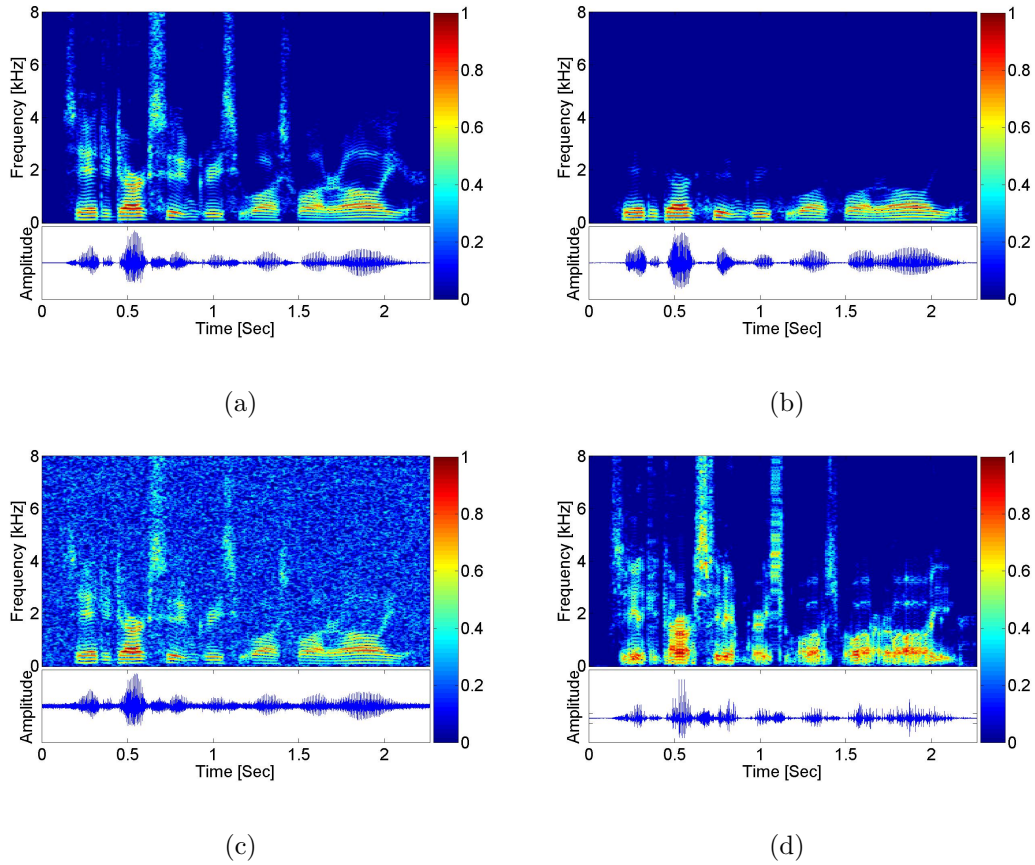


Figure 4.3: Speech spectrograms and waveforms: (a) Clean signal; (b) bone-conducted signal (LSD=2.1317); (c) noisy air-conducted signal in Gaussian noise environment (SNR=10, LSD=2.4526); (d) Speech enhanced using the LP method (LSD=1.1028).

4.8 Summary

We have presented a geometric extension method based on Nyström extension, i.e., the geometric harmonics method. However, application of the geometric harmonics method requires careful setting for the correct extension scale and condition number. A multi-scale Laplacian pyramid extension is thus utilized to circumvent the procedure for scale tuning and compared to the geometric harmonics method.

The spectrograms of the reconstructed speech have been illustrated and the LP method provides the most accurate spectrogram. The algorithms have been evaluated by computing the LSD between clean speech and reconstructed speech. The result of geometric extension methods is not influenced by the noise level and category, and they give relatively accurate reconstructed speech, especially the LP method, though at the expense of

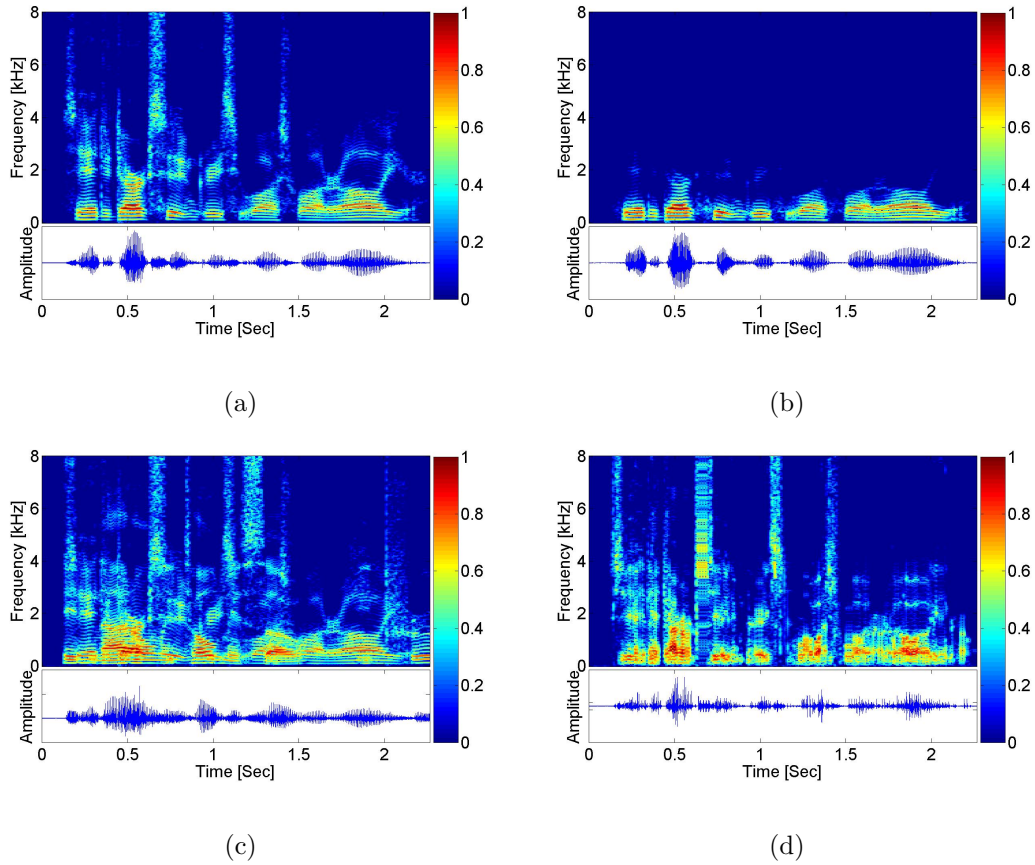


Figure 4.4: Speech spectrograms and waveforms: (a) Clean signal; (b) bone-conducted signal (LSD=2.1317); (c) noisy air-conducted signal in an interfering speaker environment (SNR=-1.7821, LSD=1.2804); (d) Speech enhanced using the LP method (LSD=1.1768).

computation time. It has also been shown that the LSD result of probabilistic approach (PA) is worse than the geometric extension method we proposed.

Chapter 5

Conclusion

5.1 Summary

In this thesis, we have presented two function extension schemes for multisensory speech enhancement, i.e., geometric harmonics and Laplacian pyramid. Clean speech is reconstructed using samples of air-conducted and bone-conducted speech signals. We introduce a model in a supervised learning framework by approximating a mapping from concatenation of noisy air-conducted and bone-conducted speech to clean speech in the short time Fourier transform domain.

The Nyström based geometric harmonics method is utilized for first describing the geometry of the concatenated features of both sensors and then extending it by superposition of the integral operator eigenvectors in discrete form. It can also be viewed as a dimensionality reduction scheme, by transforming the data of a single attribute, into a new set of coordinates that reflect its geometry and density. High dimensional data in this case is the concatenation of air-conducted and bone-conducted speech in the short time Fourier transform domain. Nonlinear mapping to the clean speech defined on this data set is learned and extended to new concatenation of speech signals by extending eigenfunctions of the affinity matrix. However, application of the geometric harmonics method requires a careful setting of the correct extension scale and condition number. A possible solution is to implement the geometric harmonics in a multi-scale manner, namely, the extension is calculated as a sum of representations in different scales.

In order to circumvent the tuning of the parameters, a multi-scale Laplacian pyramid

method is used to learn functions by pyramidal interpolation of surrounding data, where the residual of previous result enters the process as the new input with different kernel scales in each iteration. In essence, Laplacian pyramid extension is based on a kernel regression scheme where a series of Gaussian kernels with a decreasing scale are used to approximate the residual of the previous representation. Kernel regression is a non-parametric method in a least square formulation by interpolation of surrounding data according to their distances, giving the nearby samples higher weight than samples far away.

Experiments are conducted on simulated air-conducted and bone-conducted speech in Gaussian noise environments and interfering speakers, where the bone-conducted speech is a low pass filtered clean speech. Estimating the speech presence probability from the bone-conducted speech signal yields improved speech enhancement results compared to the existing spectral enhancement method. Geometric methods provide a steady reconstructed speech in spite of noise level and category, although involving distortion, enabling further noise reduction. Experimental results of the proposed methods are also compared to an existing probabilistic approach with respect to the log spectral distance, where the Laplacian pyramid method outperforms other methods.

5.2 Future Research

Geometric harmonics may also be conducted in a multi-scale manner in accordance with the Laplacian pyramid mechanism, where the extension can be viewed as approximations of residuals in a series of decreasing scales.

The relation between the iteration number in the Laplacian pyramid method and noise level in the observation can be derived to pre-determine the iteration number, rather than choosing it via trials of experiments.

Further processing of the estimated features in the STFT domain for geometric extension methods, including iterative tuning, may reduce artifacts in the reconstructed speech.

Bibliography

[1] M. J. Beal, N. Jojic, and H. Attias, "A graphical model for audiovisual object tracking," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 25, no. 7, pp. 828–836, 2003. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1206512

[2] Y. Bengio, J.-F. Paiement, P. Vincent, O. Delalleau, N. Le Roux, and M. Ouimet, "Out-of-sample extensions for lle, isomap, mds, eigenmaps, and spectral clustering," *Advances in neural information processing systems*, vol. 16, pp. 177–184, 2004. [Online]. Available: http://books.google.com/books?hl=en&lr=&id=0F-9C7K8fQ8C&oi=fnd&pg=PA177&dq=Out-of-sample+extensions+for+LLE,+isomap,+MDS,+eigenmaps,+and+spectral+clustering&ots=TGKylUPd80&sig=npGDavsVJSiHIPt6eFYGonOI-Dg

[3] J. A. Bilmes, "Graphical models and automatic speech recognition," *Journal of the Acoustical Society of America*, vol. 112, no. 5, pp. 2278–2278, 2002.

[4] C. M. Bishop and N. M. Nasrabadi, *Pattern recognition and machine learning.* springer New York, 2006, vol. 1.

[5] G. C. Burnett, J. F. Holzrichter, L. C. Ng, and T. J. Gable, "The use of glottal electromagnetic micropower sensors (gems) in determining a voiced excitation function," *The Journal of the Acoustical Society of America*, vol. 106, p. 2183, 1999. [Online]. Available: http://link.aip.org/link/?JASMAN/106/2183/6

[6] O. Cappé, "Elimination of the musical noise phenomenon with the ephraim and malah noise suppressor," *Speech and Audio Processing, IEEE Transactions on*,

vol. 2, no. 2, pp. 345–349, 1994. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=279283

[7] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 39, pp. 1–38, 1977. [Online]. Available: http://www.jstor.org/stable/10.2307/2984875

[8] L. Deng, L. J. Lee, H. Attias, and A. Acero, "A structured speech model with continuous hidden dynamics and prediction-residual training for tracking vocal tract resonances," in *Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP'04). IEEE International Conference on*, vol. 1. IEEE, 2004, pp. 557–560. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1326046

[9] L. Deng, Z. Liu, Z. Zhang, and A. Acero, "Nonlinear information fusion in multi-sensor processing-extracting and exploiting hidden dynamics of speech captured by a bone-conductive microphone," in *Multimedia Signal Processing, 2004 IEEE 6th Workshop on*. IEEE, 2004, pp. 19–22. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1436405

[10] G. A. Einicke, *Smoothing, Filtering and Prediction: Estimating the Past, Present and Future*. InTech, 2012.

[11] Y. Ephraim, "A bayesian estimation approach for speech enhancement using hidden markov models," *Signal Processing, IEEE Transactions on*, vol. 40, no. 4, pp. 725–735, 1992. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=127947

[12] Y. Ephraim, H. Lev-Ari, and W. J. Roberts, "A brief survey of speech enhancement," *The Electronic Handbook*, 2005.

[13] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *Acoustics, Speech and Signal*

Processing, IEEE Transactions on, vol. 32, no. 6, pp. 1109–1121, 1984. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1164453

[14] ——, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 33, no. 2, pp. 443–445, 1985. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1164550

[15] C. Fowlkes, S. Belongie, and J. Malik, "Efficient spatiotemporal grouping using the nystrom method," in *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, vol. 1. IEEE, 2001, pp. 231–238. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=990481

[16] M. Graciarena, H. Franco, K. Sonmez, and H. Bratt, "Combining standard and throat microphones for robust speech recognition," *Signal Processing Letters, IEEE*, vol. 10, no. 3, pp. 72–74, 2003. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1182088

[17] 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.* IEEE, 2003, pp. 73–76. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1318406

[18] H. Hermansky and N. Morgan, "Rasta processing of speech," *Speech and Audio Processing, IEEE Transactions on*, vol. 2, no. 4, pp. 578–589, 1994. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=326616

[19] J. Hershey, T. Kristjansson, and Z. Zhang, "Model-based fusion of bone and air sensors for speech enhancement and robust speech recognition," in *ISCA Tutorial and Research Workshop (ITRW) on Statistical and Perceptual Audio Processing*, 2004. [Online]. Available: http://www.isca-speech.org/archive_open/sapa_04/sap4_139.html

[20] J. Holzrichter, G. Burnett, L. Ng, and W. Lea, "Speech articulator measurements using low power em-wave sensors," *The Journal of the Acoustical Society of America*, vol. 103, pp. 622–625, 1998. [Online]. Available: http://link.aip.org/link/?JASMAN/103/622/1

[21] J. F. Holzrichter, L. C. Ng, G. J. Burke, N. J. Champagne II, J. S. Kallman, R. M. Sharpe, J. B. Kobler, R. E. Hillman, and J. J. Rosowski, "Measurements of glottal structure dynamics," *The Journal of the Acoustical Society of America*, vol. 117, pp. 1373–1385, 2005. [Online]. Available: http://link.aip.org/link/?JASMAN/117/1373/1

[22] X. Huang, A. Acero, H.-W. Hon, and R. Reddy, *Spoken language processing.* Prentice Hall PTR New Jersey, 2001, vol. 15.

[23] M. I. Jordan and Y. Weiss, "Graphical models: Probabilistic inference," MIT Press, 2002.

[24] D. Klein and C. D. Manning, "Conditional structure versus conditional estimation in nlp models," in *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10.* Association for Computational Linguistics, 2002, pp. 9–16. [Online]. Available: http://dl.acm.org/citation.cfm?id=1118695

[25] K. Kondo, T. Fujita, and K. Nakagawa, "On equalization of bone conducted speech for improved speech quality," in *Signal Processing and Information Technology, 2006 IEEE International Symposium on.* IEEE, 2006, pp. 426–431. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=4042281

[26] T. Kristjansson and J. Hershey, "High resolution signal reconstruction," in *Automatic Speech Recognition and Understanding, 2003. ASRU'03. 2003 IEEE Workshop on.* IEEE, 2003, pp. 291–296. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1318456

[27] T. T. Kristjansson, "Speech recognition in adverse environments: a probabilistic approach," Ph.D. dissertation, University of Waterloo, Waterloo, Canada, 2002.

[28] J. S. Lim and A. V. Oppenheim, "Enhancement and bandwidth compression of noisy speech," *Proceedings of the IEEE*, vol. 67, no. 12, pp. 1586–1604, 1979. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1455809

[29] Z. Liu, A. Subramanya, Z. Zhang, J. Droppo, and A. Acero, "Leakage model and teeth clack removal for air-and bone-conductive integrated microphones," in *Proc. ICASSP*, vol. 1, 2005, pp. 1093–1096.

[30] Z. Liu, Z. Zhang, A. Acero, J. Droppo, and X. Huang, "Direct filtering for air-and bone-conductive microphones," in *Multimedia Signal Processing, 2004 IEEE 6th Workshop on*. IEEE, 2004, pp. 363–366. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1436568

[31] A. McCree, K. Brady, and T. F. Quatieri, "Multisensor dynamic waveform fusion," in *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, vol. 4. IEEE, 2007, pp. IV 577–580. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=4218166

[32] P. J. Moreno, B. Raj, and R. M. Stern, "A vector taylor series approach for environment-independent speech recognition," in *Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings., 1996 IEEE International Conference on*, vol. 2. IEEE, 1996, pp. 733–736. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=543225

[33] G. Potamianos, C. Neti, J. Luettin, and I. Matthews, *Issues in Visual and Audio-Visual Speech Processing*. MIT Press Cambridge, MA, 2004, ch. Audio-visual automatic speech recognition: An overview.

[34] T. F. Quatieri, *Discrete-time speech signal processing*. Pearson Education India, 2002.

[35] T. F. Quatieri, K. Brady, D. Messing, J. P. Campbell, W. M. Campbell, M. S. Brandstein, C. J. Weinstein, J. D. Tardelli, and P. D. Gatewood, "Exploiting nonacoustic sensors for speech encoding," *Audio, Speech, and Language Processing,*

IEEE Transactions on, vol. 14, no. 2, pp. 533–544, 2006. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1597258

[36] N. Rabin and R. R. Coifman, "Heterogeneous datasets representation and learning using diffusion maps and laplacian pyramids." in *SDM*, 2012, pp. 189–199.

[37] M. S. Rahman, A. Saha, and T. Shimamura, "Low-frequency band noise suppression using bone conducted speech," in *Communications, Computers and Signal Processing (PacRim), 2011 IEEE Pacific Rim Conference on.* IEEE, 2011, pp. 520–525. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6032948

[38] M. S. Rahman and T. Shimamura, "Pitch characteristics of bone conducted speech," *Proc. of European Signal Processing Conference (EUSIPCO)*, pp. 795–798, 2010.

[39] ——, "Intelligibility enhancement of bone conducted speech by an analysis-synthesis method," in *Circuits and Systems (MWSCAS), 2011 IEEE 54th International Midwest Symposium on.* IEEE, 2011, pp. 1–4. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6026374

[40] M. Rothenberg, "A multichannel electroglottograph," *Journal of Voice*, vol. 6, no. 1, pp. 36–43, 1992. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0892199705800074

[41] M. V. Scanlon, "Acoustic sensor for health status monitoring," DTIC Document, Tech. Rep., 1998. [Online]. Available: http://oai.dtic.mil/oai/oai?verb=getRecord&metadataPrefix=html&identifier=ADA398960

[42] M. L. Seltzer, A. Acero, and J. Droppo, "Robust bandwidth extension of noise-corrupted narrowband speech." in *INTERSPEECH*, 2005, pp. 1509–1512.

[43] T. Shimamura, J. Mamiya, and T. Tamiya, "Improving bone-conducted speech quality via neural network," in *Signal Processing and Information Technology, 2006 IEEE International Symposium on.* IEEE, 2006, pp. 628–632. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=4042318

[44] T. Shimamura and T. Tamiya, "A reconstruction filter for bone-conducted speech," in *Circuits and Systems, 2005. 48th Midwest Symposium on.* IEEE, 2005, pp. 1847–1850. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1594483

[45] H. S. Shin, H.-G. Kang, and T. Fingscheidt, "Survey of speech enhancement supported by a bone conduction microphone," in *Speech Communication; 10. ITG Symposium; Proceedings of.* VDE, 2012, pp. 1–4. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6309576

[46] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection," *Signal Processing Letters, IEEE*, vol. 6, no. 1, pp. 1–3, 1999. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=736233

[47] J. Sohn and W. Sung, "A voice activity detector employing soft decision based noise spectrum adaptation," in *Acoustics, Speech and Signal Processing, 1998. Proceedings of the 1998 IEEE International Conference on*, vol. 1. IEEE, 1998, pp. 365–368. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=674443

[48] O. M. Strand, T. Holter, A. Egeberg, and S. Stensby, "On the feasibility of asr in extreme noise using the parat earplug communication terminal," in *Automatic Speech Recognition and Understanding, 2003. ASRU'03.* IEEE, 2003, pp. 315–320. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1318460

[49] A. Subramanya, Z. Zhang, Z. Liu, and A. Acero, "Multisensory processing for speech enhancement and magnitude-normalized spectra for speech modeling," *Speech Communication*, vol. 50, no. 3, pp. 228–243, 2008. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167639307001501

[50] A. Subramanya, Z. Zhang, Z. Liu, J. Droppo, and A. Acero, "A graphical model for multi-sensory speech processing in air-and-bone conductive microphones." in *INTERSPEECH*, 2005, pp. 2361–2364.

[51] T. tat Vu, M. Unoki, and M. Akagi, "An lp-based blind model for restoring bone-conducted speech," in *Communications and Electronics, 2008. ICCE 2008.*

Second International Conference on. IEEE, 2008, pp. 212–217. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=4578960

[52] T. T. Vu, K. Kimura, and M. Unoki, "A study on restoration of bone-conducted speech with mtf-based and lp-based models," vol. 10, no. 6, pp. 407–417, 2006. [Online]. Available: http://ci.nii.ac.jp/naid/80018751527/

[53] H. Wackernagel, *Multivariate geostatistics.* Springer, 2003.

[54] C. Williams and M. Seeger, "Using the nyström method to speed up kernel machines," in *Advances in Neural Information Processing Systems 13.* MIT Press, 2001, pp. 682–688. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.18.7519

[55] T. Yanagisawa and K. Furihata, "Pickup of speech signal utilization of vibration transducer under high ambient noise," *J. Acoust. Soc. Jpn*, vol. 31, no. 3, pp. 213–220, 1975.

[56] Z. Zhang, Z. Liu, M. Sinclair, A. Acero, L. Deng, J. Droppo, X. Huang, and Y. Zheng, "Multi-sensory microphones for robust speech detection, enhancement and recognition," in *Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP'04). IEEE International Conference on*, vol. 3. IEEE, 2004, pp. 781–784. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1326661

[57] 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 *Automatic Speech Recognition and Understanding, 2003. ASRU'03. 2003 IEEE Workshop on.* IEEE, 2003, pp. 249–254. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1318449

[58] M. Zhu, H. Ji, F. Luo, and W. Chen, "A robust speech enhancement scheme on the basis of bone-conductive microphones," in *Signal Design and Its Applications in Communications, 2007. IWSDA 2007. 3rd International Workshop on.* IEEE,

2007, pp. 353–355. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=4408396

[59] G. Zweig, J. Bilmes, T. Richardson, K. Filali, K. Livescu, P. Xu, K. Jackson, Y. Brandman, E. Sandness, E. Holtz *et al.*, "Structurally discriminative graphical models for automatic speech recognition-results from the 2001 johns hopkins summer workshop," in *Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on*, vol. 1. IEEE, 2002, pp. 93–96. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5743662

הדגשת דיבור בסביבה רועשת באמצעות עיבוד משולב של אותות ממיקרופון אוויר ומיקרופון עצם מוליך

מינגזי לי

הדגשת דיבור בסביבה רועשת באמצעות עיבוד משולב של אותות ממיקרופון אוויר ומיקרופון עצם מוליך

עבודת גמר

לשם מילוי חלקי של הדרישות לקבלת התואר
מגיסטר למדעים בהנדסת חשמל

מינגזי לי

הוגש לסנט הטכניון – מכון טכנולוגי לישראל
חשון תשע״ד חיפה נובמבר 2013

המחקר נעשה בהנחיית פרופ׳ ישראל כהן בפקולטה להנדסת חשמל.

תקציר

בהתבססות על מספר ערוצי הקלט, ניתן לסווג טכניקות לשיפור דיבור כאלגוריתמים חד ערוציים, ואלגוריתמים רב ערוציים. באופן כללי, למרות ששיטות רב ערוציות נוטות להיות טובות יותר מאלגוריתמים חד ערוציים, ביצועיהן ברעש שאינו סטציונרי עדיין טעון שיפור.הסיבה לכך היא שכל החיישנים הינם מסוג דומה, ובו-זמנית קולטים אותות ורעשים עם מאפיינים דומים. פתרון אחד לבעיה זו הוא עיבוד בעזרת חיישנים רבים, שבו ניתן להשתמש בסוגים שונים של חיישנים המכילים מידע משלים. פתרון זה ניתן ליישום על-ידי שילוב מיקרופון אוויר עם מיקרופוני גרון, מיקרופונים פיסיולוגיים, מכשירי סטטוסקופ, וכו'. בין כל החיישנים שאינם אקוסטיים, מיקרופון עצם מוליך, המבוסס על חיישן דמוי-פייזואלקטרי, נחקר על ידי מספר גדול של חוקרים בזכות פשטותו ועלותו הנמוכה.

בעבודה זו, אנו נבחן את בעיית שיפור הדיבור על ידי עיבוד אותות משני חיישנים שונים, מיקרופון עצם מוליך (BC) ומיקרופון אוויר. מיקרופון עצם מוליך, בזכות היותו פחות רגיש לרעשי סביבה, ניתן לשימוש כדי להשלים ולעזור לפעולתו של מיקרופון אוויר רגיל בסביבות רועשות. עם זאת, מאחר שרכיבי התדר הגבוהים של אותות מיקרופון עצם מוחלשים באופן משמעותי כתוצאה מאובדן הולכה, האיכות של אותות דיבור שנרכשו על ידי מיקרופון עצם מוליך נמוכה יחסית. מצד שני, מיקרופון האוויר קולט תחום ספקטראלי רחב של הדיבור, אך הוא פחות חסין לרעש סביבתי. לכן, ברצוננו לשפר את אות הדיבור על ידי שילוב של שני המיקרופונים והפקת אות דיבור באיכות גבוהה עם רעש רקע נמוך.

בהתאם לתפקיד של מיקרופון עצם מוליך בריבוי חיישנים, השיטות הקיימות מסווגות לשתי קטגוריות עיקריות: 1) מיקרופון עצם מוליך משמש כחיישן נוסף, או, 2) כחיישן דומיננטי. יישום של מיקרופון עצם מוליך כחיישן נוסף מסתמך על הדיוק במידע שמופק מדיבור עם הולכת עצם, מאחר והתדרים הנמוכים של הדיבור נקלטים ע"י מיקרופון עצם באופן חסין לרעש סביבה. זה מאפשר הבחנה חזקה יותר בין הקטעים ללא דיבור לקטעים עם דיבור וזיהוי מדויק יותר של

פעילות קול בסביבת יחס אות לרעש נמוכה. כמו-כן, שיטה זו יושמה לשערוך גובה הצליל (pitch) במקרים בהם מודל האות-המסנן זקוק למידע מדויק של גובה הצליל.

כאשר מיקרופון BC מנוצל כחיישן רכישה ראשי, אלגוריתמים שלומדים את פונקציית ההעברה מדיבור שהוליכו דרך האוויר לדיבור שהוליכו דרך עצם ניתן לחלק לשלוש קבוצות: 1) שוויון, 2) ניתוח וסינתזה, 3) גישה הסתברותית. באופן כללי, גישות שוויון מאמצות מסנן FIR שנגזר לפי היחס של אות דיבור ממיקרופון רגיל ללא רעש ואות דיבור שהוליכו בעצם לאחר שעבר עיבוד התחלתי במנגנון משופר. בדומה לגישות של שיוון, שיטות ניתוח וסינתזה דורשות שיטות לשיפור דיבור מקדים כדי להפיק את מסנן השחזור, המיוצג על ידי היחס של מסנן חיזוי ליניארי משני הדיבורים. גישות ההסתברות מבוססות על מודל רעש גאוסי כאשר ההערכה מתקבלת בסבירות מקסימום דמיון או מינימום שגיאה ריבועית .בהתבססות על השערות של מודלים של רעש או דיבור, אלגוריתמים מורכבים יותר הוצעו, כולל SPLICE,מודל ברזולוציה גבוהה, הרחבת רוחב פס, וכו'.

אחת מהגישות ההסתברותיות בונה את מודל הדיבור על ספקטרום עוצמה מנורמל, מה שמאפשר הקטנת הווריאציה התוך והבין דיבורית, עם מספר קטן של רכיבי גאוס. רשת בייסיאנית דינמית המייצגת את משפחת ההתפלגויות ההסתברותיות המוגדרת במונחים של גרף מכוון משמשת כדי להעריך את ההסתברות האחורית של הדיבור הנקי, בהינתן מדידות ממיקרופון אוויר ומדידות ממיקרופון הולכת עצם. התפלגות ההסתברות המשותפת של המשתנים, המיוצגת על ידי צמתים בגרף, מתקבלת על ידי לקיחת מכפלת הפונקציות של תת קבוצת צמתים מחוברים. הערכת המינימום הסופי של השגיאה הריבועית הממוצעת מחושבת על ידי הסתברות שולית. אנו מיישמים את האלגוריתם ומשווים את התוצאה עם השיטות שלנו.

במקרה שמיקרופון העצם הוא החיישן הדומיננטי, אנו מציעים לשחזר דיבור נקי בעזרת הידע של דגימות דיבור ממיקרופון עצם מוליך ומיקרופון אוויר בשיטת ההרמוניות הגיאומטרית. שיטת ההרמוניות הגיאומטרית מתארת את הגיאומטריה של נתונים בממדים גבוהים ומרחיבה את התיאורים האלה לנקודות נתונים חדשות. ניתן גם לראות בה תכנית הפחתת ממד, על ידי הפיכת הנתונים של תכונה אחת לקבוצה חדשה של קואורדינטות המשקפות את הגיאומטריה ואת הצפיפות שלה. נתונים בממדים גבוהים במקרה זה מתקבלים ע"י שרשור של אותות דיבור ממיקרופון אוויר ומיקרופון עצם בתחום הזמן-תדר (short time Fourier transform domain).

מיפוי לא לינארי לדיבור הנקי שהוגדר על מערך נתונים זה נלמד ומורחב לשרשור חדש של אותות דיבור על ידי הרחבת פונקציות עצמיות של מטריצת זיקה. עם זאת, יישום שיטת ההרמוניות הגיאומטרית דורש הגדרה זהירה לקנה המידה הנכון ומספר התנאי. פתרון אפשרי הוא ליישם את ההרמוניות הגיאומטריות באופן רב סקאלה, כלומר ההרחבה מחושבת כסכום של ייצוגים בקני מידה שונה.

שיטה הרחבה רב-סקאלה נוספת, המכונה פירמידה לפלסיאנית, עוברת השוואה לשיטת ההרמוניות הגיאומטרית. בעיקרו של הדבר, הרחבת פירמידת לפלסיאן מבוססת על תכנית נסיגת גרעין בו סדרה יורדת של גרעינים גאוסיאנים משמשת להעריך שארית מייצוג קודם. רגרסיה קרנלית היא שיטה שאינה פרמטרית בשיטת ריבועים פחותים על ידי אינטרפולציה של נתונים מקיפים על פי המרחקים שלהם, הנותנת לדגימות הסמוכות משקל גבוה יותר מאשר דגימות רחוקות.

ביצענו ניסויים באמצעות סימולציות של אותות דיבור במיקרופוני אוויר ועצם, שנערכו בסביבות רעש גאוסיאני ורמקולים מפריעים, כאשר הדיבור במיקרופון עצם סונן על ידי מסנן מעביר נמוכים. שיטות גיאומטריות מספקות שיחזור דיבור יציב ביחס לרמת רעש וסוג הרעש, אם כי האות המשוחזר הוא באיכות נמוכה יותר. תוצאות הניסוי של השיטות המוצעות מוצגות גם בהשוואה לגישה ההסתברותית הקיימת ביחס ללוג המרחק הספקטרלי, שבו שיטת פירמידת לפלסיאן טובה יותר בביצועיה משיטות אחרות.