Master Thesis

REDIAL-based Framework for Artificial Bandwidth Extension of Narrowband Speech

(REDIALに基づく狭帯域音声に対する広帯域拡張)

2014 年 2 月 6 日

Supervisor　Prof. Hirose Keikichi

電子情報学専攻

48-126418　グェンドゥックズイ

Abstract

Although humans can well perceive sounds up to 8 kHz, traditional telephone networks were designed to limit the frequency to a lower range, approximately below 3.4 kHz, in order to conserve the bandwidth and increase the number of voice streams transmittable by a transmission channel. This results in degradation of perceptual speech quality of the narrowband speech at receiving end. True wideband transmission is therefore desirable, but this requires a significant amount of cost and time, since the whole transmission chain including terminals and network elements need to be upgraded. This challenge can be overcome with Artificial Bandwidth Extension (ABE) technique. ABE is a technique that tries to recover missing low and high frequency components of the speech signal only from the narrowband speech. By integrating ABE into the current telephone networks, we can easily realize wideband transmission without modifying the networks and terminals.

A number of techniques have been proposed over the years for bandwidth extension of narrowband speech signals, including methods based on codebook mapping [1] and statistical approaches [2, 3, 4]. Most of these ABE algorithms are based on the source-filter model [5] of speech production whereby the speech signal is regarded as output of the vocal tract filter which takes excitation source signals as input. This model breaks the problem down into two subtasks: one is to extend the spectral envelope, and the other is to extend the excitation signal. The extension of spectral envelope is typically considered as the main problem of ABE since it had been shown that extension of the spectral envelope has a large effect on speech quality of the reconstructed wideband speech [6].

It is known that the Gaussian Mixture Model (GMM) [7] represents robustly the acoustic space of speech and was successfully applied to the problem of spectral transformation, especially voice conversion [8]. Based on the successes in voice conversion, in [4] an effective approach to the problem of extending the spectral envelope was proposed. In this approach, the spectral envelope of wideband speech was estimated using a GMM trained by parallel data of narrowband speech and its corresponding wideband speech. This approach showed that there was a large improvement in speech quality from the original narrowband speech to the reconstructed wideband speech. However, the gap between the reconstructed wideband and the original wideband speech was still large.

Stereo-based Piecewise LInear Compensation for Environments (SPLICE) [9], in which non-linear transformation between two feature vectors is approximated by the summation of piecewise linear transformations, is an effective and widely used method in speech

enhancement. A revised version of SPLICE, in which a discriminative model, long-span features and regularization are introduced into SPLICE, was proposed [10, 11, **?**] and has been shown to outperform the original SPLICE. This revised version was named REgularized piecewise linear mapping with DIscriminative region weighting And Long-span features (REDIAL). The objective of spectral envelope extension is to make a transformation from spectral envelope of narrowband speech to that of wideband speech. From this view of point, ABE task is very similar to the scheme used in REDIAL. Therefore, in this research we proposed a method based on the ideas of REDIAL for the ABE task.

Several experiments were carried out to examine the effectiveness of the proposed method. Objective evaluation results reported a reduction in mcel-cepstral distortion between the estimated wideband speech and the original wideband speech. Additionally, subjective evaluations also pointed out that the estimated wideband speeches of the proposed method were preferable than those estimated by other conventional methods. These results have confirmed the effectiveness of our proposed method.

Contents

List of Figures

List of Tables

Chapter 1

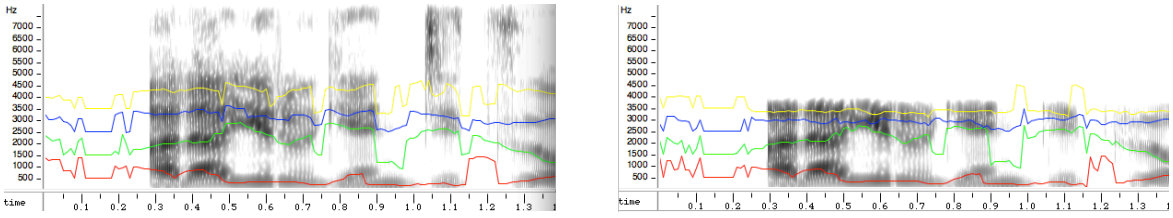
Introduction

1.1 Background

In traditional telephone networks and mobile communication systems, the speech bandwidth is typically limited to a frequency range of 300 Hz to 3.4 kHz and sampled at rate 8kHz (we call this as **narrowband speech** from now on) due to constraints of the old analogue telephone system. Limiting speech bandwidth has been shown to cause a degradation in speech quality, speech naturalness and speech intelligibility [6]. Fig. 1.1 and Fig. 1.2 show the differences in spectrogram and spectrum of wideband speech and narrowband speech. We can observed that the narrowband is missing the upper components from 3.4 kHz to 8 kHz. Meanwhile, human ear's hearing range is said to be from 20 Hz - 20 kHz [12], much wider than the frequency range of narrowband speech. Therefore, it is easy to understand that narrowband speech has poorer speech quality than the wideband.

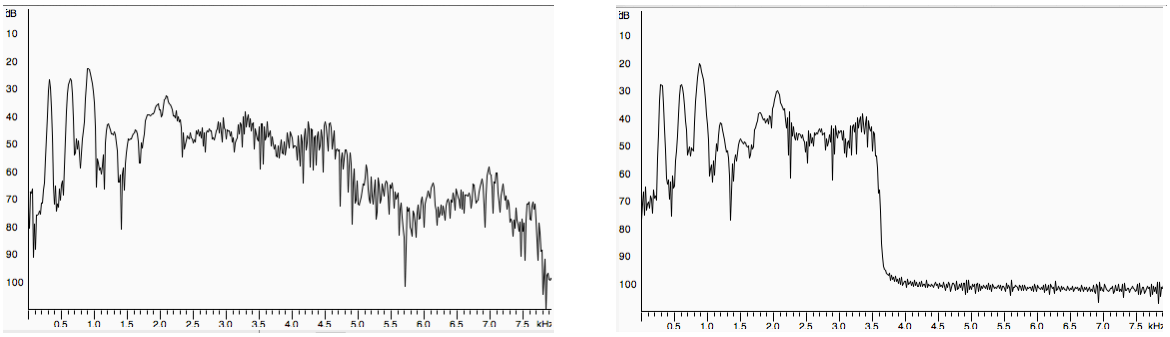
In recent years, due to the rapid development of IP networks such as Next Generation Network (NGN) [13], Long Term Evolution (LTE) [14], the high quality wideband speech transmission is beginning to become more available. In the future, true wideband speech transmission will be realized and is expected to become the main transmission media.

However, the transition phase from current network to the wideband one requires a significant amount of effort from both operators and users, since the whole transmission chain



(a) Wideband speech spectrogram (b) Narrowband speech spectrogram

Fig.1.1: Spectrogram of wideband speech (0-8 kHz)and narrowband speech (0-3.4 kHz). The narrowband speech has been up sampled for better comparision.



(a) Wideband speech spectrum (b) Narrowband speech spectrum

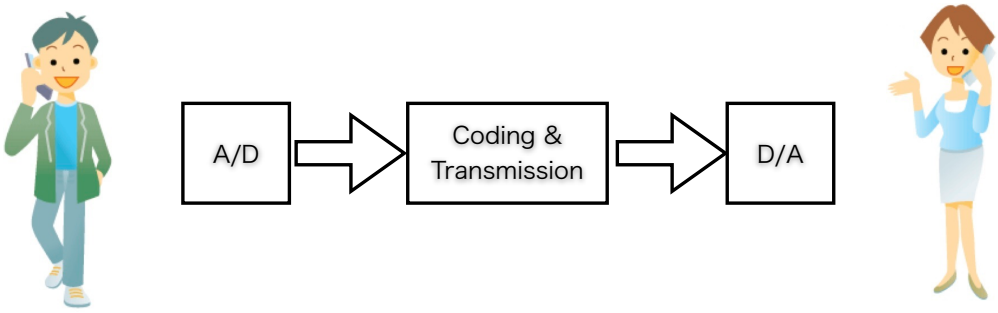
Fig.1.2: Spectrum of wideband speech (0-8 kHz)and narrowband speech (0-3.4 kHz). The narrowband speech has been up sampled for better comparision.

including terminals and network elements are required to support wideband transmission. During this long transitional period mixed telephone networks with both narrowband and wideband terminals will exist due to economical reasons. Nowadays, besides the traditional model, a lot of terminals which support wideband codec have been developed. Since wideband transmission requires both the sending end and receiving end to have wideband codec feature, the transmission between traditional model and new model of terminals will not be the wideband transmission, but the traditional narrowband one.

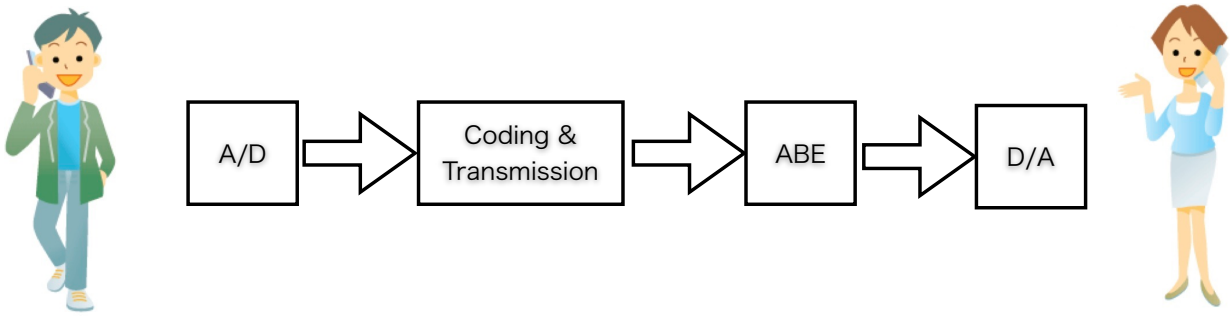
This challenge can be overcome by Artificial Bandwidth Extension (ABE) technique: missing low and high frequency components of the speech signal are recovered at the receiving end of the transmission link utilizing only the band-limited speech. The underlying assumption is that narrowband speech correlates closely with the highband signal, and hence, the higher frequency speech content can be estimated from the narrowband signal. Fig. 1.3 illustrates how ABE can be integrated into the existing telephone networks.

A number of techniques ([1], [15]-[36]) have been proposed over the years for bandwidth extension of narrowband speech signals. Most of these ABE algorithms are based on the source-filter model [5] of speech production whereby the speech signal is regarded as an excitation source signal that has been acoustically filtered by the vocal tract. This model breaks down the problem into two subtasks: one is to extend the spectral envelope, and the other is to extend the excitation signal. For the first one, several approaches have been made including, approaches based on codebook [1, 16], neural networks [17, 18, 19], linear mapping [20], Bayesian methods based on GMM [3, 21, 22], HMM [24, 2, 25] as well as joint approaches. For the latter one, several algorithms such as, non-linear characteristics [26, 27], spectral translation [28, 29], signal generators [30] have been proposed.

State-of-the-art schemes show significant improvement in quality versus narrowband speech; however a clear gap in speech quality compared with true wideband speech is still reported. Many efforts have been made to improve the estimation of highband speech by using some auxiliary information together with the narrow-band speech signals [31, 32].



(a) Traditional model of telephone transmission. In this model, the speech recovered at the receiving end is narrowband speech.



(b) New transmission model by integrating Artificial Bandwidth Extension technique. The narrowband speech transmitted from the sending end will be used to estimate the missing highband at the receiving end. This eventually results in a wideband speech.

Fig.1.3: Traditional model and revised model using ABE technique of telephone network.

Although improvements in speech quality have been made, these approaches require a modification of the network (better terminal, new codecs, etc). As ABE is a solution for realizing wideband transmission without making any changes of the existing network, in scope of this thesis, we will not consider these methods.

1.2 Objectives of the thesis

This thesis is oriented to conduct an investigation about the mechanism of ABE together with several common methods to realize it. After that, a new approach for ABE based on REgularized piecewise linear mapping with DIscriminative region weighting And Long-span features (REDIAL) [10], a method recently proposed for feature enhancement, will be presented. In details, the followings will be covered in this thesis.

1. Study the background and general framework of ABE system.
2. Introduce several typical methods for ABE system.
3. Explain the proposed REDIAL-based method for ABE.
4. Implement and evaluate the performance of the proposed method.

1.3 Organization of the thesis

The remainder of this paper is organized as follow. In Chapter. 2, an overview about the source-filter model of speech production will be introduced. After that, a general framework of Artificial Bandwidth Extension which based on this source-filter model will be presented. Moreover, several typical algorithms to realize ABE (e.g spectral translation, codebook, GMM, HMM methods) are also discussed in this chapter. In Chapter. 3, we begin with a discussion about drawbacks of conventional methods for ABE and then present a recent research which tried to deal with those drawbacks. In Chapter. 4, we further analyze the limitation of the method described in Chapter. 3. After that, we proposed a method based on REDIAL to resolve this problem. Then we explain an ABE system which will be used in our experiments. In Chapter. 5 four experiments to evaluate the effectiveness of the proposed method are introduced. Finally in Chapter. 6, we will summarize the whole works in this thesis and discuss about the future works of our research.

Chapter 2

Artificial Bandwidth Extension System

2.1 Introduction

As mentioned in Chapter. 1, most of current ABE algorithms are based on the source-filter model [5] of speech generation. In this chapter, we will present this speech generation model, following with how ABE is realized under this model.

2.2 Source-Filter Model

Fig. 2.1 show a diagram of the human speech production system. An air flow is produced from the lungs, then it is passed through other organs such as trachea, larynx, mouth, etc. When this air flow emanating from the mouth and the nostrils, a speech sound will be generated. The organs related to this speech production process are normally divided into 3 groups: the lungs, larynx and vocal tract (oral cavity, nasal cavity, pharynx). The lungs are source of power and the larynx provides periodic or noisy airflow (random noise) to the vocal tract. The vocal tract spectrally shapes this airflow and a speech sound is generated.

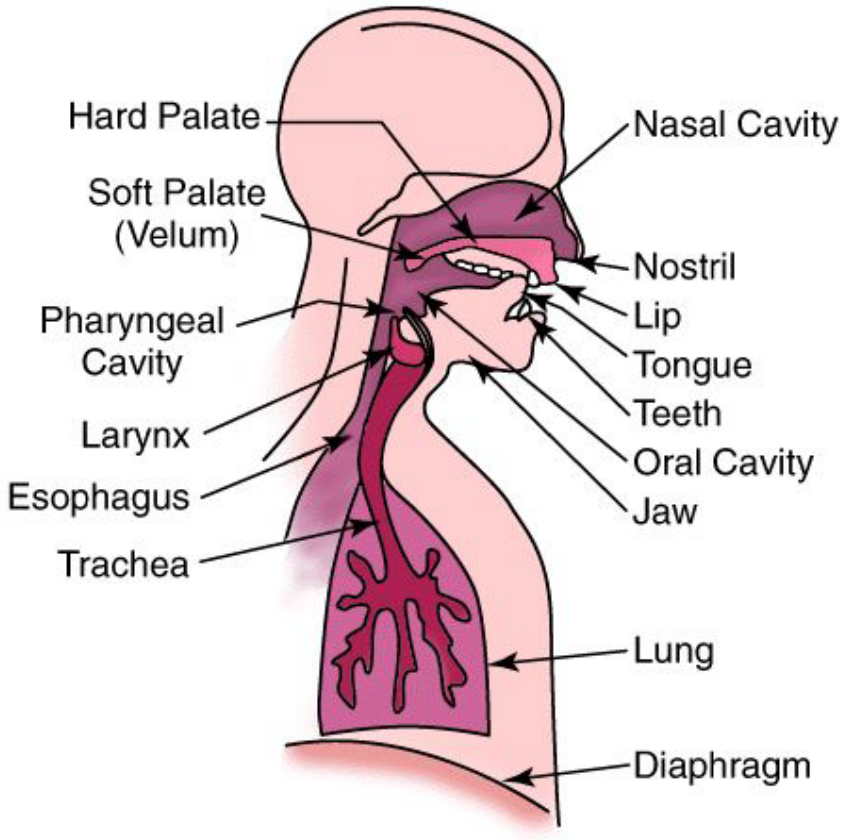


Fig.2.1: Human speech production system

The source-filter model is motivated by studies of the above process of human speech

production [5]. In this model, the speech signals are considered to consist of two parts. One part describing the excitation signal from the source and another part describing the synthesis filters. Speech signal is produced when an excitation signal is passed through the synthesis filter (see Fig. 2.2). The excitation and the filters will be described in the following sections.

2.2.1 The excitation signal

The excitation signal corresponds to the signal that could be observed directly behind the vocal chords at the larynx. For its generation, the source part of the source-filter model is differentiated between two scenarios:

- For voiced sounds, the excitation signal is modeled by a pulse train.
- For unvoiced sounds a noise generator models the excitation signal.

In reality, the excitation is typically a mix of two with one of them dominating.

2.2.2 The filter models

Typically three filters are used to model the speech production: The Glottal Pulse Model $G(z)$, The Vocal Tract Model $V(z)$ and The Radiation Model $R(z)$. Under the assumption that speech is stationary in short speech frame, these models can be considered as time invariant. The glottal pulse model is only used to model speech in the voiced case. The vocal tract model is modeling the region from the vocal chords and the glottis to the lips. The radiation model takes into account the radiation which occurs at the lips. These models combine together make a filter called Synthesis Filter $H(z)$:

$$H(z) = G(z)V(z)R(z) \tag{2.1}$$

When LP analysis is performed on a speech signal, an excitation signal and an analysis filter $A(z)$ is obtained. The synthesis filter is the inverse of the analysis filter, i.e: $H(z) = 1/A(z)$.

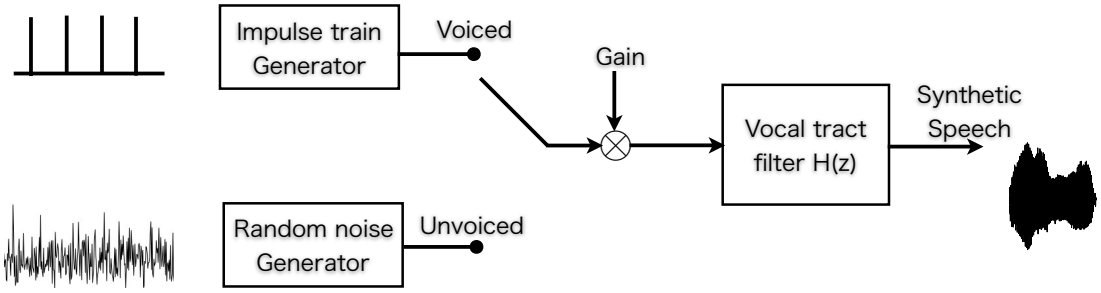


Fig.2.2: Source-filter model for human speech production system

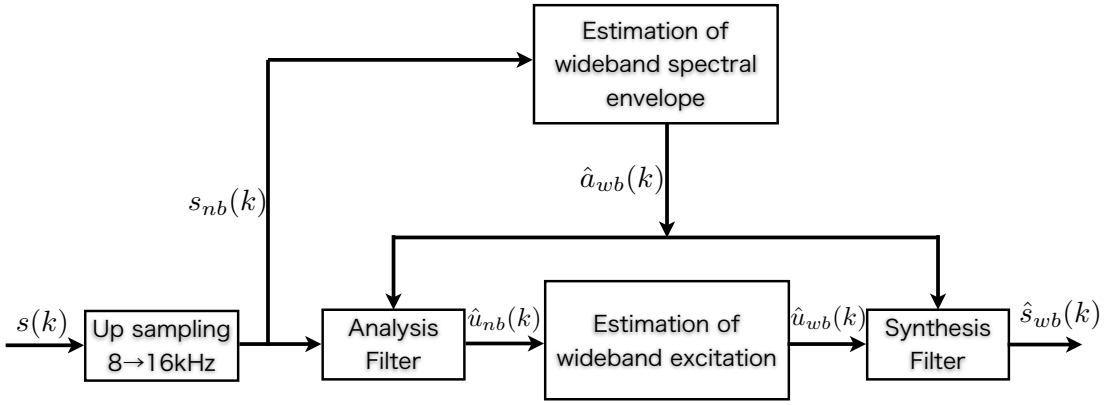


Fig.2.3: Signal flow of Bandwidth Extension algorithm

2.3 Basic scheme of Artificial Bandwidth Extension

The usage of source-filter model in approach toward ABE is motivated by its extensive use and success in the field of speech coding. By adopting this model, ABE is commonly separated into two sub-taks: one for the estimation of wideband spectral envelope and the other for extension of the excitation signal of the speech (Fig. 2.3).

The narrowband speech signal is first up sampled to 16 kHz and is analyzed to obtain feature vectors which represent the spectral envelope (LSF, MFCC, MCEP, etc).The spectral envelope of the corresponding wideband speech will be estimated using these features and statistical models trained in advance (more details in Chapter. 3). From the estimated wideband spectral envelope, coefficients of the analysis filter are obtained and will be used to extract the excitation signal of the narrowband speech. The extracted narrowband excitation signal in its turn will be used to estimate the wideband excitation signal. Since the synthesis filter is the reverse of analysis filter, we can easily achieve the synthesis filter from the estimated wideband spectral envelope. The estimated wideband speech signal is now made by driving the estimated wideband excitation signal through the estimated synthesis filter.

In summary, in order to realize ABE, it is essential to estimate the wideband excitation signal and the wideband spectral envelope. There have been several algorithms proposed for these two tasks so far. In the next sections, we will discussed in more details about these algorithms.

2.4 Estimation of Wideband Excitation

2.4.1 Estimation Using Non-Linear Characteristics

In this section, several non-linear characteristics which are appropriate for extending the narrowband excitation signal, such as Half-way Rectification, Full-way Rectification, Quadratic characteristic, Cubic characteristic, etc will be presented. It can be proved that applying a non-linear characteristic to a harmonic signal produces sub- and super-harmonics. Take quadratic characteristic as an example. Denote $u_{n/w}(k;m)$ as the excitation signal of the m^{th} frame of narrow-band signal and wide-band signal respectively. The application of a quadratic characteristic in the time domain corresponds to a convolution in the frequency domain.

$$\begin{aligned} \hat{u}_w(n) &= u_n^2(n) \\ &\leftrightarrow E_n(e^{j\Omega_k}) * E_n(e^{j\Omega_k}) && (2.2) \\ &= \sum_{i=-\infty}^{\infty} E_n(e^{j\Omega_k}) * E_n(e^{j\Omega_{k-i}}) && (2.3) \\ &= \hat{E}_w(e^{j\Omega_k}) && (2.4) \end{aligned}$$

The biggest advantage of applying non-linear characteristics for extension of the excitation signal is the production of well placed harmonics.

i) Half-way and Full-way Rectification

1. Half-way Rectification:

$$\hat{u}_w(k) = \begin{cases} 0 & (u_n(k) \leq 0) \\ u_n(k) & else \end{cases} \quad (2.5)$$

2. Full-way Retification:

$$\hat{u}_w(k) = |u_n(k)| \quad (2.6)$$

The half way rectification rectifies to alternating current by blocking the negative half wave and passing the positive half. For its part, full way rectification applies a reversion on the negative half wave of the alternating current. Both the half way and full way rectifiers produce output signal which is non-zero mean. The half-way rectifier produces even harmonics including the fundamental frequency (Fig. 2.4) and is not power conserving. Meanwhile, the full way rectifier produces even harmonics without the fundamental frequency (Fig. 2.4) and is power conserving. The disadvantages of these method is that there is a spectral gap at the cut-off point of narrowband and wideband signal.

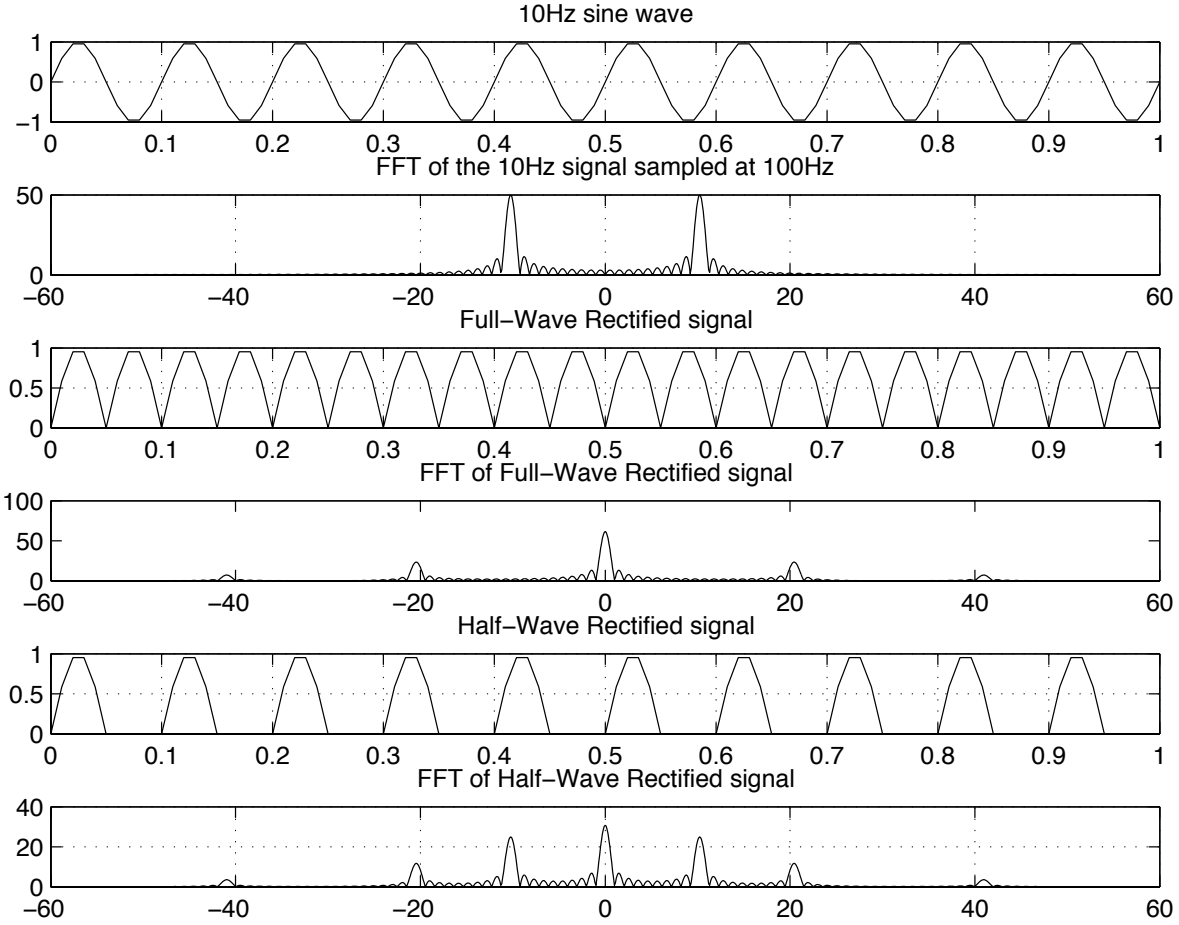


Fig.2.4: Effect of applying half and full way rectification to a 10 Hz sine wave

ii) Quadratic and Cubic Characteristic

1. Quadratic characteristic:

$$\hat{u}_w(k) = u_n^2(k) \tag{2.7}$$

2. Cubic characteristic:

$$\hat{u}_w(k) = u_n^3(k) \tag{2.8}$$

In both cases, the power of the signal is changed. The quadratic characteristic produces non-zero mean output signal, while the cubic characteristic produces zero-mean output signal if the input signal is zero-mean and symmetrically distributed. The quadratic characteristic produces the second harmonic without the fundamental frequency, while the cubic characteristic produces the third harmonic including the fundamental frequency (Fig. 2.5). The drawback of these methods is that they color the estimated excitation signal.

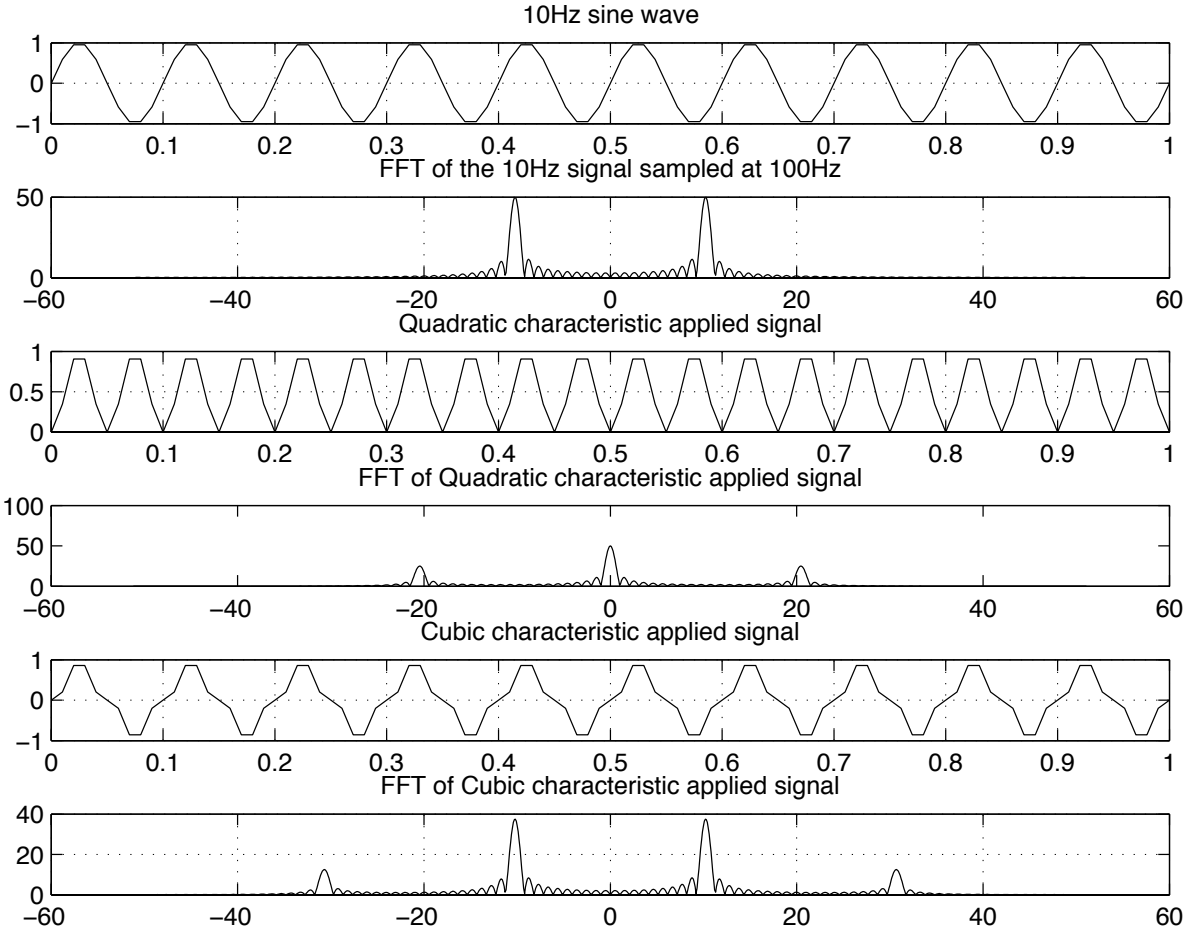


Fig.2.5: Effect of applying Quadratic and Cubic rectification to a 10 Hz sine wave

2.4.2 Estimation Using Spectral Translation

Spectral Translation (also called Spectral Modulation) is one of the most commonly used methods for estimating the missing high-frequency components of the excitation signal (Fig. 2.6) [28, 29]. The basic idea behind this approach is to "shift" the spectrum of narrowband excitation signal into the upper part of the spectrum, then combine them to make the wideband excitation signal.

$$\tilde{u}_{eb}(k) = u_{nb}(k) \times e^{j\Omega_M k} \tag{2.9}$$

$$E_{nb}(e^{j\Omega}) \times \delta(\Omega - \Omega_M) = \tilde{E}_{nb}(e^{j\Omega}) \tag{2.10}$$

which means,

$$\tilde{E}_{nb}(e^{j\Omega}) = E_{nb}(e^{j(\Omega-\Omega_M}) \tag{2.11}$$

As we can see from above equations, the multiplication operation in time domain corresponds to the convolution in frequency domain and finally results in a shift.

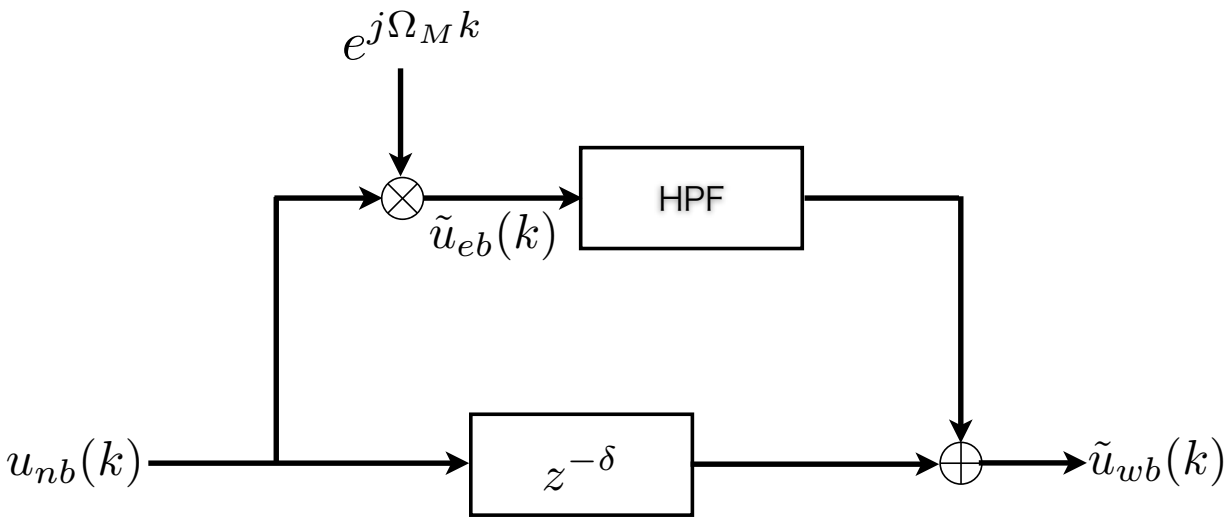


Fig.2.6: Extension of the excitation signal by modulation. The delay δ is introduced to compensate for the delay in HPF

By selecting the modulation frequency Ω_M several modulation schemes can be chosen:

1. A modulation with the Nyquist frequency, i.e., $\Omega_M = \pi$. In this case, there is a spectral gap in $\tilde{u}_{wb}(k)$ between 3.4 and 4.6 kHz. the discrete spectral components of the extended frequency band are no harmonic of the fundamental frequency.
2. To prevent the spectral gap, the modulation frequency can be chosen such that the shifted spectrum is a seamless continuation of the baseband spectrum: $\Omega_M = \Omega_{3.4} = 2\pi\frac{3.4kHz}{f_s}$, where f_s denotes the sampling rate. In general, there is a misalignment of discrete spectral components in the extension band during voiced sounds.

Informal listening tests have shown that, assuming the bandwidth extension of the spectral envelope works well, distortions of the excitation signal at frequencies above 3.4 kHz are almost inaudible. Furthermore, a misalignment of the harmonic structure of speech at high frequency does not significantly degrade the subjective quality of the enhanced speech. Therefore, the modulation with the fixed frequency of $\Omega_M = \Omega_{3.4}$ is normally used considering the balance between subjective quality and computational complexity.

2.5 Estimation of Wideband Spectral Envelope

As stated earlier, there are many algorithms for spectral envelope extension, including algorithms based on linear mapping, neural network, codebook, Bayesian methods based on GMM, HMM, etc. In following sections, three often-used algorithms, namely codebook, GMM and HMM will be discussed.

2.5.1 Codebook-based algorithm [1]

Suppose $\boldsymbol{x}, \boldsymbol{y}$ to be the feature vectors of the narrow-band and wide-band signal respectively. A joint vector $\boldsymbol{z} = [\boldsymbol{x}^\top, \boldsymbol{y}^\top]^\top$ then can be generated. The codebook denoted C_z is trained using the joint feature vectors $\boldsymbol{z}_t$ consisting of both narrowband features and wideband features.

Each entry of codebook C_z can be divided into two parts which represent the narrowband and wideband features respectively as following equation:

$$C_z(i) = [C_x(i), C_y(i)] \tag{2.12}$$

Here, $C_x(i)$ is the part of the entry describing the narrowband features and $C_y(i)$ is the part of the entry describing the wideband features.

After acquiring codebook C_z and "sub-codebooks" C_x, C_y, an estimation of $\boldsymbol{y}$ can be obtained from $\boldsymbol{x}$. This can be done as follow:

$$i^* = \arg\max_i d(\boldsymbol{x}, C_x(i)) \tag{2.13}$$

$$\hat{\boldsymbol{y}} = C_y(i^*) \tag{2.14}$$

where $d()$ is a distortion measure (Euclidean distance for example). The training of codebook C_z will be introduced in the next section.

i) Codebook training using LBG algorithm

The Linde, Buzo and Gray (LBG) is a simple and well-known algorithm for training codebook [15]. LBG algorithm is like a K-means clustering algorithm which takes a set of vectors $\boldsymbol{z}_t$ as input and generates a representative subset of vectors $C_z = \{C_z(i)|i = 1...K\}$, where $K \ll N$ is a specified parameter, as output. In practice, K is normally set to 256, 512, 1024 or even larger depends on the size of training set. The algorithm can be performed in the following steps:

1. Initialization: Define parameters of the codebook and the training.
   - K: number of vectors in the codebook (codebook size, normally 256, 512, etc)
   - $\epsilon \geq 0$: the distortion threshold
   - K initial entries to form the initial codebook $C_z^l(i), i \in \{1, 2, ..., K\}$. This can be chosen randomly from the T training features.

   At start, set iteration counter $l = 0$ and initial distortion $D_{-1} = \infty$
2. Quantization: Quantize each training vector $\boldsymbol{z}_t$ using the codebook $C_z^l(i)$ as below:

$$i^* = \arg\min_i d(\boldsymbol{z}_t, C_z^l(i)) \tag{2.15}$$

$$\boldsymbol{z}_t \mapsto Q(\boldsymbol{z}_t) = C_z^l(i) \tag{2.16}$$

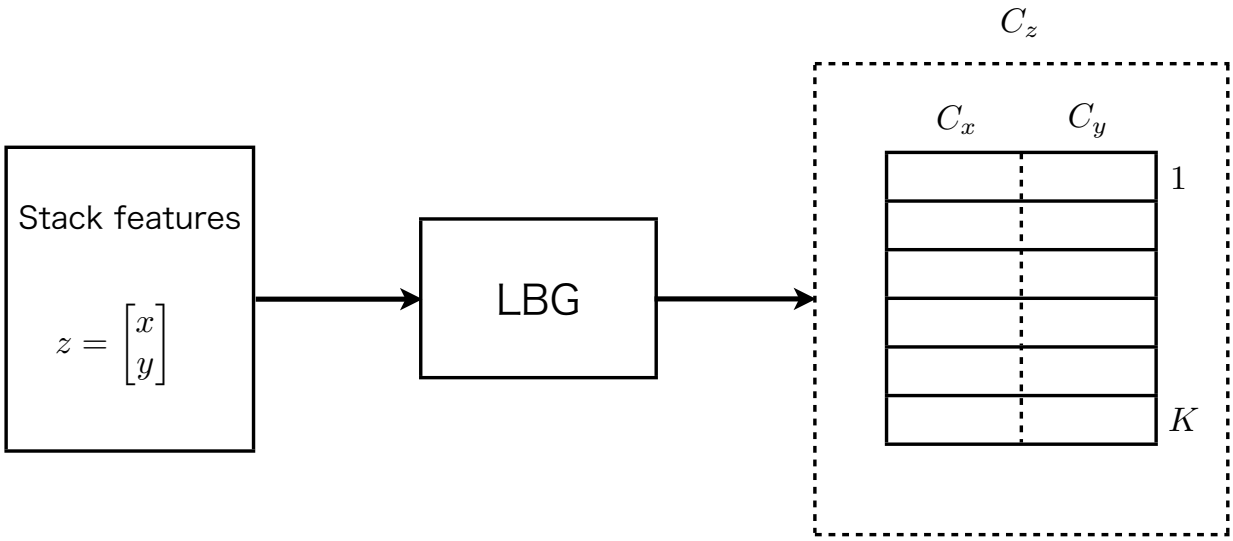


Fig.2.7: Training procedure for codebook method

3. Distortion: Calculate the average distortion between $\boldsymbol{z}_t$ and its quatized value, $Q(\boldsymbol{z}_t)$:

$$D^l = \frac{1}{T}\sum_{t=1}^{T} d(\boldsymbol{z}_t, Q(\boldsymbol{z}_t)) \tag{2.17}$$

4. Judgment Condition: Check if the distortion have decreased enough. If

$$\frac{D^{l-1} - D^l}{D^l} \leq \epsilon \tag{2.18}$$

   C_z^l will be output as the final codebook, else continue with 5.
5. Update the codebook C_z^l: For all i, calculate the mean of all the $\boldsymbol{z}_t$ quantized into the i^{th} codebook entry.

$$C_z^l(i) = \frac{1}{T}\sum_{t=1}^{T} \boldsymbol{z}_t \tag{2.19}$$

   Increase $l = l + 1$ and go to 2.

Fig. 2.7 illustrates the general framework for codebook-based method.

ii) Discussion

Codebook implementation only looks at snapshots of the speech, and not how it evolve over time. Therefore, no inter-frame modeling is done, and one could imagine that it would result in valuable information being lost. (Fig. 2.8) shows a clustering result using LBG algorithm. In Fig. 2.8, there are 259 spectral envelopes which have been clustered to a codebook entry $C_y(40)$. However, as can be seen in the figure, the shapes of the spectral envelopes in that entry are quite different. This is not good since it is desired that spectral envelopes being clustered in same entry should have similar (or almost the same) shape.

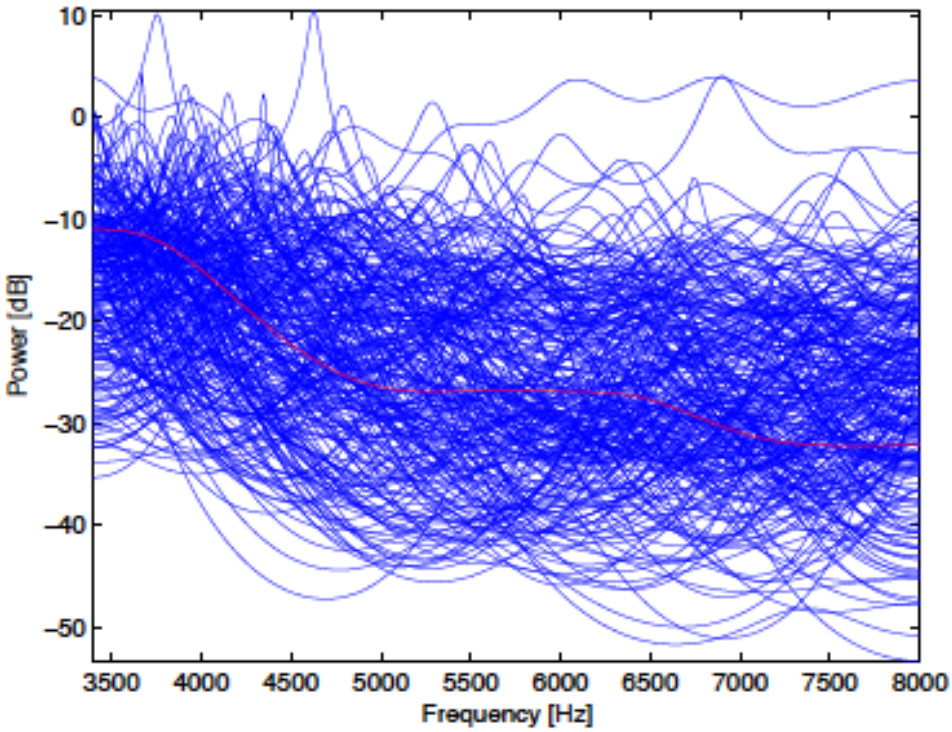


Fig.2.8: Spectral envelopes representing the extension band. Blue curves are those used in training. The red one is obtained from codebook entry $C_y(40)$

2.5.2 GMM-based algorithm

In codebook approach, spectral mapping is realized based on hard clustering and discrete mapping. Compared to the codebook approach, the GMM allows soft clustering and continuous mapping. The basic mapping algorithm was originally proposed for voice conversion. In the following section, spectral mapping based on GMM with Minimum Mean Square Error (MMSE) criterion [3] will be discussed.

i) GMM training

Let $\boldsymbol{x}_t$ and $\boldsymbol{y}_t$ be the D_x-dimensional narrow-band and D_y-dimensional wide-band feature vectors at frame t respectively. Then the joint density of vector $\boldsymbol{z}_t = [\boldsymbol{x}_t{}^T, \boldsymbol{y}_t{}^T]^T$ is modeled by a GMM as follows:

$$P(\boldsymbol{z}_t|\theta) = \sum_{m=1}^{M} \omega_m \mathcal{N}(\boldsymbol{z}_t; \boldsymbol{\mu}_m^{(z)}, \boldsymbol{\Sigma}_m^{(z)}) \tag{2.20}$$

where θ is a parameter set of GMM, which consists of weights ω_m, mean vectors $\boldsymbol{\mu}_m$ and covariance matrices $\boldsymbol{\Sigma}_m$. The GMM is trained with EM algorithm using the joint vectors in training set. The training process includes the following main steps:

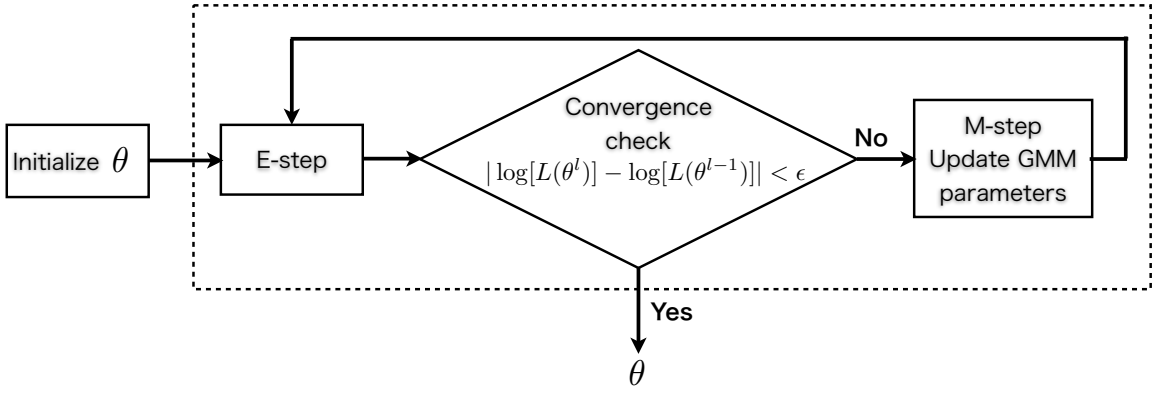


Fig.2.9: EM algorithm for training a GMM

1. E-step: Calculate the posterior probability for the every feature vectors of the training set of observations:

$$P(m; \boldsymbol{z}_t, \theta^l) = \frac{\omega_m^l \mathcal{N}(\boldsymbol{z}_t; \boldsymbol{\mu}_m^l, \boldsymbol{\Sigma}_m^l)}{\sum_{k=1}^{M} \omega_m^l \mathcal{N}(\boldsymbol{z}_t; \boldsymbol{\mu}_m^l, \boldsymbol{\Sigma}_m^l)} \tag{2.21}$$

   where l is the iteration counter.
2. Convergence check: Calculate the log-likelihood for the entire training set as below:

$$\log[L(\theta^l)] = \sum_{t=1}^{T} \log[\sum_{k=1}^{M} \omega_m^l \mathcal{N}(\boldsymbol{z}_t; \boldsymbol{\mu}_m^l, \boldsymbol{\Sigma}_m^l)] \tag{2.22}$$

   , then check if this log-likelihood is converged:

$$|\log[L(\theta^l)] - \log[L(\theta^{l-1})]| < \epsilon \tag{2.23}$$

   $\epsilon > 0$ is a pre-defined parameter. If the log-likelihood satisfies above condition, then the parameters $\omega_m^l, \mu_m^l, \Sigma_m^l$ will be output. Otherwise, go to M-step.
3. M-step: Update GMM parameters as followings:

$$\omega_m^{l+1} = \frac{1}{T} \sum_{t=1}^{T} \mathcal{N}(m; \boldsymbol{z}_t, \boldsymbol{\mu}_m^l, \boldsymbol{\Sigma}_m^l) \tag{2.24}$$

$$\mu_m^{l+1} = \frac{\sum_{t=1}^{T} \mathcal{N}(m; \boldsymbol{z}_t, \boldsymbol{\mu}_m^l, \boldsymbol{\Sigma}_m^l) \boldsymbol{z}_t}{\sum_{t=1}^{T} \mathcal{N}(m; \boldsymbol{z}_t, \boldsymbol{\mu}_m^l, \boldsymbol{\Sigma}_m^l)} \tag{2.25}$$

$$\Sigma_m^{l+1} = \frac{\sum_{t=1}^{T} \mathcal{N}(m; \boldsymbol{z}_t, \boldsymbol{\mu}_m^l, \boldsymbol{\Sigma}_m^l)(\boldsymbol{z}_t - \mu_m^{l+1})(\boldsymbol{z}_t - \mu_m^{l+1})^\top}{\sum_{t=1}^{T} \mathcal{N}(m; \boldsymbol{z}_t, \boldsymbol{\mu}_m^l, \boldsymbol{\Sigma}_m^l)} \tag{2.26}$$

The EM algorithm is summarized in Fig. 2.9.

The mean vectors $\boldsymbol{\mu}_m^{(\boldsymbol{z})}$ and covariance matrices $\boldsymbol{\Sigma}_m^{(\boldsymbol{z})}$ can be decomposed two 2 components which correspond to narrowband and wideband as follow:

$$\boldsymbol{\mu}_m^{(\boldsymbol{z})} = \begin{bmatrix} \boldsymbol{\mu}_m^{(\boldsymbol{x})} \\ \boldsymbol{\mu}_m^{(\boldsymbol{y})} \end{bmatrix}, \boldsymbol{\Sigma}_m^{(\boldsymbol{z})} = \begin{bmatrix} \boldsymbol{\Sigma}_m^{(\boldsymbol{xx})} & \boldsymbol{\Sigma}_m^{(\boldsymbol{xy})} \\ \boldsymbol{\Sigma}_m^{(\boldsymbol{yx})} & \boldsymbol{\Sigma}_m^{(\boldsymbol{yy})} \end{bmatrix} \tag{2.27}$$

ii) Estimation of wideband feature with MMSE

The estimation of wideband feature $\boldsymbol{y}$ from narrowband feature $\boldsymbol{x}$ based on MMSE criterion is as follow:

$$\hat{\boldsymbol{y}}_t = \arg\min_{\hat{\boldsymbol{y}}} E[(\hat{\boldsymbol{y}} - \boldsymbol{y}_t)^\top (\hat{\boldsymbol{y}} - \boldsymbol{y}_t) | \boldsymbol{x}_t, \theta] \tag{2.28}$$

To minimize this, the differentiation of the right side regarding $\hat{\boldsymbol{y}}$ must be 0, which means:

$$2\hat{\boldsymbol{y}} - 2\boldsymbol{E}[\boldsymbol{y}_t | \boldsymbol{x}_t, \theta] = 0 \tag{2.29}$$

Therefore,

$$\hat{\boldsymbol{y}}_t = E[\boldsymbol{y}_t | \boldsymbol{x}_t, \theta] \tag{2.30}$$

$$\hat{\boldsymbol{y}}_t = \sum_{m=1}^{M} P(m | \boldsymbol{x}_t, \theta) \boldsymbol{E}[\boldsymbol{y}_t | \boldsymbol{x}_t, \theta_m] \tag{2.31}$$

$$\hat{\boldsymbol{y}}_t = \sum_{m=1}^{M} P(m | \boldsymbol{x}_t, \theta) \boldsymbol{E}_{m,t} \tag{2.32}$$

where

$$P(m | \boldsymbol{x}_t, \theta) = \frac{\omega_m N(\boldsymbol{x}_t; \boldsymbol{\mu}_m^{(\boldsymbol{x})}, \boldsymbol{\Sigma}_m^{(\boldsymbol{xx})})}{\sum_{j=1}^{M} \omega_j N(\boldsymbol{x}_t; \boldsymbol{\mu}_j^{(\boldsymbol{x})}, \boldsymbol{\Sigma}_j^{(\boldsymbol{xx})})} \tag{2.33}$$

$$\boldsymbol{E}_{m,t} = \boldsymbol{\mu}_m^{(\boldsymbol{y})} + \boldsymbol{\Sigma}_m^{(\boldsymbol{yx})} (\boldsymbol{\Sigma}_m^{(\boldsymbol{xx})})^{-1} (\boldsymbol{x}_t - \boldsymbol{\mu}_m^{(\boldsymbol{x})}) \tag{2.34}$$

2.5.3 HMM-based algorithm [2, 24]

In this approach, the extension of spectral envelope of speech signal is based on a wideband codebook, and the codebook search is based on an HMM of the speech generation process. Each state S_i of the HMM ($i = 1, ..., N_S$) is assigned to a typical entry C_i of the codebook. For each possible state S_i of HMM, the feature $\boldsymbol{x}$ which are generated by the speech production process exhibit different statistical properties, which can be described by the following 3 parts:

1. Observation Probabilities $p(\boldsymbol{x}|S_i)$ Since x is multi-dimensional vector with continuous range of values, the observation probabilities are modeled by GMM:

$$p(\boldsymbol{x}|S_i) \approx \sum_{l=1}^{L} P_{il} N(\boldsymbol{x}; \mu_{il}, \Sigma_{il}) \tag{2.35}$$

   For each state of the HMM, one distinct GMM has to be trained using the subset of the training data.

2. Initial state probabilities $\pi_i = P(S_i)$: describes the probability that HMM resides in state S_i without incorporating any knowledge of the preceding or following states.
3. Transition Probabilities $a_{ij} = P(S_i(m+1)|S_j(m))$

Utilizing these above HMM parameter sets, an estimation of parameter vector $\boldsymbol{y}$ can be performed as follow.

$$\begin{aligned}
\tilde{\boldsymbol{y}}(m) &= E[\boldsymbol{y}(m)|\boldsymbol{X}(m)] && (2.36)\\
&= \int \boldsymbol{y}(m) p(\boldsymbol{y}(m)|\boldsymbol{X}(m)) d\boldsymbol{y}(m) && (2.37)\\
&= \sum_{i=1}^{N_S} P(S_i(m)|\boldsymbol{X}(m)) \\
&\quad . \int \boldsymbol{y}(m) p(\boldsymbol{y}(m)|S_i(m), \boldsymbol{x}(m)) d\boldsymbol{y}(m) && (2.38)\\
&= \sum_{i=1}^{N_S} E[\boldsymbol{y}(m)|S_i(m), \boldsymbol{x}(m)] P(S_i(m)|\boldsymbol{X}(m)) \\
& && (2.39)
\end{aligned}$$

where

$$\begin{aligned}
P(S_i(m)|\boldsymbol{X}(m)) &= \frac{p(S_i(m), \boldsymbol{X}(m))}{\sum_{i=1}^{N_s} p(S_i(m), \boldsymbol{X}(m))} \\
p(S_i(m), \boldsymbol{X}(m)) &= p(\boldsymbol{x}(m)|S_i(m)) p(S_i(m), \boldsymbol{X}(m-1)) \\
p(\boldsymbol{x}(m)|S_i(m)) &= \sum_{l=1}^{L} \rho_{il} N(\boldsymbol{x}(m); \mu_{\boldsymbol{x},il}, \boldsymbol{V}_{\boldsymbol{xx},il})
\end{aligned}$$

References Some other estimation rules were also proposed as follow:

ML $\tilde{\boldsymbol{y}} = \hat{\boldsymbol{y}}_{i_*}$ with $i_* = \arg\max_i p(\boldsymbol{x}(m)|S_i(m))$

MAP $\tilde{\boldsymbol{y}} = \hat{\boldsymbol{y}}_{i_*}$ with $i_* = \arg\max_i P(S_i(m)|\boldsymbol{X}(m))$

hard-MMSE $\tilde{\boldsymbol{y}} = \sum_{i=1}^{N_S} \hat{\boldsymbol{y}}_i P(S_i(m)|\boldsymbol{X}(m))$

2.6 Summary

In this chapter, the source-filter model of speech production process has been introduced. After that, a general framework of Artificial Bandwidth Extension based on this model has been presented. As in this framework, the ABE problem is divided into 2 sub-problems: Estimation of wideband excitation signal and Estimation of wideband spectral envelope. For the first problem, several methods based on non-linear characteristics or spectral modulation have been discussed. For the second problem, three common approaches (codebook, GMM, HMM) with their detailed mechanism have been presented.

Chapter 3

A previous work on ABE

3.1 Introduction

In Chapter. 2, we have discussed about the mechanism of ABE and several common methods to realize it. As revealed in previous works on ABE, the estimation of wideband spectral envelope plays more important role than the estimation of wideband excitation signal. Therefore, compared to work on estimating excitation signal, more efforts have been put on the problem of estimating spectral envelope. In Section. 2.5, three methods based on codebook, GMM and HMM for spectral envelope extension have been presented. Codebook approach bases on hard clustering and discrete mapping, while GMM and HMM approaches use soft clustering and continuous mapping. In previous works on ABE, experiments result have shown that GMM and HMM approaches have better performance than codebook approach, but no significant difference have been found between GMM and HMM approaches. In our research, we focus on GMM approach and treat it as a target for comparison.

Although the GMM-based approach described in Section. 2.5.2 is reported to have relatively high performance, it still has problem that is considered to make the performance degraded. In [4, 8], the authors have pointed out the reason is due to ignoring an inter-frames correlation. Fig. 3.1 shows the trajectories of target and and converted features. We can see that although the two trajectories have similar shape, they still have several local differences, such as those marked with ellipses in the figure. In [4], in order to realize the feature correlation between frames, an algorithm has been proposed by introducing dynamic features. The estimation of wideband features is based on the Maximum Likelihood Estimation (MLE). In the next section, the MLE-GMM-based ABE method will be described.

3.2 MLE-GMM-based Bandwidth Extension [4, 8]

Let $\boldsymbol{x} = [\boldsymbol{x}_1^\top, \boldsymbol{x}_2^\top, ..., \boldsymbol{x}_N^\top]^\top$ be the feature vectors characterizing the narrowband and $\boldsymbol{y} = [\boldsymbol{y}_1^\top, \boldsymbol{y}_2^\top, ..., \boldsymbol{y}_N^\top]^\top$ be the feature vectors characterizing the wideband speech. $\boldsymbol{X}_t = [\boldsymbol{x}_t^\top, \Delta\boldsymbol{x}_t^\top]^\top$ and $\boldsymbol{Y}_t = [\boldsymbol{y}_t^\top, \Delta\boldsymbol{y}_t^\top]^\top$ define feature vectors consisting of static and dynamic features at frame t of narrowband and wideband speech, respectively.

In the training step, we model the joint probability density of the source and the target features by a GMM as follows:

$$P(\boldsymbol{Z}_t; \theta) = \sum_{m=1}^{M} \omega_m \mathcal{N}(\boldsymbol{Z}_t; \boldsymbol{\mu}_m^{(\boldsymbol{Z})}, \boldsymbol{\Sigma}_m^{(\boldsymbol{Z})}) \tag{3.1}$$

θ defines a parameter set of GMM, which consisting of weights, mean vectors and covariance matrices. M is the total number of mixture components of GMM, and m is GMM index.

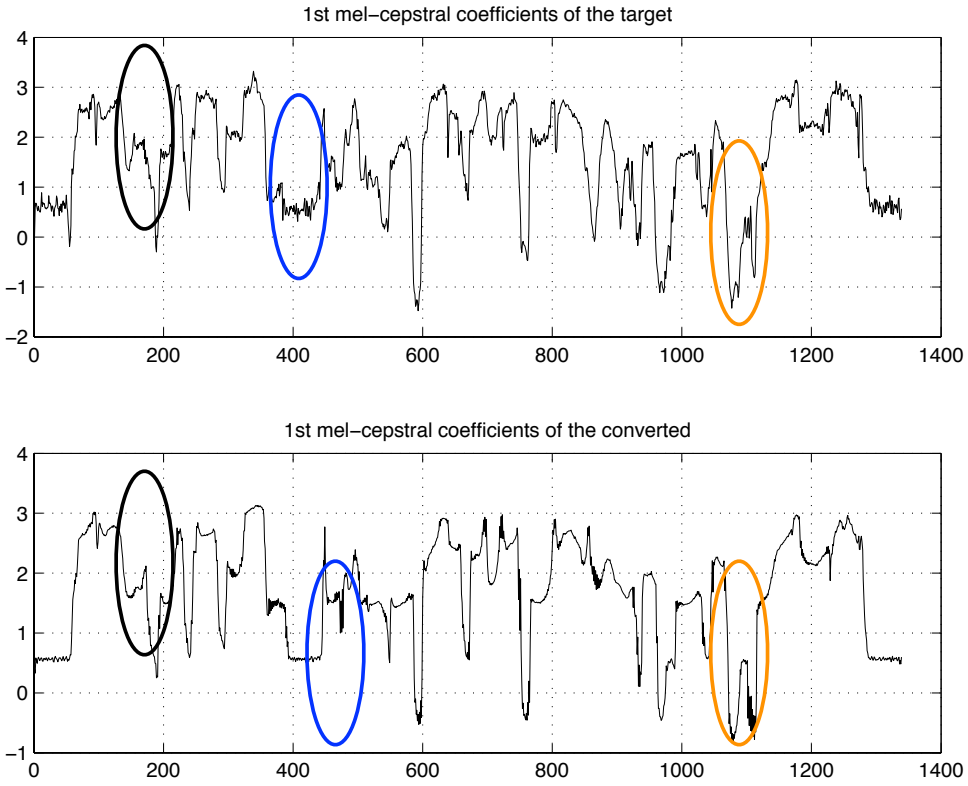


Fig.3.1: The difference of target mcep coefficients and converted mcep coefficients

The mean vectors and covariance matrices can be decomposed as below:

$$\boldsymbol{\mu}_m^{(\boldsymbol{Z})} = \begin{bmatrix} \boldsymbol{\mu}_m^{(\boldsymbol{X})} \\ \boldsymbol{\mu}_m^{(\boldsymbol{Y})} \end{bmatrix}, \boldsymbol{\Sigma}_m^{(\boldsymbol{Z})} = \begin{bmatrix} \boldsymbol{\Sigma}_m^{(\boldsymbol{XX})} & \boldsymbol{\Sigma}_m^{(\boldsymbol{XY})} \\ \boldsymbol{\Sigma}_m^{(\boldsymbol{YX})} & \boldsymbol{\Sigma}_m^{(\boldsymbol{YY})} \end{bmatrix} \tag{3.2}$$

The conditional probability $P(\boldsymbol{Y}_t^\top | \boldsymbol{X}_t^\top, m; \theta)$ is given by:

$$P(\boldsymbol{Y}_t^\top | \boldsymbol{X}_t^\top, m; \theta) = \mathcal{N}(\boldsymbol{Y}_t; \boldsymbol{E}_{m,t}^{(\boldsymbol{Y})}, \boldsymbol{D}_m^{(\boldsymbol{Y})}) \tag{3.3}$$

where

$$\boldsymbol{E}_{m,t}^{(\boldsymbol{Y})} = \boldsymbol{\mu}_m^{(\boldsymbol{Y})} + \boldsymbol{\Sigma}_m^{(\boldsymbol{YX})} \boldsymbol{\Sigma}_m^{(\boldsymbol{XX})^{-1}} (\boldsymbol{X}_t - \boldsymbol{\mu}_m^{(\boldsymbol{X})}) \tag{3.4}$$

$$\boldsymbol{D}_m^{(\boldsymbol{Y})} = \boldsymbol{\Sigma}_m^{(\boldsymbol{YY})} - \boldsymbol{\Sigma}_m^{(\boldsymbol{YX})} \boldsymbol{\Sigma}_m^{(\boldsymbol{XX})^{-1}} \boldsymbol{\Sigma}_m^{(\boldsymbol{XY})} \tag{3.5}$$

In the conversion step, firstly, we can write the time sequences of feature vectors of narrowband and wideband speech as follow:

$$\boldsymbol{X} = [\boldsymbol{X}_1^\top, \boldsymbol{X}_2^\top, ..., \boldsymbol{X}_N^\top]^\top \tag{3.6}$$

$$\boldsymbol{Y} = [\boldsymbol{Y}_1^\top, \boldsymbol{Y}_2^\top, ..., \boldsymbol{Y}_N^\top]^\top \tag{3.7}$$

A time sequence of the converted static feature vectors $\hat{\boldsymbol{y}} = [\hat{\boldsymbol{y}}_1^\top, \hat{\boldsymbol{y}}_2^\top, ..., \hat{\boldsymbol{y}}_N^\top]^\top$ is then calculated as follows:

$$\hat{\boldsymbol{y}} = \arg\max_{\boldsymbol{y}} P(\boldsymbol{m}|\boldsymbol{X}, \theta) P(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{m}; \theta)$$
$$\text{subject to} \boldsymbol{Y} = \boldsymbol{W}\boldsymbol{y} \tag{3.8}$$

where $\boldsymbol{W}$ is a matrix which maps a sequence of static features the corresponding sequence of static and dynamic features (see Fig. 3.2).

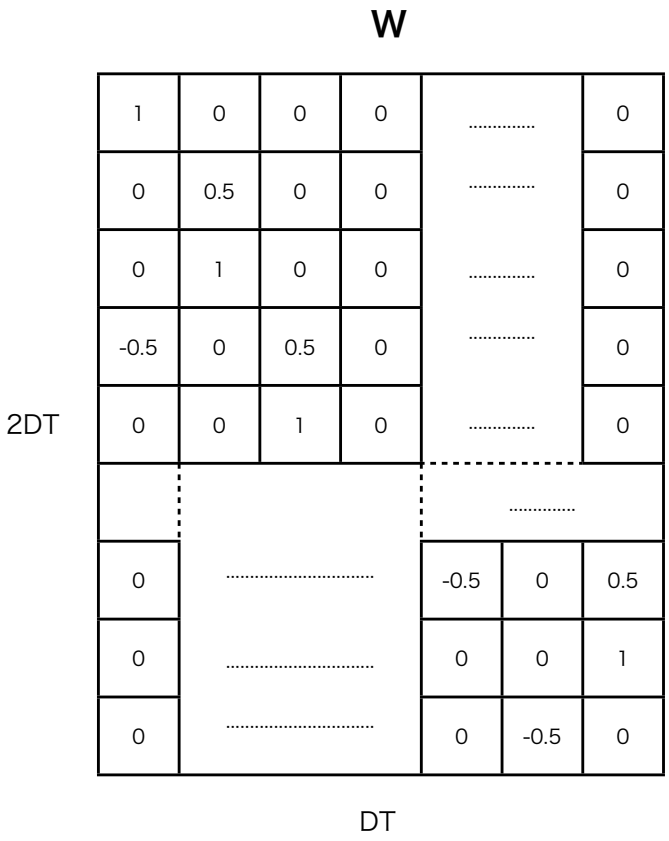


Fig.3.2: The $\boldsymbol{W}$ matrix which used to convert a sequence of static features to a sequence of static and dynamic features

This problem can be solved by EM algorithm, in which the following auxiliary function is iteratively maximized:

$$Q(\boldsymbol{Y}, \hat{\boldsymbol{Y}}) = \sum_{\text{all } m} P(m|\boldsymbol{X}, \boldsymbol{Y}, \theta) \log P(\hat{\boldsymbol{Y}}, m|\boldsymbol{X}; \theta) \tag{3.9}$$

The detailed calculation is omitted, but the first derivative of $Q(\boldsymbol{Y}, \hat{\boldsymbol{Y}})$ with respect to $\hat{\boldsymbol{Y}}$ becomes as below:

$$\frac{\partial Q(\boldsymbol{Y}, \hat{\boldsymbol{Y}})}{\partial \boldsymbol{y}} = -\boldsymbol{W}^\top \overline{\boldsymbol{D}^{(\boldsymbol{Y})^{-1}}} \boldsymbol{W}\boldsymbol{y} + \boldsymbol{W}^\top \overline{\boldsymbol{D}^{(\boldsymbol{Y})^{-1}} \boldsymbol{E}^{(\boldsymbol{Y})}} \tag{3.10}$$

To maximize $Q(\boldsymbol{Y}, \hat{\boldsymbol{Y}})$, above derivative must be 0, therefore the solution for Equation (3.8) is given by,

$$\hat{\boldsymbol{y}} = (\boldsymbol{W}^\top \overline{\boldsymbol{D}^{(\boldsymbol{Y})^{-1}}} \boldsymbol{W})^{-1} \boldsymbol{W}^\top \overline{\boldsymbol{D}^{(\boldsymbol{Y})^{-1}} \boldsymbol{E}^{(\boldsymbol{Y})}} \tag{3.11}$$

where $\overline{\boldsymbol{D}^{(\boldsymbol{Y})^{-1}}}, \overline{\boldsymbol{D}^{(\boldsymbol{Y})^{-1}} \boldsymbol{E}^{(\boldsymbol{Y})}}$ are defined as follows (see [8, 40] for more details):

$$\begin{aligned}
\overline{\boldsymbol{D}^{(\boldsymbol{Y})^{-1}}} &= \text{diag}[\overline{\boldsymbol{D}_1^{(\boldsymbol{Y})^{-1}}}, ..., \overline{\boldsymbol{D}_T^{(\boldsymbol{Y})^{-1}}}] \\
\overline{\boldsymbol{D}^{(\boldsymbol{Y})^{-1}} \boldsymbol{E}^{(\boldsymbol{Y})}} &= [\overline{\boldsymbol{D}_1^{(\boldsymbol{Y})^{-1}} \boldsymbol{E}_1^{(\boldsymbol{Y})}}, ..., \overline{\boldsymbol{D}_T^{(\boldsymbol{Y})^{-1}} \boldsymbol{E}_T^{(\boldsymbol{Y})}}] \\
\overline{\boldsymbol{D}_t^{(\boldsymbol{Y})^{-1}}} &= \sum_{m=1}^{M} \gamma_{m,t} \boldsymbol{D}^{(\boldsymbol{Y})^{-1}} \\
\overline{\boldsymbol{D}_t^{(\boldsymbol{Y})^{-1}} \boldsymbol{E}_t^{(\boldsymbol{Y})}} &= \sum_{m=1}^{M} \gamma_{m,t} \boldsymbol{D}^{(\boldsymbol{Y})^{-1}} \boldsymbol{E}_{m,t}^{(\boldsymbol{Y})} \\
\gamma_{m,t} &= P(m|\boldsymbol{X}_t^\top, \boldsymbol{Y}_t^\top; \theta)
\end{aligned} \tag{3.12}$$

3.3 Results

Fig. 3.3 shows an example of the converted trajectories of the conventional GMM method and the proposed MLE-GMM method [8]. While the converted trajectory of conventional GMM method shows inappropriate dynamic characteristics, the trajectory yielded by MLE-GMM method does not have those problems. A listening test result in [4] also pointed out that the MLE-GMM significantly outperforms the conventional GMM method.

3.4 Summary

In this chapter, a GMM-based method for estimating wideband spectral envelope with MLE criterion has been introduced. This approach was motivated by the using of static features together with their dynamic features to realize the inter-frames correlation. Significant improvement in experiment results (compared with traditional GMM method) have been reported in [4]. In our research, we will consider the MLE-GMM-based approach as a comparison target.

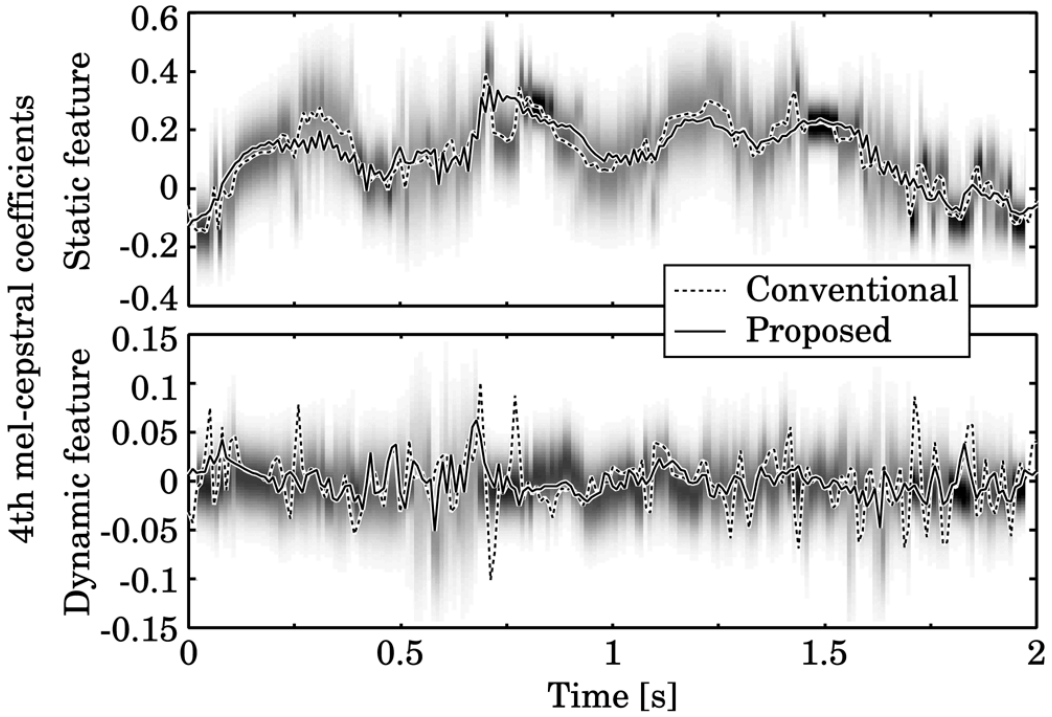


Fig.3.3: Converted trajectories by the conventional GMM and MLE-GMM methods

Chapter 4

Proposed algorithm based on REDIAL

4.1 Introduction

In previous chapters, we have presented the general framework of ABE and several approaches to the problem such as GMM, HMM method. Although these approaches are reported to regenerate wideband speech with relatively high quality, an obvious gap between the generated speech and the original speech is still recognized. Therefore, there is still space to improve the performance of ABE system.

We started our research by investigating "potential" drawbacks of conventional methods. As the statistical approaches showed much better performance than other approaches, we have decided to focus on them. As discussed in previous sections, the statistical approaches for ABE aim to make a mapping function between the narrowband and wideband features. This task is normally performed by using the narrowband (or narrowband & wideband) features to divide the feature space into sub-spaces, then estimating linear transformation functions in every sub-space. The conversion will be performed in each sub-space, and finally the estimated wideband features are obtained by summarizing all estimated features in each space. However, as the characteristics of narrowband feature space are normally different from those of wideband feature space, a mismatch in feature space probably occurs and consequently affect the accuracy of the conversion.

Speech enhancement which attempts to estimate the clean speech from an noisy input speech is a similar task to our ABE task. If we consider clean speech as wideband speech and noisy input speech as narrowband speech, the speech enhancement task becomes exactly the same as ABE task. In speech enhancement, Stereo-based Piecewise Linear Compensation for Environments (SPLICE), a statistical method works on the same mechanism as that of GMM, HMM approaches for ABE, is one of the most often-used algorithms. However, the same mismatch problem during the feature space division step was also reported [10]. To resolve this problem, in [10] proposed a new method called DIscriminative region weighting And Long-span features (REDIAL) by modifying the original SPLICE algorithm. Experiment results have shown the effectiveness of the REDIAL method. From the similarity of ABE and speech enhancement task, we hope that REDIAL will also work well in ABE task.

In this chapter, we first give an overview about SPLICE algorithm, and then describe our proposed REDIAL-based method. After that, we will introduce an ABE system which will be used in our experiments.

4.2 SPLICE algorithm for speech enhancement [9]

SPLICE is an effective and widely used approach in speech enhancement. Different from GMM approach, in which posterior probabilities of indexes of GMM of joint feature vectors were used for space division, SPLICE uses posterior probabilities of indexes of GMM of

corrupted input feature vectors.

Let $\boldsymbol{x}$ and $\boldsymbol{y}$ be N-dimensional feature vectors of clean speech and those of corrupted speech, respectively. In original SPLICE, an estimate $\hat{\boldsymbol{x}}$ of the clean speech feature is calculated as follows:

$$\hat{\boldsymbol{x}} = \sum_{k=1}^{K} p(k|\boldsymbol{y})\boldsymbol{A}_k\boldsymbol{y}', \tag{4.1}$$

where $\boldsymbol{y}' = [1, \boldsymbol{y}^\top]^\top$ is an augmented feature vector. $\boldsymbol{A}_k$ is an conversion matrix in region k which is trained as described below.

First, a K-component GMM is trained using corrupted feature vectors $\boldsymbol{y}_i$.

$$p(\boldsymbol{y}) = \sum_{k=1}^{K} \omega_k N(\boldsymbol{y}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \tag{4.2}$$

Next, the conversion matrix $\boldsymbol{A}_k$ is estimated using minimum mean square error criterion as follows:

$$\boldsymbol{A}_k = \arg\min_{\boldsymbol{A}_k} \sum_{i=1}^{I} p(k|\boldsymbol{y}_i)\|\boldsymbol{x}_i - \boldsymbol{A}_k\boldsymbol{y}_i'\|^2 \tag{4.3}$$

Define $f(\boldsymbol{A}_k)$ as below:

$$f(\boldsymbol{A}_k) = \sum_{i=1}^{I} p(k|\boldsymbol{y}_i)\|\boldsymbol{x}_i - \boldsymbol{A}_k\boldsymbol{y}_i'\|^2 \tag{4.4}$$

The first derivative of $f(\boldsymbol{A}_k)$ with respect to $\boldsymbol{A}_k$ must be equal to 0 to minimize the mean square error in Equation (4.4):

$$\frac{\partial f(\boldsymbol{A}_k)}{\partial \boldsymbol{A}_k} = 0 \tag{4.5}$$

With any symmetric matrix $\boldsymbol{W}$, we have the following formula:

$$\frac{\partial}{\partial \boldsymbol{A}}(\boldsymbol{x} - \boldsymbol{A}\boldsymbol{s})^\top \boldsymbol{W}(\boldsymbol{x} - \boldsymbol{A}\boldsymbol{s}) = -2\boldsymbol{W}(\boldsymbol{x} - \boldsymbol{A}\boldsymbol{s})\boldsymbol{s}^\top \tag{4.6}$$

Using the above formula, we have:

$$\frac{\partial\|\boldsymbol{x}_i - \boldsymbol{A}_k\boldsymbol{y}_i'\|^2}{\partial \boldsymbol{A}_k} = \frac{\partial(\boldsymbol{x}_i - \boldsymbol{A}_k\boldsymbol{y}_i')^\top \boldsymbol{E}(\boldsymbol{x}_i - \boldsymbol{A}_k\boldsymbol{y}_i')}{\partial \boldsymbol{A}_k} \tag{4.7}$$

$$= -2\boldsymbol{E}(\boldsymbol{x}_i - \boldsymbol{A}_k\boldsymbol{y}_i')\boldsymbol{y}_i'^\top \tag{4.8}$$

Hence, Equation (4.5) becomes:

$$\sum_{i=1}^{I} p(k|\boldsymbol{y}_i)\boldsymbol{E}(\boldsymbol{x}_i - \boldsymbol{A}_k\boldsymbol{y}_i')\boldsymbol{y}_i'^\top = 0 \tag{4.9}$$

$$\boldsymbol{X}\boldsymbol{P}\boldsymbol{Y}'^\top - \boldsymbol{A}_k\boldsymbol{Y}'\boldsymbol{P}\boldsymbol{Y}'^\top = 0 \tag{4.10}$$

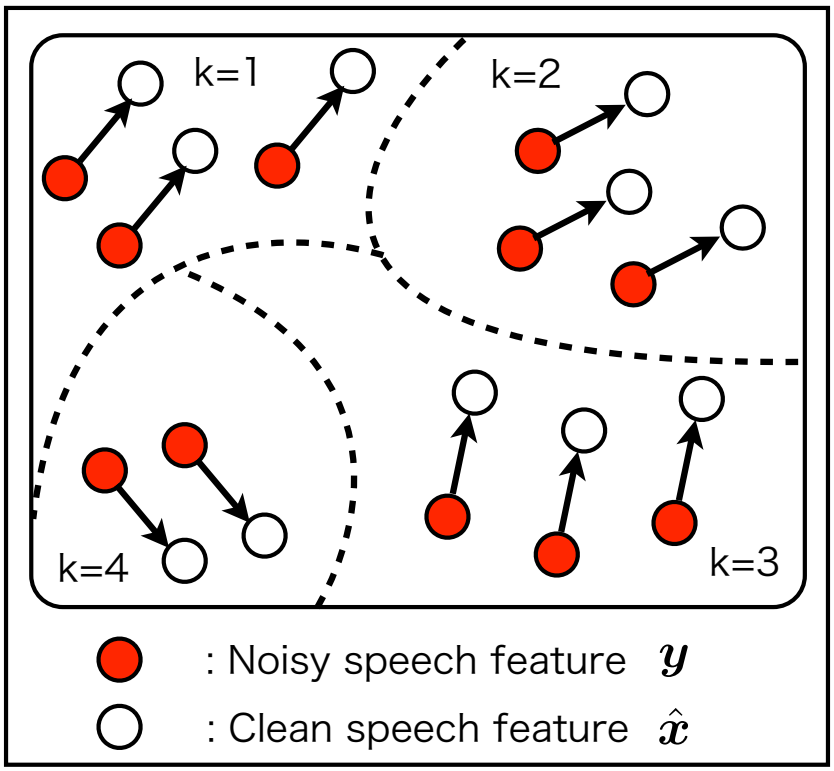


Fig.4.1: An conceptual diagram of SPLICE transformation from noisy feature $\boldsymbol{y}$ to estimated clean feature $\hat{\boldsymbol{x}}$. k is index of the sub-spaces

where,

$$\begin{aligned} \boldsymbol{X} &= [\boldsymbol{x}_1 \boldsymbol{x}_2 ... \boldsymbol{x}_I] \\ \boldsymbol{Y}' &= [\boldsymbol{y}'_1 \boldsymbol{y}'_2 ... \boldsymbol{y}'_I] \\ \boldsymbol{P} &= \mathrm{diag}\{p(k|\boldsymbol{y}_1), p(k|\boldsymbol{y}_2), ..., p(k|\boldsymbol{y}_I)\} \end{aligned} \tag{4.11}$$

Finally, solution of Equation (4.5) is:

$$\boldsymbol{A}_k = \boldsymbol{X}\boldsymbol{P}\boldsymbol{Y}'^{\top}(\boldsymbol{Y}'\boldsymbol{P}\boldsymbol{Y}'^{\top})^{-1} \tag{4.12}$$

An estimate $\hat{\boldsymbol{x}}$ of the clean speech feature is now calculated by substituting $\boldsymbol{A}_k$ into Equation (4.1). Fig. 4.1 illustrates the mechanism of SPLICE transformation.

4.3 DIscriminative region weighting And Long-span features (REDIAL) [10, 11]

REDIAL was first proposed in [10] for speech enhancement, in which a joint vector $[\boldsymbol{y}^{\top}, \hat{\boldsymbol{n}}^{\top}]^{\top}$, consisting of a corrupted feature vector $\boldsymbol{y}$ and an estimate vector $\hat{\boldsymbol{n}}$ of noise feature vector, is used instead of the corrupted vector $\boldsymbol{y}$ alone. Moreover, a discriminative

model (LDA + GMM) is introduced to space division step to calculate the posterior probabilities of the clean feature GMM using the corrupted features. The estimation of clean feature vector becomes:

$$\hat{\boldsymbol{x}} = \sum_{k=1}^{K} p(k|\boldsymbol{L}[\boldsymbol{y}^\top, \hat{\boldsymbol{n}}^\top]^\top)\boldsymbol{A}_k[1, \boldsymbol{y}^\top, \hat{\boldsymbol{n}}^\top]^\top, \tag{4.13}$$

where $\boldsymbol{L}$ is a conversion matrix of LDA trained by using joint vectors $[\boldsymbol{y}^\top, \hat{\boldsymbol{n}}^\top]^\top$ with posterior probabilities of indexes $\{k\}$ of clean GMM as their labels. The conversion matrix $\boldsymbol{A}_k$ is estimated as below:

$$\boldsymbol{A}_k = \arg\min_{\boldsymbol{A}_k} \sum_{i=1}^{I} p(k|\boldsymbol{v}_i)\|\boldsymbol{x}_i - \boldsymbol{A}_k[1, \boldsymbol{y}^\top, \hat{\boldsymbol{n}}^\top]^\top\|^2 \tag{4.14}$$

where $\boldsymbol{v}_i = \boldsymbol{L}[\boldsymbol{y}^\top, \hat{\boldsymbol{n}}^\top]^\top$ are converted vectors of LDA. By using LDA, dimensionality of feature vectors can be reduced effectively. Moreover, using clean GMM indexes as labels of LDA is expected to improve the overall performance, since the purpose of speech enhancement is to estimate feature vectors in clean space. The effectiveness of this has been shown in [10].

In addition to the method described above, the authors also considered using a joint vector of several adjacent frames instead of feature vector of only a single frame. However, to avoid the over-fitting problem that might occur since the vectors dimensionality increases, regularization was used. By concatenating adjacent frames features, the input information increases, therefore an improvement in estimation of clean feature is expected. In [11], the authors have confirmed the effectiveness of this method.

4.3.1 Proposed method: REDIAL-based Bandwidth Extension

In this research, we adopted the method explained in Section. 4.3, to solve the problem of spectral envelope extension. Its detailed procedure is explained below:

1. Extracting feature vectors $\{\boldsymbol{x}_i\}_{i=1,\ldots,I}$ of wideband speech, and $\{\boldsymbol{y}_i\}_{i=1,\ldots,I}$ of narrowband speech. Define $\boldsymbol{v}_i$ is a joint vector of several feature vectors of frames adjacent to frame i.
2. Training a GMM using wideband feature vectors $\{\boldsymbol{x}_i\}_{i=1,\ldots,I}$ and calculate $\{p(m|\boldsymbol{x}_i)\}_{i=1,..,I}$.
3. Training LDA using joint feature vectors $\{\boldsymbol{v}_i\}_{i=1,\ldots,I}$ with $\{p(m|\boldsymbol{x}_i)\}_{i=1,..,I}$ as their class labels. After obtaining the conversion matrix $\boldsymbol{L}$ of LDA, calculate converted vectors $\boldsymbol{z}_i = \boldsymbol{L}\boldsymbol{v}_i$.
4. Training a GMM using the converted vectors $\boldsymbol{z}_i$ and calculate probability $p(k|\boldsymbol{z}_i)$.
5. The linear conversion matrix $\boldsymbol{A}_k$ is estimated using a weighted minimum mean square

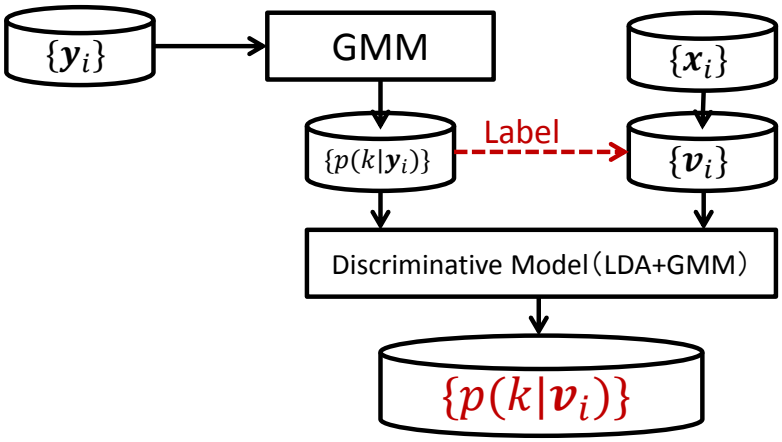


Fig.4.2: Illustration of the space division step in REDIAL method

error criterion with regularization as below:

$$\begin{aligned} \boldsymbol{A}_k &= \arg\min_{\boldsymbol{A}_k} \sum_{i=1}^{I} p(k|\boldsymbol{z}_i)\|\boldsymbol{x}_i - \boldsymbol{A}_k\boldsymbol{v}_i - \boldsymbol{\mu}_k\|^2 \\ &+ \lambda\|\boldsymbol{A}_k\|^2, \end{aligned} \tag{4.15}$$

where $\boldsymbol{\mu}_k$ is mean of component k of the GMM of wideband feature vectors and λ is regularization parameter. Solution to this problem is given by:

$$\boldsymbol{A}_k = \boldsymbol{X'}\boldsymbol{P}\boldsymbol{Y'}^{\top}(\boldsymbol{Y'}\boldsymbol{P}\boldsymbol{Y'}^{\top} + \lambda\boldsymbol{E})^{-1}, \tag{4.16}$$

where $\boldsymbol{X'}$ is the sequence of feature vectors $\boldsymbol{x}'_i = \boldsymbol{x}_i - \boldsymbol{\mu}_k$, and $\boldsymbol{Y'}$ is the sequence of joint feature vectors $\boldsymbol{v}_i$. $\boldsymbol{P}$ is a diagonal matrix given by $\boldsymbol{P} = \mathrm{diag}([p(k|\boldsymbol{z}_1), ..., p(k|\boldsymbol{z}_I)])$.

6. Finally, estimation of wideband feature vector given the narrowband one is as follow:

$$\hat{\boldsymbol{x}}_i = \sum_{k=1}^{K} p(k|\boldsymbol{z}_i)(\boldsymbol{A}_k\boldsymbol{v}_i + \boldsymbol{\mu}_k) \tag{4.17}$$

The process of calculating probabilities $p(k|\boldsymbol{z}_i)$ is normally referred to as Space division step. Fig. 4.2 gives a simple diagram for this step.

4.4 Baseline Bandwidth Extension System

4.4.1 STRAIGHT vocoder

As explained in previous chapters, the ABE task is separated to two sub-tasks, so-called estimation of wideband spectral envelope and estimation of wideband excitation signal,

based on source-filter model. As introduced in Chapter. 2 and Chapter. 3, a variety of statistical approaches have been proposed for the problem of estimating wideband spectral envelope, and have achieved successes on some level. However, the estimation of wideband excitation is still not good enough, and this leads to a converted speech with unreasonable noise. To reduce this effect, it is desired to have a method to statistically estimate the wideband excitation signal as that used for estimating spectral envelope.

Recently, a vocoder called **STRAIGHT** [41, 42] has been developed based on source-filter model and is reported to produce high quality synthetic speech. Its simplicity in analyzing and synthesizing speech has made it becoming a powerful speech research tool. STRAIGHT analyzes the speech and divide it into two parts: one is STRAIGHT spectrum (sp), and the other is mixed excitation signal which consists of F0 and aperiodic components (ap). The mixed excitation signals obtained from STRAIGHT analysis can be statistically modeled. Therefore, it is expected to be used in the same manner as spectral envelope.

In [4], the authors used STRAIGHT as analysis tool, and applied GMM-based conversion approach to both spectral envelope and aperiodic components. Listening test results indicated the high perceived quality of the estimated wideband speech. In our research, we also use STRAIGHT for analyzing and synthesizing speech signal. In the next section, we will introduce the ABE system which utilizes STRAIGHT vocoder.

4.4.2 Baseline Bandwidth Extension System

The general process of ABE is shown in Fig. 4.3. First, mel-cepstral coefficients, aperiodic components and F0 of the narrowband speech are extracted using STRAIGHT [41, 42] (*Step 1*). Aperiodic components of the wideband speech are estimated by a simple MMSE-based GMM mapping method [3] (*Step 2*). Mel-cepstral coefficients which represent the spectral envelope are estimated by performing feature transformation as described in Section. **??** and Section. 4.3.1 (*Step 2*). After that, an estimated wideband speech is generated using the extracted F0 and converted features above (*Step 3*). The estimated wideband speech is now passed through LPF and HPF to generate low-band and high-band speech signals (*Step 4*). For the input narrowband speech, we up-sample it to make an input low-band speech signal (*Step 5*). Next, the power of the estimated high-band speech signal is adjusted so that the power of the estimated low-band and input low-band speech signal are equal (*Step 6*). Finally, the wideband speech is reconstructed by adding the adjusted high-band speech signal to the input low-band speech signal (*Step 7*).

The power terms of mel-cepstral coefficients are not used in the conversion, and are estimated by a linear transformation as follows:

$$\log y = \frac{\partial_y}{\partial_x}(\log x - \mu_x) + \mu_y \tag{4.18}$$

where x, y are input and output features and $\mu_x, \mu_y, \partial_x, \partial_x$ represent means and variances of the power terms of every samples.

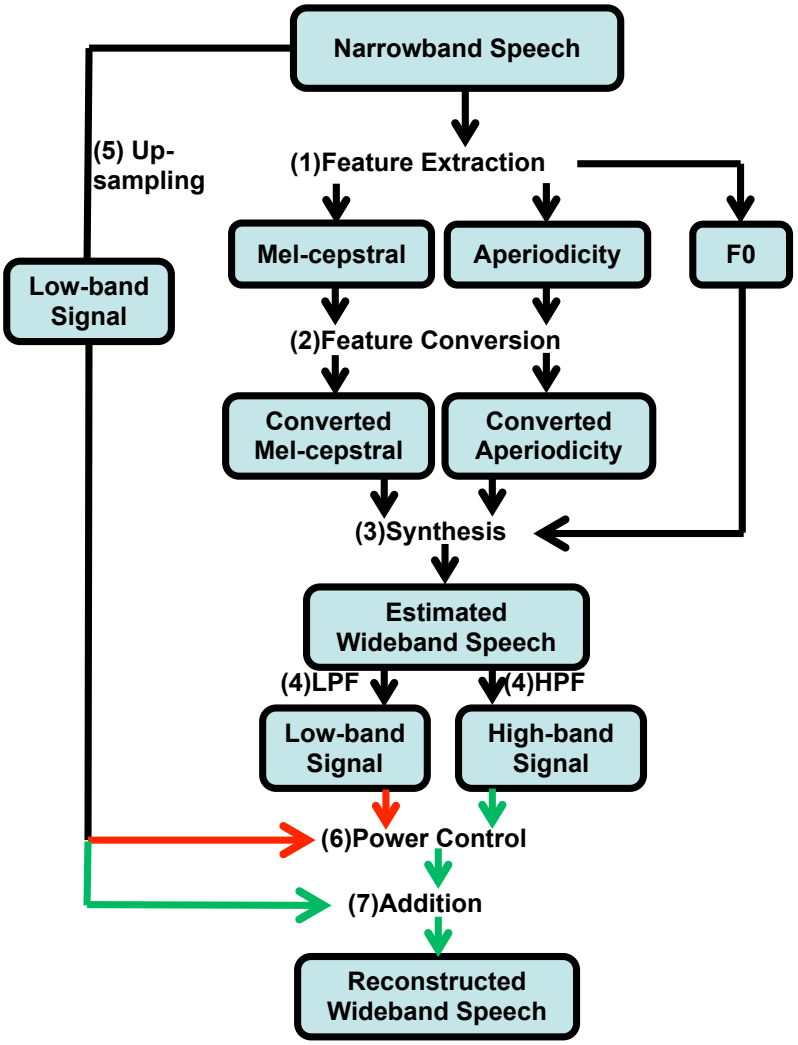


Fig.4.3: General flowchart of bandwidth extension

4.5 Summary

In this section, we have discussed about the drawback of conventional method for estimating wideband spectral envelope. Then we introduced our REDIAL-based approach to deal with this problem. Finally, a ABE system using STRAIGHT vocoder has been explained. In later experiments, we will use this system to verify the effectiveness of our proposed method.

Chapter 5

Experiments and Results

5.1 Introduction

Speech quality is a multi-dimensional term which is composed of several factors such as speech intelligibility, speech naturalness, etc. In general, the evaluation of speech quality is conducted by using objective and subjective measurements. Subjective measurements are carried out with listeners, who assess the quality of speeches from "subjective" view. When a relative number of listeners and good setting of subjective experiments are satisfied, the subjective evaluation results reflect quite accurately the speech quality. However, this is normally time consuming, expensive and labor intensive process, therefore objective measurement have been proposed as an alternative method. In objective measurement, speech quality of two speech samples are compared by calculating a distance between the two signals in either time or frequency domain. Unfortunately, objective measurements do not always show high correlation with the subjective measurements, and therefore does not always reflect perceived speech quality accurately.

In our research, we used both subjective and objective measurements to evaluate the performance of the proposed method. In the next sections, subjective and objective measurements of speech quality will be explained. Then experiments and their results will be discussed.

5.2 Subjective measurement

Several methods have been introduced for subjective measurement of speech quality such as MUltiple Stimuli with Hidden Reference and Anchor (MUSHRA) [37], Mean Opinion Score (MOS) [38], etc. In evaluation of ABE problem, MOS is often adopted. In our experiments, we also used MOS as subjective measurement.

MOS is a method described in ITU-T Recommendation P.800.1. In MOS, the speech quality is rated from 1 (bad) to 5 (excellent) by listeners. Table 5.1 shows the details of MOS scale. After obtaining scores of every individual listeners, the average MOS score for the target speech signal will be calculated. In order to achieve a reliable result for a MOS test, following requirements should be satisfied.

1. Large number of listeners: Each individual might have different evaluation even on the same speech sample. To reduce this variation, large pool of listeners is required (more than 10).
2. Controlled conditions: The subjective tests should be conducted under well controlled conditions such as quiet environment. Equipments should be high-fidelity, and if we use plural experiment sets, they (headphones, etc.) should have the same characteristics.

Table5.1: Mean Opinion Score (MOS) scale

MOS	Speech quality
5	Excellent
4	Good
3	Fair
2	Poor
1	Bad

5.3 Objective measurement

There are several measures that investigate how close the estimated envelope is to the original wideband envelope such as Mel-cepstral Distance (MCD), Log Spectral Distortion (LSD), Itakura Distance, Itakura-Saito Distance. In our experiments, we adopted the Mel-cepstral Distance (MCD). In fact, mel-cepstrum has been shown to be a compact representation of perceptually relevant speech characteristics, and the MCD measure results have also been certified to have high correlation with the subjective test results [39]. The MCD distance is defined as below:

$$MCD[dB] = \frac{10}{\ln 10}\sqrt{2\sum (mc_i^X - mc_i^Y)^2} \tag{5.1}$$

where mc^X, mc^Y are mel-cepstral coefficients of regenerated wideband speech and natural wideband speech, respectively.

5.4 ABE with speaker dependent model

5.4.1 Experiment Conditions

We conducted experiments under a speaker-dependent condition using the ATR phonetically balanced corpus [43]. The wideband speeches were 16kHz sampled speeches from subset A (training data) and subset B (evaluation data) of 4 Japanese speakers (2 males and 2 females). Each training set and evaluation set includes 50 speech samples. The narrowband speech was made by passing the corresponding wideband one through a LPF as in Fig. 5.1 (LPF specifications are described in Table 5.2), then downsampling the output.

In our experiments, we used STRAIGHT to extract mel-cepstral coefficients (spectral envelope) and F0, aperiodic components (*mixed excitation signal*). Regarding to aperiodic components, the averaged components on 3 frequency bands (0 - 1, 1 - 2 and 2 - 4 kHz) for narrowband, and those on 5 frequency bands (0 - 1, 1 - 2, 2 - 4, 4 - 6 and 6 - 8 kHz) for wideband were used.

Table5.2: LPF specifications

Stop band	3.7 - 8kHz
Transition band	3.4 - 3.7kHz
Pass band	0Hz - 3.4kHz

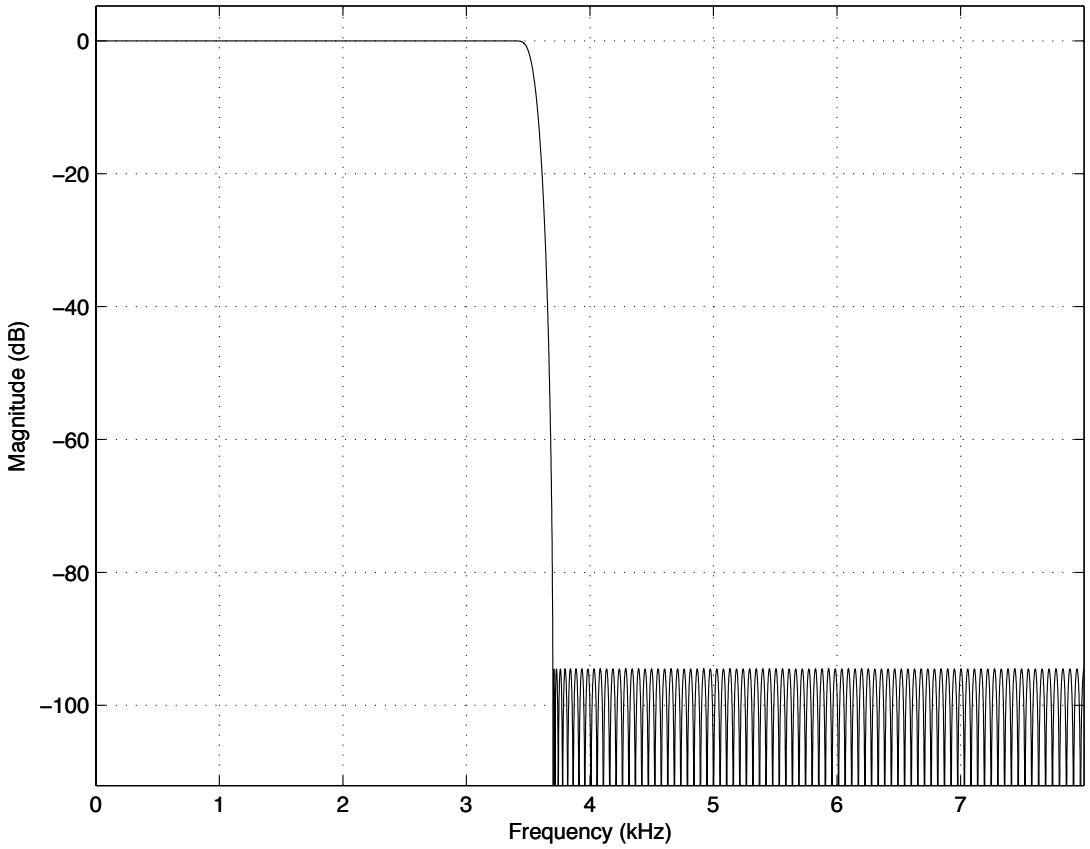


Fig.5.1: LPF magnitude response

In this paper, we adopted a simple MMSE-based GMM mapping method [3] for extension of the excitation signal. The number of mixture components of the GMM was set to 8. For extension of the spectral envelope, we used a 64-component GMM in both conventional and proposed methods.

5.4.2 Preliminary experiments

As explained in Section. 4.3, the proposed REDIAL-based approach concatenates several adjacent frames to form a feature vector and use it as the new feature. Besides, the method also introduces regularization to cope with the over-fitting problem. The performance is considered to vary with these parameters (frames number, regularization parameter).

Table5.3: Objective evaluation of REDIAL-based method considering the change in dimension of feature vectors. Dim 24 & 24 represents the case when 24-dimensional mcep vectors are used for both wideband and narrowband speeches. Dim 24 & 16 represents the case when the dimensions of wideband and narrowband features are 24 and 16 respectively.

	Speaker	ftk	fws	mmy	msh
MCD [dB]	Dim 24 & 24	1.95	1.88	1.86	1.87
	Dim 24 & 16	2.05	2.03	2.00	2.12

Therefore, it is essential to determine values giving the best performance. Regarding the mel-cepstral coefficients, there are also two parameters which require optimization: dimension of feature vectors and warping parameter which used to extract mel-cepstral coefficients.

In this section, we will perform several preliminary experiments to determine these parameters.

i) Dimension of feature vectors

In experiments described in [4], the dimensions of wideband features and narrowband features were set to 24 and 16 respectively. In our research, we performed experiments to examine if this setting is suitable for the ABE task. Specifically, we performed two kinds of experiments: in the first experiment, the dimensions of wideband and narrowband features were 24 and 16 respectively; in the second experiment, the dimensions of both wideband and narrowband features were set to 24. All other parameters and analysis condition were the same. The experiments were done using REDIAL-based method since we aim to get the best performance of this method.

Table 5.3 shows the results of objective evaluation. The results suggest that, it is better to use mel-cepstral coefficients with dimension 24 for both wideband and narrowband speeches in the proposed REDIAL-method.

ii) Warping parameters

As mentioned in SPTK manual book [1], the warping parameter α represents the phase characteristics and it is recommended to adjust this parameter in accordance with the sampling rate. In this section, we operated two experiments: one with the recommended warping parameters for each sampling rate ($\alpha = 0.42$ for wideband and $\alpha = 0.31$ for narrowband), and the other with a fixed warping parameter ($\alpha = 0.42$ for both wideband and narrowband speech).

[1] http://nyftp.netbsd.org/pub/pkgsrc/distfiles/SPTKref-3.6.pdf

Table5.4: Mel-cepstral distortion between regenerated speech and original speech when using $\alpha = 0.42$ for wideband and $\alpha = 0.31$ for narrowband speeches

	Speaker	ftk	fws	mmy	msh
MCD [dB]	Dim 24 & 24	3.59	3.70	3.51	3.43
	Dim 24 & 16	5.34	4.70	5.76	6.23

Table 5.4 shows the objective evaluation results when using $\alpha = 0.42$ for wideband and $\alpha = 0.31$ for narrowband speeches. The objective evaluation results when $\alpha = 0.42$ is used for both wideband and narrowband speech are the same in Table 5.3. Comparing these two tables, it can be concluded that using warping parameter $\alpha = 0.42$ for both wideband and narrowband features has better performance.

iii) Number of frames to be concatenated

Concatenating adjacent frames and using it as new features is expected to increase input information and consequently improve the precision of the estimation. However, the optimal number of frames to be concatenated varies depending on the contents of the tasks. In this section, we performed experiments to find out which is the optimal number for the ABE task.

According to the results of above experiments, in this experiment we used 24-dimension feature vectors for both wideband and narrowband speeches. The warping parameters for both of them were set to 0.42. Regarding the regularization parameter, we fixed it with value 0.002. We set the number of frames to be concatenated to 3, 5, 7, 9 and operated experiments for each case.

Table 5.5 shows the objective evaluation results of each case. We can observe that the performances varied dramatically between different speakers in the same case or between different cases. This is because the regularization parameters were not optimized for each speaker. However, looking at the average, we can conclude that when the number of frames to be concatenated was 5, the highest performance was achieved. Therefore, in the following experiments, we decided to use 5 as the number of concatenated frames.

iv) Regularization parameter

As a summary of 3 preliminary experiments above, to achieve the best performance of REDIAL-based method: 1. The dimension of both narrowband and wideband features should be 24; 2. The warping parameter for both narrowband and wideband speeches should be 0.42; 3. The number of frames to be concatenated should be 5. In later experiments, we always used this setting.

Now, we will investigate the optimal regularization parameters λ for each speaker in the

Table5.5: Mel-cepstral distortion between regenerated speech and original speech in cases with different numbers of frames to be concatenated

	Speaker	ftk	fws	mmy	msh
Number of frames	3	3.74	4.35	2.67	2.59
	5	3.49	3.97	2.02	2.00
	7	5.54	10.08	4.23	2.63
	9	5.44	5.73	2.80	2.83

Table5.6: Optimal regularization parameters in a speaker-dependent condition

Speaker	ftk	fws	mmy	msh
λ	0.003	0.009	0.003	0.002

training set. The regularization parameter was chosen by 5-fold cross validation: training data was divided equally into 5 subsets, then 4 subsets were used as training data and the left one was used as testing data. In general, the regularization parameter is normally small number, therefore we set the search range of regularization λ in the following set of values $\{0.001, 0.002, 0.003, 0.004, 0.005, 0.006, 0.007, 0.008, 0.009, 0.01, 0.02, 0.04, 0.08, 0.1\}$. From experiment results, the optimal regularization parameter for each speaker was determined as shown in Table 5.6.

5.4.3 Objective Evaluation

We conducted experiments using parameters optimized through the preliminary experiments above. In this section, we performed objective evaluation as explained in Section. 5.3. The objective evaluation results for the 4 speakers are shown in Table 5.7.

An approximate 50% reduction in MCD can be seen for every speaker. This demonstrates the superiority of proposed method to the conventional one.

Table5.7: Objective evaluation (Speaker-dependent): Mel-cepstral distortion between regenerated speech and original speech

	Speaker	ftk	fws	mmy	msh
MCD [dB]	GMM	3.59	3.70	3.51	3.43
	REDIAL	1.95	1.88	1.86	1.87

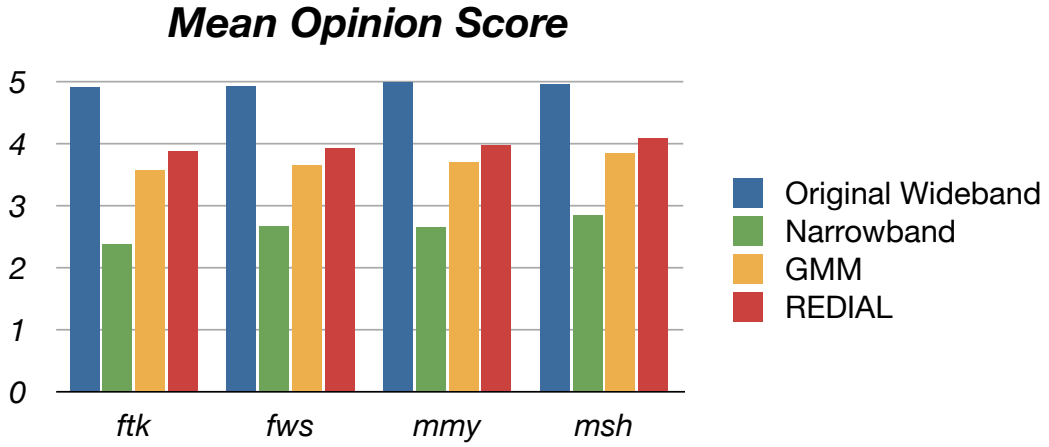


Fig.5.2: Speaker-dependent: Listening test results

5.4.4 Subjective Evaluation

Subjective evaluation was also conducted using MOS method as explained in Section. 5.2. The evaluation data of each speaker contains 10 sets of speeches. Each set consists of original wideband speech, narrowband speech, GMM-based regenerated wideband speech and REDIAL-based regenerated wideband speech. 21 listeners (15 males, 6 females; ages from 18-25) were asked to grade the speeches. All of them were asked to do experiment in quite environment and use SONY MDR-900ST headphone which has frequency response range is 5 Hz - 30 kHz. Listening test results are shown in Fig. 5.2. The reconstructed wideband speeches in both approaches showed better perceptual quality than the original narrowband. Moreover, listening test results also demonstrate that the proposed method significantly outperforms the conventional MLE-GMM approach (at significance level of 5%). The improvement was observed in male as well as in female in both objective and subjective evaluation. This suggests that our proposed method is applicable to both male and female. Fig. 5.3 shows an example of spectrograms of an original wideband speech and the resynthesizes speeches based on GMM and REDIAL approaches. Looking at areas marked with ellipses, we can see that compared to MLE-GMM-based approach, REDIAL tends to regenerate the wideband speech with higher accuracy.

5.5 ABE with speaker independent model

5.5.1 Experiment Conditions

The effectiveness of the proposed method within a speaker-dependent condition was described in the previous section. In this section, we further verify the effectiveness of proposed method in a more practical condition, speaker-independent condition, using the

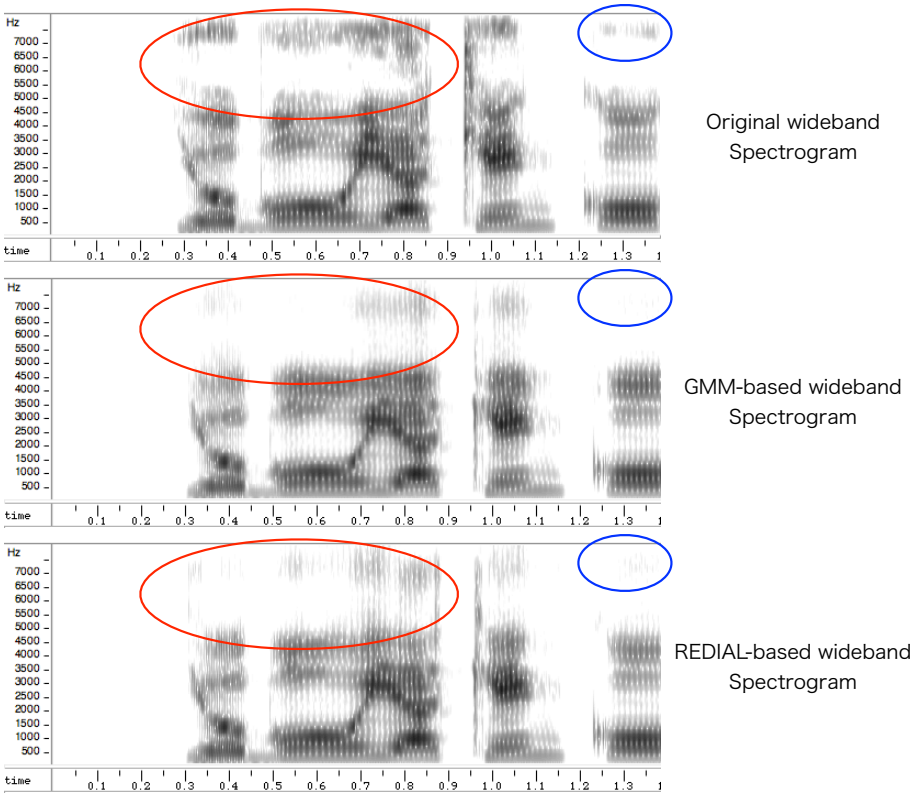


Fig.5.3: Spectrograms of an original wideband speech and its resynthesizes wideband speeches based on GMM and REDIAL-based methods in the speaker dependent model

TIMIT database [44]. The training set contains a total of 4620 utterances of 462 speakers, and the test set contains 1680 utterances of 168 speakers. Feature extraction and other analysis conditions were the same as those in speaker-dependent experiment, except the number of mixture components of GMM for spectral envelope being set to 256 instead of 64.

Similar to the speaker dependent case, we performed 8-fold cross validation to find the optimal regularization parameter λ. From the cross validation results, the parameter λ was set to 0.1. In subjective evaluation, we used 40 sets of speech samples (each contained the original wideband, narrowband, GMM-based wideband, SPLICE-based wideband and REDIAL based wideband speeches). The headphone and other experiment conditions were the same as in subjective evaluation of speaker dependent case.

5.5.2 Experiments

Results of objective evaluation and subjective evaluation of 16 listeners (12 males, 4 females; ages from 20 to 30) are shown in Table 5.8 and Fig. 5.4 respectively.

It can be concluded that the original SPLICE showed slightly better performance than the conventional GMM-based method, while the proposed REDIAL-based method signifi-

Table5.8: Objective evaluation(Speaker-independent): Mel-cepstral distortion between re-generated speech and original speech

Method	GMM	SPLICE	REDIAL
MCD[dB]	4.127	3.485	2.231

cantly outperforms both of them.

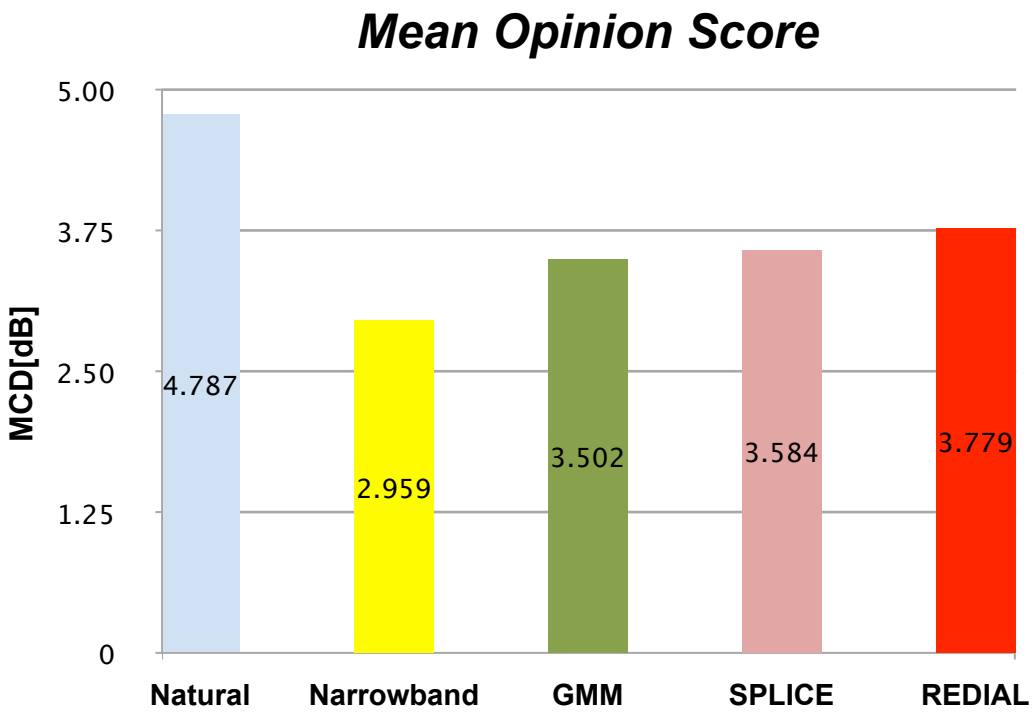


Fig.5.4: Speaker-independent: Listening test results

Similar to the speaker-dependent case, in this subjective evaluation we also observed a remarkable improvement in speech quality of reconstructed wideband as compared to the original narrowband speech in all of three methods. More importantly, with the proposed method we achieved reconstructed wideband speech with significantly better speech quality compare to conventional GMM and SPLICE-based methods (at significance level of 5%). An example of spectrograms of an original wideband speech and the resynthesizes speeches based on GMM and REDIAL approaches is shown in Fig. 5.5. A similar tendency in the difference of accuracy of the two approaches was also observed in this example.

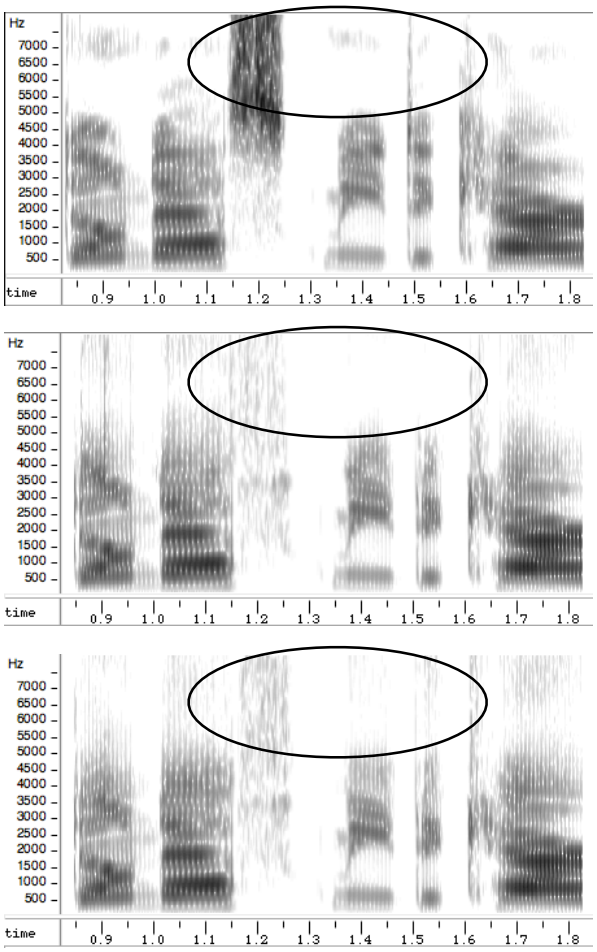


Fig.5.5: Spectrograms of an original wideband speech and its resynthesizes wideband speeches based on GMM and REDIAL-based methods in the speaker independent model

5.6 REDIAL-bases approach with dynamic features

As explained in Chapter. 3.2, the dynamic features were used together with the static features in the MLE-GMM-based approach. In the experiments described above, we used dynamic features for the MLE-GMM-based approach but not used them for the proposed REDIAL-based method. Therefore, in this section we conducted experiments including dynamic features into the proposed method to examine how these features affect the performances.

Firstly, for the speaker dependent case, we set the regularization parameter λ to 0.001, 0.01 and 0.1, and conducted the objective evaluations. For comparison, we conducted experiment without using dynamic features with λ being set to 0.1 in this case.

The objective evaluation results of the two types of experiments are shown in Table 5.9 and Table 5.10 respectively.

Secondly, we also performed experiments using dynamic features with REDIAL-based

Table5.9: Mel-cepstral distortion in speaker dependent experiments with ATR database: Utilizing dynamic features with the proposed REDIAL-based method

Spk/α	ftk	fws	mmy	msh
0.001	2.25	5.04	5.03	5.05
0.01	2.44	5.04	5.02	5.05
0.1	2.45	5.05	5.02	5.05

Table5.10: Mel-cepstral distortion in speaker dependent experiments with ATR database: Without utilizing dynamic features with the proposed REDIAL-based method

Spk/α	ftk	fws	mmy	msh
0.001	2.49	3.96	2.03	2.00

method in speaker independent model with TIMIT database. The experiment conditions are the same with those in speaker dependent model case. Table 5.11 and Table 5.12 show objective evaluation results of experiments with dynamic features and without dynamic features.

From the above results, it can be concluded that utilizing dynamic features with REDIAL-based method has degraded its performance. In the conventional MLE-GMM method, a usage of dynamic features was proposed to deal with the inter-frame correlation problem. In REDIAL-based method, this problem is resolved, since the method concatenates adjacent frames. Therefore, in REDIAL method, even if dynamic features are introduced, the performance might not be improved. From the results we could see that dynamic features do not work well with the proposed method, and hence in experiments from now on, we will not use them.

Table5.11: Mel-cepstral distortion in speaker indedependent experiments with TIMIT database: Utilizing dynamic features with the proposed REDIAL-based method

α	MCD
0.001	2.295
0.01	2.294
0.1	2.294

Table5.12: Mel-cepstral distortion in speaker dependent experiments with ATR database: Without utilizing dynamic features with the proposed REDIAL-based method

α	MCD
0.1	2.231

5.7 Experiments when training data number varies

In all experiments with speaker dependent model described above, the number of training data for each speaker was 50. As from both objective and subjective evaluations results, the regenerated wideband speeches were in high quality compared to the original narrowband one. However, we wondered how the performance might change when the number of training data decreases. In this section, we performed experiments in speaker dependent model with number of training data set to 10, 20, 30, 40. Besides changing the training data number, we also changed the mixture number of GMM used in both conventional and proposed methods to investigate the effects of GMM features on the performance.

The experiments results are shown in Fig. 5.6 (for the conventional the MLE-GMM method), and Fig. 5.7 (for the proposed REDIAL method). It can be observed that in the MLE-GMM case, when the number of training data is fixed, the performance became better when GMM mixture increased. Similarly when the GMM mixtures were constant, MCD decreased when the number of training data increased. The best results were found in case with 64-component GMM and 50 training data. Meanwhile, in the REDIAL case, no similar trend was observed. This might be due to the fact that regularization parameters were not well optimized (they were fixed). Overall, except some aberrant cases, the most ineffective scenario was found to be with 16-component GMM and 10 training data.

The objective evaluations described above have shown that decreasing the number of training data and GMM mixtures degraded the performance of the estimation. For subjective evaluation, I personally conducted an informal listening test to compare the original narrowband speech, the estimated wideband speeches in optimized condition (i.e 50 training data, 64-component GMM), and the regenerated wideband speeches with small number of training data and GMM mixtures (i.e 10 training data, 20 training data with 16-component GMM). It was revealed that, though the experiment was preliminary, the artificial wideband speeches were still more preferred than the original narrowband speech. Moreover, the difference between speeches estimated under the worst condition (i.e 10 training data, 16-component GMM) and those estimated under the best condition were extremely minor. This result indicated that for speaker dependent case, we can use a smaller corpus with smaller GMM mixtures to perform the estimation.

5.8 Summary

In this chapter, we have discussed about objective and subjective evaluations which used to assess speech quality. After that, we have described experiments under a variety of conditions, such as speaker dependent condition and speaker independent condition. The experiments results have pointed out that our proposed method for ABE significantly outperformed the conventional MLE-GMM method.

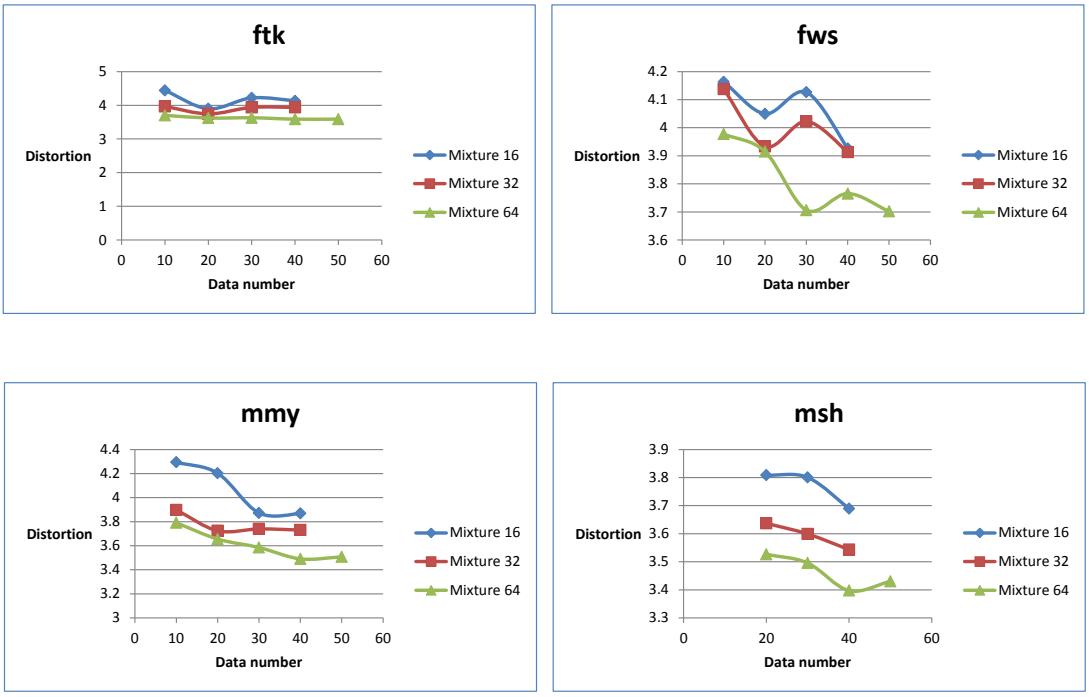


Fig.5.6: Mel-cepstral distortion between the resynthesizes wideband speech using the MLE-GMM-based method and original wideband speech. In these experiments, the number of training data and GMM mixtures are varied.

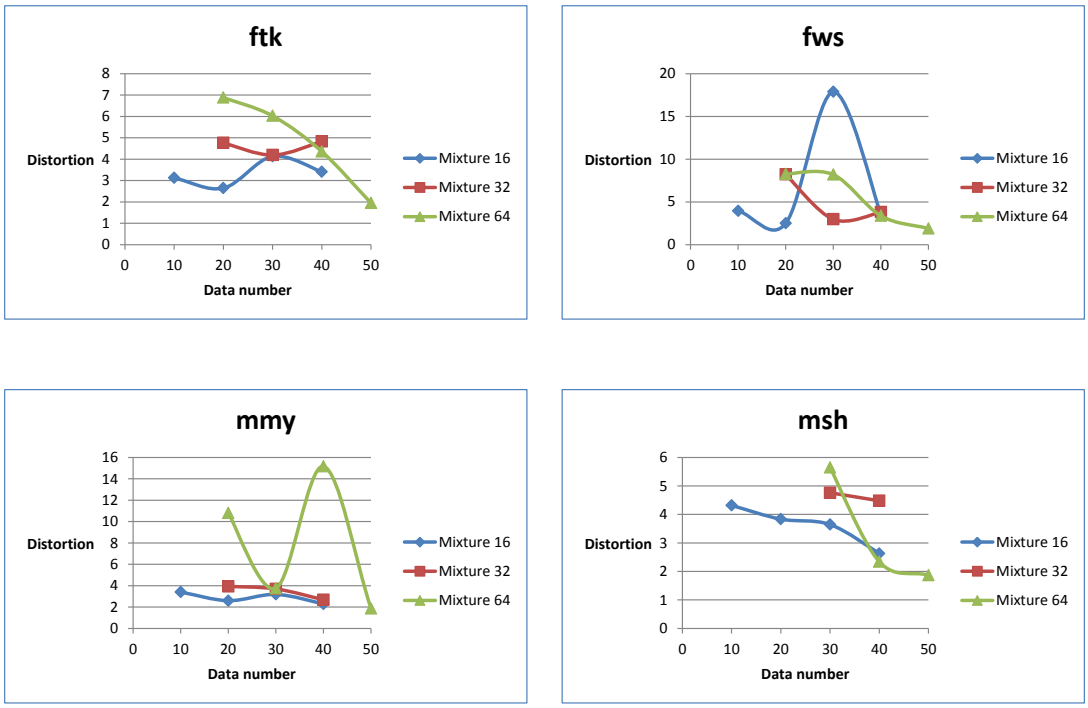


Fig.5.7: Mel-cepstral distortion between the resynthesizes wideband speech using the proposed method and original wideband speech. In these experiments, the number of training data and GMM mixtures are varied.

Chapter 6

Conclusions

6.1 Conclusions

Artificial Bandwidth Extension (ABE) is one of important technologies in tele-communication today due to the need of improving the quality of the transmitted speech signal. Based on the source-filter model, ABE can be separated into two sub-tasks, so called Estimation of wideband spectral envelope and Estimation of wideband excitation signal. For each task, numerous algorithms have been proposed and were shown to give promising results. Especially, statistical algorithms such as GMM have been verified to have better performance than others. However, as mentioned in Chapter. 3 and Chapter. 4, these approaches still have several limitations such as the ignorance of inter-frame correlations, and mismatch during the space division process. In order to solve these problems, we proposed a new approach based on REDIAL.

In this approach, a discriminative model (LDA + GMM) was used in order to utilize information of the target wideband features for the feature space division process. This is expected to be able to reduce the mentioned mismatch problem. Additionally, feature vectors which are generated by concatenating several adjacent frames were used instead of using only single frame. This increases the input information, and also considers the inter-frame relation. Therefore an improvement in transformation accuracy is also expected. Moreover, we also introduced a regularization of the transformation matrix to avoid the over fitting problem.

In Chapter. 5, we have conducted several experiments under various experiment conditions. We first performed experiments under the speaker dependent condition using speech samples of 4 speakers from ATR database. Objective and subjective experiment results have confirmed the effectiveness of our proposed method on every speaker. After that, we conducted experiments under the speaker independent condition suing TIMIT database. Similar to previous experiments, the results also indicated the superiority of our proposed method to the conventional methods. In the next experiments, we considered the performance of our proposed method when introducing dynamic features. Finally, we performed experiments with the number of training data and mixtures of GMM vary to observe their effects on the performance of the proposed method.

6.2 Future works

There are three works which we intend to implement in the future.

6.2.1 Estimation of highband spectral envelope

As explained in Section. 4.4.2, in current framework of our ABE system, the wideband spectral envelope is first estimated from the narrowband spectral envelope, and then the narrowband part of the estimated spectral envelope is replaced by the original narrowband

spectral envelope. Although this procedure was proved to achieve estimated speech with high quality, its process still looks unnatural. We think that, it is able to directly estimate the highband spectral envelope from the narrowband one by considering the narrowband one as an adding condition. We are actively working on this problem.

6.2.2 Speaker adaptation in ABE

It is normally difficult to collect data of every speaker to make a speaker dependent model. Therefore, in general a speaker independent model is often used. However, as seen in the experiment results, the performance of the system in the condition of speaker independent is inferior as compared with that in speaker dependent condition. To deal with this problem, in speech recognition, several speaker adaptation techniques such as MAP, MLLR, etc, which use a small number of data to adapt the speaker independent data, have been proposed. These methods have been verified to improve the recognition rate. In the future, we plan to apply speaker adaptation techniques to our ABE problem to confirm if they also work with this kind of task.

6.2.3 ABE in noisy environment

All of the experiments we have done so far were in ideal condition, which means no noise was considered. As ABE is proposed to cope with problem in real environment, it must be able to work in real condition. In other research, we are working on how to use noisy data as adaptation data of the model to make a better model in noisy environment. Several results have been achieved from this research. Therefore, in the future we plan to applying results from this research to the current research of ABE. This is relatively arduous task but once it is realized, it is expected to be used in a variety of real applications.

Acknowledgement

この２年間に渡り本研究を進めるにあたって，親身に指導して頂きました指導教員である広瀬啓吉教授並びに峯松信明教授・齋藤大輔助教に深く感謝を致します．先生方の的確な助言無しには，ここまで修士論文を完成させることは出来なかったと思います．また，日頃の研究室活動を様々な面で支えて下さった高橋登技官，秘書の折茂結実子さん，池上恵さんに心より感謝致します．

そして，本研究を進めていく上で，元博士課程の鈴木雅之氏[1]・博士課程の柏木陽佑氏及び橋本浩弥氏には数多くの鋭いご指摘やご意見を頂きました．深く感謝致します．

また，同期の池島純君・尾崎洋輔君・正木大介君・毛利圭佑君・寺井真君は研究・勉強だけでなく日常生活に関してもいつも相談にのってくれて有り難い気持ちでいっぱいです．

共に研究に励んだ先輩・同期・後輩の皆様のおかげで楽しく，実りのある研究室生活を送ることが出来ました．深く感謝致します．

そして，ここに名前を全て列挙することが出来ませんが，私が関わった全ての方に感謝の意を述べたいと思います．

最後に，生まれてから今日まで私を支えてくれた家族に感謝致します．また，日頃の生活を支えてくれた，妻の Nguyen Thuy Duong に感謝致します．

どうもありがとうございました．

6th February 2014
Nguyen Duc Duy

[1]現 IBM Research - Tokyo

References

[1] N. Enbom and N. B. Kleijn, "Bandwidth expansion of speech based on vector quantization of the mel frequency cepstral coefficients," 1999 IEEE Workshop on Speech Coding Proceedings, pp.171-173, 1999.

[2] P. Jax, P. Vary, "Artificial bandwidth extension of speech signals using MMSE estimation based on a hidden Markov model," Proceedings of the ICASSP 2003, pp.680-683, 2003.

[3] K.-Y. Park and H. S. Kim, "Narrowband to wideband conversion of speech using gmm based transformation," IEEE International Conference on Acoustics, Speech, and Signal Processing, 2000. ICASSP ' 00. Proceedings., pp.1843-1846, 2000.

[4] W. Fujitsuru, H.Sekimoto, T.Toda, H.Saruwatari, K.Shikano, "Bandwidth Extension of Cellular Phone Speech Based on Maximum Likelihood Estimation with GMM," Proc. 2008 RISP International Workshop on Nonlinear Circuits and Signal Processing (NCSP'08), pp.283-286, March 2008.

[5] L. R. Rabiner and R.W. Schafer, "Digital processing of speech signals," Prentice Hall, 1978.

[6] P. Jax, "Enhancement of Bandlimited Speech Signals: Algoriths and Theoretical Bounds," PhD thesis, Institut fur Nachrichtengerate und Datenverarbeitung, 2002.

[7] D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," IEEE Trans. Speech and Audio Processing, Vo1.3, no. 1, pp. 72-83, 1995.

[8] T. Toda, A.W. Black, K. Tokuda, "Voice Conversion Based on Maximum-Likelihood Estimation of Spectral Parameter Trajectory," *IEEE Trans. ASLP*, Vol. 15, No. 8, pp. 2222-2234, 2007.

[9] J. Droppo, L.Deng, A. Acero, "Evaluation of SPLICE on the Aurora 2 and 3 Tasks," *Proc. ICSLP*, pp. 29-32, 2002.

[10] M. Suzuki, T. Yoshioka, S. Watanabe, N. Minematsu, K. Hirose, "MFCC enhancement using joint corrupted and noise feature space for highly non-stationary noise environments," *Proc. ICASSP*, pp. 4109-4112, 2012.

[11] M. Suzuki, T. Yoshioka, S. Watanabe, N. Minematsu, K. Hirose, "Feature enhancement with Joint Use of Consecutive Corrupted and Noise Feature Vectors with Discriminative Region Weighting," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 21, Issue 10, pp. 2172 - 2181, 2013-10.

[12] , Cutnell, John D. and Kenneth W. Johnson. Physics. 4th ed. New York: Wiley, 1998: 466.

[13] ITU-T, http://www.itu.int/ITU-T/ngn/

[14] "An Introduction to LTE, " 3GPP LTE Encyclopedia. Retrieved December 3, 2010.

[15] Y. Linde, A. Buzo, R.M. Gray, "An algorithm for vector quantizer design", IEEE TRANSACTIONS ON COMMUNICATIONS, Vol. 28, No. 1, 1980.

[16] J. Epps andW. Holmes, "A new technique for wideband enhancement of coded narrowband speech," 1999 IEEE Workshop on Speech Coding Proceedings, pp.174-176, 1999.

[17] A.Uncini, F. Gobbi, F. Piazza, "Frequency Recovery of Narrow-Band Speech Using Adaptive Spline Neural Networks," In Proc. ICASSP, Phoenix, AZ, USA, May 1999.

[18] D.Zaykovskiy and Bernd Iser, "Comparison of Neural Networks and Linear Mapping in an Application for Bandwidth Extension," In Proc. 10th International Conference on Speech and Computer (SPECOM), pp.695-698, Patras, Greece, October 2005.

[19] A. Shahina and B. Yegnanarayana, "Mapping Neural Networks for Bandwidth Extension of Narrowband Speech," In Proc. Interspeech, pp.1435-1438, Pittsburgh, PA, USA, September 2006.

[20] D.Zaykovskiy, "On the Use of Neural Networks for Vocal Tract Transfer Function Estimation," Master ' s thesis, University of Ulm, Ulm, 2004.

[21] H. Pulakka, U. Remes, S. Yrttiaho, K.J. Palomaki, M. Kurimo, P. Alku, "Low-frequency bandwidth extension of telephone speech using sinusoidal synthesis and Gaussian mixture model," Proc. INTERSPEECH, pp.1181-1184, Florence, Italy, August 2011.

[22] H. Pulakka, U. Remes, K.J. Palomaki, M. Kurimo, P. Alku, "Speech bandwidth extension using Gaussian mixture model based estimation of the highband mel spectrum," Proc. ICASSP, pp.5100-5103, Prague, Czech Republic, May 2011.

[23] B. Iser and G. Schmidt, "Bandwidth extension of telephone speech," Technical Report 2, Termic SDS Research, June 2005.

[24] P. Jax and P. Vary, "Wideband extension of telephone speech using a hidden markov model," 2000 IEEE Workshop on Speech Coding, pp.133-135, 2000.

[25] P. Jax and P. Vary, "Artificial bandwidth extension of speech signal," ICASSP, 2003.

[26] H. Carl and U. Heute, "Bandwidth Enhancement of Narrow-Band Speech Signals," In Proc. EUSIPCO, vol. 2, pp.1178-1181, Edinburgh, Scotland, UK, September 1994.

[27] U. Kornagel, "Improved Artificial Low-Pass Extension of Telephone Speech," In Proc. International Workshop on Acoustic Echo and Noise Control (IWAENC), pp.107-110, Kyoto, Japan, September 2003.

[28] U. Kornagel, "Spectral Widening of the Excitation Signal for Telephone-Band Speech Enhancement," In Proc. International Workshop on Acoustic Echo and Noise Control (IWAENC), pp.215-218, Darmstadt, Germany, September 2001.

[29] W.R. Gardner, R.C. Hardie, J.A. Fuemmeler, "Techniques for the Regeneration of Wideband Speech from Narrowband Speech," EURASIP J. Appl.Signal. Process., vol. 2001, no. 4, pp.266-274, 2001.

[30] G. Miet, A. Gerrits, J.C. Valiere, "Low-Band Extension of Telephone-Band Speech, In Proc. ICASSP, vol. 3, pp.1851-1854, Istanbul, Turkey, June 2000.

[31] B. Geiser, P. Vary, "Backwards compatible wideband telephony in mobile networks: Celp watermarking and bandwidth extension," ICASSP 2007, April 2007.

[32] P.Bauer, T.Fingscheidt, "A Statistical Framework for Artificial Bandwidth Extension Exploiting Speech Waveform and Phonetic transcription," 17th European Signal Processing Conference, pp.1839-1843, 2009.

[33] A.H. Nour-Eldin, P. Kabal, "Mel-frequency cepstral coefficient-based bandwidth extension of narrowband speech," INTERSPEECH 2008, pp.53-56, 2008.

[34] A.H. Nour-Eldin, P. Kabal, "Combining frontend-based memory with MFCC features for Bandwidth Extension of narrowband speech," ICASSP 2009, pp.4001-4004, 2009.

[35] A.H. Nour-Eldin, P. Kabal, "Memory-Based Approximation of the Gaussian Mixture Model Framework for Bandwidth Extension of Narrowband Speech," INTERSPEECH 2011, pp.1185-1188, 2011.

[36] K. Kalgaonkar, M.A. Clements, "Sparse probabilistic state mapping and its application to speech bandwidth expansion," ICASSP 2009, pp.4005-4008, 2009.

[37] International Telecommunications Union, ITE-R Recommendation BS.1534: "Method for the subjective assessment of intermediate quality level of coding systems (MUSHRA)," 2003.

[38] ITU-T Recommendation P.800.1, "Mean Opinion Score (MOS) terminology," 2003.

[39] Kubichek, R.F., "Mel-cepstral distance measure for objective speech quality assessment," Communications, Computers and Signal Processing, 1993., IEEE, Vol. 1, pp. 125-128, May 1993.

[40] K. Tokuda , T. Kobayashi , T. Masuko , T. Kobayashi , T. Kitamura, "Speech parameter generation algorithms for HMM-based speech synthesis," *in Proc. ICASSP*, pp. 1315-1318, Jun. 2000.

[41] H. Kawahara, J. Estill, O. Fujimura, "Aperiodicity extraction and control using mixed mode excitation and group delay manipulation for a high quality speech analysis, modification and synthesis system STRAIGHT," *MAVEBA 2001*, September 2001.

[42] H. Kawahara, I. Masuda, A. Cheveigne, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a reptitive structure in sounds," *Speech Communication*, 27, pp.187-207, 1999.

[43] A. Kurematsu, K. Takeda, Y. Sagisaka, S. Katagiri, H. Kuwabara, K. Shikano, "ATR Japanese Speech Database as a Tool of Speech Recognition and Synthesis," Speech Communication, 9, 357-363 (1990).

[44] L. D. Consortium, "Timit acoustic-phonetic continuous speech corpus," CD-ROM, ISBN 1-58563-019-5.

Publications

Journal papers

[1] D.D.Nguyen, M.Suzuki, N.Minematsu, K.Hirose, "Wideband Re-synthesis of Narrowband Speech using Discriminative Piecewise Linear Transformation," Journal of Research Institute of Signal Processing, vol.17, no.4, pp.131-134 (2013-7).

International conferences

[2] D.D.Nguyen, M.Suzuki, N.Minematsu, K.Hirose, "Wideband Re-synthesis of Narrowband Speech using Discriminative Piecewise Linear Transformation," Proc. RISP International Workshop on Nonlinear Circuits, Communication and Signal Processing (NCSP), pp.1-4 (2013-3).

[3] D.D.Nguyen, *et al*, "Artificial bandwidth extension based on regularized piecewise linear mapping with discriminative region weighting and long-span features," Proc. INTERSPEECH, pp.3453-3437 (2013-8).

Domestic conferences

[4] グェンドゥックズイ, 吉岡拓也, 峯松信明, 広瀬啓吉, "特徴量強調における教師なし話者適応に関する検討," 情報処理学会音声言語情報処理研究会資料，2012-SLP-94(23)，pp.1-6 (2012-12).

[5] グェンドゥックズイ, 鈴木雅之, 峯松信明, 広瀬啓吉, "識別的な区分的線形変換を用いた狭帯域音声に対する帯域拡張," 日本音響学会春季講演論文集，3-10-16，pp.793-796 (2013-3).

[6] グェンドゥックズイ, 鈴木雅之, 峯松信明, 広瀬啓吉, "REDIAL を用いた狭帯域音声の帯域復元," 電子情報通信学会音声研究会資料，SP2013-7，pp.37-42 (2013-5).