

2014

# Noise-Robust Voice Conversion

Trang Thi Minh Tran  
ttmt001@bucknell.edu

Follow this and additional works at: [https://digitalcommons.bucknell.edu/masters\\_theses](https://digitalcommons.bucknell.edu/masters_theses)

---

## Recommended Citation

Tran, Trang Thi Minh, "Noise-Robust Voice Conversion" (2014). *Master's Theses*. 115.  
[https://digitalcommons.bucknell.edu/masters\\_theses/115](https://digitalcommons.bucknell.edu/masters_theses/115)

This Masters Thesis is brought to you for free and open access by the Student Theses at Bucknell Digital Commons. It has been accepted for inclusion in Master's Theses by an authorized administrator of Bucknell Digital Commons. For more information, please contact [dcadmin@bucknell.edu](mailto:dcadmin@bucknell.edu).

I, Trang Tran, do grant permission for my thesis to be copied.



# NOISE-ROBUST VOICE CONVERSION

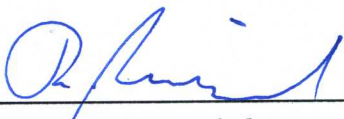
by

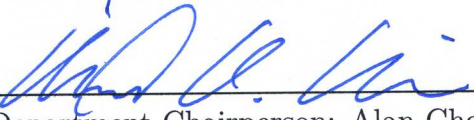
Trang Tran

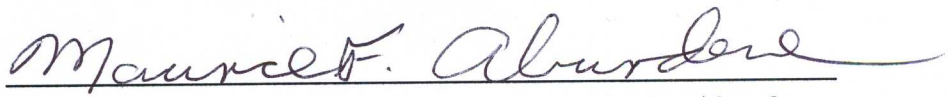
A Master's Thesis

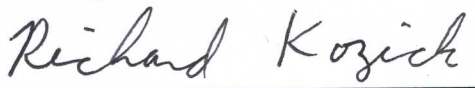
Presented to the Faculty of  
Bucknell University  
In Partial Fulfillment of the Requirements for the Degree of  
Master of Science in Electrical Engineering

Approved:

  
\_\_\_\_\_  
Adviser: Robert Nickel

  
\_\_\_\_\_  
Department Chairperson: Alan Cheville

  
\_\_\_\_\_  
Engineering Thesis Committee Member: Maurice Aburdene

  
\_\_\_\_\_  
Engineering Thesis Committee Member: Richard Kozick

\_\_\_\_\_  
April, 2014

# Acknowledgments

I would like to thank my advisor, Professor Robert Nickel, for being an incredible mentor over the past two years. Thank you for introducing me to *speech processing*, teaching me speech processing research with so much enthusiasm, and being always supportive and resourceful. Your mentorship helped me grow both research-wise and personally, and I am now much more confident to pursue a career in academia.

I would also like to thank my thesis committee members, Professor Maurice Aburdene and Professor Richard Kozick, who have been my mentor and advisor since my undergraduate years at Bucknell. Thank you for your continuing support over the past seven(!) years. I really appreciate you being on this committee and giving me valuable feedback.

To everyone in the Electrical and Computer Engineering Department, thank you for being so supportive of everything I do. Thank you for your advice regarding my career (and life) and for your constant encouragements. Special thanks to Judy Harris for always being so helpful (and for the chocolates!), you made my life a lot easier.

To Bucknell University, thank you for the amazing seven years. Studying abroad and having the freedom to learn what I'm excited about was very important to me – and I was able to do just that. Thank you for the education (and food and shelter), for helping me afford the education, and for all the friends I've made here.

Last but not least, I would like to thank my parents, for your unconditional love and support. Thank you for encouraging me to keep going, even though it means

I am not home very much. To my sister, grandmother, my family, and my friends both here and in Viet Nam, I thank you for always being there for me.

# Table of Contents

<b>List of Tables</b>	<b>v</b>
<b>List of Figures</b>	<b>vi</b>
<b>Abstract</b>	<b>viii</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Motivation . . . . .	1
1.2 Summary of Current Research . . . . .	3
1.2.1 Speech Enhancement . . . . .	3
1.2.2 Voice Conversion . . . . .	5
1.3 Summary of Thesis and Outline . . . . .	7
<b>2 Speech Enhancement</b>	<b>9</b>
2.1 Speech Enhancement Overview . . . . .	9
2.2 Filtering-Based Speech Enhancement . . . . .	12
2.2.1 Spectral Subtraction Methods . . . . .	12
2.2.2 Statistical Filtering Methods . . . . .	15
2.3 Filtering-based Methods Experiments . . . . .	17
2.3.1 Experiment Description . . . . .	17

2.3.2	Experimental Results . . . . .	19
2.4	Inventory-Based Speech Enhancement . . . . .	27
2.4.1	Feature Extraction . . . . .	29
2.4.2	System Training . . . . .	30
2.4.3	Inventory Search and Enhancement . . . . .	31
<b>3</b>	<b>Voice Conversion</b>	<b>33</b>
3.1	Voice Conversion Overview . . . . .	33
3.2	The Standard Model . . . . .	34
3.2.1	Feature Extraction . . . . .	35
3.2.2	Conversion Model Construction . . . . .	37
3.2.3	Target Speech Synthesis . . . . .	39
<b>4</b>	<b>Noise-Robust Voice Conversion</b>	<b>41</b>
4.1	System Description . . . . .	41
4.1.1	Feature Extraction . . . . .	44
4.1.2	Inventory Design and System Training . . . . .	45
4.1.3	Inventory Based Voice Conversion . . . . .	46
4.2	Experiment Description and Results . . . . .	48
4.3	Future Work . . . . .	51
4.3.1	Speech Enhancement . . . . .	51
4.3.2	Voice Conversion . . . . .	53
<b>5</b>	<b>Conclusion</b>	<b>54</b>
	<b>Bibliography</b>	<b>56</b>



# List of Tables

2.1	Algorithm Definitions for Spectral Subtraction Implementation . . . . .	21
2.2	Results of Spectral Subtraction Methods . . . . .	22
2.3	Results of Spectral Subtraction Methods with Loizou's Code . . . . .	23
2.4	Algorithm Definitions for Statistical Filtering Methods . . . . .	24
2.5	Results of Statistical Filtering Methods . . . . .	25
2.6	Algorithm Definitions for Windowing Effects Comparison . . . . .	25
2.7	Results of Windowing Effects Comparison Experiment . . . . .	26

# List of Figures

2.1	Block Diagram of a Speech Enhancement System (Parts of figure modified from [1]) . . . . .	11
2.2	Block Diagram of the Inventory-style Enhancement System . . . . .	28
3.1	Block Diagram of a Conventional Voice Transformation System . . . . .	34
4.1	Block Diagram of the Noise-Robust Voice Conversion Procedures . . . . .	43
4.2	Response counts of a <i>Comparison Category Rating</i> test with white noise after ITU-T recommendation P.800 [2]. . . . .	50
4.3	Response counts of a <i>Comparison Category Rating</i> test with jet cockpit noise after ITU-T recommendation P.800 [2]. . . . .	51

# Abstract

A persistent challenge in speech processing is the presence of noise that reduces the quality of speech signals. Whether natural speech is used as input or speech is the desirable output to be synthesized, noise degrades the performance of these systems and causes output speech to be unnatural. Speech enhancement deals with such a problem, typically seeking to improve the input speech or post-processes the (re)synthesized speech. An intriguing complement to post-processing speech signals is voice conversion, in which speech by one person (source speaker) is made to sound as if spoken by a different person (target speaker). Traditionally, the majority of speech enhancement and voice conversion methods rely on parametric modeling of speech. A promising complement to parametric models is an inventory-based approach, which is the focus of this work. In inventory-based speech systems, one records an inventory of clean speech signals as a reference. Noisy speech (in the case of enhancement) or target speech (in the case of conversion) can then be replaced by the best-matching clean speech in the inventory, which is found via a correlation search method. Such an approach has the potential to alleviate intelligibility and unnaturalness issues often encountered by parametric modeling speech processing systems. This work investigates and compares inventory-based speech enhancement methods with conventional ones. In addition, the inventory search method is applied to estimate source speaker characteristics for voice conversion in noisy environments. Two noisy-environment

voice conversion systems were constructed for a comparative study: a direct voice conversion system and an inventory-based voice conversion system, both with limited noise filtering at the front end. Results from this work suggest that the inventory method offers encouraging improvements over the direct conversion method.

# Chapter 1

## Introduction

### 1.1 Motivation

Speech processing is an area in signal processing that specifically deals with speech signals. Manipulation of speech signals includes: speech recognition (interpreting speech for machine use), speaker identification (recognizing the identity of a speaker correctly), speech synthesis (producing speech from text), among many others. In all these technologies, it is desirable to have high quality speech, whether speech is used as input or it is being produced by the machine. Such speech is not always available, due to a persistent challenge in speech processing research: the presence of noise.

*Speech enhancement* deals with improving the quality and intelligibility of speech signals degraded by noise. On the input side, the need to enhance speech signals arises in many scenarios: communication over cellular/radio systems nearly always suffers from background noise; speech recognition systems rely heavily on undistorted speech signal inputs; digital hearing aids still call for speech-selective enhancement to relieve the user from fatigue due to their loss of auditory focus ability.

On the output side, synthesized speech signals usually suffer from lack of intelligi-

bility and naturalness, making them perceptually displeasing to human listeners. On the one hand, this problem can be alleviated by refining the details of pre-processing and main subsystems (by using higher quality recording devices, for example). On the other hand, one can also add a post-processing subsystem that specifically deals with the already-synthesized speech. The latter approach might be more effective in many cases such as when one does not have access to the main subsystems or when these subsystems require independent developments (distant collaboration, for example).

Such speech post-processing methods can be further complemented by *voice conversion*: modifying one's speech to sound in a way that's different from the original speech, but preserving the textual content. In the more general case, *voice conversion* (or *voice transformation*) seeks to make a speech signal uttered by a *source* speaker sound as if uttered by a *target* speaker. In many concatenative speech synthesis systems, voice conversion allows for a more economical way of generating new voices instead of having to create a completely new database for the desired voice. Besides speech synthesis, voice conversion can also be used to protect one's identity in sensitive cases such as witness testimonies. In addition, numerous applications of voice conversion can be found in the entertainment industry: creating new voices for cartoon characters, dubbing foreign language films so that the translator can sound like the original actor etc.

While the majority of speech enhancement and voice conversion methods rely on parametric modeling of speech signals (statistical filtering for speech enhancement and Gaussian Mixture Models for voice conversion), a promising alternative/complement to the existing methods is *inventory-based* (or *corpus-based*) processing. In inventory-based speech processing systems, an inventory of clean speech signals can be used to replace the degraded speech or used as the target speech. This approach has

the potential to reduce the so-called *musical noise* and alleviate intelligibility and unnaturalness issues often encountered by traditional speech processing systems. This work investigated the inventory approach to speech enhancement and voice conversion in noisy environments.

## 1.2 Summary of Current Research

### 1.2.1 Speech Enhancement

One of the simplest and oldest classes of speech enhancement algorithms is the spectral subtraction method and its variations. These methods are based on the assumption that noise is additive, therefore speech can be enhanced by subtracting the estimated noise from the noisy signal. This method was first proposed by Weiss *et al.* [3] in 1975, where the subtraction was done in the correlation domain. Later, Boll [4] implemented spectral subtraction in the Fourier domain. Based on the fact that most additive noise affects certain frequencies more severely than others, many researchers proposed spectral subtraction methods that are frequency-band dependent. An example of such approach is the multiband spectral subtraction algorithm by Kamath and Loizou [5].

Another popular class of speech enhancement algorithms consists of methods based on the statistical estimation framework. These methods seek an estimate of the clean signal, given the noisy observations and an assumed probabilistic model of speech and noise. One of the earliest works in this area is the maximum likelihood approach for estimating the Fourier transform coefficients of the clean signal by McAulay and Malpass [6]. Ephraim and Malah [7] followed shortly after with a minimum mean square error (MMSE) and the Log-MMSE approach to magnitude spectrum estimation, which has proven successful in many applications [8]. Within

the statistical estimation framework, much attention has been dedicated to the estimation of the noise level and the estimation of the short-time magnitude spectrum of the clean signal. Martin [9], for example, proposed a noise power spectral density estimation scheme using minimum statistics while Cohen [10] had several works on noise estimation using minima-controlled recursive averaging. On the spectral magnitude estimation side, the work by Ephraim and Malah [11] following a decision-directed approach in 1985 remains one of the most successful to date.

Subspace algorithms for speech enhancement offer yet another perspective on the model of speech and noise. Rooted in linear algebra, subspace methods assume clean signals are confined to a subspace of the noisy Euclidean space. Consequently, these methods seek to decompose the noisy signal space into a clean subspace and a noise subspace. The clean signal can then be approximated by nulling the component belonging in the noise subspace [12]. Variations in the subspace methods are inspired by the variations in the decomposition schemes. For example, Ephraim and Van Trees [11] followed the eigenvalue decomposition (EVD) approach, while Dendrinos *et al.* [13] followed the singular value decomposition (SVD) approach.

All the methods mentioned above work well for stationary noise, but have been reported to perform less effectively in nonstationary noise [14]. In addition, there is usually a trade-off between noise suppression and speech distortion: speech signals that are aggressively filtered often suffer from psychoacoustically unpleasant artifacts such as musical noise [15]. Recent works to alleviate these distortion issues include cepstral smoothing by Breithaupt *et al.* [16] and over-attenuated spectral component regeneration by Ding *et al.* [17].

Further attempts to reduce distortion in speech enhancement have been implemented successfully in an alternative approach known as the inventory-based (or corpus-based) methods. In such a scheme, the noisy signal is not merely filtered, but



resynthesized from optimally chosen clean segments in a prerecorded inventory. Xiao and Nickel [18] first proposed such an approach in 2010, which was refined by Nickel *et al.* in 2013 [15]. Ming *et al.* [14] investigated corpus-based enhancement in non-stationary noise in particular, using the longest segments of clean speech identified. In addition to potentially generating an artifact-free enhanced speech signal, these methods have the advantage of being noise-independent, since the principal focus lies in finding the best segments in the *clean* inventory.

Most recently, Tseng *et al.* [19] proposed an enhancement scheme that combines statistical filtering and dictionary learning, named the Sparsity-based Wiener plus Dictionary Learning (SWDL). Earlier this year, Xu *et al.* [20] proposed a regression-based speech enhancement framework using deep neural networks (DNNs), which was reported to significantly reduce musical artifacts. Despite the numerous successes by these methods, much is left to be investigated since the quality of speech enhancement systems depend on many factors such as the application in question and computational costs.

### 1.2.2 Voice Conversion

Automatic voice conversion usually relies on two fundamental components: (1) a parametric encoding of the underlying sounds that allows for a faithful analysis/resynthesis of speech signals, and (2) a mapping that converts parameters of a source speaker’s voice into corresponding parameters of a target speaker’s voice. Examples for parameterizations that have been successfully used in the past include sinusoidal models [21] and Harmonic plus Noise Models (HMN) [22]. Very frequently cited is, in addition, the STRAIGHT parameterization introduced by Kawahara *et al.* in 2008 [23, 24]. STRAIGHT, which is also used in this work, refers to Speech Transformation

and Representation using Adaptive Interpolation of weiGHTed spectrum.

A variety of methods have been employed for feature vector conversion mappings, ranging from elementary vector quantization techniques from the early days of voice transformation [25] to very sophisticated mappings proposed in recent years. A conversion based on neural networks was proposed by Desai *et al.* in 2010 [26], a partial least squares regression was considered by Helander *et al.* in 2010 [27] and later refined in 2012 [28], a Gaussian Mixture Model (GMM) in combination with a noisy channel model was proposed by Saito *et al.* in 2012 [29], and Nirmal *et al.* suggested the use of GMMs in combination with radial basis functions in 2013 [30].

Traditional approaches of voice conversion rely on parameter mappings that operate instantaneously on a frame-by-frame basis. Significant improvements can be obtained by considering mappings that take the temporal evolution of feature vectors into account as well. Hidden Markov Models (HMMs) were successfully applied in this context by Duxans *et al.* [31], Nose and Kobayashi [32], and most recently by Percybrooks *et al.* [33].

In lieu of many other feature conversion techniques, the most widely used tools for feature mappings today are still Gaussian mixture model, as considered by Kain in 2001 [21] and Ohtani *et al.* in 2006 [34] for example. A significant problem of GMMs, however, can be found in their tendency to “oversmooth” the estimated parameter representation (see Toda *et al.* [35]). This problem has led a number of researchers to pursue Frequency Warping (FW) and/or Amplitude Scaling (AS) methods, either in conjunction with GMMs as proposed by Toda *et al.* in 2001 [36], or in lieu of GMMs as proposed by Godoy *et al.* in 2012 [37] and Erro *et al.* in 2013 [38].

Modern voice conversions systems have become able to convincingly transform the identity of a speaker. Two significant challenges, however, still remain at the forefront of study: (1) the generation of “natural” sounding converted speech [22],

and (2) the robustness of conversion procedures in noisy environments as considered by Takashima *et al.* in 2012 [39]. As discussed in the previous paragraph, due to the “oversmoothing” of many GMM based algorithms, the naturalness of the converted speech is often compromised. An approach that proved fruitful in enhancing the naturalness of text-to-speech systems is provided by concatenative speech synthesis. In *concatenative synthesis* a speech signal is produced by concatenating appropriately chosen “units” from a prerecorded voice inventory of the *target* speaker. The employment of a corpus-based concatenative approach to voice conversion was explored by Dutoit *et al.* in 2007 [40]. Related is also the work in unit selection by Shuang *et al.* from 2008 [41]. We are focusing on the inventory approach in [15] mainly because of the available speech enhancement framework, allowing for convenient extension to voice conversion.

### 1.3 Summary of Thesis and Outline

We are considering an alternative approach to improve noise robustness of voice conversion via a *concatenative analysis* approach on the *source* speaker. The work is motivated by the success of inventory-based enhancement schemes as proposed by Xiao and Nickel in 2010 [18] and later refined by Nickel *et al.* in 2013 [15]. In noisy environments reliable estimation of the “true” underlying parameterization of speech is difficult, and hence the performance of conventional voice conversion schemes tends to degrade significantly. In our work we employ a STRAIGHT feature set [23] and a simple instantaneous GMM based conversion mapping.

This thesis is organized as follows: **Chapter 2** describes several speech enhancement algorithms that were studied in preparation for this project. In particular, we implemented a subset of standard filtering-based and the inventory-based speech

enhancement algorithms. The quality of these systems are compared and remaining challenges are noted. **Chapter 3** describes the most popular voice conversion system: the Gaussian Mixture Model-based voice conversion. Details are given regarding how to construct such a system. **Chapter 4** describes our proposed system: voice conversion in noisy environments using an inventory approach for source analysis. Results of our experiments are presented and discussed. Future work to potentially deal with remaining challenges of both aspects (speech enhancement and voice conversion) are also outlined. **Chapter 5** provides final remarks on this work.

# Chapter 2

## Speech Enhancement

### 2.1 Speech Enhancement Overview

A speech enhancement system aims to improve the quality of an input speech signal degraded by noise. Though the specific enhancement methods vary, the basic building blocks of a speech enhancement system can be described as follows.

We begin by constructing a model for speech signals corrupted by *additive* noise. Let us denote the underlying *clean* speech signal  $z[n]$  and the corrupting noise  $v[n]$ , our *observed* (noisy) signal is then  $x[n] = z[n] + v[n]$ . A speech enhancement system typically consists of three subsystems: *analysis* of the input signal  $x[n]$ , *enhancement* of this noisy signal in the parameter domain (such as frequency) and *resynthesis* of the enhanced signal. Signal *analysis* takes the noisy input and divides it into overlapping segments of a specified length (typically 20-ms segments with 50% overlap). One can also apply different windows (Hamming, Hann) to each segment or leave each as-is, which corresponds to applying a rectangular window. The purpose of such operation is to reduce boundary effects and spectral leakage when converting samples from the time domain to the parameter domain (frequency). Typically, the parameter of choice

is the coefficients of the discrete Fourier transform, so the fast Fourier transform (FFT) is then performed frame-wise, resulting in a new set of frequency-domain samples. We call the matrix of these samples' magnitudes  $X[k, n]$  where  $k$  denotes the  $k^{\text{th}}$  frequency bin and  $n$  denotes the  $n^{\text{th}}$  frame from the segmentation process. Each enhancement method can be represented by an *adaptive gain function*  $G[k, n]$ , which depends on the type of algorithm under study. Several such basic gain functions will be briefly described in Section 2.2. The gain function matrix is applied element-by-element to the noisy magnitude matrix  $X[k, n]$  to estimate the magnitude spectrum of the enhanced signal. We denote the *enhanced* frequency-domain magnitude matrix  $Y[k, n] = X[k, n] \cdot G[k, n]$ . The resulting magnitude spectrum is then combined with the stored noisy phase spectrum to arrive at the overall enhanced spectrum  $Y'[k, n]$ . The inverse FFT is performed on  $Y'[k, n]$ , yielding the time-domain enhanced signal matrix. Depending on the type of window applied originally, the *enhanced* signal  $y[n]$  can be obtained by concatenating the weighted-and-overlapped samples in each segment.

Note that while reconstruction of the final signal employed the enhanced magnitude spectrum, the phase spectrum was taken from the noisy signal. This approach, however, has yielded relatively good results since the human ear is less sensitive to the phase component of speech signals. The enhancement process is summarized in the block diagram of Figure 2.1.

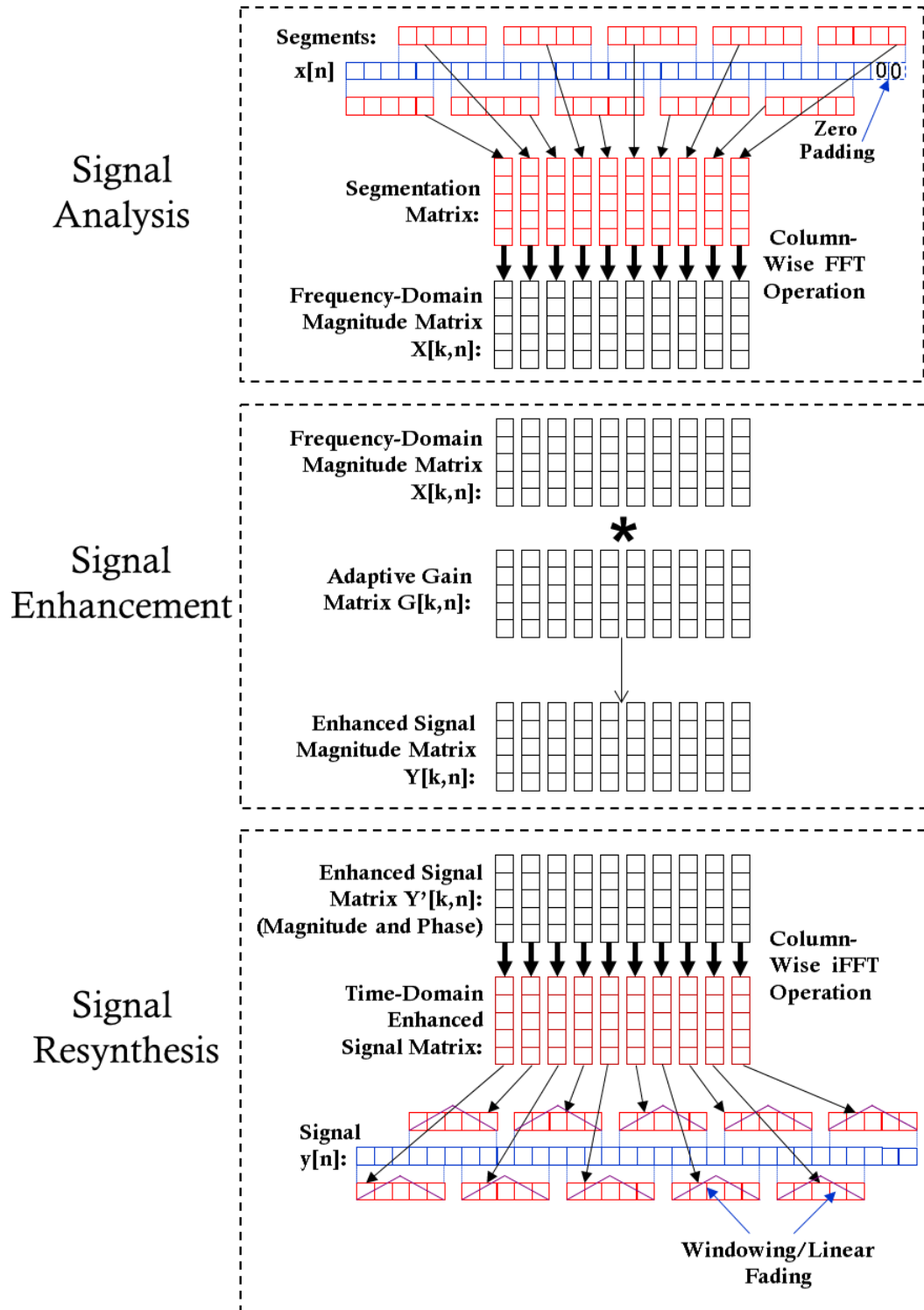


Figure 2.1: Block Diagram of a Speech Enhancement System  
(Parts of figure modified from [1])

## 2.2 Filtering-Based Speech Enhancement

Several fundamental speech enhancement methods are considered in the following sections. Our main reference was the text *Speech Enhancement, Theory and Practice* [42] by Loizou. Specifically, we cover the Spectral Subtraction and Statistical Filtering methods, along with their variations.

### 2.2.1 Spectral Subtraction Methods

The spectral subtraction (SS) algorithm is historically one of the first algorithms proposed for acoustic noise reduction [42, Ch. 5]. Assuming additive noise that is uncorrelated to speech, one can reasonably estimate the clean signal spectrum by subtracting the approximated noise spectrum from the noisy one. There are multiple variations to spectral subtraction methods, two of which are presented here.

- Simple Spectral Subtraction (SSS):

The gain function for basic spectral subtraction is shown in Equation 2.1. Since we are operating on the magnitude spectrum, negative values of the enhanced magnitude spectrum  $Y[k, n]$  are meaningless. The simplest solution for this problem is to floor negative difference spectrum values to 0.

$$G[k, n] = \begin{cases} \sqrt{1 - \frac{\widehat{D}^2[k, n]}{X^2[k, n]}} & \text{if } X[k, n] > \widehat{D}[k, n] \\ 0 & \text{otherwise} \end{cases} \quad (2.1)$$

where  $\widehat{D}[k, n]$  denotes the average magnitude spectrum of the noise. In this simple method, we assume that the first few frames (typically 40) of the signal contain only noise. Therefore  $\widehat{D}[k, n]$  can be found by averaging the spectra of these first frames.



Even though noise is reasonably suppressed in this simple approach, considerable distortion of speech arises. The effect of *musical noise*, the irritating and unnatural sound heard in the background, is especially pronounced in this situation. This type of speech distortion is aggravated by flooring the negative difference spectrum values in the final enhanced spectrum [42, Ch. 5].

- Multiband Spectral Subtraction (MBSS):

A multiband spectral subtraction method was shown to perform significantly better than most other SS approaches [42, Ch. 5]. A multiband method is based on the fact that noise affects speech signals differently in different frequency bands. In addition to taking into account this non-uniformity of noise corruption, MBSS pre-processes the noisy magnitude spectrum according to Equation 2.2, which turned out to play a significant part in the quality enhancement of the final speech.

$$\bar{X}[k, n] = \sum_{i=-M}^M W_i X[k, n - i] \quad (2.2)$$

This *smoothing* process essentially computes a weighted spectral average over  $M$  preceding and  $M$  succeeding frames for a certain segment;  $M$  is usually limited to 2. The weights  $W_i$  are empirically [42, Sec. 5.6] set to

$$W_i = [0.09, 0.25, 0.32, 0.25, 0.09].$$

Finally, the enhanced magnitude spectrum is the result of multiplying this *smoothed* magnitude spectrum with the gain function in Equation 2.3. The MBSS gain function is different for each frequency band  $i$ .

$$G_i[k, n] = \begin{cases} \sqrt{1 - \alpha_i \cdot \delta_i \frac{\widehat{D}^2[k, n]}{\overline{X}^2[k, n]}} & \text{if } G_i[k, n] > \beta^{1/2} \text{ and } b_i \leq k \leq e_i \\ \beta^{1/2} & \text{otherwise} \end{cases} \quad (2.3)$$

where  $b_i$  and  $e_i$  are the beginning and ending frequency bins of the  $i^{\text{th}}$  frequency band,  $\alpha_i$  is the oversubtraction factor of the  $i^{\text{th}}$  band and  $\beta$  is the spectral floor parameter, typically set to 0.002 as suggested in [42, Sec. 5.6].  $\delta_i$  is the additional band-subtraction factor that can be individually set for each frequency band to customize the noise removal process. Loizou [42, Sec. 5.6] provides more details on choosing the optimal values for these parameters.

Further masking of musical noise can also be achieved by reintroducing a small amount of noise to the enhanced spectrum as in Equation 2.4.

$$\overline{\overline{Y}}^2[k, n] = Y^2[k, n] + 0.05 \cdot \overline{X}^2[k, n] \quad (2.4)$$

Although this last step seems counter-intuitive in the noise-reduction sense, human listeners turn out to prefer a slightly noisier signal to one with musically distorted speech. This observation indicates that algorithmic optimization of signal quality does not necessarily lead to the corresponding result in the *psychoacoustic* sense. Human perception of speech quality (and our limited understanding in this area) therefore furthers the challenges in speech enhancement optimization.

### 2.2.2 Statistical Filtering Methods

In statistical filtering methods, the speech enhancement problem is posed in a statistical estimation framework [42, Ch. 7]. The noisy magnitude spectrum is our set of observations, from which we try to obtain an estimate for the underlying parameters - the clean speech magnitude spectrum. As with SS methods, we can also write these statistical estimators in forms of gain functions to operate on the noisy magnitude spectrum. Loizou [42, Ch. 7] provides an overview of various statistical filtering enhancement algorithms, with the log-Minimum Mean Square Error (MMSE) estimator recommended as having superior performance.

The MMSE/log-MMSE models belong to a subset of the Bayesian estimation approach, where we make use of the *a priori* probability density function (PDF) of the estimation parameter  $Z[k, n]$ . However, measuring the true probability distribution of the speech Fourier transform coefficients is difficult because speech signals are not stationary [42, Ch. 7]. Ephraim and Malah [43] proposed a statistical model that makes two assumptions: (1) The Fourier transform coefficients (real and imaginary parts) have a Gaussian PDF with zero mean and time-varying variance; (2) The Fourier transform coefficients are statistically independent and, hence, uncorrelated. Despite these unrealistic assumptions, the resultant models have proven useful in practice.

Derivations in [42, Ch. 7] show that both MMSE and log-MMSE estimators' gain functions can be written as  $G[k, n] = F(\xi[k, n], \gamma[k, n])$ , where  $\xi[k, n]$  and  $\gamma[k, n]$  are referred to as the *a priori* and *a posteriori* SNR estimates, respectively. The *a posteriori* SNR  $\gamma[k, n]$  can be obtained from Equation 2.5.

$$\gamma[k, n] = \min \left( \frac{X^2[k, n]}{\widehat{D}^2[k, n]}, 40 \right) \quad (2.5)$$

where  $\widehat{D}^2[k, n]$  is the variance of the noise magnitude spectrum, again estimated using the first few frames of the signal. However, in these calculations, a small value of  $\widehat{D}^2[k, n]$  can cause unrealistically high values in  $\gamma[k, n]$ . To avoid this, our implementation sets a higher bound for  $\gamma[k, n]$  at 40, as suggested by Loizou [42, Ch. 7].

Estimation of the *a priori* SNR is another essential component of the statistical filtering estimators. This is no easy task since we only have access to the noisy speech signal. Several methods have been proposed to deal with this problem, among which the decision-directed approach by Ephraim and Malah [43] has proven useful. Equation 2.6 provides the standard estimation formula for  $\xi[k, n]$ .

$$\xi[k, n] = \max \left[ a \frac{Y^2[k, n-1]}{\widehat{D}^2[k, n-1]} + (1-a) \max[\gamma[k, n] - 1, 0], \xi_{min} \right] \quad (2.6)$$

where  $\xi_{min}$  is the minimum value allowed for  $\xi[k, n]$ . A value of  $\xi_{min} = -15$  dB was suggested by Cappe [44]. As can be seen from Equation 2.6, the decision-directed approach to estimating the *a priori* SNR is superior in that it takes into account previous frames' enhanced spectra. The initial conditions recommended for  $\xi[k, 0]$  are:

$$\xi[k, 0] = a + (1-a) \max[\gamma[k, 0] - 1, 0] \quad (2.7)$$

where the  $\max(\cdot)$  operator ensures non-negativity and  $a = 0.98$  was empirically suggested.

Both our preliminary statistical filtering enhancement methods use the *a priori* SNR estimation described above and the gain functions for each algorithm are presented in Equations 2.8 and 2.10 below.

- MMSE Estimator

$$G[k, n] = \frac{\sqrt{\pi}}{2} \frac{\sqrt{\nu[k, n]}}{\gamma[k, n]} e^{(-\frac{\nu[k, n]}{2})} \left[ (1 + \nu[k, n]) I_0 \left( \frac{\nu[k, n]}{2} \right) + \nu[k, n] I_1 \left( \frac{\nu[k, n]}{2} \right) \right] \quad (2.8)$$

where

$$\nu[k, n] = \frac{\xi[k, n]}{1 + \xi[k, n]} \gamma[k, n] \quad (2.9)$$

and  $I_0(\cdot), I_1(\cdot)$  are modified Bessel functions of zero and first order, respectively.

- log-MMSE Estimator

$$G[k, n] = \frac{\xi[k, n]}{\xi[k, n] + 1} \exp \left\{ \frac{1}{2} \int_{\nu[k, n]}^{\infty} \frac{e^{-t}}{t} dt \right\} \quad (2.10)$$

where the integral is known as the exponential integral and can be evaluated numerically [42, Ch. 7].

## 2.3 Filtering-based Methods Experiments

We now describe the experiments and results from implementing the filtering based speech enhancement methods. Although Loizou [42] provides MATLAB codes of these methods, we wrote our own implementations to (1) understand better the workings of these methods and (2) allow for modifications and expansions if needed.

### 2.3.1 Experiment Description

In our study, we have access to an inventory of prerecorded speech utterances by a set of individuals. A group of five random utterances by the same speaker are

concatenated into one long (10-15 sec) signal stream, which forms the signal  $z[n]$ . In addition, we are provided with an inventory of noise sounds (e.g. White Gaussian, Pink, Speech Babble) that we use as our additive noise signal  $v[n]$ . Our resulting noisy signal  $x[n]$  becomes the input to the enhancement algorithms, which produces the *enhanced* signal  $y[n]$ . All these speech signals are sampled at 16 kHz and are assumed to be band-limited between 50 Hz and 8 kHz [15].

The database of clean speech available to us is the `CMU_ARCTIC` database from the Language Technologies Institute at Carnegie Mellon University<sup>1</sup>. This set consists of sample utterances from 7 speakers, chosen to form a corpus with large phonetic content, diverse speech patterns and accents [15]. Five utterances of each speaker were concatenated to form a long stream of clean signal. For each speaker, ten such streams are formed to be used as clean speech data. We concatenate such a stream of utterances because of the *adaptive* nature of the speech enhancement algorithms. Using a longer speech signals allows the enhancement filters to adapt accordingly, thus yielding better quality enhanced signals.

The type of additive noise employed in these preliminary experiments was White Gaussian noise, taken from the `NOISEX` database from the Institute for Perception-TNO, The Netherlands Speech Research Unit, RSRE, UK<sup>2</sup>. This noise was added to clean speech data at signal-to-noise (SNR) ratios of 5 dB and 10 dB, under consideration of respective *active speech level* (ASL) [45]. In this way, our noisy signals are produced with SNR levels strictly with respect to *active speech*, whereas the overall SNR of the signal is slightly lower due to silent periods within the signal. In other words,  $\text{SNR}_{\text{speech}} = \text{SNR}_{\text{signal}} - 10 \log_{10}(\text{ASL})$  where  $0 < \text{ASL} < 1$ . The resulting noisy signal  $x[n]$  was the input to our various speech enhancement systems.

---

<sup>1</sup>This corpus is available at [http://festvox.org/cmu\\_arctic/](http://festvox.org/cmu_arctic/).

<sup>2</sup>This noise corpus is available at [http://spib.rice.edu/spib/select\\_noise.html](http://spib.rice.edu/spib/select_noise.html).

In these experiments, two primary classes of enhancement methods were considered: spectral subtraction methods and statistical filtering methods. In both cases, *noise estimation* was an important component. Our experiments assumed speech absence in the first few frames of the segmentation matrix; hence we used averages extracted from these frames as our estimated noise magnitude spectrum  $\hat{D}[k, n]$ . Loizou in [42, Ch. 9] details several improved approaches to noise estimation that were not implemented in these preliminary experiments.

### 2.3.2 Experimental Results

Following the initial set up in Section 2.3.1, algorithms described in Section 2.2.2 were applied on the pre-mixed noisy signal  $x[n]$  at SNR levels of 5 dB and 10 dB with ASL considerations. Our speech samples were from two speakers in the CMU\_ARCTIC database, whose identifiers were BDL (US English male) and SLT (US English female). The speech analysis portion of our experiments had the following common parameters: 20 ms frame length for signal segmentation with 50% overlap and rectangular windowing. While other types of windowing can be used, the rectangular window is the simplest to implement and has the narrowest main-lobe width frequency response (which is desirable for a fast transition from the pass-band to the stop-band from an FIR filter design point of view) [46]. For noise estimation, we averaged the magnitude spectrum values from the 40 first frames (assumed silent). For resynthesis, we used linear cross fading to concatenate our final signal. Other experimental parameters were algorithm-specific and will be defined below.

As a companion resource to [42], Loizou provided us with his MATLAB implementations of the enhancement methods discussed above. For verification and comparison purposes, Loizou’s equivalent algorithms were also applied on our noisy signals. There

are, however, two important differences between our implementation and Loizou’s: (1) Loizou’s uses the Hamming window for signal segmentation and the overlap-add method for reconstructing the enhanced signal; (2) Loizou’s algorithms incorporate voice activity detection (VAD) [47] (See also [42, Sec. 11.2]) for noise estimation. Instead of using a fixed average of the noise spectrum, the VAD updates the noise estimate by assigning more “noise weight” to frames assumed speech absent, thus potentially yielding a better noise estimate overall.

We present here four sets of experiments: (1) implementation of our spectral subtraction methods; (2) implementation of Loizou’s spectral methods; (3) implementation of statistical filtering methods (both ours and Loizou’s) and (4) implementation of our algorithms modified for studying the effects of windowing.

Quality assessment can be done using subjective listening tests and/or objective quality measures. Subjective evaluation relies on having a group of listeners rate the quality of the enhanced speech according to a certain quantitative scale [42, Ch. 10]. Subjective tests, however, are time-consuming and expensive. Consequently, we need mathematical/algorithmic-based measures to evaluate the enhancement methods, at least in the learning/early implementation stages. These “objective” measures quantify quality by measuring the “distance” between the original and processed signals [42, Ch. 10]. Clearly, these objective measures need to correlate well with subjective listening tests to be valid since the ultimate judge of our signal quality is the human ear.

Four such objective measures were chosen for all our algorithm evaluations: the *Perceptual Evaluation of Speech Quality* (PESQ) [48] and three *Composite* measures for *Signal Distortion* (CSIG), *Background Noise Distortion* (CBAK) and *Overall Quality* (COVL) [42]. The PESQ measure is highly correlated with subjective listening tests [48] and therefore is considered one of the more reliable objective quality



measures. The *Composite* measure is achieved by combining multiple objective measures. Since different objective measures capture different characteristics of the distorted signal, combining them can yield significant gains in correlation [42, Ch. 10]. Our specific composite measures above were computed via linear combinations of standard objective measures including the PESQ, the *Cepstrum Distance Measure* [49] and the *Frequency-Weighted Segmental SNR* [50].

### 1. Implementation of Spectral Subtraction Methods:

Our first experiment compares the performances of spectral subtraction methods. In particular, we compare the results from simple spectral subtraction (SSS) and multiband spectral subtraction (MBSS). Table 2.1 defines algorithm parameters of each method.

Table 2.1: Algorithm Definitions for Spectral Subtraction Implementation

ALG.	Description	Windowing	Reconstruction	Algorithm Parameters	Noise Estimation
1	SSS	Rectangular	Linear fading	n/a	40 first frames' average
2	MBSS	Rectangular	Linear fading	8 bands; linear spacing	40 first frames' average
3	MBSS	Rectangular	Linear fading	8 bands; log spacing	40 first frames' average
4	MBSS	Rectangular	Linear fading	8 bands; mel spacing	40 first frames' average

Table 2.2 presents the results of this experiment. As a reference, the four quality measures were also computed for the noisy signal. Higher scores for an algorithm

compared to those for the noisy signal indicate quality improvement.

For both noise levels, it is clear that the MBSS method is superior to the simple SS, regardless of the way frequency bands are divided. Within MBSS variations, the objective measures seem to agree that the log-band division performs the best. However, psychoacoustically little difference was perceived among these enhanced signals.

Table 2.2: Results of Spectral Subtraction Methods

10 dB SNR	PESQ		CSIG		CBAK		COVL	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Noisy Signal	1.576	0.261	1.361	0.380	2.426	0.191	1.387	0.399
ALG 1	1.906	0.224	1.022	0.099	2.172	0.156	1.084	0.182
ALG 2	1.978	0.183	1.667	0.427	2.586	0.1139	1.801	0.311
ALG 3	1.986	0.187	1.794	0.426	2.568	0.1208	1.862	0.311
ALG 4	1.958	0.189	1.661	0.432	2.574	0.1176	1.787	0.316
5 dB SNR	PESQ		CSIG		CBAK		COVL	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Noisy Signal	1.228	0.244	1.021	0.057	1.961	0.186	1.099	0.122
ALG 1	1.561	0.228	1.0000	0.0000	1.908	0.139	1.019	0.070
ALG 2	1.565	0.195	1.175	0.197	2.286	0.109	1.270	0.263
ALG 3	1.568	0.204	1.239	0.259	2.254	0.117	1.309	0.285
ALG 4	1.533	0.207	1.171	0.194	2.268	0.115	1.258	0.259

## 2. Implementation of Loizou’s Spectral Subtraction Methods:

To verify that our SS methods were implemented appropriately, we obtained the same quality measures for the MBSS methods provided by Loizou. Table 2.3 shows that Loizou’s implementations yielded higher values for the same quality

measures. This is due to the two significant differences in his code discussed above – windowing/reconstruction as well as VAD. However, Loizou’s results agree with ours in the trend of objective quality comparison, i.e. again we see the log-frequency spacing method achieving highest scores.

Regarding musical noise, MBSS was superior to SSS, though slight distortions still seem to be present. Even more masking via Equation 2.4 can be achieved by adding noise to the enhanced signal. This, however, would likely decrease our objective quality scores as well as make the background noise louder, due to the trade-off between noise suppression and perceived signal naturalness.

Table 2.3: Results of Spectral Subtraction Methods with Loizou’s Code

10 dB SNR	PESQ		CSIG		CBAK		COVL	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Noisy Signal	1.576	0.261	1.361	0.380	2.426	0.191	1.387	0.399
ALG 2	1.992	0.314	1.913	0.639	2.598	0.238	1.939	0.489
ALG 3	2.033	0.280	2.131	0.542	2.595	0.205	2.065	0.421
ALG 4	2.001	0.310	1.995	0.642	2.593	0.232	1.985	0.488
5 dB SNR	PESQ		CSIG		CBAK		COVL	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Noisy Signal	1.228	0.244	1.021	0.057	1.961	0.186	1.099	0.122
ALG 2	1.364	0.370	1.273	0.299	2.129	0.269	1.308	0.326
ALG 3	1.408	0.371	1.390	0.409	2.133	0.260	1.373	0.390
ALG 4	1.363	0.373	1.324	0.348	2.124	0.268	1.333	0.351

### 3. Implementation of Statistical Filtering Methods:

Our next experiment compared statistical filtering methods. The algorithm

parameters are presented in Table 2.4 and their results shown in Table 2.5. We include Loizou’s implementation results in the same table for reference.

Table 2.4: Algorithm Definitions for Statistical Filtering Methods

ALG.	Description	Windowing	Reconstruction	Noise Estimation
1	MMSE	Rectangular	Linear fading	40 first frames’ average
2	MMSE (Loizou’s)	Hamming	Overlap and Add	VAD
3	log-MMSE	Rectangular	Linear fading	40 first frames’ average
4	log-MMSE (Loizou’s)	Hamming	Overlap and Add	VAD

At both SNR levels, we see improvement in objective scores from the simple MMSE algorithm to the log-MMSE method. Again, Loizou’s implementation yielded higher scores, but was reassuring in that it confirms the comparative trend between the two measures. Compared to SS methods above, all statistical filtering methods received higher objective scores. Musical noise also seems to be less pronounced in the enhanced signals.

#### 4. Effects of Windowing:

Our final experiment considered the effects of windowing/reconstruction method on the enhanced signal. Table 2.6 presents our parameter definitions and Table 2.7 shows the results of this experiment.

Examination of Table 2.7 shows windowing does affect the objective quality of the enhanced signal. More importantly, resynthesis of the enhanced signal

Table 2.5: Results of Statistical Filtering Methods

10 dB SNR	PESQ		CSIG		CBAK		COVL	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Noisy Signal	1.576	0.261	1.361	0.380	2.426	0.191	1.387	0.399
ALG 1	1.936	0.178	1.887	0.361	2.559	0.129	1.874	0.268
ALG 2	2.289	0.161	2.321	0.514	3.020	0.100	2.291	0.346
ALG 3	2.170	0.142	1.982	0.297	2.604	0.137	2.024	0.219
ALG 4	2.493	0.099	2.491	0.434	3.127	0.072	2.471	0.272
5 dB SNR	PESQ		CSIG		CBAK		COVL	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Noisy Signal	1.228	0.244	1.021	0.057	1.961	0.186	1.099	0.122
ALG 1	1.613	0.226	1.318	0.330	2.293	0.128	1.380	0.325
ALG 2	1.932	0.179	1.651	0.496	2.664	0.087	1.763	0.349
ALG 3	1.878	0.210	1.372	0.339	2.369	0.127	1.543	0.303
ALG 4	2.130	0.165	1.813	0.455	2.769	0.073	1.935	0.319

Table 2.6: Algorithm Definitions for Windowing Effects Comparison

ALG.	Description	Windowing	Reconstruction	Noise Estimation
1	log-MMSE	Rectangular	Linear fading	40 first frames' average
2	log-MMSE	Hamming	Linear fading	40 first frames' average
3	log-MMSE	Hamming	Weighted Overlap-Add	40 first frames' average
4	log-MMSE (Loizou's)	Hamming	Overlap-Add	VAD

needs to be consistent with the windowing pattern in the segmentation stage. For example, linear fading reconstruction is “geometrically” consistent with rectangular segmentation because of the weights applied to tailing samples. Similarly, a Hamming-shaped segmentation would be approximately consistent with simply adding overlapping samples at resynthesis rather than weighing them before adding. Listening tests agree with these objective scores: while ALG 1, 3 and 4 showed little perceptual difference, ALG 2 produced a clearly more distorted speech signal.

Table 2.7: Results of Windowing Effects Comparison Experiment

10 dB SNR	PESQ		CSIG		CBAK		COVL	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Noisy Signal	1.576	0.261	1.361	0.380	2.426	0.191	1.387	0.399
ALG 1	2.170	0.142	1.982	0.297	2.604	0.137	2.024	0.219
ALG 2	2.061	0.206	1.883	0.362	2.402	0.155	1.913	0.297
ALG 3	2.142	0.164	1.973	0.297	2.414	0.124	2.004	0.231
ALG 4	2.493	0.099	2.491	0.434	3.127	0.072	2.471	0.272
5 dB SNR	PESQ		CSIG		CBAK		COVL	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Noisy Signal	1.228	0.244	1.021	0.057	1.961	0.186	1.099	0.122
ALG 1	1.878	0.210	1.372	0.339	2.369	0.127	1.543	0.303
ALG 2	1.809	0.276	1.339	0.344	2.224	0.182	1.473	0.363
ALG 3	1.864	0.223	1.367	0.322	2.225	0.148	1.533	0.305
ALG 4	2.130	0.165	1.813	0.455	2.769	0.073	1.935	0.319

To summarize, our preliminary experiments show superior performance of statistical filtering methods compared to spectral subtraction ones. Within statistical

filtering algorithms, the log-MMSE method yielded promising objective scores. In addition, our implementations showed consistent score patterns with those by Loizou, even though our scores were always lower. This suggests noise estimation plays an important role in achieving good enhancement results (again, at least in the objective mathematical sense).

Informal listening tests correlate well with these standard objective measures. Although both the MBSS and the log-MMSE methods showed significant improvement to the noisy speech, there is still a considerable amount of background noise. In cases where noise was significantly suppressed, speech distortion became more pronounced. This is the typical trade-off regarding the quality of the resynthesized signal and remains a challenging topic for speech enhancement research.

## 2.4 Inventory-Based Speech Enhancement

To further improve speech enhancement without compromising quality and intelligibility, alternative methods employing an *inventory* of clean speech signals have been proposed [18, 15, 14]. The noisy signals in this approach are not merely filtered, but resynthesized based on the undistorted, pre-recorded speech waveforms. The advantage of this “waveform matching” scheme lies in the fact that not only spectral magnitude but also spectral phase is estimated. Successful implementations of this *inventory-style* enhancement scheme include the work of Xiao *et al.* [18], and later refined by Nickel *et al.* in [15].

The block diagram for an inventory-style enhancement procedure following [15] is shown in Figure 2.2. Compared to the standard enhancement procedure presented in Figure 2.1, the significant modifications lie in the addition of a VAD mechanism in speech analysis and the enhancement subsystem. In this approach, the enhancement

subsystem consists of three components: (1) a conventional log-MMSE estimator; (2) a VAD mechanism and (3) an inventory search procedure. The log-MMSE filter is implemented as described in Section 2.2.2. The VAD block is employed according to Sohn *et al.* [47]; it activates the inventory search subsystem *only* during intervals of assumed speech presence. The VAD mechanism also controls the stream selection and normalization in the post-processing block. In addition, similar to Loizou’s implementations, the VAD mechanism is used in the noise estimation subroutine of the log-MMSE estimator to improve its performance. The following sections focus on the inventory-based enhancement subsystem components, which consists of feature extraction, system training and inventory search.

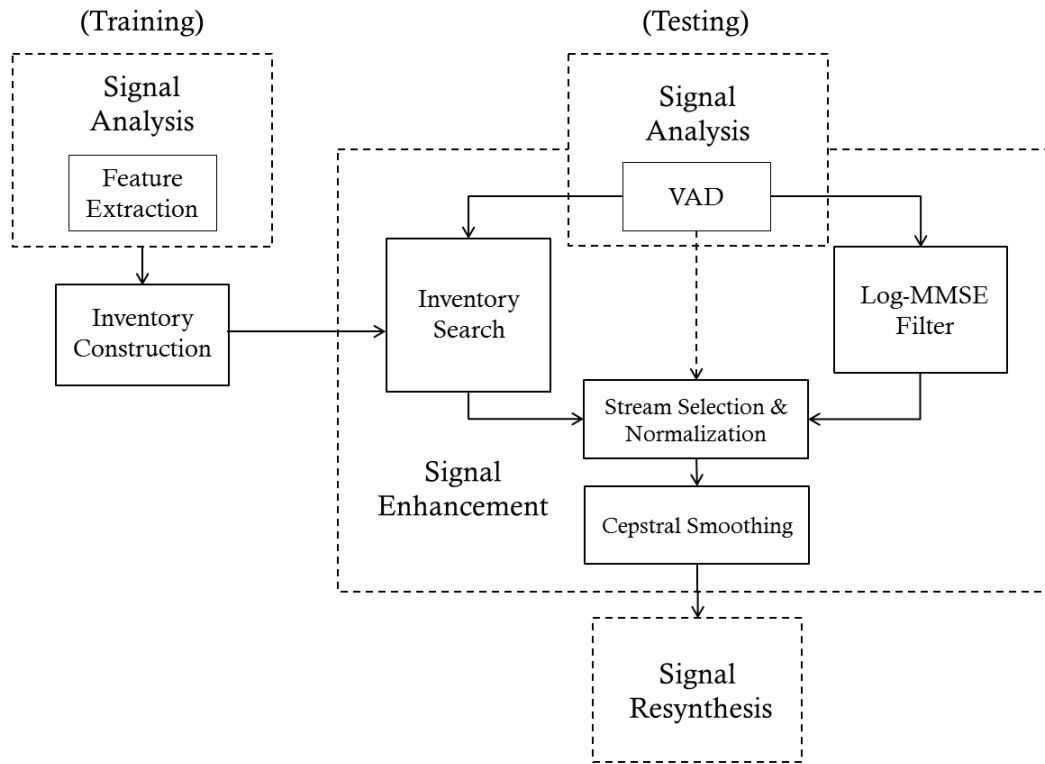


Figure 2.2: Block Diagram of the Inventory-style Enhancement System



### 2.4.1 Feature Extraction

The goal of *feature extraction* is to represent a speech signal in an efficient way that reflects the signal's most important characteristics. In particular, an ideal feature extractor maps a speech signal  $z[n]$  to a feature vector  $\mathbf{z}$  (or a set of feature vectors  $\mathbf{z}[i]$ ) while dramatically reducing data dimensionality and retaining most of the information relevant to the system's task. Since processed speech signals eventually are received by humans, psychoacoustically motivated features have proven to be most successful [51]. One such set of features widely used are the *Mel-Frequency Cepstral Coefficients* (MFCCs), whose definition and computation process is described in [51]. Another advantage of the MFCCs is their robustness against additive noise [18, 15], which makes them a desirable option for not only speech enhancement but also speech recognition applications. Our inventory-based speech enhancement system uses MFCCs as our features for this reason.

The MFCC feature extraction process follows a slightly simpler procedure than the one described in [18, 15]. Following a pre-emphasis filter, we compute 13-dimensional MFCC vectors  $\hat{\mathbf{c}}_{\mathbf{z}}[i]$  and apply a sinusoidal lifter to these coefficients according to the recommendations by Young *et al.* [52]. We then augment these 13 MFCCs with their  $\Delta$  and  $\Delta\Delta$  coefficients after [15] to arrive at 39-dimensional feature vectors:

$$\mathbf{c}_{\mathbf{z}}[i] = [ \hat{\mathbf{c}}_{\mathbf{z}}^T[i] \quad \Delta\hat{\mathbf{c}}_{\mathbf{z}}^T[i] \quad \Delta\Delta\hat{\mathbf{c}}_{\mathbf{z}}^T[i] ]^T. \quad (2.11)$$

The  $\Delta$  coefficients for a certain frame are the differences between the corresponding coefficients in the neighboring frames; i.e.  $\Delta\hat{\mathbf{c}}_{\mathbf{z}}[i] = \hat{\mathbf{c}}_{\mathbf{z}}[i + 1] - \hat{\mathbf{c}}_{\mathbf{z}}[i - 1]$ . The  $\Delta\Delta$  coefficients are obtained in a similar manner, where the subtraction process is done on the  $\Delta$  coefficients:  $\Delta\Delta\hat{\mathbf{c}}_{\mathbf{z}}[i] = \Delta\hat{\mathbf{c}}_{\mathbf{z}}[i + 1] - \Delta\hat{\mathbf{c}}_{\mathbf{z}}[i - 1]$ . In [15], the authors additionally implemented cepstral mean subtraction for improved robustness.

## 2.4.2 System Training

The system training stage attempts to establish a library of probability models for clean speech. Given an undistorted signal  $s[n]$ , signal segmentation and VAD are first applied to retain only the non-silent frames of the speech signal. The feature extraction mechanism described in Section 2.4.1 is then applied, where we obtain the MFCC features for each speech segment. Each speech segment  $\mathbf{s}[i]$  and its corresponding feature vector  $\mathbf{c}_s[i]$  can then be assigned to a unique *waveform* set  $\mathbb{S}_q$  and *feature* set  $\mathbb{C}_q$  containing frames with similar phonetic characteristics. The categorization of segments into clusters in this work was done using available transcription files in our experimental database (the CMU\_ARCTIC database). If a phonetic transcription is not available one may also use the unsupervised clustering method as in [18]. Once the phonetic groups have been established, we can train a *Gaussian Mixture Model* (GMM) for each group. For example, in this work there were 40 phonetic clusters and a GMM with 3 mixtures and diagonal covariance was trained on each  $\mathbb{C}_q$  to yield 40 PDF models

$$\mathcal{C}_q(\hat{\mathbf{c}}_s[i]) = \sum_{k=1}^3 \alpha_{\mathbf{c}_{sk},q} \cdot \mathcal{N}_{\mathbb{R}^{39}}(\hat{\mathbf{c}}_s[i]; \mu_{\mathbf{c}_{sk},q}, \Sigma_{\mathbf{c}_{sk},q}) \quad (2.12)$$

for  $q = 1, 2, \dots, 40$  with the weights  $\alpha_{\mathbf{c}_{sk}}$ , the mean vectors  $\mu_{\mathbf{c}_{sk},q}$  and the covariance matrices  $\Sigma_{\mathbf{c}_{sk},q}$  of mixtures  $k$  in cluster  $q$ . In [15], the authors further classified each of the 40 clusters into 3 subclusters to account for coarticulation effects. In that case, the inventory would consist of 120 Gaussian PDF models. GMMs can be computed numerically by the *Expectation-Maximization* (EM) algorithm [53, 54, 55], which is the approach we followed.

We can also establish a state transition model by tallying observed transitory patterns. Specifically, each time state  $q_i$  transitions to state  $q_j$ , we add a count to cell

$[i, j]$  in our 40-by-40 transition matrix  $\mathbf{P}$ . These counts are then normalized by the total count in each row to yield the probabilities of each state transition. The cluster PDFs  $\mathcal{C}_q(\hat{\mathbf{c}}_s)$  for  $q = 1, 2, \dots, 40$  and the matrix  $\mathbf{P}$  constitute the *Hidden Markov Model* (HMM) that will be used in the temporal development estimation of the enhancement process [15].

### 2.4.3 Inventory Search and Enhancement

Once the inventory is constructed, noisy test signals  $x[n]$  can be processed by the enhancement system. All frames are passed through the log-MMSE branch, yielding streams  $\mathbf{y}^F[i]$ ; only frames  $\mathbf{x}[i]$  flagged as non-silent are subjected to the feature extraction process as described in 2.4.1, resulting in MFCC features  $\mathbf{c}_{\text{cx}}[i]$ . These parameters can then be used to compute likelihood values  $\lambda_q[i]$  for each phonetic class  $q$  via Equation 2.12. We jointly optimize the probabilities of these likelihood values and those in  $\mathbf{P}$  via a *Viterbi algorithm* [15] to arrive at the most likely frame  $q^*[i]$  for the present voice-active frame  $\mathbf{x}[i]$ .

As a further improvement to this proposed enhancement scheme, Nickel *et al.* in [15] pursued a *local* multipath Viterbi search where not one but *multiple* state candidates were passed into the inventory search. For example, three state estimates  $q_1[i]$ ,  $q_2[i]$  and  $q_3[i]$  are chosen for each frame  $\mathbf{x}[i]$ . The search for the best matching waveform is then performed on the *merged* inventory collections  $\mathbb{S}_{q_{1,2,3}}[i]$ .

To find the optimal waveform within a chosen phonetic cluster, we follow the matched filter approach in [18]. Specifically, this correlation search seeks the optimal segment  $\hat{\mathbf{s}}$  in inventory group  $\mathbb{S}_q$  that maximizes the normalized inner product between

the present frame of interest  $\mathbf{x}[i]$  and the (to be) chosen frame  $\hat{\mathbf{s}}[i]$ , i.e.

$$\hat{\mathbf{s}}[i] = \operatorname{argmax}_{\mathbf{s} \in \mathcal{S}_q} \frac{\mathbf{x}^T[i] \cdot \mathbf{s}}{\|\mathbf{x}[i]\| \cdot \|\mathbf{s}\|} \quad (2.13)$$

The best-fitting inventory frames are then normalized to match the energy of the corresponding replaced frames  $\mathbf{x}[i]$  in the log-MMSE branch, resulting in the inventory-branch estimate  $\mathbf{y}^I[i]$ .

$$\mathbf{y}^I[i] = \frac{\mathbf{x}^T[i] \cdot \hat{\mathbf{s}}[i]}{\|\mathbf{x}[i]\| \cdot \|\hat{\mathbf{s}}[i]\|} \quad (2.14)$$

The VAD indicator then combines streams  $\mathbf{y}^I[i]$  and  $\mathbf{y}^F[i]$  to produce the final estimate  $\mathbf{y}[i]$ :

$$\mathbf{y}[i] = \begin{cases} \mathbf{y}^I[i] & \text{if frame } i \text{ is flagged as "voice active"} \\ \mathbf{y}^F[i] & \text{otherwise} \end{cases} \quad (2.15)$$

Finally, cepstral smoothing as proposed by [16] and [56] can be applied before signal resynthesis. The cepstral smoothing method was reported to reduce musical artifacts due to pitch and phase mismatches at frame boundaries [15]. Both objective measures and subjective listening tests in [15] support this claim for various noise levels and noise types. The inventory-based enhancement method performs better than filtering methods especially in non-stationary noise cases such as Babble noise.

# Chapter 3

## Voice Conversion

### 3.1 Voice Conversion Overview

*Voice conversion (VC)* attempts to make an utterance spoken by a source speaker sound as if it was spoken by a target speaker. Analogous to speech enhancement, VC systems typically consist of three main components: source signal *analysis*, signal *conversion*, and target signal *synthesis*. As in speech enhancement, a source signal  $x[n]$  is also segmented into overlapping frames (typically 20-ms segments with 50% overlap). These frames are then parametrized as appropriate source features  $\mathbf{f}_x[i]$ , unlike in typical speech enhancement systems where FFTs are commonly used. The reason for this, as will be clear later, is that these source features are to be mapped to target speaker features  $\hat{\mathbf{f}}_y[i]$  via a pre-trained model, which cannot be easily obtained when the feature vectors are of too high a dimension. At the final step, the target features are converted back to the time domain for synthesis of target speech  $y[n]$ .

The major difference between voice conversion and speech enhancement lies in the *voice conversion model*. As reviewed in Section 1.2, many schemes to arrive at this model are under study, among which the most commonly used approach is the

*Gaussian Mixture Model*. Our work focuses on this approach.

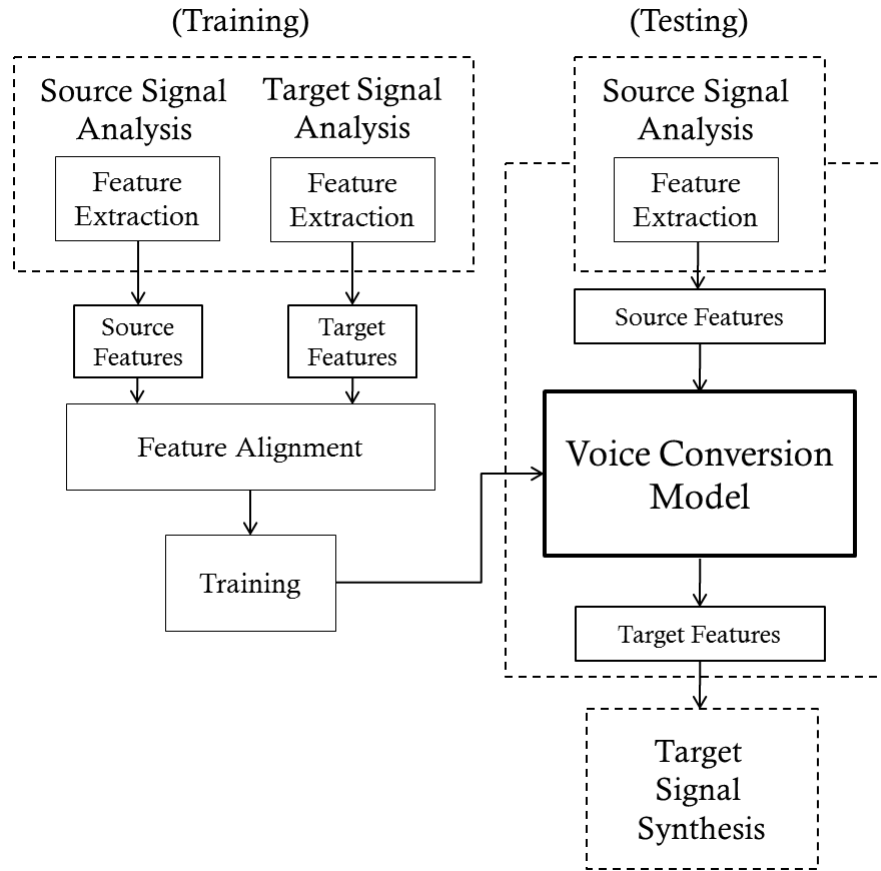


Figure 3.1: Block Diagram of a Conventional Voice Transformation System

## 3.2 The Standard Model

A block diagram for a standard voice conversion system is shown in Figure 3.1. In the training stage, we need access to sample speech sounds from both the *source* and the *target* speakers. In addition, preferably the utterances by both speakers have matching textual content, since the basic idea is to parametrize source and target speeches and then estimate a *joint* Gaussian Mixture Model of source and target feature probabilities. A continuous probabilistic transformation function can then be

implemented using the trained joint GMM [21].

In the testing/conversion stage, we perform analysis on the source speech, extracting the new feature vectors to be mapped to target feature vectors. The details involved in *feature extraction*, *conversion model construction*, and *target speech synthesis* are presented in the following sections.

### 3.2.1 Feature Extraction

Similar to the inventory-based approach for speech enhancement, feature extraction is necessary to describe input speech signals in terms of efficient parameters. However, in voice transformation, features other than MFCCs are often used. Examples of such features are the *linear prediction coefficients* (LPCs), their close relatives *Line Spectral Frequencies* (LSFs), and more recently TANDEM-STRAIGHT related features developed by Kawahara *et al.* [23]. In this work, we used a combination of such features; a brief description of each feature is provided below.

The LPCs, in essence, are the coefficients of the Wiener filter that attempts to predict the present input value based on  $p$  past input values, which also determines the order of the filter. In view of the speech production source-filter model, the LPCs model the filter that represents effects of the vocal tract [51, 57]. LPCs can be computed numerically from an incoming speech signal using the autocorrelation method and the Levinson-Durbin algorithm [58].

LSFs are an alternative representation of LPCs, found by solving for the roots of the polynomials  $P(z)$  and  $Q(z)$  given by:

$$\begin{aligned} P(z) &= A(z) - z^{-(p+1)}A(z^{-1}) \\ Q(z) &= A(z) + z^{-(p+1)}A(z^{-1}) \end{aligned}$$

where  $A(z) = 1 + \sum_{k=1}^p a_k z^{-k}$  with LPC values  $a_k$  and  $p$  is the order of the LPC filter. Compared to LPCs, LSFs are reported to possess superior interpolation properties and quantization robustness in low bit rate speech coding applications [57, 59].

STRAIGHT (Speech Transformation and Representation using Adaptive Interpolation of weiGHTed spectrum) is a speech analysis, modification and synthesis system, first developed by Kawahara [60] and later refined by Kawahara *et al.* [23]. The most current version of the system, TANDEM-STRAIGHT is a toolbox that allows the extraction of numerous speech features, among which we focused on the so-called STRAIGHT spectrum, pitch  $F_0$ , and aperiodicity values. The most notable difference between STRAIGHT and conventional speech analysis systems lies in the aperiodicity estimation, which takes into account the fact that speech sounds are not strictly periodic due to movements of articulators and fluctuations of the excitation source [23]. STRAIGHT is widely used in the voice conversion research community and is reported to produce high-quality resynthesized speech [24]. For this reason, we incorporated a subset of features extracted by this toolbox in our work.

Efficient feature extraction is important in training voice conversion models. Preferably, we would like relevant features to also have relatively low dimensions. The reason for this is that higher dimensional feature vectors require more data and lead to numerical problems during training [54]. The STRAIGHT spectrum extracted by the toolbox is not appropriate for direct parameterizations and training, since it has comparable dimensionality to the FFTs and therefore are difficult to train on. Consequently, we use only the pitch  $F_0$  and aperiodicity information extracted by STRAIGHT for our training data.

The overall feature extraction process for constructing a voice conversion model is outlined as follows. First, we compute LPCs of order  $p$  and then convert them to LSF features  $\mathbf{f}_{\text{LSF}}[i]$ . We also save  $Q[i]$ , the residual mean square error of the



$p^{\text{th}}$  coefficient from the LPC computation process.  $Q[i]$  is necessary to correctly reconstruct the energy of the LPC spectrum from corresponding LSFs, which will be used in the synthesis process. We then used the STRAIGHT toolbox to get the pitch estimate  $F_0$  and 2 aperiodicity values  $\mathbf{ap}^T[i]$ . These parameters are then augmented into our overall *LSF-STRAIGHT* feature vector, which is  $(p + 4)$ -dimensional:

$$\mathbf{f}_z[i] = [ \mathbf{f}_{\text{LSF}}^T[i] \ F_0[i] \ \mathbf{ap}^T[i] \ Q[i] ]^T \quad (3.1)$$

In our work,  $p = 16$ , making the feature vectors 20-dimensional. Were we to use FFTs or STRAIGHT spectrum frames, each feature vector would have been 257-dimensional, which makes training not feasible.

### 3.2.2 Conversion Model Construction

A *Voice Conversion Model* attempts to predict output features  $\hat{\mathbf{f}}_y[i]$  from input features  $\mathbf{f}_x[i]$  via a joint PDF of trained input, output, and a corresponding regression function [21]. According to Kain *et al.* [61], modeling the joint density rather than source density alone should lead to more judicious allocation of mixtures for the transformation function estimation, though obviously the cost of computation is higher.

After training features of the source  $\mathbf{f}_{\text{sx}}[i]$  and target  $\mathbf{f}_{\text{sy}}[i]$ , each with dimension  $L$ , have been computed, we shift and normalize each feature vector such that each vector has zero mean and unit variance to avoid numerical problems during GMM training. In particular, the *normalized* feature vectors  $\tilde{\mathbf{f}}_{\text{sx/y}}[i]$  are calculated as:

$$\tilde{\mathbf{f}}_{\text{sx/y}}[i] = \alpha_{\text{sx/y}} \circ \mathbf{f}_{\text{sx/y}}[i] - \bar{\mathbf{f}}_{\text{sx/y}} \quad (3.2)$$

where  $\circ$  denotes element-wise multiplication with

$$\boldsymbol{\alpha}_{\text{sx}/\text{y}} = \sqrt{\frac{1}{\text{Var}\{\mathbf{f}_{\text{sx}/\text{y}}\}}} \quad (3.3)$$

and

$$\bar{\mathbf{f}}_{\text{sx}/\text{y}} = \boldsymbol{\alpha}_{\text{sx}/\text{y}} \circ \text{E}\{\mathbf{f}_{\text{sx}/\text{y}}\} \quad (3.4)$$

The normalization parameters  $\boldsymbol{\alpha}_{\text{sx}/\text{y}}$ ,  $\bar{\mathbf{f}}_{\text{sx}/\text{y}}$  are stored so that one can easily revert back to nominal values from normalized features and vice versa.

Next, we align source and target features linearly across matching phonemes, such that each frame of the source utterance has a corresponding frame from the target utterance. Here we used the available transcription in our database. These feature vectors can then be stacked in joint  $2L$ -dimensional vectors  $\mathbf{f}_{\text{xy}}[i] = \begin{bmatrix} \tilde{\mathbf{f}}_{\text{sx}}^T[i] & \tilde{\mathbf{f}}_{\text{sy}}^T[i] \end{bmatrix}^T$ . Using these aligned and normalized joint features, we can train a GMM with  $K$  mixtures and full/diagonal/scalar covariance matrices to obtain a joint density function:

$$\mathcal{F}(\hat{\mathbf{f}}_{\text{xy}}[i]) = \sum_{k=1}^K \alpha_{\text{xy}k} \cdot \mathcal{N}_{\mathbb{R}^{2L}}(\hat{\mathbf{f}}_{\text{xy}}[i]; \mu_{\text{xy}k}, \boldsymbol{\Sigma}_{\text{xy}k}) \quad (3.5)$$

In this work,  $L = 20$ ,  $K = 12$  and full covariance matrices were used.

The feature conversion function was derived following the continuous probabilistic transform method in [62]. Given source feature  $\mathbf{f}_{\text{x}}[i]$  and the trained GMM for joint features  $\mathbf{f}_{\text{xy}}[i]$  in Equation 3.5, the corresponding target feature can be estimated by evaluating the conditional expectation; i.e.  $\hat{\mathbf{f}}_{\text{y}}[i] = \text{E}\{\mathbf{f}_{\text{y}}[i] \mid \mathbf{f}_{\text{x}} = \mathbf{f}_{\text{x}}[i]\}$ . The transformation function is a weighted sum of linear models, where the weights depend on the

probability of the input being in a particular class [21]:

$$\hat{\mathbf{f}}_{\mathbf{y}}[i] = \mathcal{T}(\mathbf{f}_{\mathbf{x}}[i]) = \sum_{k=1}^K (W_k \cdot \mathbf{f}_{\mathbf{x}}[i] + b_k) \cdot p(k|\mathbf{f}_{\mathbf{x}}[i]) \quad (3.6)$$

where

$$p(k|\mathbf{f}_{\mathbf{x}}[i]) = \frac{\alpha_{\mathbf{xy}k} \cdot \mathcal{N}_{\mathbb{R}^L}(\mathbf{f}_{\mathbf{x}}[i]; \mu_k^X, \Sigma_k^{XX})}{\sum_{n=1}^K \alpha_{\mathbf{xy}n} \cdot \mathcal{N}_{\mathbb{R}^L}(\mathbf{f}_{\mathbf{x}}[i]; \mu_n^X, \Sigma_n^{XX})} \quad (3.7)$$

$$W_k = \Sigma_k^{YX} (\Sigma_k^{XX})^{-1} \quad (3.8)$$

$$b_k = \mu_k^Y - \Sigma_k^{YX} (\Sigma_k^{XX})^{-1} \mu_k^X \quad (3.9)$$

$$\Sigma_{\mathbf{xy}k} = \begin{bmatrix} \Sigma_k^{XX} & \Sigma_k^{XY} \\ \Sigma_k^{YX} & \Sigma_k^{YY} \end{bmatrix} \quad (3.10)$$

and

$$\mu_{\mathbf{xy}k} = \begin{bmatrix} \mu_k^X \\ \mu_k^Y \end{bmatrix} \quad (3.11)$$

The transformation function in Equation 3.6 assumes (1) Gaussian distribution of source features, and (2) source and target features are jointly Gaussian.

### 3.2.3 Target Speech Synthesis

Target speech is synthesized from the transformed feature vectors  $\hat{\mathbf{f}}_{\mathbf{y}}[i]$ . Depending on the type of features used, the synthesis process essentially decodes the transformed parameters back to a speech signal. In our work, the STRAIGHT toolbox allowed synthesis from the LSF-STRAIGHT features described in Section 3.2.2. To be con-

sistent with the input syntax of the toolbox, we first use the toolbox to create a STRAIGHT object that stores source signal parameters necessary for synthesis. After feature conversion, we replace relevant elements of source features  $\mathbf{f}_x[i]$  in the STRAIGHT object with corresponding converted elements of  $\hat{\mathbf{f}}_y[i]$ , i.e. the converted pitch estimate  $\hat{F}_0[i]$  and aperiodicity parameters  $\widehat{\mathbf{ap}}[i]$ . Target LPC features are computed from the converted  $\hat{\mathbf{f}}_{\mathbf{LSF}}[i]$  and a new LPC spectrum can be obtained from these new LPCs and the converted residual error  $\hat{Q}[i]$ . We replace the STRAIGHT spectrum in the current STRAIGHT object by this LPC spectrum.

# Chapter 4

## Noise-Robust Voice Conversion

While the majority of voice conversion systems rely on a clean source input, such ideal conditions might not be realistic. To the best of my knowledge, limited research has been conducted on voice conversion in noisy environments [39]. Our research investigated a noisy voice conversion approach with inventory-based analysis of the source signal.

An immediate solution to the problem of voice conversion in noisy conditions is to concatenate a speech enhancement subsystem with a standard voice conversion system. In this study such a system was our reference for comparative experiments. Alternatively, we propose a method that follows the inventory-style processing algorithm to indirectly estimate the underlying features of the source speech. Both the reference system and our proposed system are described in the following sections.

### 4.1 System Description

We assume to have access to an inventory of undistorted speech signals from our source and our target speaker with matching content. Inventory utterances are properly

energy equalized via ITU-T P.56 [63] and concatenated into one long signal stream  $s[n]$  for each speaker. The observed noisy utterance by the source speaker  $x[n] = z[n] + v[n]$  is the input to our voice conversion system, where  $z[n]$  denotes the underlying clean signal and  $v[n]$  denotes additive noise. Our inventory signals are band-limited between 50 Hz and 4 kHz and sampled at 8 kHz. We lowered our upper band cut-off frequency (and correspondingly reduced the sampling rate) to reduce the computational cost involved in training joint source-target features (See Section 4.1.2).

All subsystems in our proposed procedure employ signal segmentations of 20-ms Hamming-windowed frames with 50% overlap. We use  $\mathbf{x}[i]$  to denote the  $i^{\text{th}}$  frame from segmentation, i.e.

$$\mathbf{x}[i] = [x[80 \cdot i] \ x[80 \cdot i + 1] \ \dots \ x[80 \cdot i + 159]]^T. \quad (4.1)$$

Frame symbols for other signals such as  $\hat{\mathbf{z}}[i]$  are defined analogously.

A conceptual block diagram for the proposed conversion system and the employed reference system is shown in Figure 4.1. First, the noisy source-speaker signal is pre-processed by a standard Log-MMSE filter with a decision-directed approach after Ephraim and Malah [7], resulting in signal frames  $\hat{\mathbf{z}}[i]$ . This filtered signal is then processed by two separate voice conversion systems: the *direct conversion* branch, which serves as our reference system, and the *inventory conversion* branch, which represents the proposed procedure. Target speech signals are synthesized in both branches with the TANDEM-STRAIGHT toolbox [24].

In the *direct conversion* branch, LSF-STRAIGHT features are extracted from the pre-processed frames  $\hat{\mathbf{z}}[i]$  as described in Section 3.2.1. We denote the source-speaker feature vectors as  $\mathbf{f}_{\mathbf{z}}[i]$ . These features are then converted to target-speaker features  $\hat{\mathbf{f}}_{\mathbf{DI}}[i]$  through a pre-trained Gaussian Mixture Model (GMM) as described

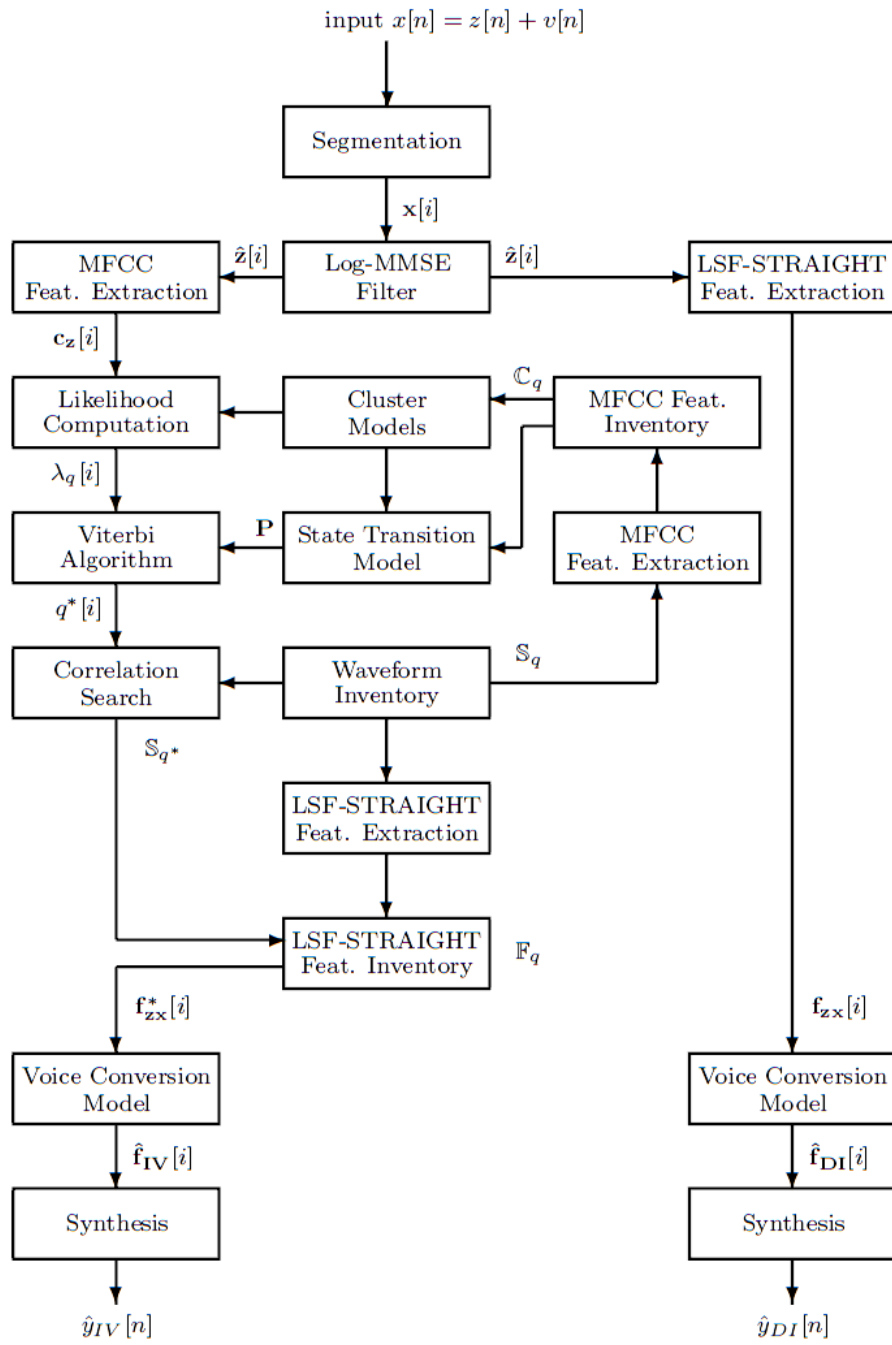


Figure 4.1: Block Diagram of the Noise-Robust Voice Conversion Procedures

in Section 3.2.2. We use this directly converted signal  $\hat{y}_{DI}[n]$  as the output of our reference system for a comparative performance analysis.

In the *inventory conversion* branch, instead of directly computing LSF-STRAIGHT features from  $\hat{\mathbf{z}}[i]$ , we employ the inventory search procedure described in Section 4.1.3 to find the optimal feature frames available in our inventory; these features will be the input to the pre-trained GMM. The converted features, denoted by  $\hat{\mathbf{f}}_{IV}[i]$ , are used to synthesize the proposed target-speaker speech signal  $\hat{y}_{IV}[n]$ .

#### 4.1.1 Feature Extraction

In our proposed procedure, two sets of features need to be extracted: Mel-Frequency Cepstral Coefficients (MFCCs) for the inventory search procedure and LSF-STRAIGHT features for the conversion procedure. We need both types of features because MFCCs are more robust to determine the phonetic cluster membership of a frame under noisy conditions, whereas the LSF-STRAIGHT features of clean frames allow us to re-synthesize high quality speech via the TANDEM-STRAIGHT toolbox.

The MFCC feature extraction process follows the procedure described in Section 2.4.1: we augment 13-dimensional MFCC vectors  $\hat{\mathbf{c}}_{\mathbf{z}}[i]$  with the usual  $\Delta$  and  $\Delta\Delta$  coefficients after [15] to arrive at 39-dimensional feature vectors:

$$\mathbf{c}_{\mathbf{z}}[i] = [ \hat{\mathbf{c}}_{\mathbf{z}}^T[i] \quad \Delta\hat{\mathbf{c}}_{\mathbf{z}}^T[i] \quad \Delta\Delta\hat{\mathbf{c}}_{\mathbf{z}}^T[i] ]^T. \quad (4.2)$$

The LSF-STRAIGHT feature extraction process follows the procedure in Section 3.2.1: using the STRAIGHT toolbox [24], we compute the pitch estimate  $F_0[i]$  and 2 aperiodicity parameters  $\mathbf{ap}[i]$  for each frame. For LSF features, we first compute the LPCs of order 16 and then convert them to LSF features  $\mathbf{f}_{\mathbf{LSF}}[i]$ , saving the residual mean square error  $Q[i]$  of the 16<sup>th</sup> coefficient from the LPC computation process. The



resulting LSF-STRAIGHT feature vectors are of dimension 20:

$$\mathbf{f}_z[i] = [ \mathbf{f}_{\text{LSF}}^T[i] \ F_0[i] \ \mathbf{ap}^T[i] \ Q[i] ]^T \quad (4.3)$$

As described in Section 3.2.2, we use the LPC spectrum to synthesize output speech signal in lieu of the STRAIGHT spectrum because the former allows for a significantly lower-dimensional parametrization for feature training.

### 4.1.2 Inventory Design and System Training

Corresponding to the two sets of features described in Section 4.1.1, two training procedures are necessary: one for the inventory search subsystem with MFCCs and another for the conversion subsystem with the LSF-STRAIGHT features.

We design our *waveform inventory* by dividing clean source utterances  $s[n]$  into collections  $\mathbb{S}_q$  of phonetically similar segments. After all silent parts of  $s[n]$  are removed, the remaining segments of  $s[n]$  are categorized into one of 40 phonetic classes ( $q = 1, 2, \dots, 40$ ). In addition to the *waveform inventory*  $\mathbb{S}_q$ , we construct our *feature inventory*  $\mathbb{C}_q$  of MFCC features belonging to the same phonetic class. The training process for the waveform and MFCC inventories follows the simplified method described Section 2.4.2: A GMM with 3 mixtures and diagonal covariance is trained on each  $\mathbb{C}_q$  to yield 40 PDF models

$$\mathcal{C}_q(\hat{\mathbf{c}}_s[i]) = \sum_{k=1}^3 \alpha_{\text{csk}} \cdot \mathcal{N}_{\mathbb{R}^{39}}(\hat{\mathbf{c}}_s[i]; \mu_{\text{csk},q}, \Sigma_{\text{csk},q}) \quad (4.4)$$

for  $q = 1, 2, \dots, 40$  with the weights  $\alpha_{\text{csk}}$ , the mean vectors  $\mu_{\text{csk},q}$  and the covariance matrices  $\Sigma_{\text{csk},q}$  of mixtures  $k$  in cluster  $q$ . In addition, we construct  $\mathbf{P}$ , the state transition probability matrix by counting the number of observed transitions from

a particular state and dividing it by the total number of occurrences of each state transition.

Analogously, we construct a *feature inventory*  $\mathbb{F}_q$ ,  $q = 1, 2, \dots, 40$ , for LSF-STRAIGHT features that are phonetically similar based on the database transcriptions. At this point, we train a different GMM for the voice conversion process following Section 3.2.2: we compute all training LSF-STRAIGHT features  $\mathbf{f}_{\text{sx}}[i]$  and  $\mathbf{f}_{\text{sy}}[i]$  for the source and target speakers, respectively. We also shift and normalize each feature vector such that each vector has zero mean and unit variance as in Equations 3.2 through 3.4. We align source and target features linearly across matching phonemes, such that each frame of the source utterance has a corresponding frame from the target utterance. These feature vectors can then be stacked in joint 40-dimensional vectors  $\mathbf{f}_{\text{xy}}[i] = \left[ \tilde{\mathbf{f}}_{\text{sx}}^T[i] \ \tilde{\mathbf{f}}_{\text{sy}}^T[i] \right]^T$ . We train a GMM with 12 mixtures and full covariance matrices on these joint features  $\mathbf{f}_{\text{xy}}[i]$  and obtain a joint density function:

$$\mathcal{F}(\hat{\mathbf{f}}_{\text{xy}}[i]) = \sum_{k=1}^{12} \alpha_{\text{xy}k} \cdot \mathcal{N}_{\mathbb{R}^{40}}(\hat{\mathbf{f}}_{\text{xy}}[i]; \mu_{\text{xy}k}, \Sigma_{\text{xy}k}) \quad (4.5)$$

### 4.1.3 Inventory Based Voice Conversion

For voice conversion, our system consists of the *inventory search* procedure and the *feature conversion and synthesis* procedure.

#### Inventory Search

The inventory search procedure follows the search procedure in Section 2.4.3. The Log-MMSE enhanced (pre-processed) signal frames  $\hat{\mathbf{z}}[i]$  are subjected to an extraction of MFCC features  $\mathbf{c}_{\mathbf{z}}[i]$ . The likelihood value for each class  $q$  is calculated via

$\lambda_q[i] = \mathcal{C}_q(\mathbf{c}_z[i])$  as in equation 4.4. With these likelihood estimates and the established state transition matrix  $\mathbf{P}$ , we find the most likely phonetic class  $q^*[i]$  associated with the current frame via the Viterbi algorithm. Within the selected cluster inventory  $\mathbb{S}_{q^*}$ , we search for the best waveform representation for  $\hat{\mathbf{z}}[i]$  with the normalized matched filter approach described in [15]. Knowing the location of this best waveform in our inventory  $\mathbb{S}_{q^*}$ , we can find the corresponding best matching LSF-STRAIGHT feature  $\mathbf{f}_{\mathbf{zx}}^*[i]$  in our feature inventory  $\mathbb{F}_{q^*}$ .

### Feature Conversion and Synthesis

The feature conversion procedure follows the process described in Section 3.2.2: we follow the continuous probabilistic transform method in [62]. Given source feature  $\mathbf{f}_{\mathbf{zx}}[i]$  and the trained GMM for joint features  $\mathbf{f}_{\mathbf{xy}}[i]$ , the target feature can be estimated by the transformation function  $\hat{\mathbf{f}}_{\mathbf{zy}}[i] = \mathcal{T}(\mathbf{f}_{\mathbf{zx}}[i])$  as in Equations 3.6 through 3.11.

Note that in the *inventory conversion* branch we use the feature mapping  $\hat{\mathbf{f}}_{\mathbf{IV}}[i] = \mathcal{T}(\mathbf{f}_{\mathbf{zx}}^*[i])$  while in the *direct conversion* branch we use the feature mapping  $\hat{\mathbf{f}}_{\mathbf{DI}}[i] = \mathcal{T}(\mathbf{f}_{\mathbf{zx}}[i])$ .

Target speech is then synthesized from the respective transformed feature vectors via the TANDEM-STRAIGHT toolbox, as described in Section 3.2.3. We first use the toolbox to extract STRAIGHT features and create a STRAIGHT object that stores source signal parameters necessary for synthesis. After inventory search and feature conversion, we replace relevant elements of source features  $\mathbf{f}_{\mathbf{zx}}[i]$  in the STRAIGHT object with corresponding converted elements of  $\hat{\mathbf{f}}_{\mathbf{zy}}[i]$ , i.e. the converted pitch estimate  $\hat{F}_0[i]$  and aperiodicity parameters  $\widehat{\mathbf{ap}}[i]$ . Target LPC features are computed from the converted  $\hat{\mathbf{f}}_{\mathbf{LSF}}[i]$  and a new LPC spectrum can be obtained from these new LPCs and the converted residual error  $\hat{Q}[i]$ . We replace the STRAIGHT spectrum in the current STRAIGHT object by this LPC spectrum.

## 4.2 Experiment Description and Results

We analyzed the performance of the proposed method with voice recordings from the CMU\_ARCTIC database, mentioned in Section 2.3.1. Two of the 7 speakers were used to study the performance and feasibility of the proposed voice conversion scheme. As our source speaker we employed the *US English* male speaker with the identifier BDL and as a target speaker we chose the *US English* female speaker with the identifier SLT. The two speaker sets contain a minimum of 1132 phonetically balanced English utterances each. Most utterances are between one and four seconds long. The content of the recorded sentences was specifically designed to cover a large variety of articulatory gestures for each speaker. Full phonetic transcriptions of all utterances with (roughly) 40 elementary phonetic units per speaker are available. The datasets of the two chosen speakers (BDL and SLT) were divided into two strictly disjoint sets: a *training* set of 1000 utterances, which constituted our respective speech inventory, and a *testing* set which encompassed all remaining utterances. All data was appropriately resampled to a 4kHz bandwidth with an 8kHz sampling rate. Additive white noise and jet cockpit noise was taken from the NOISEX database from the Institute for Perception-TNO, The Netherlands Speech Research Unit, RSRE, UK [64]. The noise was added to the source speaker *testing* data at a signal-to-noise ratio (SNR) of 10 dB under consideration of the respective active speech level after ITU recommendation ITU-T P.56 [63]. System training was performed with the procedures outlined in Section 4.1.2.

To evaluate the performance of the proposed procedure *perceptually* we designed a *Comparison Category Rating* (CCR) test after ITU-T recommendation P.800 [2] with 21 non-expert human listeners. As a reference we employed the Log-MMSE enhanced LSF-STRAIGHT method that is termed *direct conversion* in Section 4.1.

The subjects were presented with 20 one-to-one comparisons between the proposed method and the competing direct method for each of the two considered noise types (white noise and jet cockpit noise). Subjects were asked to rate the overall quality of the proposed inventory based method in comparison to the reference method on a seven point scale. Ratings could be entered as +3 (Much Better), +2 (Better), +1 (Slightly Better), 0 (About the Same), -1 (Slightly Worse), -2 (Worse), and -3 (Much Worse). No particular instruction was given to the subjects on *how* to judge speech “quality”. Prior to the experiment subjects were merely exposed to *Modulated Noise Reference Unit* (MNRU) example signals after ITU standard P.810 [65] as an aid to explain the test procedure.

Histograms of the response counts across all subjects and all test utterances for each of the two considered noise types are shown in Figure 4.2 and Figure 4.3. The majority of subjects rated the inventory based method as *Slightly Better* in each case. Approximately 21 % of the subjects assigned a *Better* and *Much Better* score for the white noise case, whereas approximately 30 % of the subjects assigned a *Better* and *Much Better* score for the jet cockpit noise case. For white noise the average score across all responses was 0.77 with a standard deviation of 0.98 and for jet cockpit noise the average was 1.00 with a standard deviation of 1.02. Most subjects had a strong preference for the output of the proposed method over the reference method.

Lastly, we also conducted an ABX test, in which A and B were utterances by either the source or the target speaker. The content of the utterances in X were the same as those in A and B. Subjects were presented with utterances A, B, and X and then asked to rate the converted utterance X on a scale of 1 to 5. A 1 indicated more similarity to the source speaker and a 5 indicated more similarity to the target speaker. For the white noise case, the average score of all subject ratings was 4.05 with a standard deviation of 0.77. For the jet cockpit noise case, the average score was

3.84 with a standard deviation of 0.77. Across both noise cases, the average score was 3.95 with a standard deviation of 0.77. All subjects reported that converted speech was clearly identifiable as the target speaker's voice, but some gave lower scores due to imperfections in prosodic characterization.

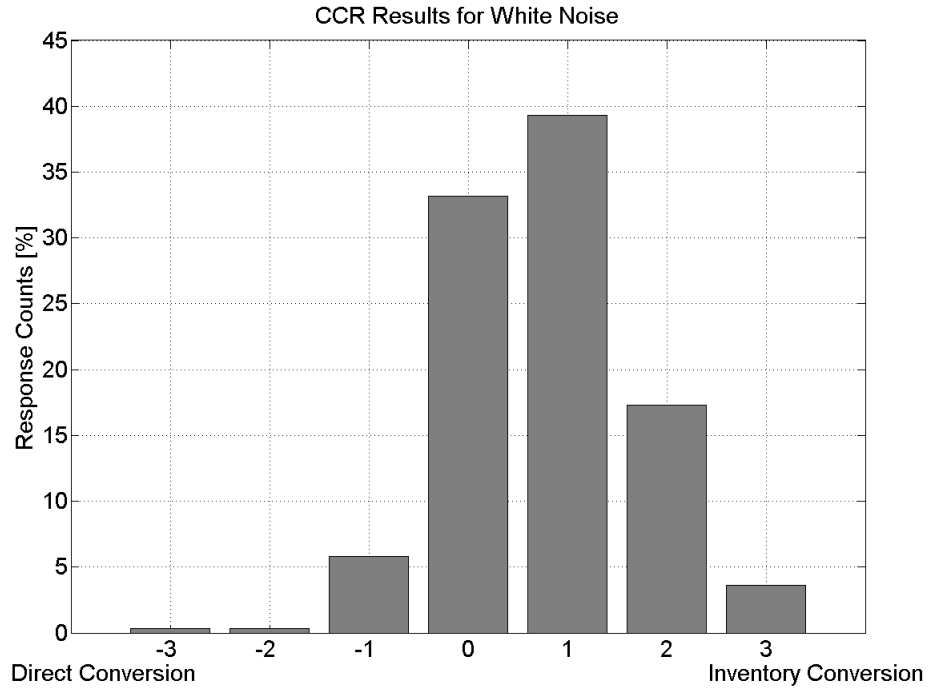


Figure 4.2: Response counts of a *Comparison Category Rating* test with white noise after ITU-T recommendation P.800 [2].

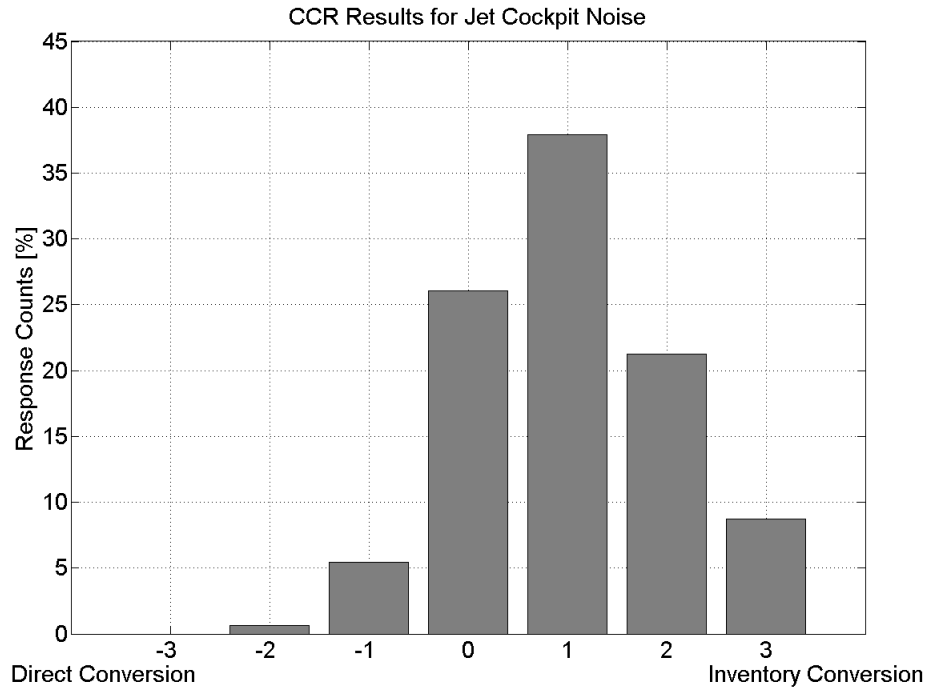


Figure 4.3: Response counts of a *Comparison Category Rating* test with jet cockpit noise after ITU-T recommendation P.800 [2].

## 4.3 Future Work

As mentioned in Section 4.2, our proposed system yielded encouraging but unideal results. Since we are dealing with both speech enhancement and voice conversion, possible improvements exist in both aspects.

### 4.3.1 Speech Enhancement

For a long time, speech enhancement researchers have focused on optimizing only the magnitude spectrum estimation. In fact, in the *MMSE sense*, the noisy phase is the optimal estimator and does not affect amplitude estimation [42, Ch. 7]. However, we have seen that optimization in the mathematical sense does not guarantee optimal

perceived speech quality. Therefore, the role of *phase estimation* in speech enhancement has been potentially underestimated [66]. In [67], Gerkmann and Krawczyk argue that clean speech phase provides additional information that can be exploited for an improved amplitude estimation. They derive a MMSE optimal estimator for the clean speech spectral amplitude when the spectral phase is given. This estimator was shown to potentially improve the PESQ measure by 0.5 in babble noise as compared to state-of-the-art amplitude estimators [67].

In addition to phase estimation, we are also considering slight modifications to our inventory search procedure. In particular, as described in Section 2.4.3, the inventory search at the moment seeks the clean segment with optimal correlation values to the noisy segment of interest. This approach has potential boundary issues when concatenating segments not originally adjacent to each other. To take into account boundary effects, we can introduce *concatenation cost* to our inventory search procedure, defined as the correlation value between the overlapping samples from frames to be concatenated. The optimal segment will be chosen based on both the optimal correlation value and concatenation cost.

Most recent research in speech enhancement as described in Section 1.2 include a combination of Wiener filtering and dictionary learning by Tseng *et al.* [19] and deep neural networks (DNN) by Xu *et al.* [20]. The authors claimed to successfully reduce musical artifacts [20] and intelligibility issues [19]. These improvements can be incorporated in our system depending on how much restructuring these modifications require.



### 4.3.2 Voice Conversion

On the voice conversion side, a promising modification to our conversion model is the *Bilinear Frequency Warping Plus Amplitude Scaling* (BLFW+AS) approach by Erro *et al.* [38]. The main difference of this approach lies in the conversion model function. In particular, Erro *et al.* propose using a bilinear warping matrix in place of the weighing matrix  $W_k \cdot p(k|(\cdot))$  in Equation 3.6 and a similar bias vector  $b_k \cdot p(k|(\cdot))$  with modified estimation during training [38]. The authors reported results in quality of speech synthesized comparable to other state-of-the-art voice conversion methods despite the relative simplicity. Because of the similarity in the system flow of this approach to ours, we are considering implementation of BLFW+AS in our system.

Another possible modification to our system is using a different set of features and a promising candidate are the *pitch bases* of voiced segments LPC residuals. This approach is motivated by Nickel and Oswal’s work on optimal pitch bases expansions [68] in 2003. By signal “residual” we mean the output of the voiced speech signal after the inverse LPC filter. In essence, the pitch bases are obtained by (1) collecting LPC residuals of voiced segments of a certain speaker, (2) time-aligning these residuals by pitch events, and (3) performing principal component analysis (such as singular value decomposition) on the collected pitch waveforms. The singular values resulting from this operation can be used as additional features for our voice conversion system. Since the LPCs/LSFs encode most of the information about the spectral envelope, we hope to use the pitch bases to encode spectral details (or residuals). Informal listening tests of this method gave encouraging results considering the relative simplicity of the approach.

# Chapter 5

## Conclusion

This thesis presented a voice conversion system for operation in noisy environments. In constructing such a system, we studied essential methods in speech enhancement as well as voice conversion. Traditional enhancement methods (spectral subtraction and statistical filtering) were implemented and compared to a new inventory-based method. The inventory-based method outperformed most traditional methods in quality of resynthesized speech.

For voice conversion, a standard Gaussian Mixture Model-based system was studied. This popular approach was integrated with the STRAIGHT toolbox to develop a voice conversion system with noisy input considerations. In particular, source signal parameters were indirectly estimated by searching for the best matching clean feature in an established inventory. Two noisy-environment voice conversion systems were constructed for a comparative study: a direct voice conversion system and an inventory-based voice conversion system, both with limited noise filtering at the front end. Listening tests indicated that the proposed inventory-based conversion system slightly outperformed the direct conversion system. Improvements to our system can be incorporated in both the enhancement and the conversion aspects.

This thesis was one of the first attempts to solve a problem of limited attention so far: voice conversion in noisy environments. Consequently, our work was one of the first to integrate both speech enhancement and voice conversion in one system. Our implementation can therefore be used as a research platform for further work in this area. Considering the complexity of this system, however, the quality of converted signals was less than ideal. Future work may therefore have to focus on improvements and/or alternative approaches. The encouraging results do indicate, nevertheless, that the inventory-based method holds its merit in both speech enhancement and voice conversion.

# Bibliography

- [1] R.M. Nickel. Spectral Analysis. Lecture Slides, 2012.
- [2] ITU-T (International Telecommunication Union). *Methods for Subjective Determination of Transmission Quality*. Recommendation P.800, 1996.
- [3] M. R. Weiss, E. Aschkenasy, and T. W. Parsons. Study and development of the INTEL technique for improving speech intelligibility. Technical report, 1975.
- [4] S. Boll. Suppression of acoustic noise in speech using spectral subtraction. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 27(2):113–120, Apr 1979.
- [5] Sunil Kamath and Philipos Loizou. A multi-band spectral subtraction method for enhancing speech corrupted by colored noise. In *Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on*, volume 4, pages IV–4164–IV–4164, May 2002.
- [6] R. McAulay and M. Malpass. Speech enhancement using a soft-decision noise suppression filter. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 28(2):137–145, Apr 1980.
- [7] Y. Ephraim and D. Malah. Speech enhancement using a minimum mean square

- error log-spectral amplitude estimator. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 33:443–445, April 1985.
- [8] Yi Hu and P.C. Loizou. Subjective comparison of speech enhancement algorithms. In *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*, volume 1, pages I–I, May 2006.
- [9] R. Martin. Noise power spectral density estimation based on optimal smoothing and minimum statistics. *Speech and Audio Processing, IEEE Transactions on*, 9(5):504–512, Jul 2001.
- [10] I. Cohen. Noise spectrum estimation in adverse environments: improved minima controlled recursive averaging. *Speech and Audio Processing, IEEE Transactions on*, 11(5):466–475, Sept 2003.
- [11] Y. Ephraim and H.L. Van Trees. A signal subspace approach for speech enhancement. *Speech and Audio Processing, IEEE Transactions on*, 3(4):251–266, Jul 1995.
- [12] Yi Hu and P.C. Loizou. A generalized subspace approach for enhancing speech corrupted by colored noise. *Speech and Audio Processing, IEEE Transactions on*, 11(4):334–341, July 2003.
- [13] M. Dendrinos, S. Bakamidis, and G. Carayannis. Speech enhancement from noise: A regenerative approach. *Speech Communication*, 10(1):45 – 57, 1991.
- [14] Ji Ming, R. Srinivasan, and D. Crookes. A corpus-based approach to speech enhancement from nonstationary noise. *Audio, Speech, and Language Processing, IEEE Transactions on*, 19(4):822–836, May 2011.

- [15] R. M. Nickel, R. F. Astudillo, D. Kolossa, and R. Martin. Corpus-based speech enhancement with uncertainty modeling and cepstral smoothing. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(5):983–997, 2013.
- [16] C. Breithaupt, T. Gerkmann, and R. Martin. Cepstral Smoothing of Spectral Filter Gains for Speech Enhancement Without Musical Noise. *Signal Processing Letters, IEEE*, 14(12):1036–1039, 2007.
- [17] Huijun Ding, I.Y. Soon, and Chai-Kiat Yeo. Over-attenuated components regeneration for speech enhancement. *Audio, Speech, and Language Processing, IEEE Transactions on*, 18(8):2004–2014, Nov 2010.
- [18] X. Xiao and R.M. Nickel. Speech Enhancement With Inventory Style Speech Resynthesis. *Audio, Speech, and Language Processing, IEEE Transactions on*, 18(6):1243–1257, 2010.
- [19] Hung-Wei Tseng, S. Vishnubhotla, Mingyi Hong, Jinjun Xiao, Zhi-Quan Luo, and Tao Zhang. A novel single channel speech enhancement approach by combining wiener filter and dictionary learning. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pages 8653–8657, May 2013.
- [20] Yong Xu, Jun Du, Li-Rong Dai, and Chin-Hui Lee. An experimental study on speech enhancement based on deep neural networks. *Signal Processing Letters, IEEE*, 21(1):65–68, Jan 2014.
- [21] A. Kain. *High Resolution Voice Transformation*. PhD thesis, OGI School of Science & Engineering, Oregon Health and Science University, 2001.

- [22] Y. Stylianou. Voice Transformation: A survey. In *Proceedings of ICASSP*, pages 3585–3588, 2009.
- [23] H. Kawahara, M. Morise, T. Takahashi, R. Nisimura, T. Irino, and H. Banno. TANDEM-STRAIGHT: A temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, F0, and aperiodicity estimation. In *Proceedings of ICASSP*, pages 3933–3936, 2008.
- [24] H. Kawahara, T. Takahashi, M. Morise, and H. Banno. Development of exploratory research tools based on TANDEM-STRAIGHT. In *Annual Summit and Conference of the Asia-Pacific Signal and Information Processing Association (APSIPA ASC)*, 2009.
- [25] M. Abe, S. Nakamura, K. Shikano, and H. Kuwabara. Voice conversion through vector quantization. In *Proceedings of ICASSP*, pages 655–658, New York, 1988.
- [26] S. Desai, A. W. Black, B. Yegnanarayana, and K. Prahallad. Spectral mapping using artificial neural networks for voice conversion. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(5):954–964, 2010.
- [27] E. Helander, T. Virtanen, J. Nurminen, and M. Gabbouj. Voice conversion using partial least squares regression. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(5):912–921, 2010.
- [28] E. Helander, H. Silen, T. Virtanen, and M. Gabbouj. Voice conversion using dynamic kernel partial least squares regression. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(3):806–817, 2012.
- [29] D. Saito, S. Watanabe, A. Nakamura, and N. Minematsu. Statistical voice con-

- version based on noisy channel model. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(6):1784–1794, 2012.
- [30] J. Nirmal, P. Kachare, S. Patnaik, and M. Zaveri. Cepstrum liftering based voice conversion using RBF and GMM. In *International Conference on Communications and Signal Processing (ICCSP)*, pages 570–575, 2013.
- [31] H. Duxans, A. Bonafonte, A. Kain, and J. Van Santen. Including dynamic and phonetic information in voice conversion systems. In *International Conference on Spoken Language Processing*, 2004.
- [32] T. Nose and T. Kobayashi. Speaker-independent HMM-based voice conversion using adaptive quantization of the fundamental frequency. *Speech Communication*, 53(7):973–985, 2011.
- [33] W. Percybrooks, E. Moore, and C. McMillan. Phoneme independent HMM voice conversion. In *Proceedings of ICASSP*, pages 6925–6929, 2013.
- [34] Y. Ohtani, T. Toda, H. Saruwatari, and K. Shikano. Maximum likelihood voice conversion based on GMM with STRAIGHT mixed excitation. In *International Conference on Spoken Language Processing*, pages 2266–2269, 2006.
- [35] T. Toda, A. W. Black, and K. Tokuda. Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(8):2222–2235, 2007.
- [36] T. Toda, H. Saruwatari, and K. Shikano. Voice conversion algorithm based on Gaussian mixture model with dynamic frequency warping of STRAIGHT spectrum. In *Proceedings of ICASSP*, volume 2, pages 841–844, 2001.



- [37] E. Godoy, O. Rosec, and T. Chonavel. Voice conversion using dynamic frequency warping with amplitude scaling, for parallel or nonparallel corpora. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(4):1313–1323, 2012.
- [38] D. Erro, E. Navas, and I. Hernaez. Parametric voice conversion based on bilinear frequency warping plus amplitude scaling. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(3):556–566, 2013.
- [39] R. Takashima, T. Takiguchi, and Y. Ariki. In *IEEE Spoken Language Technology Workshop (SLT)*.
- [40] T. Dutoit, A. Holzapfel, M. Jottrand, A. Moinet, J. Perez, and Y. Stylianou. Towards a voice conversion system based on frame selection. In *Proceedings of ICASSP*, volume 4, pages IV513–IV516, 2007.
- [41] Z. Shuang, F. Meng, and Y. Qin. Voice conversion by combining frequency warping with unit selection. In *Proceedings of ICASSP*, pages 4661–4664, 2008.
- [42] P.C. Loizou. *Speech Enhancement, Theory and Practice*. CRC Press, 1st edition, 2007.
- [43] 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*, 32(6):1109–1121, 1984.
- [44] O. Cappe. Elimination of the musical noise phenomenon with the Ephraim and Malah noise suppressor. *Speech and Audio Processing, IEEE Transactions on*, 2(2):345–349, 1994.
- [45] Objective Measurement of Active Speech Level. ITU-T (International Telecomm. Union), 1993. Rec. P. 56.

- [46] Sanjit K. K. Mitra. *Digital Signal Processing: A Computer-Based Approach*. McGraw-Hill Higher Education, 3rd edition, 2008.
- [47] Jongseo Sohn, Nam Soo Kim, and Wonyong Sung. A statistical model-based voice activity detection. *Signal Processing Letters, IEEE*, 6(1):1–3, 1999.
- [48] A.W. Rix, J.G. Beerends, M.P. Hollier, and A.P. Hekstra. Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs. In *Acoustics, Speech, and Signal Processing, 2001. Proceedings. (ICASSP '01). 2001 IEEE International Conference on*, volume 2, pages 749–752 vol.2, 2001.
- [49] N. Kitawaki, H. Nagabuchi, and K. Itoh. Objective quality evaluation for low-bit-rate speech coding systems. *Selected Areas in Communications, IEEE Journal on*, 6(2):242–248, 1988.
- [50] J.M. Tribolet, P. Noll, B. McDermott, and R.E. Crochiere. A study of complexity and quality of speech waveform coders. In *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '78.*, volume 3, pages 586–590, 1978.
- [51] R.M. Nickel. Feature - Automatic speech character identification. *Circuits and Systems Magazine, IEEE*, 6(4):10–31, 2006.
- [52] S. Young, G. Evermann, D. Kershaw, G. Moore, J. Odell, D. Ollason, V. Valtchev, and P. Woodland. The HTK book. *Cambridge University Engineering Department*, 3, 2002.
- [53] Jeff Bilmes. A Gentle Tutorial of the EM Algorithm and its Application to

- Parameter Estimation for Gaussian Mixture Mode and Hidden Markov Models. Tutorial, 1998.
- [54] Lawrence Rabiner and Biing-Hwang Juang. *Fundamentals of Speech Recognition*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1993.
- [55] M. D. Srinath and P. K. Rajasekaran. *An Introduction to Statistical Signal Processing with Applications*. New York: John Wiley & Sons, 1979.
- [56] C. Breithaupt, T. Gerkmann, and R. Martin. A novel a priori SNR estimation approach based on selective cepstro-temporal smoothing. In *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, pages 4897–4900, 2008.
- [57] Sadaoki Furui. *Digital Speech Processing, Synthesis, and Recognition*. Marcel Dekker, 2nd edition, 2000.
- [58] S. L. Marple. *Digital Spectral Analysis With Applications*. Prentice Hall, Australia, Sydney, 1987.
- [59] Kuldeep K. Paliwal. Interpolation properties of linear prediction parametric representations. In *EUROSPEECH*. ISCA, 1995.
- [60] H. Kawahara. Speech representation and transformation using adaptive interpolation of weighted spectrum: vocoder revisited. In *Acoustics, Speech, and Signal Processing, 1997. ICASSP-97., 1997 IEEE International Conference on*, volume 2, pages 1303–1306 vol.2, Apr 1997.
- [61] A. Kain and M.W. Macon. Spectral voice conversion for text-to-speech synthesis. In *Acoustics, Speech and Signal Processing, 1998. Proceedings of the 1998 IEEE International Conference on*, volume 1, pages 285–288 vol.1, May 1998.

- [62] Y. Stylianou, O. Cappe, and E. Moulines. Continuous probabilistic transform for voice conversion. *IEEE Transactions on Speech and Audio Processing*, 6(2):131–142, 1998.
- [63] ITU-T (International Telecommunication Union). *Objective Measurement of Active Speech Level*. Recommendation P.56, 1993.
- [64] H. J. M. Steeneken and F. W. M. Geurtsen. Description of the RSG-10 Noise Database. *Technical Report IZF 1988-3, TNO Institute for Perception, Soesterberg, The Netherlands*, 1988.
- [65] ITU-T (International Telecommunication Union). *Modulated Noise Reference Unit (MNRU)*. Recommendation P.810, 1996.
- [66] Kuldip Paliwal, Kamil Wójcicki, and Benjamin Shannon. The importance of phase in speech enhancement. *Speech Commun.*, 53(4):465–494, April 2011.
- [67] T. Gerkmann and M. Krawczyk. MMSE-Optimal Spectral Amplitude Estimation Given the STFT-Phase. *Signal Processing Letters, IEEE*, 20(2):129–132, 2013.
- [68] R.M. Nickel and S. P. Oswal. Optimal pitch bases expansions in speech signal processing. In *Signals, Systems and Computers, 2004. Conference Record of the Thirty-Seventh Asilomar Conference on*, volume 2, pages 1885–1889 Vol.2, Nov 2003.