JAIST Repository
https://dspace.jaist.ac.jp/

Title	Quality improvement of HMM-based synthesized speech based on decomposition of naturalness and intelligibility using non-negative matrix factorization
Author(s)	Dinh, Tuan Anh
Citation	
Issue Date	2016-03
Type	Thesis or Dissertation
Text version	author
URL	http://hdl.handle.net/10119/13625
Rights	
Description	Supervisor:Masato Akagi, 情報科学研究科, 修士

JAIST
JAPAN
ADVANCED INSTITUTE OF
SCIENCE AND TECHNOLOGY

Japan Advanced Institute of Science and Technology

Quality improvement of HMM-based synthesized speech based on decomposition of naturalness and intelligibility using non-negative matrix factorization

By Dinh Anh Tuan

A thesis submitted to
School of Information Science,
Japan Advanced Institute of Science and Technology,
in partial fulfillment of the requirements
for the degree of
Master of Information Science
Graduate Program in Information Science

Written under the direction of
Professor Masato Akagi

March, 2016

Quality improvement of HMM-based synthesized speech based on decomposition of naturalness and intelligibility using non-negative matrix factorization

By Dinh Anh Tuan (1410030)

A thesis submitted to
School of Information Science,
Japan Advanced Institute of Science and Technology,
in partial fulfillment of the requirements
for the degree of
Master of Information Science
Graduate Program in Information Science

Written under the direction of
Professor Masato Akagi

and approved by
Professor Masato Akagi
Professor Jianwu Dang
Associate Professor Masashi Unoki

February, 2016 (Submitted)

Copyright © 2016 by Dinh Anh Tuan

Contents

1 Introduction **2**
1.1 Motivation . . . 2
1.2 Related works . . . 3
1.3 Purpose . . . 4

2 HMM-based text-to-speech **6**
2.1 Background on HMM-based text-to-speech . . . 6
2.1.1 Text-to-speech synthesis definition . . . 6
2.1.2 The linguistic specification . . . 6
2.1.3 Statisical parametric models for speech synthesis . . . 7
2.2 Methods in quality improvement of HMM-based text-to-speech . . . 11
2.2.1 Improving parameter generation algorithm . . . 12
2.2.2 Ultilizing voice conversion techniques . . . 14

3 Asymmetric bilinear model using non-negative matrix factorization **17**
3.1 Definition of asymmetric bilinear model . . . 17
3.2 Find appropriate acoustic feature . . . 18
3.2.1 Fundamental frequency . . . 18
3.2.2 Formant related features: LSF, LPC, PLP . . . 18
3.2.3 Fine structure related features: cepstrum, MFCC, MCC . . . 20
3.2.4 Experiment and Discussion . . . 20
3.3 Proposed asymmetric bilinear model using non-negative matrix factorization 22
3.3.1 Experiment and Discussion . . . 23
3.4 Scheme of applying asymmetric bilinear model . . . 27

4 Evaluation and Discussion **30**
4.1 Preference test . . . 30
4.2 Modified rhyme test . . . 31

5 Conclusion **35**
5.1 Summaries . . . 35
5.2 Contributions . . . 36
5.3 Future works . . . 37

Chapter 1

Introduction

1.1 Motivation

Speech synthesis is a great invention of speech processing area. A speech synthesis or text-to-speech system receives text as input then output speech waveform. In other words, it converts written language to spoken language. The systems usually comprise 2 parts: "front end" and "waveform generation." Front end part convert text into linguistic specification. Linguistic information consists of phonetic information and linguistic information which is necessary for synthesizing waveform speech. The part is language-dependent. The linguistic information is used by waveform generation to generate speech waveform. The part is mostly language-independent.

The system has a wide variety of application from speech-to-speech translation to story teller system, and speech generation devices for the handicapped. The goal of speech synthesis is building natural and intelligible synthesized voices. To achieve the goal, there are two popular approaches for text-to-speech: concatenation speech synthesis and hidden Markov model (HMM) based speech synthesis.

Concatenation speech synthesis simply stores the speech corpus. The corpus is indexed using linguistic specification. In other words, the stored data is labelled for an efficient process of searching and extracting appropriate parts to concatenate in synthesis. The use of index is analogous to using book-index to search for all the occurrences of a required linguistic specification. In general, the labelling consists of phonetic information and linguistic information. The retrieval process is not easy because some desired specification may not be available in the corpus. In the selection, the best available sequence of units are chosen from amongst the many slightly mis-matched parts to concatenate. There are speech waveform or more suitable representation for concatenation such as LPC in the corpus.

By concatenating real speech-segments to generate synthesized speech, the generated speech is human-sounding. Many commercial applications utilize concatenation speech synthesis system. To ensure the quality of synthesized speech, a huge speech database is required to cover all possible linguistic specification.

The disadvantage of concatenation speech synthesis is requirement of large and hight

quality speech database recorded in a consistent environment to build a new synthesized speech. The requirement makes the technique costly and hard to apply in limited data condition such as building personal voice for the sufferers from speech disorder or building voices with non-native accent in speech-to-speech translation.

In recent year, statistical approach for speech synthesis receives huge attention from researchers in speech processing area. There is no stored speech in the approach. Instead, a model , which mainly is hidden Markov model (HMM), is fitted to the speech database during the training process, then, is stored.

The idea of utilizing HMM to represent speech in speech recognition is acquired for speech synthesis. It's not certain that HMM is an appropriate model to represent human speech. Since the availability of training method such as expectation-maximization, and the efficient of searching algorithm such as Viterbi algorithm, HMM become a very powerful tool. Context-dependent phonemes are represented by the models. Then, they are indexed using linguistic specification An appropriate sequence of context-dependent models is selected. Then, it is used to generate speech parameters in synthesis process. This is not a trivial selection because there can be missing models. It is necessary to create a model for any required linguistic specification by sharing parameters with sufficiently similar models. This process is the same as selecting slightly mis-matched segments in a concatenation text-to-speech system.

The advantage of HMMs in text-to-speech is the capability of modelling statistical distribution of not only speech parameters but also the rate of changes of speech parameters. Therefore, synthesized speech with arbitrary content can be generated with high intelligibility. In addition, the training process of the HMMs requires smaller footprint than concatenation technique. Therefore, HMM based text-to-speech is a state-of-the-art technique in building new voice in limited data condition.

However, the generated parameters tend to be close to the mean of Gaussian distribution. The averaging phenomenon is unavoidable in training of statistical model.

Averaging many frames of speech will widen formant bandwidths and reduce spectral envelope's dynamic range. The phenomenon produces over-smooth trajectories and over-smooth spectral envelopes in synthesis phase. As a result, the synthesized speech is muffled and not natural [1]. Adjusting generated parameters to have the same variances found in parameters of natural speech is a critical topic in HMM-based text-to-speech, especially in limited data condition.

1.2 Related works

Many efforts have been spent on improving the naturalness of synthesized speech. There are two main approaches for solving the problem.

In the first approach, an objective evaluation for smoothness of synthesized speech is used to control the over-smoothing effect on generated parameters. In [12], global variance (GV) was proposed as a measurement of smoothness in generated parameters. GV is defined as the variance of Mel-cepstral coefficients in time domain. The generated parameters are adjusted so that its variance increase to the variance of natural speech

parameters. With each Mel-cepstral coefficient, there is one GV value.

In [13], modulation spectrum (MS) was proposed as an extension of GV in controlling smoothness of generated parameters. MS can capture the dynamic change of Mel-cepstral coefficients in time domain. MS is defined as Fourier transformation of Mel-cepstral coefficients in time domain. With each Mel-cepstral coefficient, there is a vector of MS with several elements. In the approach, parameter generation process becomes a joint optimization of HMMs and likelihoods of objective evaluations such as GV and MS. While the optimization of HMMs has close-form solution, the joint optimization typically has no close-form solution.

In second approach, voice conversion techniques are utilized in post processing manner to transform over-smooth spectra to natural one. In [14], local linear transformation (LLT) combined with modified restricted temporal decomposition (TD) [19] was proposed to convert over-smooth spectra to natural spectra. First, speech spectra is decomposed into LSF. By decompose LSF sequence of synthesized speech using TD, we obtain event targets associating to speaker information and event functions corresponding to phonetic information [23]. By assuming that synthesized speech belongs to a different speaker from natural speech, event targets of synthesized speech are transformed to those of natural speech. The transformation is done using LLT. In [15], over-smooth speech spectra is directly converted into natural one using Deep Neural Network. In addition, simple post-filters were efficiently used to adjust the variance of generated speech parameters [21]. Utilizing the objective evaluation of smoothness such as GV and MS, the techniques assume that objective evaluation of synthesized speech and natural speech follow 2 different statistical distributions (with different means and variances). Using means and variances of the 2 distributions, a filter can scale the dynamic of generated parameters the dynamic of natural one. The second approach benefits close-form solution from optimizing HMMs in parameter generation phase.

1.3 Purpose

Although conventional voice conversion techniques can convert bad speech spectra to natural one, they unexpectedly violate the acceptable intelligibility of synthesized speech. To improve naturalness without violating intelligibility, naturalness and intelligibility is decomposed using asymmetric bilinear model [16].

In the [17], A voice conversion technique based on decomposing style and content was proposed. An asymmetric bilinear was utilized with Singular Value Decomposition (SVD) to decompose style, which is speaker information, and content, which is phonetic information, from Linear Spectral Frequency (LSF) observations of source speakers' speech. By assuming that phonetic information is the same when different speaker say the same sentence, a small amount of target speakers' speech with corresponding phonetic information from source speaker is used to find speaker information of target speaker. Lastly, combining phonetic information extracted from source speaker for target sentence and speaker information of target speaker, we can obtain target sentence spoken in target speaker style.

To apply asymmetric bilinear model in naturalness and intelligibility decomposition, there are two problems need to be solved. In the first problem, finding efficient acoustic feature strongly associating to naturalness is very important. LSF is a way to represent speech spectra emphasizing on formants. Formant is important in perceiving speaker characteristic, thus LSF is widely used in voice conversion application. However, it's not certain whether formants are enough for perceiving naturalness.

In the second problem, finding appropriate constrain is very important for decomposing naturalness and intelligibility. SVD generates negative values in factorized matrices. The negative values in naturalness matrix indicate unrealistic subtraction of intelligibility factors. Non-negativity constrain needs to be examined in the task of decomposing naturalness and intelligibility. To do so, SVD is replaced by non-negative matrix factorization (NMF) [20]. Since NMF only permits additive combination, this leads to a parts-based representations. Whilst factorized matrices learnt using SVD typically do not have physical interpretation, parts learnt using NMF do have physical meaning.

The thesis is organized in following order. In Chapter 2, background knowledge on HMM-based speech synthesis is explained in detail. Moreover, state-of-the-art techniques for quality improvement of synthesized speech are reviewed. Chapter 3 provide basics on asymmetric bilinear model, then two problems of applying asymmetric bilinear model in decomposing naturalness and intelligibility are tackled. In Section 3.2, finding efficient acoustic feature is discussed. An experiment was conducted by exchanging several kinds of acoustic features between a pair of natural speech and synthesized speech of the same sentence. Moreover, non-negativity constrain is introduced into asymmetric bilinear model using non-negative matrix factorization (NMF) [20] to decompose naturalness and intelligibility. Physical interpretation of factorized matrices revealed by using NMF is discussed in Section 3.3. In the section, conducted experiments provided concrete evidences to prove the strong relation of factorized matrices strongly to naturalness and intelligibility. In Chapter 4, we conducted subjective experiments evaluate proposed method in improving naturalness and preserving intelligibility of synthesized speech. Moreover, the performance of proposed model using NMF is compared with other methods in the task of improving naturalness of synthesized speech. Chapter 5 summarizes the results of all the thesis and future works.

Chapter 2

HMM-based text-to-speech

2.1 Background on HMM-based text-to-speech

The chapter reviews background knowledge on HMM-based speech synthesis [1].

2.1.1 Text-to-speech synthesis definition

Text-to-speech (TTS) system is an automatic conversion of written to spoken language. The system receives input text and generates speech waveform. A TTS system usually consists of two essential parts as in Figure 2.1. The first part converts text into "linguistic specification" which is then used by the second part to generate speech waveform. The front end is dependent of language, while the waveform generation is mostly language-independent.

The process of text-to-speech is a sequence of processes to transform written language into spoken language. In a different perspective, speech synthesis can be viewed as vocoding, in which some compact representations of speech signal are considered. In Figure 2.2, parametrised speech is stored and then can be retrieved and processed to generate desired speech waveform. Either speech data, or statistical models trained from speech data can be used as stored form.

2.1.2 The linguistic specification

Linguistic specification is the input of synthesis phase. This includes a phoneme sequence, and supra-segmental information such as the linguistic information of desired speech. In other words, the linguistic specification consists of every aspects associating to acoustic

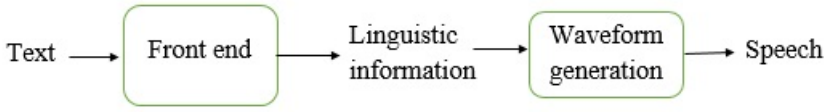

Figure 2.1: Overview of typical TTS system

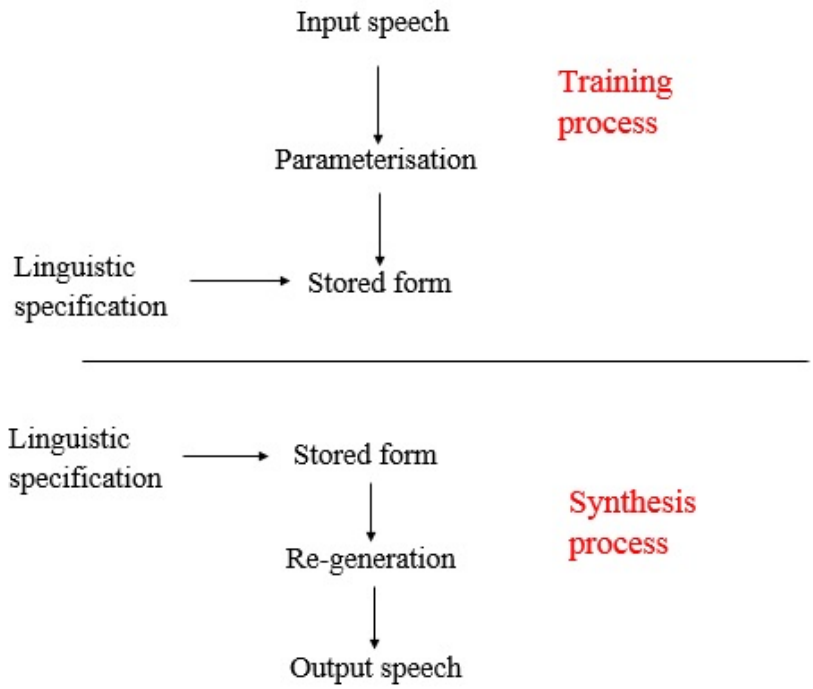

Figure 2.2: Speech synthesis viewed as a vocoder

realisation of output utterance. Let consider an example with the vowel of word "speech". The linguistic specification consists of all information about how to spell the vowel. In other words, it summarises all the information about context in which this vowel occurs. For instance, necessary contextual factors consist of preceding bilabial unvoiced plosive affecting the formant sequences in the vowel, and the fact that a mono-syllabic word contains the vowel (which emphasizes the difference of the vowel's duration).

The context includes information about current word and current utterance, such as neighbouring words, phonemes, phrases and the prosodic pattern, and could be extended to the surrounding utterances and even to paralinguistic factors such as the listener's identity or the emotion of speakers. However, most current system only consider factors within utterances for practical reasons. Table 2.1 lists typical contextual factors considered in regular text-to-speech system.

Linguistic information can form a rather long list. When number of different values for each factor is considered, the permutation produces a vast number of different contexts. However, it doesn't mean all factors will be effective all of the time. In fact, it's hoped that a small number of factor will be significant at any given moment. This reduces the amount of linguistic specification to a more manageable number.

2.1.3 Statisical parametric models for speech synthesis

HMM-based speech synthesis is a model-based approach to text-to-speech, in which statistical parametric models are learnt from speech data. The model is parametric because

statics is used in the model to describe those parameters. For instance, means and variances of probability density functions are used to represent the distribution of those parameters found in training speech data. Historically, the success of the Hidden Markov Model (HMM) for speech recognition triggered research field of statistical speech synthesis. It's not certain that HMM is a natural model of speech. However, HMM is very powerful since the availability of training algorithms such as Expectation-Maximisation, and computationally efficient search algorithms such as Viterbi algorithm. There are different ways to evaluate the model. Subjective tests are used in speech synthesis while word error rates is used in speech recognition. The choice of evaluation method depends on appropriate configuration. Parametrisation of speech signal (the "observations" of the model) and the choice of modelling unit are 2 important aspects of the configuration. Because a context-dependent phoneme is typically a modelling unit, this means selecting contextual factors have to be considered. In table 2.2, some differences in model configuration between speech recognition and speech synthesis are summarised.

Table 2.1: An example of context factors included in linguistic specification

Previous and next phonemes
Segment position in syllable
Syllable position in word and phrase
Word position in phrase
Stress/accent/length of current/previous/next syllables
Distance from stressed/accented syllable
Previous/current/next word POS
Previous/current/next phrase length
End tone of phrase
Utterance length calculated in syllables/words/phrases

Signal representation

In TTS system, speech signal is parametrised as a set of vocoder parameters per frame using a vocoder. Each frame typically has 40-60 parameters to represent spectral envelope, the value for F0 (fundamental frequency), and 5 parameters for aperiodic excitation. A vocoder can be used for speech parametrisation. The process is called encoding. In the synthesis phase, vectors are generated using HMM, then used to generate speech waveform using the vocoder.

In general, we can use any vocoder for HMM-based speech synthesis. However, statical modelling process of parameters leads to better performance of some vocoder than others. The necessary operations required in statistical modelling include averaging of vocoder parameters in training, and interpolation and extrapolation of the values in synthesis. So, the vocoder parameters must be stable under such operations and not lead to unstable values. For example, a widely used vocoder in HMM-based speech synthesis is

Table 2.2: Comparing HMM configurations between recognition and synthesis

	recognition	synthesis
observations	12-parameter represented spectral envelope	60-parameter represented spectral envelope plus source features
modelling unit	triphone, considering previous and next phoneme	full context, including previous two and next two phonemes, and features listed in Table 2.1
duration model	state self-transitions	explicit parametric model of state duration
parameter estimation	Baum-Welch	Baum-Welch, or Trajectory Training
decoding	Viterbi search	no
generation	no	Maximum-likelihood algorithm

STRAIGHT (Speech Transformation and Representation using Adaptive Interpolation of weiGHTed spectrum) [28].

Terminology

Not HMMs but HSMMs: It's important to clarify that the most widely used model in statistical speech synthesis is not HMM. The duration model in the HMM is simplistic. The system requires a better model of duration for high-quality speech synthesis. Once we add an explicit duration model to the HMM, it is no longer a Markov model. The model now becomes "semi-Markov" because an explicit model of duration within each state is not Markov. We call the model Hidden Semi-Markov Model (HSMM). Wherever HMM-based text-to-speech is discussed, it means HSMM speech synthesis.

Lables and context: The linguistic specification is a complex, structured representation; it may consist of lists, trees and other linguistic-useful structures. In HMM-based text-to-speech system, speech is generated from a linear sequence of models. Each model associates to a specific linguistic unit. Thus, Flattening the structured linguistic specification into a linear label-sequence is important. To do so, all other linguistic information is attached to the phoneme tier in the linguistic specification. This produces a linear chain of context-dependent phonemes. The corresponding sequence of models can be retrieved based on the *full context labels*. From the sequence of HMMs, speech waveform is produced.

Statics, deltas and delta-deltas: The parameters are only requirement reconstruct speech using a vocoder. But, to produce human-sounding speech with HMM-based speech synthesis, it requires not only modelling statistical distribution of speech parameters,

but also modelling their rate of change in time domain. The vocoder parameters taken into consideration are *static coefficient*, and their first-order derivatives such as *delta coefficients*. In fact, *delta-delta coefficients* or acceleration is modelled to gain further benefit. We stack the 3 types of parameter together into a single vector (an observation) for the modelling.

Training

Figure 2.4 shows an overview of training and synthesis phase in HMM-based text-to-speech. In training phase, HMMs are trained using labelled data and audio data. The labels are *full context labels* and timing for each phoneme.

Synthesis

The process of synthesis is shown in Figure 2.4 as the lower part. First, the system analysis input text to produce a sequence of full context labels. A long chain of states is formed by attaching together all models associating to this sequence of labels. From this chain of models, the following algorithm is used to generate speech parameters. The generated parameters are used to fit in vocoder to generate speech waveform.

Generating the parameters from model: The models are used to generate a sequence of observations using maximum likelihood concept. First, we describe a naive way and see that it produces unnatural parameter trajectories. Then we introduce actual method. "Parameter" refers to the output of the model. It does not refer *model* parameters such as means and variances of the Gaussian distribution.

Duration: The duration, the amount of parameter frames to be generated by each model state, is determined in advance for both naive method and the MLPG algorithm. The means of explicit duration distribution for each state are used as state duration.

Naive method for parameter generation: The most likely observations from each state are generated. The observations consist only static parameters. The most likely observation is certainly the mean of the Gaussian distribution in that state. As a result, piecewise constant parameter trajectories are generated. The trajectories change value abruptly at each state transition. This is not natural. MLPG algorithm is used to solve the problem.

The maximum likelihood parameter generation (MLPG) algorithm: One crucial aspect of the parameter trajectories is missed by naive method in natural speech. It only considers *static* parameters' statistical properties. But in human speech, both absolute value of the vocoder and speed need to be considered. *Delta* coefficients need to be considered. In fact, statistical properties of the *delta-delta* coefficients can also be considered. In figure 2.3, MLPG algorithm is described. The chain of HMMs are constructed by concatenating the models associating to the full context label sequence predicted from

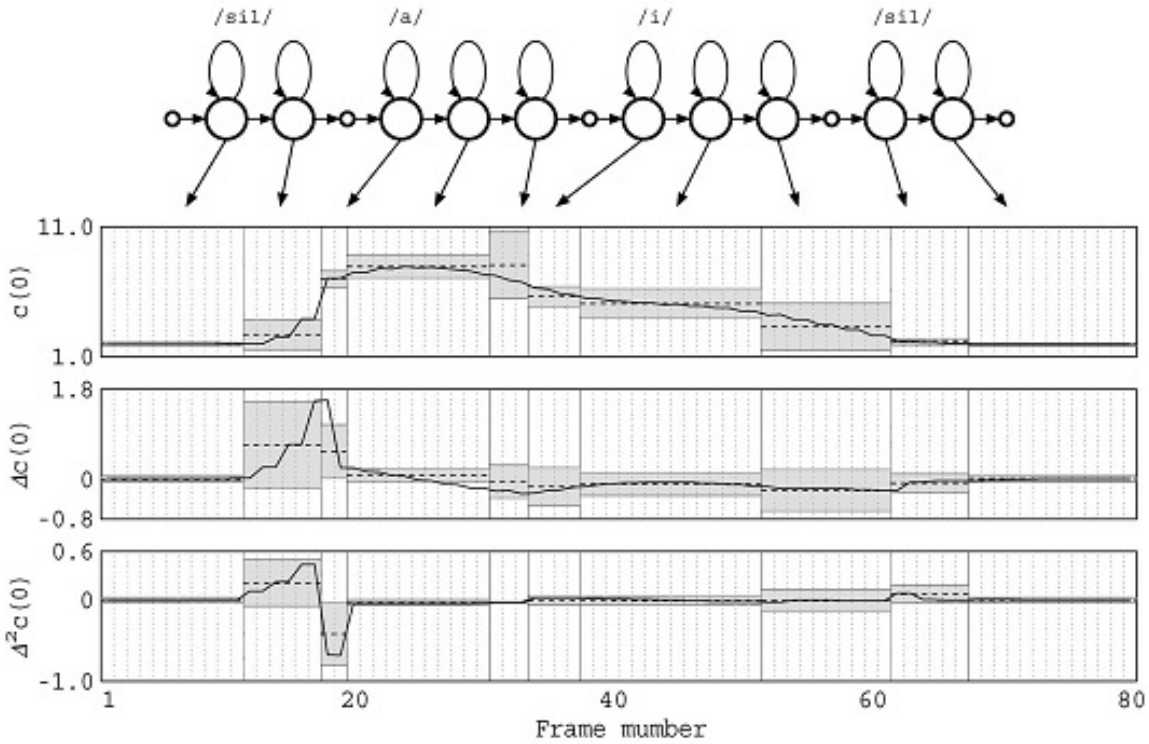

Figure 2.3: Generated parameters from a sentence-level HMM composed of phoneme-level HMMs for /a/ and /i/. The mean and standard deviation, respectively, of a Gaussian pdf at each state are described by dashed lines

text using the front end. Then a state sequence is selected using the duration model. In other words, how many frames will be generated from each state in the model is determined. The sequence of output distributions for each state is shown in the Figure 2.3. The most likely sequence of output distribution for static, delta and delta-delta distributions is found by MLPG. The 0-th cesptral coefficient $c(0)$ is shown in the figure, all generated parameters follows the same principal.

2.2 Methods in quality improvement of HMM-based text-to-speech

In the section, state-of-the art techniques for naturalness improvement are reviewed in detail. There are two main approaches: improving parameter generation algorithm using objective evaluation for smoothness of speech, and using voice conversion technique to convert over-smooth spectra to natural one.

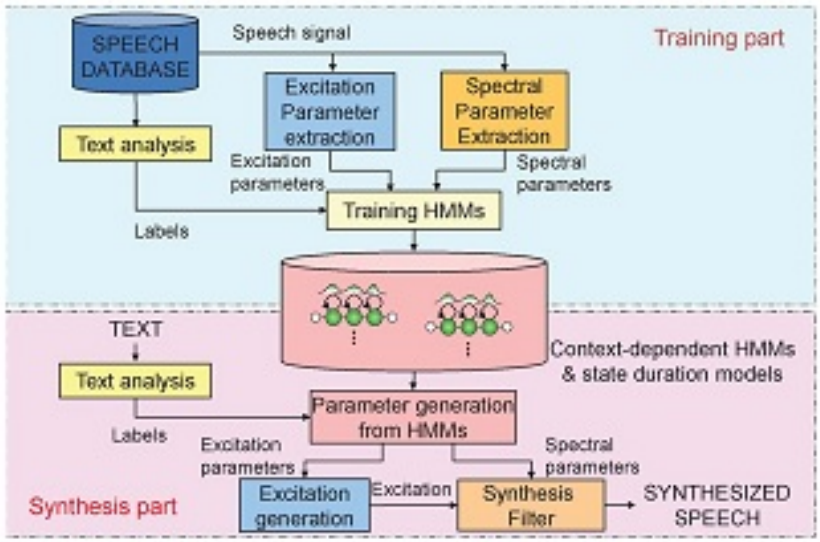

Figure 2.4: Overview of the HMM-based speech synthesis system [3]

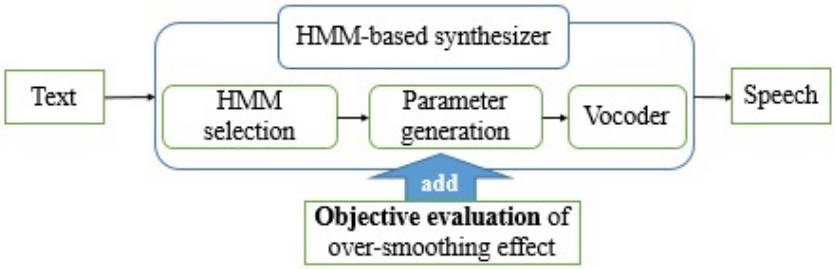

Figure 2.5: Improve parameter generation algorithm

2.2.1 Improving parameter generation algorithm

The speech parameters generated from HMMs tend to be close to means of Gaussian distribution. As a result, generated parameters are over-smooth. To alleviate the phenomenon, objective evaluations for smoothness of parameter sequences are introduced to parameter generation process shown in Figure 2.5. The parameter generation becomes a joint optimization of HMMs and likelihoods of objective evaluations. Many kinds of objective evaluation are examined. Two main kinds of them proposed in [12] and [13] are global variance (GV) and modulation spectrum (MS).

Parameter generation algorithm considering global variance

Assume $c_k = [c_{1k}, c_{2k}, ..., c_{Dk}]^\top$, k=1,2,. . .,T with D is order of MFCC, and T is number of frames. the GV of the vectors is defined:

$$\mathbf{v}(\mathbf{C}) = [v_1, v_2, \ldots, v_D]^\top , \tag{2.1}$$

$$v_d = \frac{1}{T}\sum_{k=1}^{T}\left(c_{kd} - \bar{c}_d\right)^2, \tag{2.2}$$

$$\bar{c}_d = \frac{1}{T}\sum_{\tau=1}^{T} c_{\tau d}. \tag{2.3}$$

In the method, static feature sequences are determined considering not only the output probability of the static and dynamic feature vectors but also this of the GV. In other words, the following criterion is maximized, which is based on a product of the two output probabilities, with respect to the static feature sequence $\mathbf{C}$,

$$L = log\left\{p(\mathbf{O}|\mathbf{Q},\lambda)^w.p(\mathbf{v}(\mathbf{C})|\lambda_v)\right\} \tag{2.4}$$

where $p(\mathbf{v}(\mathbf{C})|\lambda_v)$ is modelled by a single Gaussian distribution. Model parameters λ_v include the mean vector and covariance matrix for the distribution of GV. This Gaussian model λ_v and the HMMs λ are independently trained from the speech corpus. A balance between the two probabilities are controlled by weight w.

Parameter generation algorithm considering modulation spectrum

While a GV is scaler value for temporal scaling of the parameter trajectory in each feature dimension, the MS explicitly represents the temporal fluctuation as a vector. When we sum over all modulation frequency except bias, it's equal to the GV [21]. MS $\mathbf{s}(\mathbf{c})$ of the parameter trajectory $\mathbf{c}$ is defined as follows:

$$\mathbf{s}(\mathbf{c}) = \left[\mathbf{s}(1)^\top, \ldots, \mathbf{s}(d)^\top, \ldots, \mathbf{s}(D)^\top\right]^\top, \tag{2.5}$$

$$\mathbf{s}(d) = \left[s_d(0), \ldots, s_d(m), \ldots, s_d(M-1)\right]^\top, \tag{2.6}$$

$$s_d(m) = R^2_{d,m} + I^2_{d,m}, \tag{2.7}$$

$$= \left(\sum_{t=0}^{T} c_t(d) cos\, kt\right)^2 + \left(\sum_{t=0}^{T} c_t(d) sin\, kt\right)^2, \tag{2.8}$$

where length of Discrete Fourier Transform (DFT)is $2M$, $k = \frac{-\pi m}{M}$ is a modulation frequency. The MS likelihood is defined as $\mathcal{N}(\mathbf{s}(\mathbf{c}); \mu_s, \mathbf{\Sigma}_s)$ where μ_s and $\mathbf{\Sigma}_s$ are a DM-by-1 mean vector and a DM-by-DM covariance matrix, respectively. The objective function L_s is defined as follows:

$$L_s = log\mathcal{N}(\mathbf{Wc}; \mu_q, \mathbf{\Sigma}_q)^w \mathcal{N}(\mathbf{s}(\mathbf{c}); \mu_s, \mathbf{\Sigma}_s) \tag{2.9}$$

where balance between the HMM and MS likelihoods is controlled by MS weight w.

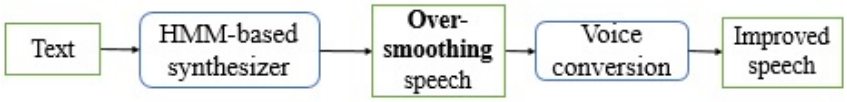

Figure 2.6: Utilizing voice conversion techniques

In the approach, parameter generation process is joint optimization of HMMs and GV (and MS) likelihoods. The optimization of HMMs without considering GV (and MS) has close-form solution. But the joint optimization typically has no close-form solution. To benefit the close-form solution from optimizing HMMs, the next approach uses voice conversion techniques to transform over-smooth parameters to natural one.

2.2.2 Ultilizing voice conversion techniques

In the approach, over-smoothing spectra of synthesized speech is transformed to natural on using voice conversion technique. To do so, a parallel data of over-smoothing spectra and natural spectra is prepared. A transformation function is learned from the parallel data. Then, the transformation function is apply to generated parameters of HMM-based synthesized speech in post-processing manner shown in Figure 2.6.

A post-filter to modify the modulation spectrum in HMM-based speech synthesis

In the case, transformation function is a scaling function applied modulation spectrum domain to scale the over-smooth modulation spectrum to natural one. MS $\mathbf{s}(\mathbf{c})$ of the parameter trajectory $\mathbf{c}$ is calculated as:

$$\mathbf{s}(\mathbf{c}) = \left[\mathbf{s}(1)^\top, \ldots, \mathbf{s}(d)^\top, \ldots, \mathbf{s}(D)^\top\right]^\top, \tag{2.10}$$

$$\mathbf{s}(d) = \left[s_d(0), \ldots, s_d(m), \ldots, s_d(M-1)\right]^\top, \tag{2.11}$$

where $s_d(m)$ is the m-th MS of the d-th dimension of the parameter sequence $[c_1(d), ..., c_t(d), ..., c_T(d)]^\top$, m is a modulation frequency index, M is one half number of the Discrete Fourier Transform (DFT) length.

The following filter is applied to the generated speech parameter sequence $\mathbf{c}$:

$$s'_d(m) = (1-k)s_d(m) + k\left[\frac{\sigma^{(N)}_{d,m}}{\sigma^{(G)}_{d,m}}\left(s_d(m) - \mu^G_{d,m}\right) + \mu^N_{d,m}\right] \tag{2.12}$$

where k is a post-filter emphasis coefficient valued between 0 and 1. If k=0, the filtered sequence will be the same as the non-filtered sequence. When k=1, the MS will be adjusted to be close to the MS of natural speech parameter sequences. The filtered parameter sequence is calculated using the MS and frequency phase characteristics of the parameter sequence. MS and frequency phase characteristics are calculated before filtering.

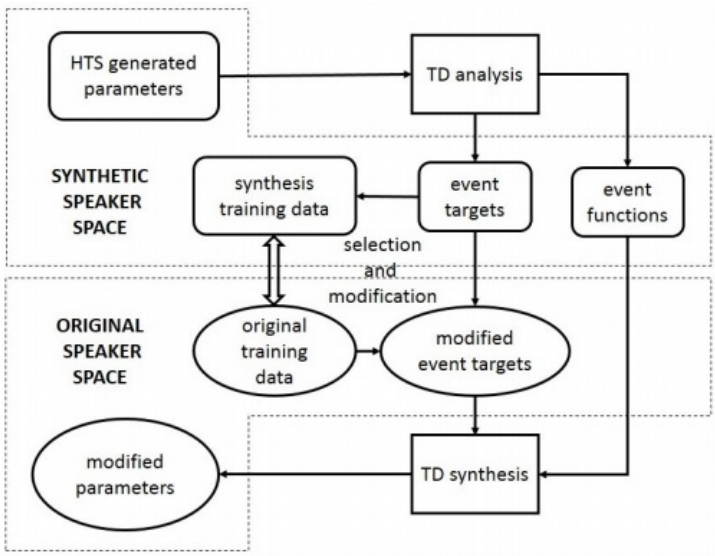

Figure 2.7: Combining local linear transform and temporal decomposition [14]

Combining local linear transformation and temporal decomposition to improve voice quality of HMM-based speech synthesis

The framework of the approach is described in Figure 2.7. A parallel data should be built up. The dataset is constructed by using HTS[1]. With the guide of labels, this synthesized speech is aligned to the original speech. Before conversion, spectral parameters are first decomposed into a sequence of event targets and overlapping event functions with TD as follows:

$$\hat{y}(n) = \sum_{k=1}^{K} \mathbf{a}_k \phi_k(n), 1 \leq n \leq N \tag{2.13}$$

where $\mathbf{a}_k$ is the k-th event target. $\phi_k(n)$, the k-th event function, describes the temporal evolution from the k-th target to the next. $y(n)$ is approximated by $\hat{y}(n)$. N is the total number of frames in the analysed speech.

Event functions correspond to the content or intelligibility of speech. Event targets, which are context-independent, associate to voice quality or speaker's style [23]. Since intelligibility of speech synthesized by HTS is acceptable, event functions are kept ensure the transformation continuous. On the other hand, event targets are converted from synthesized speaker space to the human speaker space using voice conversion method. Local linear transformation (LLT) method was utilized to avoid over-smoothing effect. In LLT, a set of neighbours for each source vector (synthetic target vector) is selected and the transformation between these vectors and their aligned target vectors (human target parameters) is calculated as follows:

$$\mathbf{N}^s \mathbf{W} = \mathbf{N}^o \tag{2.14}$$

where $\mathbf{N}^s$ and $\mathbf{N}^o$ are the aligned synthetic and natural training data. The local regression

$\mathbf{W}$ is obtained by solving the follows using least square method:

$$\mathbf{W} = ((N^s)^T N^s)^{-1} (N^s)^T N^o \tag{2.15}$$

Then, the transformation $\mathbf{W}$ is applied to convert the synthetic event target by the following equation:

$$(\mathbf{a}_{conv})^T = (\mathbf{a}^s)^T \mathbf{W} \tag{2.16}$$

In TD synthesis part, new spectral parameters are constructed by combining the modified event targets with the preserved event functions. The parameters are in the natural speaker space.

Chapter 3

Asymmetric bilinear model using non-negative matrix factorization

3.1 Definition of asymmetric bilinear model

The model was proposed in [16]. In a symmetric model the style s, which is voice characteristic of speaker, and the content c, which is phonetic/spectral content, are represented as parameter vectors denoted a^s and b^c having dimensionality I and J respectively. Let K-dimensional vector y^{sc} be an observation vector in style s and content class c. The vector y^{sc} can be represented as a bilinear function of a^s and b^c as follows:

$$y_k^{sc} = \sum_{i=1}^{I} \sum_{j=1}^{J} w_{ijk} a_i^s b_j^c \tag{3.1}$$

In this formula, elements of the style, content and observation vectors are denoted as i, j and k. The interaction between the style and content factors is described using the term w_{ijk}, which is independent of both style and content. We derive asymmetric bilinear models from the symmetric bilinear models by allowing the interaction terms w_{ijk} to vary with the style. This leads to a more flexible style description. Equation 3.1 becomes:

$$y_k^{sc} = \sum_{i,j} w_{ijk}^s a_i^s b_j^c \tag{3.2}$$

Varying interaction term w_{ijk} and style to make the style-specific terms in Equation 3.2 into

$$a_{jk}^s = \sum_{i} w_{ijk}^s a_i^s \tag{3.3}$$

gives

$$y_k^{sc} = \sum_{j} a_{jk}^s b_j^c \tag{3.4}$$

Denoting by A^s the $K \times J$ matrix with entries a_{jk}^s we can rewrite the Equation 3.4 as

$$y^{sc} = A^s b^c \tag{3.5}$$

The interpretation of the a^s_{jk} terms can be a style specific linear mapping from the content-space to the observation-space. In the research, a^s_{jk} denotes naturalness and b^c denotes intelligibility.

In [17], the observations is LSF vectors. LSF is a way of representing speech spectra, which emphasizes formant information. We are not sure if the formant information is enough for perceiving naturalness. Finding an efficient acoustic feature which strongly relates to naturalness is important.

3.2 Find appropriate acoustic feature

In the section, an experiment was conducted to find an appropriate acoustic feature strongly related to naturalness. The appropriate acoustic feature is transformed to improve naturalness of synthesized speech, while other intelligibility-related acoustic features are preserved. Several acoustic features are prepared. They are fundamental frequency F0, formant related features such as LSF, LPC and PLP; fine structure related coefficients such as cepstrum, MFCC and MCC.

3.2.1 Fundamental frequency

The fundamental frequency of a sound (the rendering of a vowel)which has vocal cord vibration is taken as the lowest measurable frequency in the spectrum of the sound as it unfolds in time [24]. In practice, this corresponds to the vocal cords' rate of vibration. Fundamental frequency is a concept in acoustic and therefore in acoustic phonetics, but there is a corresponding abstract concept in phonology. In phonology we are often concerned with the way the fundamental frequency of a stretch of speech is perceived to change; thus we might perceive a rise in pitch towards the end of a sentence, signalling a question, or a fall in pitch towards the end, signalling a statement. The phenomenon of perceived changes in pitch is called intonation. Because intonation is a perceived concept and because, it is relative in nature, it is completely abstract and cannot therefore be measured in a waveform or spectrogram. What we can do, however, is measure the physical acoustic signal which gave rise to a perceived intonation pattern. When we do this we are measuring a physical correlate of what is perceived - the physical correlate of a cognitive phenomenon. The correlates rarely match in a one-to-one, or linear way. We cannot actually measure intonation, only measure its correlate of changing fundamental frequency.

3.2.2 Formant related features: LSF, LPC, PLP

Linear Prediction Code (LPC) methods are widely used various area of speech processing such as speech coding, speech recognition, speech synthesis, speaker verification and recognition, and for speech storage. LPC methods provide accurate estimates of speech parameters, and work efficiently. The basic idea of Linear Prediction is to closely approx-

imate current speech samples as a linear combination of past samples as follows:

$$s(n) = \sum_{k=1}^{p} a_k s(n-k) \tag{3.6}$$

For period signals with period N_p, it is obvious that:

$$s(n) \approx s(n - N_p) \tag{3.7}$$

LP uses p ($p \ll N_p$) most recent values of $s(n)$ to estimate $s(n)$ by linearly predicting its value. LP determines the predictor coefficients a_k by minimizing the sum of squared differences between the linearly predicted samples and the actual ones.

Another set of parameters, that is as similarly informative as the LPC coefficient, are Line Spectral Frequency (LSF) proposed in [25]. The LP polynomial

$$A(w) = 1 - \sum_{k=1}^{p} a_k w^{-k}$$

can be represented as

$$A(w) = \frac{1}{2}\left[P(w) + Q(w)\right]$$

All the zeros of $p(w)$ and of $q(w)$ are on the unit circle when all the zeros of $A(w)$ are inside the unit circle Moreover, the zeros of $p(w)$ and $q(w)$ are intertwined. The zeros are uniquely specified by their angles, because they are on the unit circle. These angles, representing frequencies, are called the LSF. The LSF is a useful representation of the all-pole filter because:

- LSF coefficients are a very homogeneous set.
- LSF coefficients quantize well.
- LSF coefficients interpolate well.

Perceptual linear predictive (PLP) was proposed in [26]. The technique derives an estimation of the auditory spectrum using 3 concepts from the psychophysics of hearing:

1. Critical-band spectral resolution
2. Equal-loudness curve
3. Intensity-loudness power law

An autoregressive all-pole model is used to approximate the auditory spectrum. A 5th-order all-pole is effective in suppressing speaker-dependent details of the auditory spectrum. PLP analysis is more consistent with human hearing in compare to conventional linear predictive (LP).

LP analysis can capture the information on resonance frequencies. In other words, LP analysis emphasizes formants on spectra. LSF analysis is widely applied in voice conversion techniques because it can capture the speaker-dependent information. However, PLP analysis tends to remove all fine harmonic structure. As a result, it can capture well speaker-independent information.

3.2.3 Fine structure related features: cepstrum, MFCC, MCC

Cepstral analysis is a modelling of speech based on the use of cepstrum, defined as the inverse Fourier transform of the logarithm of the Fourier transform module.

$$signal\ cepstrum = FT\left\{log\left[FT^{-1}\left(signal\right)\right]\right\}$$

Mel-Frequency Cepstral Coefficients (MFCC) is a popular feature extraction techniques based on frequency domain using the Mel scale. The scale is based on the human ear scale. MFCC is derived from the Fast Fourier Transform of a windowed short-time signal. The difference of MFCC from the real cepstral is the use of non-linear frequency scale based on the behaviour of the auditory system. Moreover, these coefficients are robust and reliable regardless of speakers and recording conditions. MFCC analysis is an audio feature extraction technique extracting parameters from the speech similar to ones utilized by humans for hearing speech, and de-emphasizes all other information.

Firstly, the speech signal is separated into frames including an arbitrary number of samples. Smooth transition from frame to frame is ensured using overlapping of frames. Frequency components of a signal in the time-domain is extracted using Fast Fourier Transform for each frame. Then, the logarithm Mel-Scaled filter bank is applied to each frame of frequency components. This scale is approximately linear up to 1 kHz, and logarithmic at greater frequencies. The following equation describes the relation between frequency and Mel scale:

$$M(f) = 1125\, ln\left(\frac{1+f}{700}\right) \tag{3.8}$$

In Mel-scale filter bank, the higher frequency filters have greater bandwidth than the lower frequency filters. However, their temporal resolutions are the same. Lastly, Discrete Cosine Transformation is calculated to the outputs from filter bank. MFCC analysis is an biased estimator of log spectrum. On other hand, MCC proposed in [27] is an unbiased estimator of log spectrum. The estimated spectral envelope by MCC passes through all peak of speech spectra.

3.2.4 Experiment and Discussion

The purpose of the experiment is to find out an efficient acoustic feature vector strongly related to naturalness. To do so, with a certain kind of acoustic feature, feature values are exchanged between a pair of HMM-based synthesized speech and natural speech. If exchanging drastically increases naturalness of synthesized speech, the acoustic feature is strongly related to naturalness

There are 3 steps in the procedure shown in Figure 3.3. In the first step, several kinds of acoustic features are prepared. They are fundamental frequency F0, peak related parameters such as linear prediction coefficient (LPC) w/wo residual power, LSF w/wo residual power, and perceptual linear prediction (PLP), fine structure related coefficients such as Mel-frequency cepstral coefficient (MFCC)[29], Mel-cepstral coefficient (MCC)($\gamma = 0, \alpha = 0.42$ for 16 kHz speech)[30] and cepstrum. To examine one kind

of acoustic features, the feature sequences are exchanged between a pair of synthesized speech and natural speech of the same sentence. Exchanging acoustic feature means improving naturalness of synthesized speech and decreasing quality of natural speech. If quality of natural speech decreases and naturalness of synthesized speech increases a lot after exchanging, the kind of acoustic feature strongly relates to naturalness. In the experiment, one utterance for one natural speech sentence is synthesized by HTS[1]. The HMM-based synthesized speech is aligned to original natural-speech using label to ensure the same phoneme durations. STRAIGHT vocoder was used to analyse speech. Frame length is 10 ms, frame-shift is 1 ms. It decomposes speech into a spectral envelope, F0, and aperiodicity. The STRAIGHT-based spectral envelope is further encoded into LPC, LSF, MFCC, MCC, PLP, and cepstrum. The order of LSF is 16. The order of PLP is 16. The order of MFCC is 16. The order of MCC is 39. After the step, 18 new stimuli are obtained as in Table 3.1. Adding natural speech (denoted as I1) and HMM-based synthesized speech (denoted as I2), we have 20 stimuli.

Table 3.1: Stimuli in experience

Group	Stimuli	Natural speech with **acoustic feature** of HMM-based synthesized speech	Stimuli	HMM-based synthesized speech with **acoustic feature** of natural speech
A	A1	Cepstrum	A2	Cepstrum
B	B1	F0	B2	F0
C	C1	LPC	C2	LPC
D	D1	LPC with power	D2	LPC with power
E	E1	LSF	E2	LSF
F	F1	LSF with power	F2	LSF with power
G	G1	MCC	G2	MCC
H	H1	MFCC	H2	MFCC
J	J1	PLP	J2	PLP

In the second step, naturalness of obtained stimuli is compared using Scheffe's method of paired comparison [18] to sort them based on naturalness. There are six participants (non native English speakers with fluent English level, and good hearing ability). Each participant listened to 380 pairs of stimuli. With each pair, they compare naturalness of stimuli using five grades shown in Figure 3.1 from -2 (A is more natural), 0 (comparable), +2 (B is more natural) in A-B comparison.

Lastly, the efficient acoustic feature is decided by looking for the one that improves naturalness of synthesized speech the most. Experimental results are shown in Figure 3.2. Two-tailed t-test shows that these results are statically significant at a 95% confidence level. These experimental results also indicate that exchanging MCC values improves naturalness of synthesized speech the most (I2 to G2). Exchanging LSF does not sig-

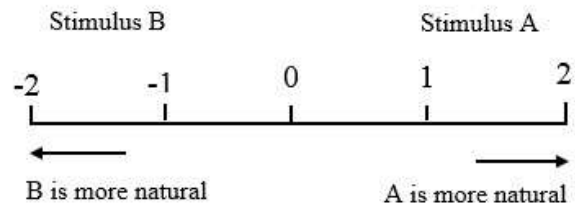

Figure 3.1: Scheffe pair comparison test scale

nificantly improve naturalness (I2 to E2). In frequency domain, fine structure is more important than formant in perceiving naturalness. MCC is the most suitable acoustic feature in improving naturalness.

Although MCC can represent the fine structure in frequency domain, it cannot represent the dynamics of spectra in time domain. In recent years, modulation spectrum becomes a popular concept in capturing the fine structure of speech spectra in time domain. In the paper, modulation spectrum (MS) of MCC sequences $\mathbf{c}_k = [c_{1k},\, c_{2k},\, ...,\, c_{Dk}]^\top$, $k = 1, 2, ..., T$, in which D is the order of cepstral analysis and T is the number of frames, is utilized to capture the over-smoothing effect in both time-frequency domain of speech spectra. Spectral analysis of a speech utterance produces a matrix $R = [\mathbf{c}_1,\, \mathbf{c}_2,\, ...,\, \mathbf{c}_T]$ of size $D \times T$. The time trajectory of cepstral coefficient d is calculated as $\mathbf{r}_d = [c_{d1},\, c_{d2},\, ...,\, c_{dT}]$, $d = 1, 2, ..., D$. The MS of trajectory r_d is defined as:

$$M(d, f) = |FT\,[\mathbf{r}_d]|$$

where f denotes the modulation-frequency bin, defined by the number of points in the Fourier analysis. The number of points in the FFT must be greater than the maximum number of frames T of an utterance. The MS of each utterance is calculated for each cepstral coefficient. Using ABM, MS of synthesized trajectories is modified to be closer to the modulation characteristics of natural trajectories.

3.3 Proposed asymmetric bilinear model using non-negative matrix factorization

In the section, an asymmetric bilinear model is proposed to factorize naturalness and intelligibility of synthesized speech. Figure 3.4 shows an overview of proposed model. An important point of the model is forming a parallel data of S voices (can be natural voices or HTS voices) with N sentences is prepared. The data for each sentence (in one cell) is modulation spectrum of MCC sequences. NMF is used to factorize the parallel data into naturalness and intelligibility. In parallel data of S HTS voices, the variation of speech quality is presented in columns. The speech quality is different among the HTS voices and it's assumed to strongly relate to naturalness. Likewise, the variation of phonetic information is presented in rows and it's assumed to strongly relate to intelligibility. The columns of Matrix 1 in Figure 3.3 summarize the parallel data's vertical structure which is

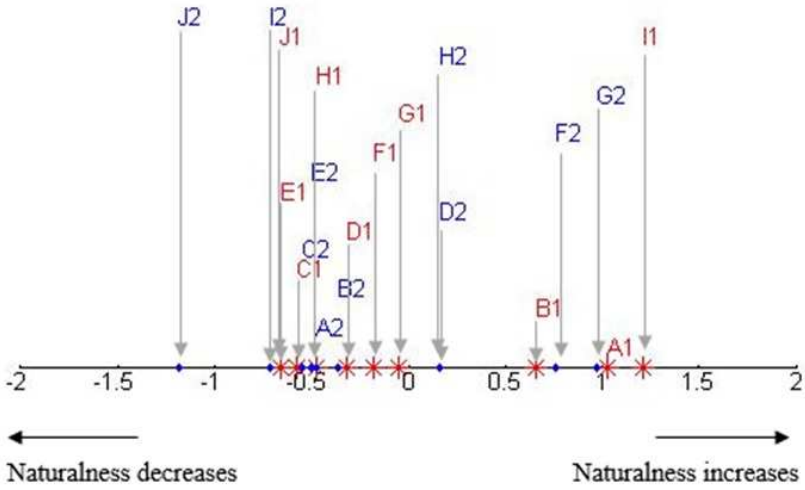

Figure 3.2: Result of Scheffe pair comparison test within 95% confidence interval

assumed to relate to naturalness. Likewise, the rows of Matrix 2 do so for parallel data's horizontal structure which is assumed to relate to intelligibility. We are not sure the relation between Matrix 1 and naturalness, Matrix 2 and intelligibility. In next sections, physical interpretation of Matrix 1 and Matrix 2 will be explained.

3.3.1 Experiment and Discussion

Investigation experiment for the relation between Matrix 1 and speech naturalness

In the experiment, we prove the strong relation of Matrix 1 and naturalness of speech. To do so, naturalness matrix of synthesized speech and naturalness matrix of actual speech are exchanged. Exchanging Matrix 1 aims to decrease naturalness of actual speech and increase naturalness of synthesized speech. If it happens, the naturalness matrix strongly relates to naturalness of speech.

There are 3 steps in the experiments:

1. Decompose natural speech and synthesized speech of target sentences using proposed method.

2. Exchanging Matrix 1 between HTS voice and actual voice of target sentences

3. Comparing naturalness of obtained stimuli.

The first step is shown in Figure 3.5. In the step, a parallel data of S (S=3) HTS voices are used. The speech is first decomposed into fundamental frequency, spectral envelope and aperiodicity using STRAIGHT. Frame length is 10 ms. Frame shift is 5 ms. Spectral envelope is further decomposed into MCC. Order of MCC is 49. To obtain modulation spectrum FFT length is 2048. The voices are trained from 3 CMU dataset (RMS, SLT,

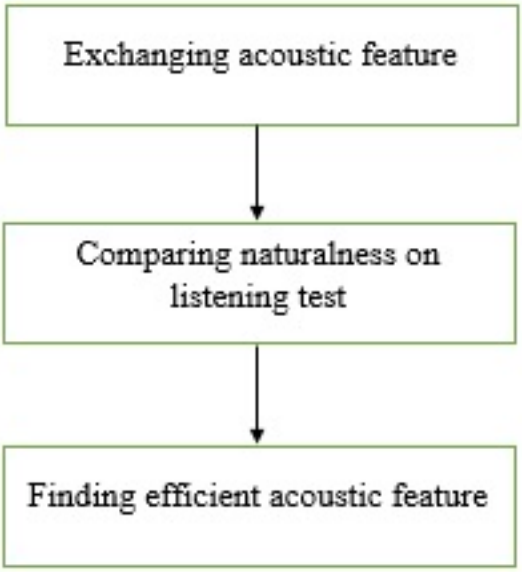

Figure 3.3: Experiment procedure to find appropriate acoustic feature

Table 3.2: 4 kinds of stimuli obtained by exchanging Matrix 1

Stimuli	Components
A	(1) + (2)
B	(1) + (4)
C	(3) + (2)
D	(3) + (4)

CLB). There are 5 training sentences which are synthesized with the guide of label files of their original speech. There are 18 target sentences. The original speech of the HTS voices are used to form a parallel data of 3 natural voices.

In step 2, The Matrix 1 (first components) in factorized matrices are exchanged between HTS SLT speech and actual SLT speech. Obtained stimuli are shown in Table 3.2.

In step 3, we compare naturalness of the 4 kind of stimuli using preference test. Each has 18 utterances. In total, there are 216 pairs of utterances. Participants listened to each pair and answered which stimulus is more natural. There are ten participants (8 non-native and 2 native English speakers) with normal hearing ability.

In Figure 3.6, A denotes human speech while D denotes HMM-based synthesized speech. After exchanging Matrix 1, preference score of A decreases to this of C. Preference score of D increases to this of B. This indicates the strong relation of Matrix 1 to naturalness of speech. By observing Matrix 1, we can see what happened when exchanging the matrices. Figure 3.7 shows an example of naturalness matrix. When observing the high-order coefficient area, we can see that the magnitude of the coefficient in natural speech

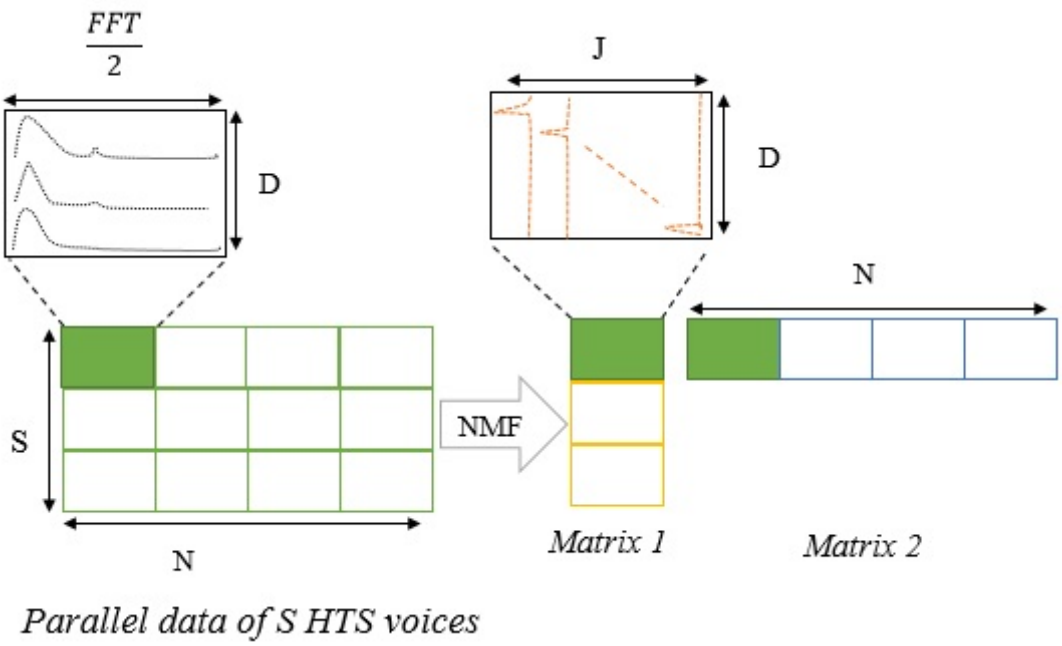

Figure 3.4: Asymmetric bilinear model

is higher than this of synthesized speech. By exchanging the Matrix 1, the magnitude of high-order coefficient is emphasized. The high-order coefficient represents fine-structure of speech spectra. Exchanging Matrix 1 restore fine-structure of speech spectra. It improves naturalness of synthesized speech.

One interesting phenomenon in the experiment is the different between native English speaker and non-native English speaker. Our desired result is C is comparable to D while B is comparable to A. This means Matrix 1 contains all naturalness of speech. The results of native English speaker are the same as desired results. The difference lies in how the listeners define naturalness of speech.

Investigation experiment for the relation between Matrix 2 and intelligibility

In the experiment, we prove the strong relation of Matrix 2 to intelligibility. To do so, Matrix 2 of synthesized speech and Matrix 2 of actual speech are exchanged. Exchanging Matrix 2 matrices aims to decrease intelligibility of actual speech and increase intelligibility of synthesized speech. If it happens, the Matrix 2 strongly relates to intelligibility.

There are 3 steps in the experiments:

1. Decompose naturalness and intelligibility from HTS voices and actual voices of target sentences using proposed method

2. Exchanging intelligibility between HTS voice and actual voice of target sentences

3. Comparing intelligibility of obtained stimuli using modified rhythm test.

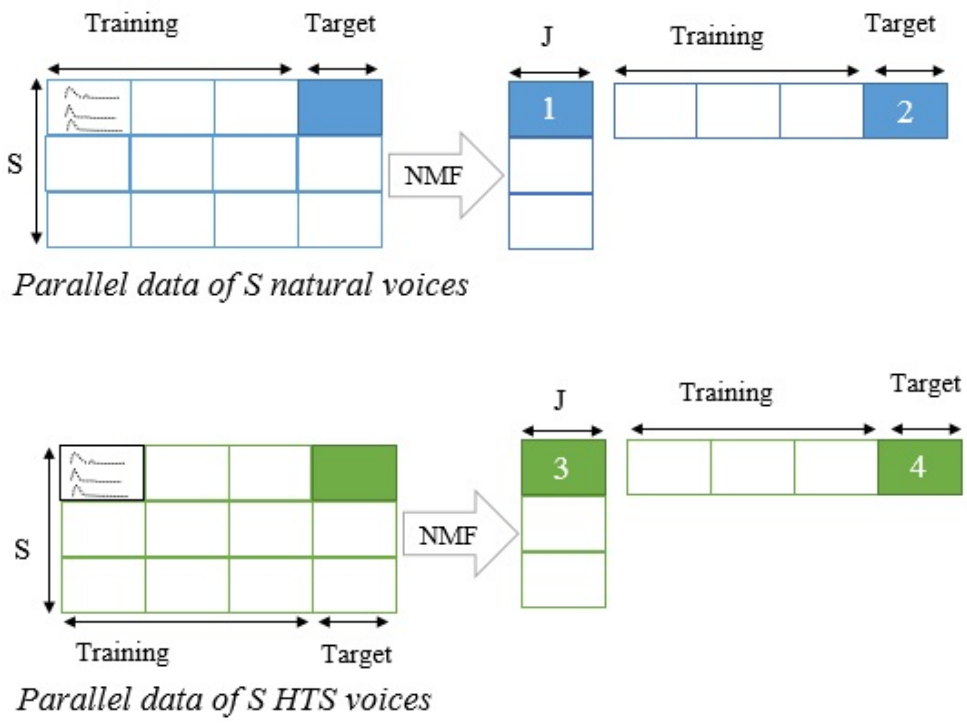

Figure 3.5: Experiment to investigate physical interpretation of factorized matrices using proposed method

The process of doing the experiment is the same as previous experiment with naturalness-matrix. With each target sentence, we compare intelligibility of 4 kinds of stimuli as shown in Table 3.2. In Figure 3.8, A denotes human speech while D denotes HMM-based synthesized speech. The word correctness of A decreases to B and word correctness of D increase to C. This indicates that Matrix 2 strongly relating to intelligibility.

The experimental results provide evidences for the strong relation between Matrix 1 and Matrix 2 in asymmetric bilinear model to speech naturalness and intelligibility. Therefore, proposed method can decompose naturalness and intelligibility of speech.

Physical interpretation of factorized matrices by NMF

Figure 3.7 shows the differences between naturalness matrix of HMM-based synthesized speech and that of natural speech. These columns of naturalness matrix contain information about magnitude of cepstral coefficients. The differences become clear with high-order cepstral coefficients. The magnitude of natural speech's cepstral coefficients is stronger than that of HMM-based synthesized speech. High-order cepstral coefficients can be used to control fine structure of speech spectra. Controlling fine-structure of speech spectra can manipulate naturalness of speech. By emphasizing the magnitude of cepstral coefficients, especially in high order region, the naturalness of synthesized speech is improved.

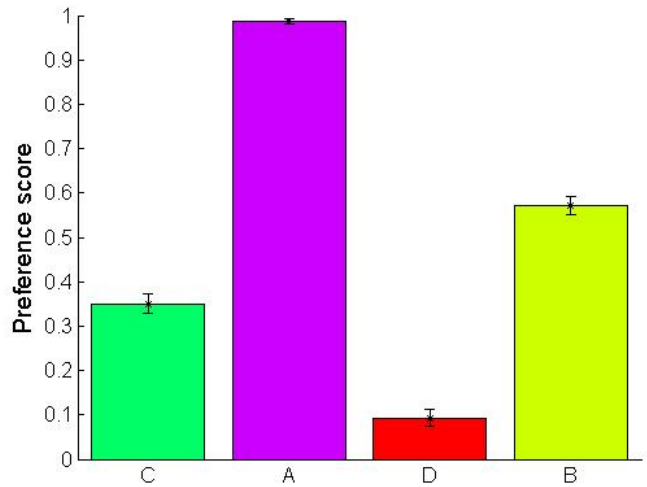

Figure 3.6: Preference scores with 95% confidence interval

3.4 Scheme of applying asymmetric bilinear model

In the section, the process of applying ABM in naturalness improvement is described. The process consists of 3 major steps shown in Figure 3.9:

1. Separating naturalness and intelligibility.
2. Obtaining naturalness of actual speech.
3. Reconstructing speech.

The goal of step 1 is to obtain acceptable intelligibility from parallel data of synthesized voices to preserve it. Naturalness factor and intelligibility factor are factored from the data using non-negative matrix factorization (NMF). Each cell of the parallel data of synthesized speech is MS of one utterance. In Figure 3.9, $\frac{FFT}{2}$ denotes half of length of FFT for MS, D is order of MCC, and S denotes number of HTS [1] (S $\geq$ 2). There are 1 target utterance. Number of training sentences can be as small as 5. J is model dimensionality chosen as $J = S \times D$.

In step 2, naturalness of actual speech A^s is obtained using a small data of actual speech y^{sc} and corresponding intelligibility set C obtained from step 1. We can derive the desired naturalness A^s by minimizing the total squared error over actual speech data,

$$E = \sum_{c \in C} ||y^{sc} - A^s b^c||^2$$

Lastly, naturalness of actual speech A^s and intelligibility of synthesized speech are combined to obtained an improved version of synthesized speech. Intelligibility is preserved even in synthesized speech.

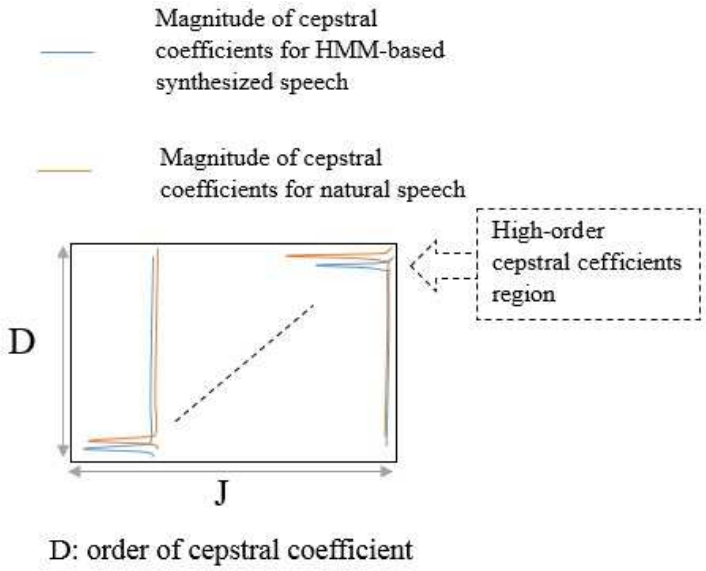

Figure 3.7: Differences between naturalness matrix of HMM-based synthesized speech and that of natural speech

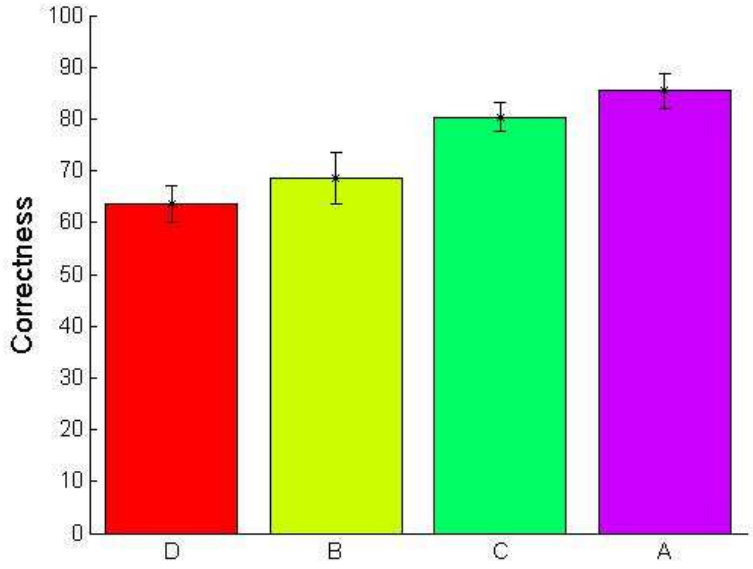

Figure 3.8: Word correctness with 95% confidence interval

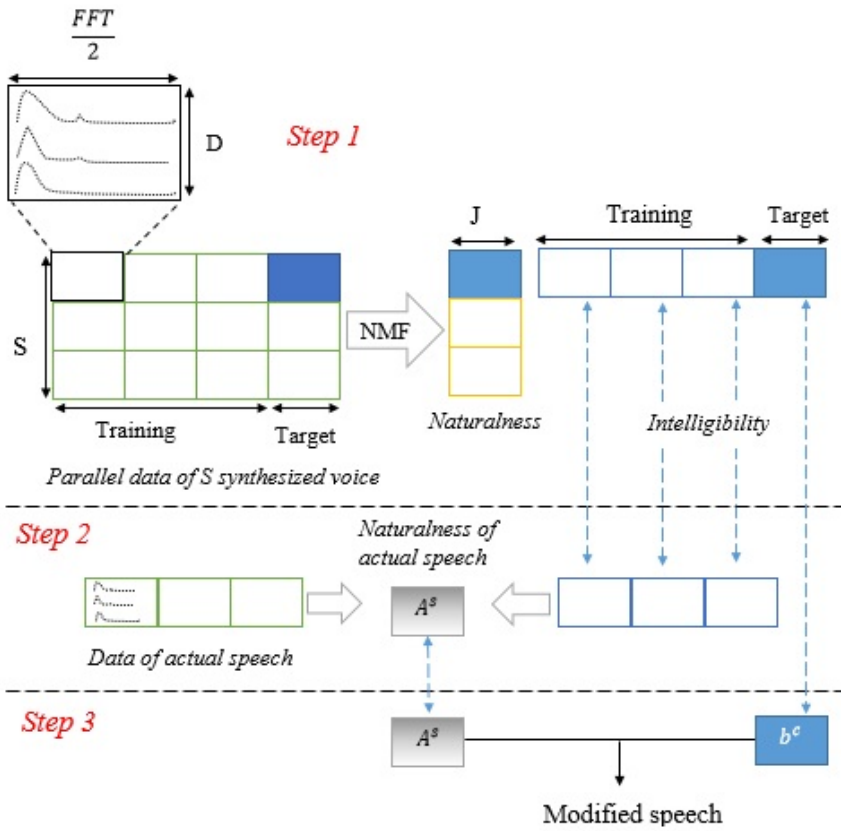

Figure 3.9: Scheme of applying ABM

Chapter 4

Evaluation and Discussion

In the section, preference test and modified rhyme test are conducted to evaluate the naturalness and intelligibility of re-synthesized speech using proposed method. The objectives are showing naturalness is improved, and intelligibility is preserved.

4.1 Preference test

In the test, 2 synthesized voices are trained using 2 CMU dataset (SLT and RMS). $S = 2$. $D = 49$. Sample rate is 16 kHz. Frame length is 10 ms. Frame-shift is 5 ms. $\frac{FFT}{2} = 2048$. Ten samples are synthesized for each voice. Different methods are applied on the samples to improve the samples. They are GV method [12], MS method [21], proposed method. Both limited data condition and large data condition are taken into consideration. In both cases, proposed method is trained using only 5 sentences. In limited data condition, only 5 natural sentences are used to train MS. In the case, GV cannot be trained. In large data condition, 500 natural sentences are used for training GV method, and MS method. There are 11 participants. Ten participants are non-native English speaker with fluent English level. One speaker is native. They listened to pairs of stimuli A-B, and decided which is more natural. Natural speech is explained as human-like speech. There are 240 pairs of stimuli in limited data condition and 400 pairs in large data condition. Before participants did main test, they took a practice session to get familiar with testing procedure.

Figure 4.1 shows score of proposed method denoted as NMF is greater than default ABM denoted as SVD. It indicates that proposed ABM outperforms default ABM in improving naturalness, in limited data condition. Moreover, proposed method outperforms GV method in limited data condition. With GV method, the MS of synthesized speech is scaled using means and variances of MS of natural speech. The values are miscalculated with limited data of natural speech.

In Figure 4.2, the score of proposed method is a little bit smaller than other methods. It's because no constrains were used other than non-negative matrix factorization (NMF) technique. In recent years, NMF is improved a lot by integrating different kinds of constrains in variety of tasks. In next works, sufficient constrains will be investigate to

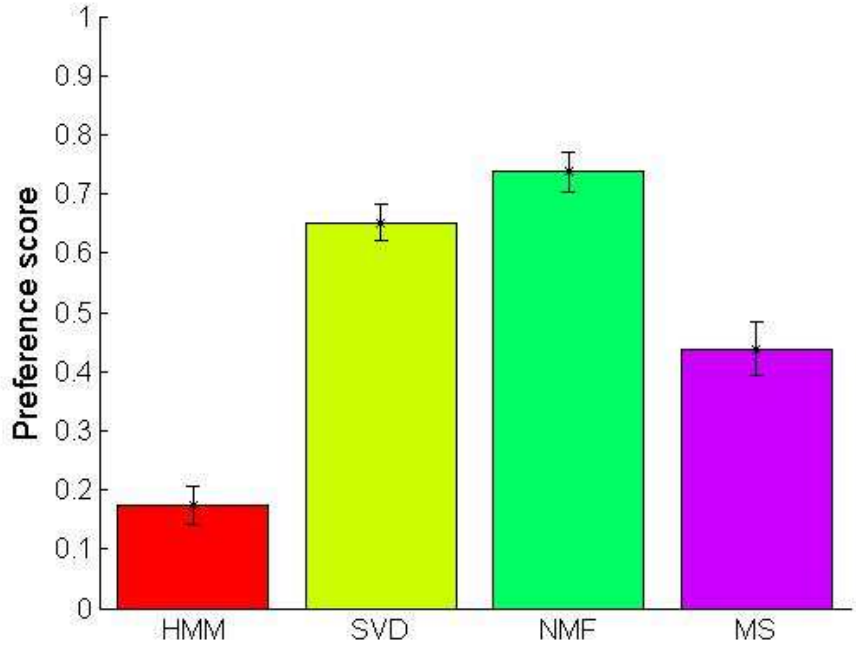

Figure 4.1: Preference scores in limited data condition with 95% confidence interval

improve the performance of proposed method in limited data condition.

4.2 Modified rhyme test

In the test, 2 synthesized voices are trained using 2 CMU dataset (SLT and RMS). $S = 2$. $D = 49$. Sample rate is 16 kHz. Frame length is 10 ms. Frame-shift is 5 ms. $\frac{FFT}{2} = 2048$. Three hundred words [22] are synthesized. Different methods are applied on the sound segments to improve the samples. They are GV method [12], MS method [21], proposed method. Both limited data condition and large data condition are taken into consideration. In limited data condition, only 5 natural sentences are used to train improvement methods for synthesized speech. In the case, GV cannot be trained. In large data condition, 500 natural sentences are used for GV method, and MS method. There are 10 participants. There are 8 non-native English speakers and 2 native English speakers. For each method, each speaker listened to 50 words randomly chosen from 300 words. Then, participants choose the right word from six candidate words.

Figure 4.3 shows that correctness of proposed method (NMF) is equal to the correctness of HMM-based synthesized speech (HMM) in limited data condition. From Figure 4.4, we can see the same trend in large data condition. The intelligibility is preserved with proposed method.

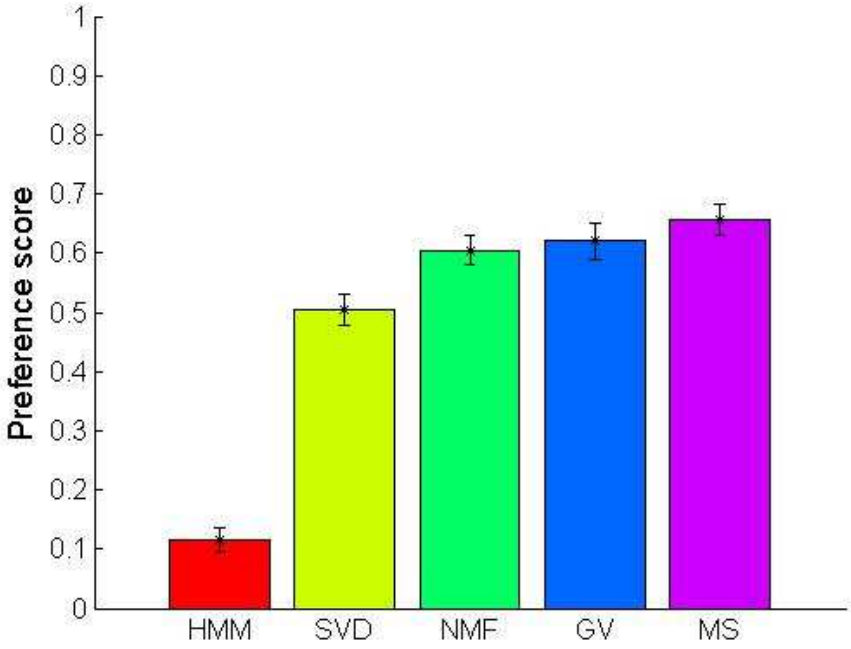

Figure 4.2: Preference scores in large data condition with 95% confidence interval

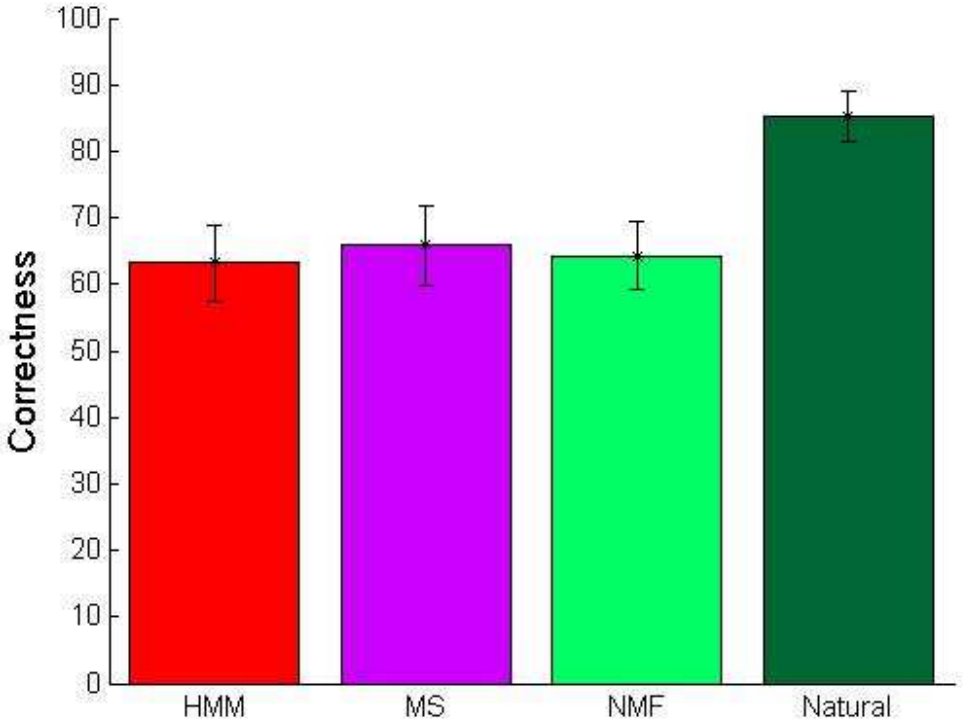

Figure 4.3: MRT correctness in limited data condition with 95% confidence interval

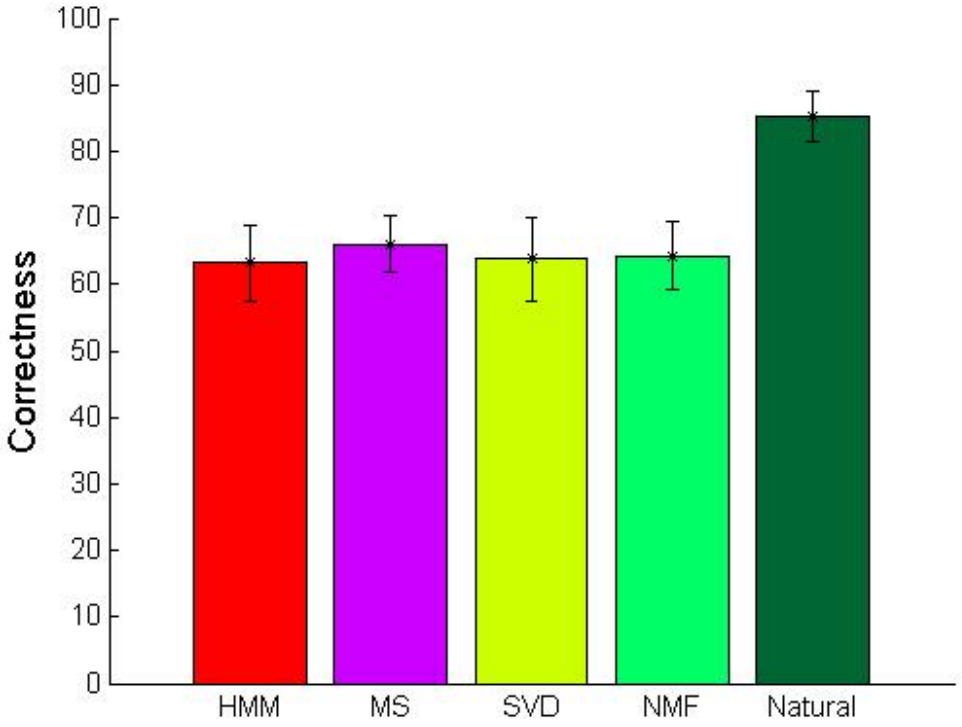

Figure 4.4: MRT correctness in large data condition with 95% confidence interval

Chapter 5

Conclusion

5.1 Summaries

The ultimate goal of the research is improving naturalness of HMM-based synthesized speech without violating its acceptable intelligibility. To do so, naturalness and intelligibility are decomposed using asymmetric bilinear model. In order to apply asymmetric bilinear model in the task of decomposing naturalness and intelligibility, two main problems are addressed and tackled. The first problem is finding efficient acoustic feature strongly associating to naturalness. The naturalness associated acoustic feature will be improved, whist the intelligibility relating features will be preserved. The second problem is finding appropriate constrains for naturalness and intelligibility decomposition.

In the first problem, an experiment is conducted to find an efficient acoustic feature strongly relating to naturalness. In the experiment, several kinds of acoustic features were prepared. They are fundamental frequency F0, formant-related features: LPC, LSF, PLP and fine structure-related coefficients: MFCC, MCC. One pair of natural utterance and synthesized utterance of the same sentence is prepared. For each examined acoustic feature, feature sequences are exchange between natural speech and synthesized one. The purpose of exchanging is to reduce the quality of natural speech and improve naturalness of synthesized speech. If the phenomenon happens strongly with considered acoustic feature, the acoustic feature strongly relates to naturalness. To compare naturalness of stimuli obtained after exchanging different kinds of acoustic feature, Scheffe pair comparison listening test was conducted. Experimental results shows that the preference score of synthesized speech increase the most after exchanging MCC. The preference score of actual speech changed in opposite direction after exchanging MCC. It indicates MCC as the most-related acoustic feature to naturalness. The same phenomenon happened with exchanging LSF. However, it is not as strong as exchanging MCC. It indicates that fine structure is more important than formant information in perceiving naturalness of speech. MCC is a representation of speech spectra in frequency domain. Over-smooth effect of synthesized speech happens in both frequency and time domain. To make the feature becomes more efficient in both frequency and time domain, modulation spectrum (MS) of Mel-cepstral coefficient is utilized.

In second problem, non-negativity constrain is introduced to asymmetric bilinear model using NMF. Experiments are conducted to examine the physical interpretation of factorized matrices using NMF. In the experiment, asymmetric bilinear model using NMF was applied to natural speech, then to HMM-based synthesized speech. Then, we exchange the naturalness matrices of naturalness speech and synthesized speech and re-synthesize them Experimental results show that exchanging naturalness matrices reduces the naturalness of actual speech and improve naturalness of synthesized speech. It indicates that naturalness matrix strongly relates to naturalness of speech.

The columns of naturalness matrix are important. They contain the information about magnitude of cepstral coefficients. In region of high order cepstral coefficients, the magnitudes of the coefficients in natural speech is higher than those of coefficients in synthesized speech. Exchanging naturalness matrices means emphasizing high-order coefficients in synthesized speech. This enhances the fine-structure of speech spectra in synthesized speech.

An MRT was also conducted to investigate to relation of intelligibility matrix with speech intelligibility. Experimental results shows that the intelligibility of actual speech is reduced a lot; whilst the intelligibility of synthesized speech is increased a lot after exchanging. It indicates a strong relation of intelligibility matrix and speech intelligibility.

Finally, proposed method using NMF is compared with default asymmetric bilinear model using SVD in the task of improving naturalness without violating intelligibility of HMM-based synthesized speech. Moreover, proposed method is compared with other methods using GV and MS to show how well its performance is. All of the methods are tested in limited data condition (with only 5 training actual speech) and large data condition.

In limited data condition, the preference score of proposed method is higher than those of other methods. It indicates that proposed method outperforms other methods in improving naturalness in limited data condition. In large data condition, the preference score of proposed methods are comparable with other methods. It indicates competitive proposed method in large data condition.

Conducted MRT test in limited data condition shows that proposed method using NMF and default asymmetric bilinear model using SVD have the same word correctness with HMM-based synthesized speech. It indicates that using asymmetric bilinear model can preserve the acceptable intelligibility of synthesized speech. The same trend happens in large data condition in which intelligibility of HMM-based synthesized speech is preserve after improving naturalness with proposed method.

5.2 Contributions

In the research, naturalness of HMM-based synthesized speech is improved based on naturalness and intelligibility decomposition.

Firstly, an efficient acoustic feature strongly relating to naturalness was figured out as MCC. To improve the efficiency of MCC in capturing fine structure in both frequency domain and time domain, modulation spectrum of MCC sequences in time domain was

utilized.

Secondly, an asymmetric bilinear model using non-negative matrix factorization was proposed. The non-negativity constrain is proved to be efficient in decomposing naturalness and intelligibility. Moreover NMF also revealed physical interpretation of factorized matrices. Naturalness matrix contains information about magnitude of cepstral coefficient. Intelligibility matrix contains weights of the coefficients. Naturalness improvement means improving magnitude of high-order cepstral coefficient in naturalness matrix to enhance the fine-structure of speech spectra.

5.3 Future works

Although proposed method outperforms other methods in limited data condition, there is still many room to improve proposed method in large data condition. The parameters for NMF need to be optimized. Fine tuning for the parameters of proposed method is necessary work.

Moreover, although non-negativity constrain was proved to be efficient, finding other constrains to improve the performance of proposed method in large data condition is necessary work. Other constrains need to be considered to ensure good and stable performance of proposed methods.

Acknowledgement

I would first like to thank my thesis advisor Professor Masato Akagi of the School of Information Science at Japan Advanced Institute of Science and Technology. The door to Prof. Akagi office was always open whenever I ran into a trouble spot or had a question about my research or writing. He consistently allowed this paper to be my own work, but steered me in the right the direction whenever he thought I needed it.

I would also like to acknowledge Associate Professor Masashi Unoki of the School of Information Science at Japan Advanced Institute of Science and Technology as the second reader of this thesis, and I am gratefully indebted to his for his very valuable comments on this thesis.

Finally, I must express my very profound gratitude to my parents for providing me with unfailing support and continuous encouragement throughout my years of study and through the process of researching and writing this thesis. This accomplishment would not have been possible without them. Thank you.

Author

Dinh Anh Tuan

Bibliography

[1] H. Zen, K. Tokuda and W. Black, Statistical parametric speech synthesis Speech Comm., vol. 51, no. 11, pp. 1039-1064, 2009.

[2] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, T. Kitamura; Speech parameter generation algorithms for HMM-based speech synthesis; 2000.

[3] S. King; A beginner's guide to statistical parametric speech synthesis; June 2010.

[4] K. Tokuda, T. Masuko, N. Miyazaki, and T. Kobayashi; Multi-space probability distribution HMM; IEICE Trans, Inf. Syst., vol. E85-D, no. 3, pp. 455-464, 2002.

[5] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X.-Y. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland; The Hidden Markov Model Toolkit (HTK) Version 3.4; 2006

[6] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura; Duration modeling for HMM-based speech synthesis; in Proc. Int. Conf. Spoken Lang. Process. 1998, pp. 29-32

[7] H. Zen, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura; Hidden semi-Markov model based speech synthesis systems; IEICE Trans. INf. Syst., vol. E90-D, no. 5, pp. 825-834,2007

[8] K. Tokuda, H. Zen, and A. W. Black; An HMM-based speech synthesis system applied to English; in Proc. IEEE Speech synthesis Workshop, 2002, pp. 227-230

[9] J. J. Odell; The use of context in large vocabulary speech recognition; Ph.D. thesis, Queens College, Cambridge, UK, 1995.

[10] K. Shinoda and T. Wantanabe; MDL-based context-dependent sub-word modelling for speech recognition; J. Acoust. Soc. Jpn, vol. 21, no. 2, pp. 79-86, 2000.

[11] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura; Simultaneously modelling of spectrum, pitch, and duration in HMM-based speech synthesis; Proc. Eurospeech, 1999. pp. 2347-2350.

[12] T. Toda and K. Tokuda. A speech parameter generation algorithm considering global variance for HMM-based speech synthesis. IEICE Trans, Vol. E90-D, No. 5, pp. 816-824, 2007.

[13] S. Takamichi, T. Toda, A. W. Black, and S. Nakamura. Parameter generation algorithm considering modulation spectrum for HMM-based speech synthesis. In ICASSP, pp. 4210-4214, 2015.

[14] Y. Jiao, X. Xie, X. Na, M. Tu; Improving voice quality of HMM-based speech synthesis using voice conversion method; in ICASSP, pp. 7964-7968, 2014.

[15] L. H. Chen, T. Raitio, C. V. Botinhao, J. Yamagishi, Zhen-Hua Ling. DNN-based stochastic post-filter for HMM-based speech synthesis; in Interspeech, pp. 1954-1958, 2014.

[16] J. Tenenbaum, W. Freeman; separating style and content with bilinear models; Neural Computation; pp. 1247-1283, 2000.

[17] V. Popa, J. Nurminen, M. Gabbouj; A novel technique for voice conversion based on style and content decomposition with bilinear models. In Interspeech, pp 2655-2658, 2009.

[18] H. Scheffe, An analysis of variance for paired comparisons, Journal of the American Statistical Association, vol. 37, pp. 381-400, 1952.

[19] P.C. Nguyen, T. Ochi, and M. Akagi; Modified restricted temporal decompostion and its application to low rate speech coding; IEICE Transactions on Information and Systems, vol. E86-D, no. 3,2003.

[20] Lee, D.D. and Seung, H.S.; Learning the Parts of Objects by Non-negative Matrix Factorization; Nature, pp. 788791, 1999.

[21] S. Takamichi, T. Toda, G. Neubig, S. Sakti and S. Nakamura; A post-filter to modify the modulation spectrum in HMM-based speech synthesis. In ICASSP, pp. 290-294,2014

[22] A. S. House, C. E. Williams, M. H. L. Hecker, and K. D. Kryter; Articulation-testing methods: consonantal differentiation with a closed-response set; Journal of Acoustic Society American 37; pp. 158-166; 1965.

[23] P. N. Binh and M. Akagi; Efficient modelling of temporal structure of speech for applications in voice transformation; Interspeech, pp. 294-299, 2009.

[24] M. Tatham, K. Morton; A guide to speech production and perception; p. 18. 2011.

[25] F. Itakura; Line spectrum representation of linear predictor coefficients of speech signals; J. Acoust. Soc. Amer., vol. 57, 1975.

[26] H. Hermansky; Perceptual linear predictive (PLP) analysis of speech; J. Acoust. Soc. Amer., vol. 87, 1990.

[27] S. Imai; Cepstral analysis synthesis on the mel frequency scale; pp. 93-96, ICASSP 83.

[28] H. Kawahara, I. Masuda-Katsue, A. de Cheveigne, Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous frequency-based F0 extraction: Possible role of a –repetitive structure in sounds, J. Speech Communication, vol. 27, pp. 187–207, 1999.

[29] Y. Stylianou, O. Cappe, E. Moulines, "Continuous probabilistic transform for voice conversion", IEEE Trans. Audio, Speech, Lang. Process., vol. 6 pp. 131–142, 1998.

[30] K. Tokuda, T. Masuko, S. Imai; Mel–generalized cepstral analysis – a unified approach to speech spectral estimation, Proc. ICSLP, pp. 1043–1046, 1994.