NAIST-IS-DD1361007

DOCTORAL DISSERTATION

Acoustic modeling and speech parameter generation for high-quality statistical parametric speech synthesis

Shinnosuke Takamichi

March 14, 2016

Graduate School of Information Science
Nara Institute of Science and Technology

A DOCTORAL DISSERTATION
submitted to Graduate School of Information Science,
Nara Institute of Science and Technology
in partial fulfillment of the requirements for the degree of
Doctor of ENGINEERING

Shinnosuke Takamichi

Thesis Committee:

Professor Satoshi Nakamura	(Supervisor)
Professor Yuji Matsumoto	(Co-supervisor)
Professor Tomoki Toda	(Nagoya University)
Professor Alan W. Black	(Carnegie Mellon University)
Assistant Professor Sakriani Sakti	(Co-supervisor)
Assistant Professor Graham Neubig	(Co-supervisor)

Acoustic modeling and speech parameter generation for high-quality statistical parametric speech synthesis*

Shinnosuke Takamichi

Abstract

Speech is one of the natural ways for people to communicate, and speech synthesis is a technique to synthesize a speech waveform through a computer. The field of speech synthesis studies includes Text-To-Speech (TTS) synthesis in which speech is generated from arbitrary text and Voice Conversion (VC) that converts the input speech into the another speech having desired non-/para-linguistic information. Speech synthesis is used in many applications that are very helpful for human-to-human and human-to-computer communication, such as a speech-to-speech translation and spoken dialogue systems.

Thanks to developments in machine learning techniques and computational environments, statistical approaches have come into the main stream of recent speech synthesis research. Although many state-of-the-art methods have been proposed, Hidden Markov Model (HMM)-based TTS and Gaussian Mixture Model (GMM)-based VC have gained popularity thanks to their solid mathematical foundation. However, they have a drawback; the quality of the synthetic speech they produce is not high. That is, their synthetic speech often sounds muffled and can be easily distinguished from natural speech. There are three main reasons causing this problem; parameterization errors in the analysis/synthesis stage, inaccurate modeling in the training stage, and over-smoothing in the synthesis stage. This thesis mainly addresses the latter two reasons, which are more critical than the parameterization errors.

In conventional HMM-based TTS and GMM-based VC, some speech segments are averaged in order to construct the acoustic models, i.e., HMMs and GMMs. Consequently, this modeling loses information on the individual speech segments, and speech parameters generated from the models are also averaged. To address these inaccuracies, this thesis introduces ideas of unit selection synthesis, which directly uses speech waveform segments. The individual speech segments are first

*DOCTORAL DISSERTATION, Graduate School of Information Science, Nara Institute of Science and Technology, NAIST-IS-DD1361007, March 14, 2016.

modeled with different acoustic models that are robust to the unseen input features (called *rich context models*), and then they are used to construct a mixture model (called *Rich context GMM (R-GMM)*). Preserving the information of the individual speech segments in this way yields higher-quality speech than that of basic HMM-based TTS and GMM-based VC. Moreover, the proposed method has the capability of directly utilizing the mathematical foundations of HMM-based TTS and GMM-based VC.

The over-smoothing effect is a main cause of quality degradation in the synthesis stage. This phenomenon is one in which the generated speech parameters are overly smoothed compared with the natural speech parameters. One promising approach to alleviating the over-smoothing effect is to extract a specific feature to quantify the effect and to generate speech parameters so that their corresponding features become more similar to those of natural speech parameters. Although the Global Variance (GV) is a well-known such feature, the gap in quality between natural and the synthetic speech it produces is still large. This thesis introduces a new feature more sensitively correlated to the over-smoothing effect than the GV, the *Modulation Spectrum (MS)*. The MS of a speech parameter sequence is defined as the power spectrum of the sequence, and can be regarded as a mathematical extension of the GV. This thesis also proposes a MS-based post-filter that modifies the MS of the generated speech parameters. Because the process is performed separately from the basic HMM-based TTS and GMM-based VC, the post-filter has the capability of not only improving the quality of synthetic speech but also having the portability to apply to various speech synthesis systems.

We further propose joint optimization algorithms for the basic acoustic models, HMMs and GMMs, and a statistical model of the MS. The proposed MS-based post-filter improves the MS criterion, but the basic training and generation criteria using HMMs and GMMs are degraded because the post-filtering process completely ignores them. To address this issue, the algorithms jointly optimize these criteria. This thesis first integrates the MS criterion into the speech parameter generation to directly alleviate the over-smoothing effect in the synthesis stage. The objective function is iteratively updated to generate speech parameters. Furthermore, this thesis integrates the MS criterion into the training stage to perform high-quality and computationally-efficient speech synthesis. The training algorithm trains trajectory HMMs and GMMs under the MS constraint. It makes it possible to directly utilize the computationally-efficient basic parameter generation algorithm, while compensating the MS of the generated speech parameters. We conducted several experimental evaluations on the proposed methods. The results demonstrate that (1) our generation algorithm achieved higher quality than that of the conventional generation algorithm considering the GV, and (2) our training algorithm achieved the best quality while preserving computational

efficiency, among the several training algorithms tested.

Keywords:

text-to-speech synthesis, voice conversion, hidden Markov model, Gaussian mixture model, acoustic modeling, speech parameter generation

高音質な統計的パラメトリック音声合成のための音響モデリング法と音声パラメータ生成法*

高道 慎之介

内容梗概

音声は，人間にとって基本的なコミュニケーションツールのひとつである．音声合成は，コンピュータにより音声を人工的に合成する技術であり，本論文では，任意のテキストから音声波形を合成するテキスト音声合成（Text-To-Speech: TTS）技術と，入力音声波形を異なる音声情報を持つ音声波形に変換する声質変換（Voice Conversion: VC）技術を指す．この技術は，音声対話や音声翻訳等の音声コミュニケーションシステムを初めとする広い応用を見据え，活発に研究されている．

機械学習と計算環境の発達により音声合成に対する多様な統計的手法が提案される中，本論文で取り扱う隠れマルコフモデル（Hidden Markov Model: HMM）に基づくテキスト音声合成（HMM 音声合成）と混合正規分布モデル（Gaussian Mixture Model: GMM）に基づく声質変換（GMM 声質変換）は，数理モデルとしての強力性・柔軟性から，強く支持されている．一方で，これらの手法において合成される音声の音質は，自然音声と比較して著しく劣化する．その要因は，分析・生成部におけるパラメータ表現のエラー，学習部における不正確な音響モデリング，また，生成部における過剰な平滑化の３要因に分類されるが，本論文では特に，後者の２要因に焦点を当て，合成音声の音質改善に取り組む．

学習部における音響モデリングの問題点の一つは，統計モデリングにおける平均化処理により，個々の音声波形の情報が消失する点である．これに対し本論文では，直接的に音声波形を利用する素片選択合成法の考えを導入する．提案法では，各音声波形の情報を，未知の入力情報に対して頑健な統計モデル（Rich context model）として保持し，更に，複数の統計モデルから一つの混合モデル（Rich context GMM: R-GMM）を構築する．これにより，従来の HMM 音声合成と GMM 声質変換の柔軟性を保持しつつ，より高音質な音声を合成可能となる．

次に，生成される音声パラメータの過剰な平滑化は，合成部における音質劣化の主要因である．平滑化現象を定量化する特徴量を生成パラメータから抽出し，それを自然音声パラメータの特徴量に近づく様に補償することで，この過剰な平滑化の問題は緩和される．従来，系列内変動（Global Variance: GV）と呼ばれ

*奈良先端科学技術大学院大学 情報科学研究科 博士論文, NAIST-IS-DD1361007, 2016 年 3 月 14 日.

る特徴量が広く利用されてきたが，未だに，過剰な平滑化の影響は大きい．本論文では，過剰な平滑化を一層効果的に定量化する新たな特徴量として，変調スペクトル（Modulation Spectrum: MS）を導入する．MS は，音声パラメータ系列のパワースペクトルとして定義され，GV の拡張形としてみなされる特徴量である．本論文では更に，生成パラメータ系列の MS を補償するポストフィルタを提案する．このフィルタ処理は，通常の HMM 音声合成・GMM 声質変換から独立した処理であるため，MS による音質改善効果のみならず，他の音声合成方式への容易な移植性をもたらす．

我々は更に，ポストフィルタ処理において個別に考慮されていた，音響モデル（HMM と GMM）と MS の統計モデルの基準を，同時に最適化するアルゴリズムを提案する．まず，生成部における過剰な平滑化を直接的に緩和するため，音声パラメータ生成アルゴリズムに対して，MS の統計モデルを組み込む．新たに定義された目的関数を反復的に最大化することで，高音質な音声パラメータ系列を生成する．また，高音質かつ高速な音声合成を可能にする手法として，学習部に対して MS を組み込む手法を提案する．生成される音声パラメータ系列の MS を補償するように音響モデルを学習するため，従来の高速な音声パラメータ生成法を利用しながらも合成音声の音質改善が可能となる．実験的評価では，(1) 提案するパラメータ生成法は、従来の GV を考慮したパラメータ生成法を超える音質改善効果をもたらすこと，(2) 提案する学習法は，高速な音声合成能力を保持しつつ従来の学習法を超える音質改善効果をもたらすことを示す．

キーワード

音声合成，声質変換，隠れマルコフモデル，混合正規分布モデル，音響モデリング，音声パラメータ生成

Acknowledgements

まず，研究室活動全般に渡りご指導下さった中村 哲教授に深く感謝致します．先生からは，知能コミュニケーション研究室の創設から 5 年間，研究者としての在り方から，指導者としての考え方，人生訓に至るまでの幅広いご教授を頂戴しました．また，身に余るほどのご支援をいくつも承りました．感謝致します．

奈良先端科学技術大学院大学の松本 裕治教授には，本論文の副指導教員を引き受けて頂き，また，本論文を製錬するためにコメントして頂きました．感謝致します．

名古屋大学の戸田 智基教授（前 奈良先端科学技術大学院大学 准教授）には，知能コミュニケーション研究室の音声合成グループにて，数え切れぬ程，ご指導頂きました．思えば，入学以前に拝読した，戸田先生の論文の技術に憧れ，本学へ入学しました．そこから 5 年間の手厚いご指導とメンターシップに感謝致します．

Carnegie Mellon University の Alan W. Black 教授には、本論文の後半部分の研究についてご指導頂きました。拙い英語コミュニケーション能力しか持たぬ私に、丁寧かつ貴重な指導を下さった事を感謝致します。

奈良先端科学技術大学院大学の Sakriani Sakti 助教と Graham Neubig 助教からは，普段の研究報告から他分野からの貴重なコメントを継続的に頂きました．また，若手研究者としての立場として多くのアドバイスを頂きました．感謝致します．

奈良先端科学技術大学院大学の松田 真奈美 秘書には，日々の生活，また，研究室内外に向けた活動を厚くサポートして頂いたことを感謝致します．

鹿野 清宏 奈良先端科学技術大学院大学名誉教授には，修士論文までの副指導教員を引き受けて頂いたことを感謝致します．

情報通信研究機構 ユニバーサルコミュニケーション研究所の木俵 豊 研究所長，河井 恒 音声コミュニケーション研究室室長，及び，志賀芳則 同研究室主任研究員からは，貴研究室でのインターンシップ及び短時間研究員としての機会とご指導を頂戴しました．特に河井室長と志賀研究員からは，linux の使い方も知らずプログラムも書けなかった私に手厚いご指導を承りました．感謝いたします．

博士後期課程の間、私の研究に関する訪問講演を行いました。講演依頼を快く引き受けて下さった、University of Edinburgh の Simon King 教授、Cambridge University の Mark J. F. Gale 教授、Google London の Heiga Zen 博士、大阪大学の川村 新 准教授に感謝致します。また、訪問講演や学会でお会いした国内・国外の若手研究者の皆様に対して、皆様と共に新たなる未来を創造できることを感謝致します。

日本音響学会の関連では，若手を活性化する活動に参加させて頂きました．貴重な機会を下さった，北陸先端科学技術大学院大学の森川 大輔 助教，並びに，森川助教の率いる日本音響学会 学生・若手フォーラムの皆様に感謝致します．

長岡技術科学大学の島田 正治 名誉教授，穂刈 治英 技術職員，またサウンドコミュニケーション研究室の諸先輩方には，学部生だった私の人間性と信号処理を徹底的に叩き直して頂いた．皆様に鍛えて頂いたからこそ，研究成果の基盤があります．皆様に感謝致します．また，島田 名誉教授，穂刈技術職員には，大学

卒業後も変わらずご指導いただきました．改めて感謝致します．
　知能コミュニケーション研究室の学生の皆様にも数え切れぬ程お世話になりました．尊敬する諸先輩方には，その研究への姿勢のみならず，先輩の斯くあるべきを教えて頂きました．同期諸君とは，私事・公事で互いに支えあったのみならず，知能コミュニケーション研究室第 1 期生として，研究室の立ち上げに共に立ち向かいました．我ながら未熟な人間であったが，後輩諸君はついてきてくれました．スペースの都合から全員の名前を記載しないが，皆様に感謝致します．
　最後に，本論文は家族と友の支え無しに完成しませんでした．友に感謝するとともに，博士課程を中退しようとした私に「自分の信じる道を貫いて欲しい」と言ってくれた両親，父・高道 修二と，亡き母・高道 啓子への最大限の感謝をもって，全ての皆様への感謝の言葉とします．

Contents

List of Figures

List of Tables

Chapter

1

Introduction

1.1 General background

Objects in children's dreams are bestowed with magical powers and no boundaries (or limitations) exist between them, unlike objects in the real world. I believe that research is an action toward blurring such boundaries between objects and realizing the dreams. The target of this thesis is speech, and the boundaries we want to remove are ones between humans and computers and ones between human beings.

Speech is one of a natural medium for people to communicate. It is used for not only delivering verbal information but also conveying non-verbal information, such as emotions, characteristics, speaker individuality, and so on. Speech plays the most important role in human communication. Moreover, thanks to the development of Text-To-Speech (TTS) synthesis [4], which is a technique to synthesize speech from arbitrary text, computers can now speak, and many different speech-based applications have been deployed as aids in human-to-human and human-to-computer communication, including speech-to-speech translation and spoken dialogue systems. In addition, Voice Conversion (VC) [5], which is a technique to convert para-/non-linguistic information while keeping the linguistic information unchanged, has an important role in such applications. VC makes it possible to augment speech production beyond the physical constraints and limitations of an individual human being, such as his/her skills of expression and language knowledge.

Developments in machine learning and computational environments have enabled speech synthesis systems including TTS and VC to be automatically constructed using pre-recorded data. Generally, this type of speech synthesis is called *corpus-based speech synthesis* [4]. The corpus-based approach has yielded dramatic improvements in synthetic speech quality because researchers have been able to share the common knowledge, findings, and corpora. There are two main synthesis techniques; unit selection synthesis (sample-based speech synthesis) and statistical parametric speech synthesis. Unit selection synthesis directly uses acoustic inventories selected from a speech corpus for synthesizing speech waveforms. One of its main advantages is that high-quality speech keeping the original voice characteristics can be synthesized by concatenating natural speech segments. Here, the characteristics of the generated speech are fully dependent on original voices.

On the other hand, statistical parametric speech synthesis [6], which utilizes statistical models, was established in the 1990s [7, 5]. The input parameters (textual parameters for TTS and speech parameters for VC) are first extracted from the input information in an analysis stage, and then the relationship between the input parameters and output speech parameters are represented using the statistical models in a training stage. Speech parameters corresponding to the input parameters are generated from the trained models, and finally, the speech

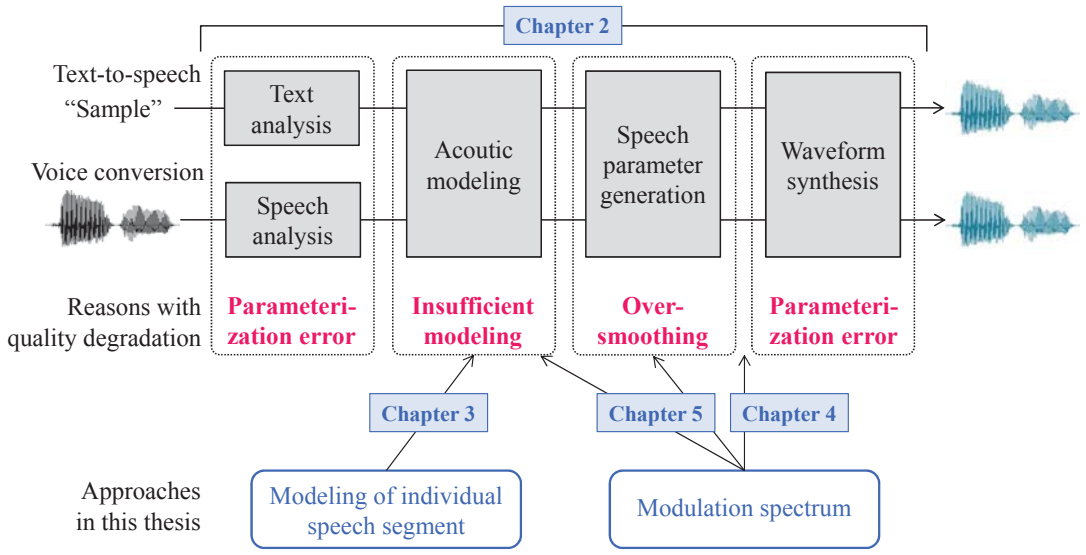


Figure 1. Problem definition and outline of the of this thesis. This thesis mainly deals with the problems of inaccurate modeling and the over-smoothing effect. The parameterization error is addressed in the appendices.

waveform is synthesized in a synthesis stage. Many state-of-the-art methods have been proposed, but HMM-based TTS [8] and GMM-based VC [9] are widely used in speech communication systems, thanks to their stability, strong mathematical foundation, and flexibility. They have advantages in controlling the characteristics of synthetic speech and having much more flexibility in how they can be used in comparison with unit selection synthesis.

However, the drawback of HMM-based TTS and GMM-based VC is the poor quality of the synthetic speech they produce. Their speech often sounds muffled, it can be easily distinguished from natural speech. There are three main reasons for this problem [6] as shown in Fig. 1: parameterization error in the analysis/synthesis stage [10, 11, 12], inaccurate modeling in the training stage [13, 14], and an over-smoothing effect in the synthesis stage [15, 16].

1.2 Thesis scope

This thesis addresses quality improvements of synthetic speech in HMM-based TTS and GMM-based VC, mainly focusing on the latter two factors of the quality degradation, which are often more critical than the parameterization error.

1.2.1 Better acoustic modeling that preserves information of individual speech parameters

Statistical models are used in statistical parametric speech synthesis. Inaccurate modeling here indicates that the modeling lacks several important aspects. For example, a model may (1) lose the relationship between input and output parameters: i.e., the acoustic model does not transmit the input parameters important for speech production. It may also put (2) unrealistic constraints on the output speech: it produces speech that is improbable. Finally, it loses (3) the information of the individual speech parameters. Unlike unit selection synthesis that utilizes individual speech parameters, statistical modeling averages several speech parameters are averaged, and loses information on the individual speech parameters.

In an attempt to exploit the excellent quality of unit selection synthesis, some hybrid methods have been proposed that combine HMM-based speech synthesis and unit selection synthesis. Maximum Likelihood (ML)-based unit selection synthesis [17] is a hybrid method to improve the quality of synthetic speech. Suitable waveform segments are searched for and taken from a speech corpus to maximize HMM likelihood. The use of waveform segments dramatically improves speech quality compared with that in HMM-based TTS. A hybrid approaches having more flexibility than unit selection is to develop rich context models that represent the individual waveform segments with probability distributions of individual speech component parameters [18]. In the synthesis stage, the probability distributions are jointly selected for all speech parameter components. Although these methods yield significant improvements in quality, they lose the flexibility of the original HMM-based TTS because their formulation is completely different.

This thesis proposes ML-based speech parameter generation methods using rich context models as hybrid methods that preserve the flexibility of the HMM-based TTS. The trained rich context models are used for constructing a *Rich context GMM (R-GMM)*. Furthermore, we extend this idea to GMM-based VC. In the synthesis stage, given the input parameters, the speech parameter sequence is iteratively generated by using R-GMMs. Because the proposed methods share the formulation of the original HMM-based TTS and GMM-based VC, they have more flexibility than the other hybrid methods. For initializing the iterative generation, we generate a less-smoothed but highly discontinuous initial speech parameter sequence from over-trained acoustic models and then refine out the discontinuities through iterative generation.

1.2.2 Better speech parameter generation using a metric quantifying the over-smoothing effect

Speech parameters generated from trained statistical models tends to be over-smoothed compared with natural speech parameters. One promising approach to alleviating the over-smoothing effect is to extract a specific feature to quantify the effect and to generate speech parameters so that their corresponding features become more similar to those of natural speech parameters. One widely known example of such a feature is the Global Variance (GV) [19, 9], which is defined as a second-order moment of the speech parameter sequence. Considering the GV during the speech parameter generation has been shown to alleviate the over-smoothing effect and significantly improve the quality of the synthetic speech. However, despite that the GV-based metric is widely used [20, 21, 22], its use in the parameter generation tends to generate artificial sounds [21] and the quality gap between natural and synthetic speech is still large.

This thesis introduces a new feature more sensitively correlated to the over-smoothing effect than the GV, the *Modulation Spectrum (MS)*. The MS of a speech parameter sequence is defined as the power spectrum of the sequence. The linear-scaled MS is a second order moment of the parameter sequence, the same as the GV, and can be regarded as a mathematical extension of it. This thesis additionally proposes three methods using the MS.

MS-based post-filter A post-filter is the simplest way to compensate for over-smoothing effect. It is performed after the standard speech parameter generation, and it filters the generated speech parameters. The approach proposed in this thesis modifies the generated speech parameter sequence so that its MS becomes more similar to that of natural speech. The post-filter modifies the MS utterance by utterance and can be automatically constructed using natural speech and synthetic speech as training data. This utterance-level post-filter is further extended to a segment-level post-filter to modify the MS segment by segment in order to generate parameters without much delay. Because the post-filtering process is independent from the original speech synthesis process, it can be applied to a variety of speech synthesis systems.

Speech synthesis integrating the MS The post-filtering approach is effective but still limited because it possibly causes adverse effects due to it completely ignores the basic criteria, i.e., the HMMs or GMMs. Moreover, integrating the MS into speech synthesis procedures is straightforward to apply various useful techniques such as model training and adaptation. Here, we propose to integrate the MS into the HMM-based TTS and GMM-based VC.

First, we propose a speech parameter generation algorithm considering the MS. The algorithm generates speech parameter trajectories by maximizing a novel

objective function consisting of the basic criterion and the MS likelihoods. The MS likelihood works as a penalty term to make the MS of the generated parameters close to that of natural ones. Although the algorithm recovers the MS of the generated speech parameters, it loses the basic computationally-efficient generation ability.

Consequently, we proposes a training algorithm as yet another MS approach to improve the speech quality while preserving the computationally-efficient generation capability. Here, we implement trajectory HMM training [23] for GMM-based VC, which is a training algorithm consistent with basic computationally-efficient generation. Then, we integrate the MS into the trajectory training for both HMM-based TTS and GMM-based VC. The HMMs or GMMs are trained to recover the MS of the speech parameters generated from them. Because this training algorithm is consistent to the basic generation algorithm and takes account of the MS in training, the basic generation method can be used without MS compensation in synthesis. In addition, the training algorithm makes it possible to perform input-parameter-dependent modeling of the MS.

1.3 Rest of this thesis

The rest of this thesis is organized as follows (see also Fig. 1.).

Chapter 2: We explain the basic frameworks of speech synthesis. After reviewing unit selection synthesis and statistical parametric speech synthesis, we describe acoustic modeling and speech parameter generation algorithm of HMM-based TTS and GMM-based VC. We also describe the conventional approaches of better for better acoustic modeling and speech parameter generation, which are hybrid approaches combining unit selection synthesis including conventional rich context modeling for HMM-based TTS, trajectory HMM training, and global variance.

Chapter 3: We propose statistical sample-based speech synthesis using rich context models. After applying rich context modeling to GMM-based VC, we reformulate the models as R-GMMs. We describe the iterative generation algorithm and its better initialization to generate speech parameters from the R-GMMs. Moreover, we discuss the proposed methods with the conventional hybrid approaches in term of their flexibility, then, we demonstrate quality improvements had by them in comparison with basic HMM-based TTS and GMM-based VC.

Chapter 4: We introduce the MS to quantify the over-smoothing effect observed in the generated speech parameters. First, an utterance-level MS-based post-filter is first proposed for the spectral, F_0, and HMM-state duration; then, its

process is localized at the segment level. We discuss the mathematical relationship between the GV and MS and describe an experimental evaluation confirming the quality gain had by using MS-based post-filters.

Chapter 5: We integrate the MS into the HMM-based TTS and GMM-based VC. First, we propose the speech parameter generation algorithm considering the MS, then, we propose an MS-constrained trajectory model training in HMM-based TTS and GMM-based VC. The these methods are compared with the conventional generation and training algorithms as to their effectiveness.

Chapter 6: We summarize this thesis and discuss the future directions of research.

Chapter

2

Speech synthesis

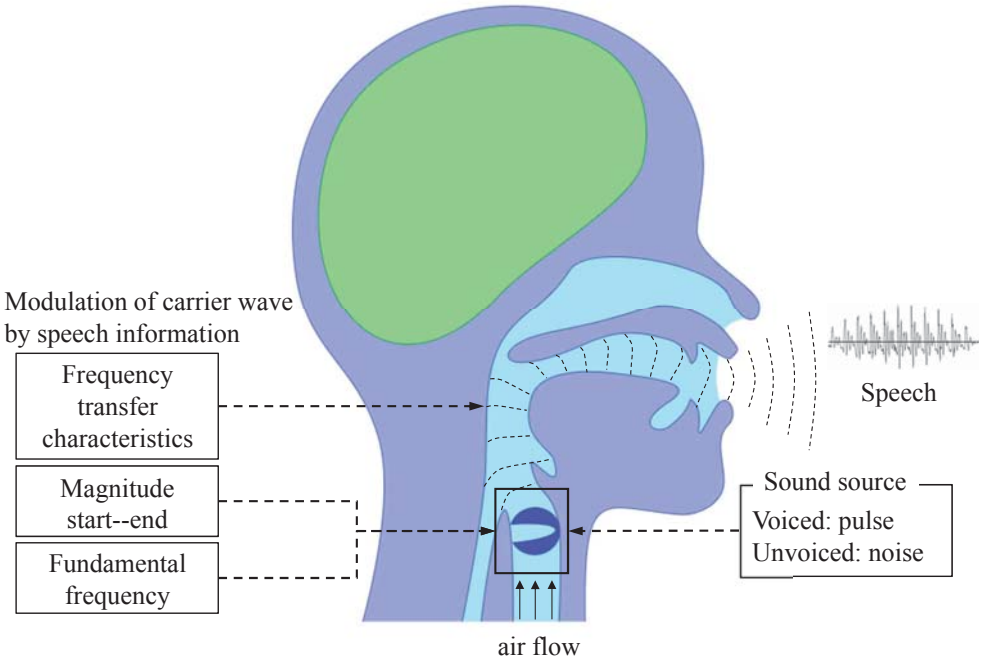


Figure 2. Outline of speech production process by human being [1]. The voiced excitation signal is given as a simple pulse signal in this figure, but in this thesis, it is given as the signal mixing the pulse (periodic) and noise (aperiodic) signals.

2.1 Introduction

A message a human being wants to produce is first translated into movements of articulators and organs. As shown in Fig. 2, air-flowing from the lungs generates vocal excitation signals containing periodic (by vocal cord vibration) and aperiodic (by turbulent noise) components. By filtering the source signals with time-varying vocal tract transfer functions controlled by the articulators, their frequency characteristics are modulated. Finally, the filtered source signals are emitted. The TTS [4] procedure mimics these actions with a computer. The produced speech waveform contains time-varying components that control the linguistic/para-linguistic features and time-invariant non-linguistic components that incorporate physical characteristics such as the vocal tract length and shape of the vocal cords. The VC [24] procedure dissociates these components and converts the physical constraint components into others.

The TTS and VC processes are different in terms of their input type (see Fig. 3.), a discrete-to-continuous value conversion for TTS and continuous-to-continuous value conversion for VC, but some of their internal processes are the same. Researchers in the past tried to synthesize speech based on their individual

Figure 3. Speech synthesis techniques used in this thesis; Text-To-Speech (TTS) and Voice Conversion (VC). TTS and VC do not use the same input types, but they share in common certain internal processes.

rules [25], while modern speech synthesis systems usually take a corpus-based (or data-driven) approach [4]. The training corpus consists a collection of pairs of input raw texts and output speech waveforms for TTS, and input/output speech waveforms for VC[1]. The corpora approach has yielded dramatic improvements of speech synthesis because it enables researchers to share the common knowledge, findings, and corpora.

Currently, there are two main approaches to speech synthesis; unit selection synthesis (also called sample-based speech synthesis) and statistical parametric speech synthesis. Unit selection synthesis [28, 29, 4, 30] synthesizes speech corresponding to input text by concatenating small segments of speech waveform stored the training corpora. One of the main advantages of concatenating natural speech segments is that it creates high-quality speech keeping the original voice characteristics [31]. However, the characteristics of the generated speech are fully dependent on the original voices.

On the other hand, statistical parametric speech synthesis [6] utilizes statistical models trained to fit the training corpora. It was established in the 1990s [7, 5], and has been used for about a decade. Nowadays, many technologies have been studied within this basic framework, including speech synthesis using Hidden Markov Models (HMMs) [8], Gaussian Mixture Models (GMMs) [9], Classification And Regression Trees (CART) [32], and kernel regression [33, 34]. Whereas unit selection synthesis directly uses waveform segments or natural speech parameter segments to synthesize a speech waveform, statistical parametric speech synthesis collects statistics from the speech parameter segments and utilizes them to generate the speech parameters used in waveform synthesis.

Many methods have been proposed, but HMM-based TTS [8] and GMM-based VC [9] have the gained popularity thanks to its stability, mathematical foundation and flexibility. These statistical modelings and synthesis frameworks make it possible to build small footprint synthesizers [35], adapt existing voices to other target voices by using only a small amount of speech data [36, 37], and flexibly control the voice characteristics of synthetic speech [38, 39, 40]. Moreover, the knowledge and techniques are easily applied from other research areas, such

[1] Corresponding raw texts are also required for text-dependent VC which converts voice through speech-to-text encoding and text-to-speech decoding processes, such as [26, 27], but we won't discuss such approaches in this thesis.

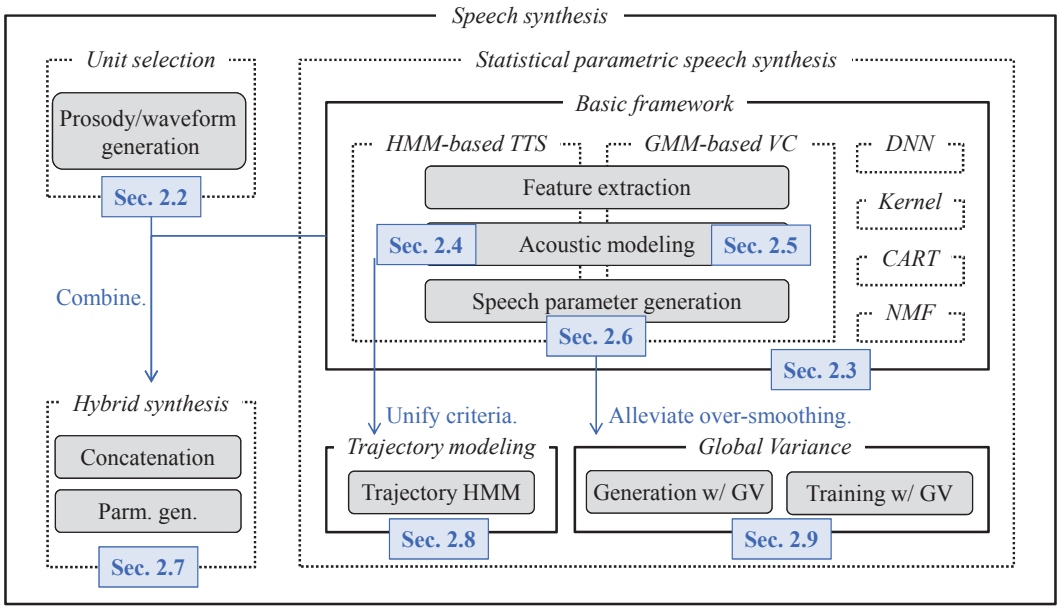


Figure 4. The rest of Chapter 2.

as HMM-based speech recognition [41] and GMM-based speaker verification [42]. In addition, HMM-based TTS and GMM-based VC can be easily combined with Deep Neural Net (DNN)-based speech synthesis [14, 43] which is a powerful but as yet unstable approach. On the other hand, a serious drawback of these methods compared with unit selection synthesis is the poor quality of the synthetic speech they produce.

In this chapter, we first describe the basic speech synthesis frameworks and review conventional approaches to high-quality speech synthesis. The rest of this chapter is organized as shown in Fig. 4. **Section 2.2** and **Section 2.3** reviewed the two approaches to speech synthesis, i.e., unit selection synthesis and statistical parametric speech synthesis. Acoustic modelings for HMM-based TTS and GMM-based VC are described in **Section 2.4** and **Section 2.5**. The generation of speech parameters from HMMs and GMMs is described in **Section 2.6**. **Section 2.7** and **Section 2.9** explain the conventional approaches for better acoustic modeling and speech parameter generation. The *hybrid approach* in **Section 2.7** introduces ideas of unit selection synthesis into the acoustic modeling using HMMs. *Trajectory modeling* in **Section 2.8** is a way to remove inconsistencies between training and synthesis in HMM-based TTS. The *Global Variance (GV)* in **Section 2.9** quantitatively captures the over-smoothing effect in the synthesis stage. **Section 2.10** is a summary of this chapter.

2.2 Unit selection synthesis

Unit selection synthesis directly uses acoustic inventories selected from a speech corpus for synthesizing a speech waveform. After predicting the target information, the speech parameter sequence or waveform segments are selected so as to minimize a defined cost. The speech waveform is synthesized by concatenating these selected segments. Although unit selection synthesis can produce high-quality speech by directly using speech segments, the voice characteristics are fully dependent on the original speech included in the acoustic inventories. Note that text/speech analysis stages are done not only for statistical parametric speech synthesis but also unit selection synthesis (we explain them in **Section 2.3**).

2.2.1 Target generation

The target information is predicted from the input. In TTS, prosodic features such as the F_0 contour, power contour, and phoneme duration are predicted from the contextual information corresponding to the input text[2]. Fujisaki's model [44] effectively represents the F_0 contour. This model decomposes the F_0 contour into two components, i.e., a phrase component that decreases gradually toward the end of a sentence and an accent component that increases and decreases rapidly at each accent phrase. The data-driven approach, e.g., HMM-based TTS [8], is also used to generate the target information [45]. For VC, input speech parameters are used as the target information [30]. To eliminate the difference between the input and output speech parameters, the input speech parameters are modified using methods such as Vocal Tract Length Normalization (VTLN) and pitch difference normalization [46].

2.2.2 Waveform segment selection

In waveform segment selection, an optimum set of waveform segments is selected from a speech corpus by minimizing the degradation in perceived naturalness caused by various factors, e.g., prosodic differences, spectral differences, and a mismatches of phonetic environments. A target cost and a concatenation cost are often used as standard selection measures as shown in Fig. 5. The optimum set is selected to minimize a cost function $C^{(\mathrm{us})}$ summarizing the target cost and the concatenation cost as follows:

$$C^{(\mathrm{us})} = \sum_{n=1}^{N} w_{\mathrm{t}}^{(n)} C_{\mathrm{t}}^{(\mathrm{us})}\left(t_n, u_n\right) + \sum_{n=2}^{N} w_{\mathrm{c}}^{(n)} C_{\mathrm{c}}^{(\mathrm{us})}\left(u_{n-1}, u_n\right), \tag{1}$$

[2] Consequently, this stage for TTS is called *prosody generation*.

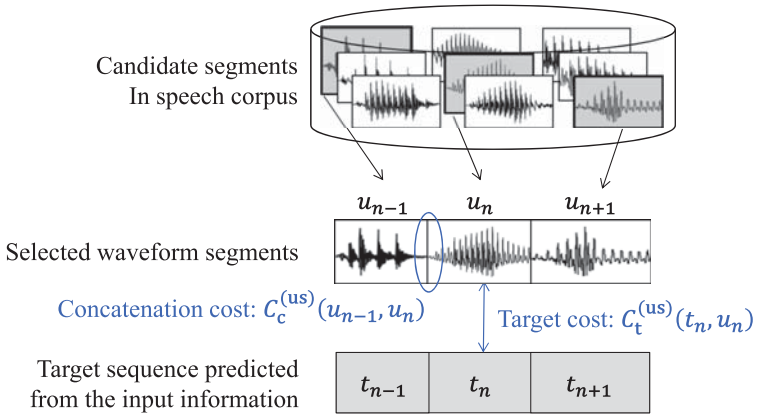


Figure 5. Waveform segment selection in unit selection synthesis. Small waveform segments are selected to minimize the weighted sum of the target costs and concatenation costs.

where t_n and u_n are the n-th target and candidate waveform segments, respectively, $C_{\mathrm{t}}^{(\mathrm{us})}(t_n, u_n)$ and $C_{\mathrm{c}}^{(\mathrm{us})}(u_{n-1}, u_n)$ are respectively the target cost function evaluating the difference between t_n and u_n and the concatenation cost function evaluating the discontinuity at a joint point between u_{n-1} and u_n, while $w_{\mathrm{t}}^{(n)}$ and $w_{\mathrm{c}}^{(n)}$ are respectively the weight of target and concatenation cost function for the n-th segment. The target cost captures the degradation in naturalness arising from prosodic differences, spectral differences, and differences between phonetic environments. The concatenation cost is an estimate of the quality of a joint point between consecutive waveform segments, and this cost function captures the degradation caused by concatenating waveform segments. The weight of each cost function is often determined manually on the basis of the results of perceptual experiments [47]. The sum of these two costs is minimized using a dynamic programming search.

2.2.3 Waveform synthesis

After waveform segment selection, an output speech waveform is synthesized by concatenating the selected waveform segments. However, if the prosody of the selected waveform segments is different from the predicted target information, the naturalness of the synthetic speech is degraded. This degradation can be alleviated by various methods such as Time-Domain Pitch-Synchronous OverLap-Add (TD-PSOLA) [48]. Unit selection synthesis generally needs a larger training corpus to alleviate the quality degradation caused by the signal processing.

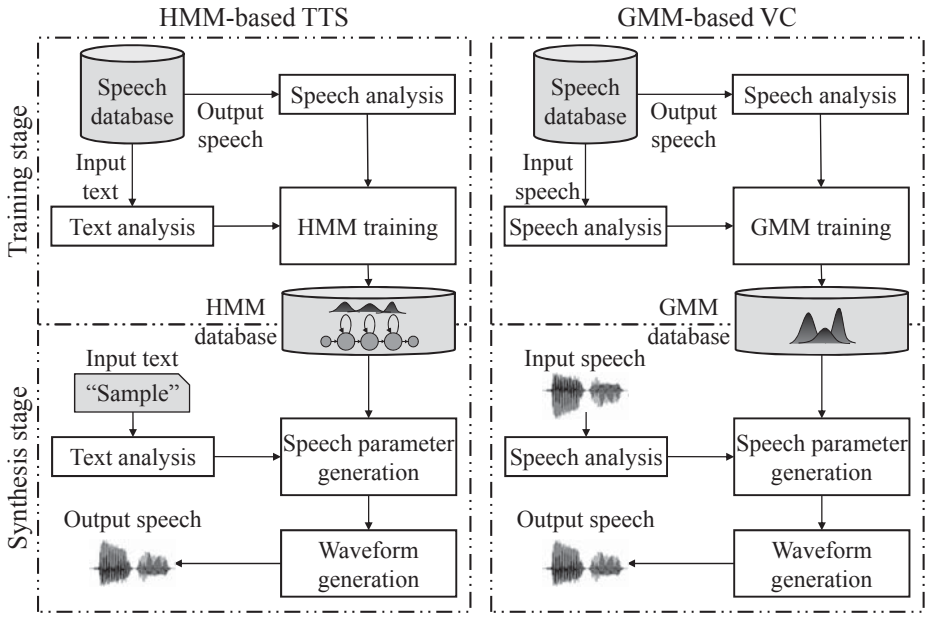


Figure 6. Statistical parametric speech synthesis procedures, e.g., HMM-based TTS and GMM-based VC.

2.3 Statistical parametric speech synthesis

Whereas unit selection synthesis directly utilizes the acoustic inventories, the speech waveforms of a speech corpus are first parameterized with text or speech analyzers, and then, instead of selecting a speech waveform, we select the statistical models trained to represent the relationship between the input and output features. There are three main modules: the *text/speech analysis* module, *training* module, and *speech parameter generation* module, as shown in Fig. 6.

2.3.1 Text analysis

The target language in TTS has an individual language system that controls the speech waveform, and the language-dependent contextual factors should be extracted from the text (for example, Japanese [49][3], English [51], and Chinese [52]). There is a variety of prosodic and duration systems, for example, for tone languages such as Chinese, pitch accent languages such as Swedish and Japanese, and intonation languages such as English [53, 54]. Japanese and English are also classified as mora-timed and stress-timed languages, respectively [55]. An example of contextual factors for Japanese TTS is shown in Fig. 7. The Japanese contex-

[3] The most popular text analyzer for Japanese is MeCab [50].

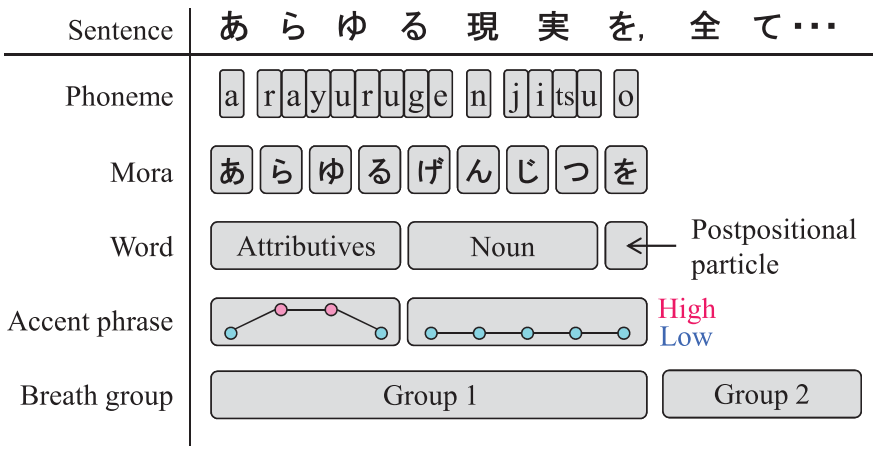


Figure 7. Example of contextual factors typically used in HMM-based Japanese TTS. In addition to the kinds of phoneme, and part-of-speech, their numbers within one phrase or utterance are also used as contextual factors.

tual factors include phoneme, mora, accent type, additionally, word, and breath group. These hierarchical contextual factors are composed into the phoneme level[4]. Consequently, there is an enormous number of combination of contextual factors (called as context label), and basically, one context labels appears only one time in the training corpora. This sparsity problem can be alleviated with the tree-based context clustering described in **Section 2.4** or dimensional reduction approaches [56, 57, 58][5]. For a variety of prosodies, additional context labels are used e.g., the autosegmental-metrical model (AM) model [53], ToBI (Tones and Break Indices) labels ([62] for English and [63] for Japanese), para-linguistic features [64, 65], and mixed-language features [66, 67].

2.3.2 Speech analysis

One of the aims of speech analysis is to dissociate the vocal tract characteristics and excitation characteristics, and to efficiently represent them. The speech signals are first windowed with a window function; then, their spectral parameters and excitation parameters are estimated. Fig. 8 shows an example of the observed power spectra and speech parameters including the spectral parameters (spectral envelopes) and excitation parameters (detailed spectra). According to the source-filter model, the speech signals are represented as convolutions of spectral param-

[4] **Section A.1** describes the contextual factors of Indian languages that we designed for Blizzard Challenge 2015.

[5] The incremental TTS [59, 60, 61] faces a similar problem. It aims at starting delivery of the synthetic speech before the full sentence context becomes available, e.g. while a user is still typing the text to vocalize.

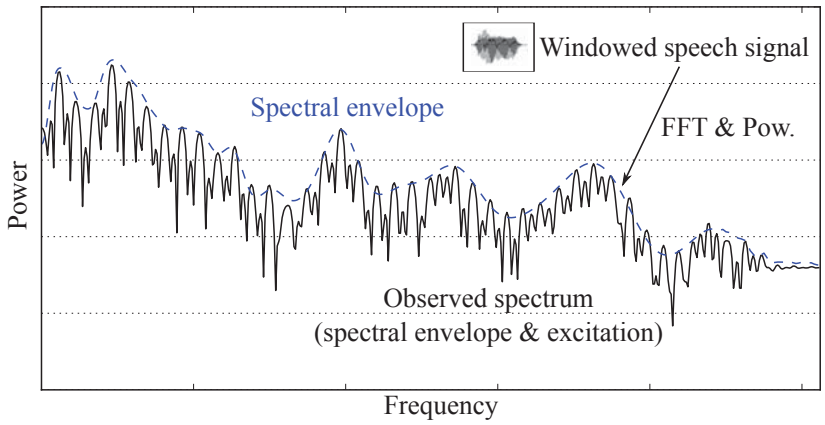


Figure 8. Examples of speech parameters extracted from the windowed raw speech signal. The observed spectra consist of vocal tract characteristics and excitation characteristics, and the spectral envelope corresponds to the vocal tract characteristics. For clear illustration, we draw the observed spectrum of speech synthesized by the STRAIGHT system instead of that of raw speech. FFT indicates the Fast Fourier Transform.

eters and excitation parameters. In mixed excitation, the excitation parameters are further decomposed into periodic factors (a.k.a., fundamental frequency or F_0) and aperiodic factors (a.k.a., aperiodicity). Whereas the spectral parameters and aperiodicity are the continuous variables, F_0 is a multiple-dimensional feature as shown in Fig. 9. 1-dimensional F_0 value is observed at the voiced frames (V) and the 0-dimensional feature is observed at the unvoiced frames (U).

Basically, a spectral parameter is a high-dimensional feature[6]. The most-used method to reduce dimensionality is to use the mel-cestral coefficient (used in this thesis) or mel-generalized coefficient [68], which considers the perceptual effects of the lower frequency components. An alternative approach is a data-driven one to extract the efficient parameters [69, 69, 43, 70]. Also, articulatory parameters, such as such as Liljencrants-Fant (LF) and Fujisaki models, are used for better modeling and generation of speech parameters [13, 71, 72, 73].

In order to dissociate and represent the vocal tract characteristics and excitation characteristics, this thesis uses the STRAIGHT analysis-synthesis system [10, 74], which performs the F_0-adaptive spectral envelope extraction and mixed excitation modeling. The STRAIGHT system is often used in speech synthesis [75, 12][7]. [76, 77] in HMM-based TTS.

[6] For example, in a 1024-tap DFT, the number of dimensions is 513.

[7] **Section A.3** investigates the speech analysis-synthesis systems STRAIGHT and WORLD. Whereas the STRAIGHT system is patented software, the WORLD system is BSD-licensed.

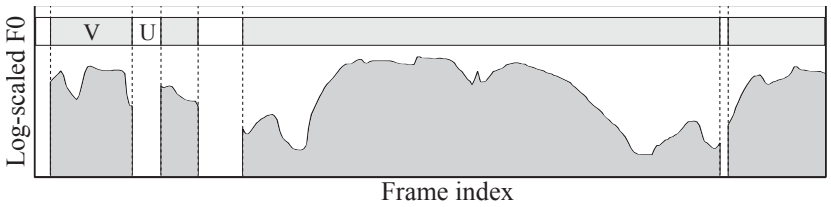


Figure 9. Examples of observed F_0 contours. Voiced frames (V) have an actual 1-dimensional F_0 value, and unvoiced frames (U) have a 0-dimensional value (discrete variable).

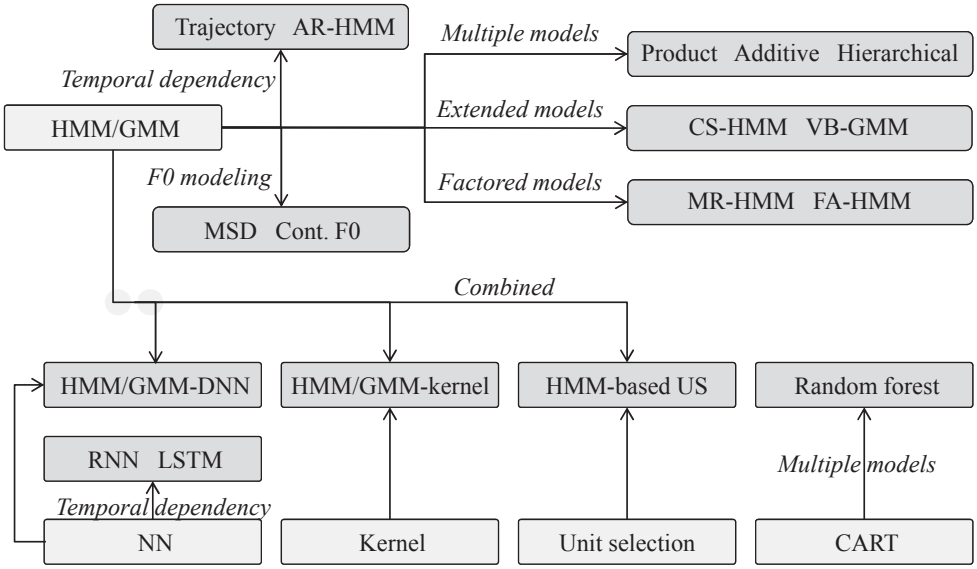


Figure 10. Acoustic modelings and their developments in statistical parametric speech synthesis.

2.3.3 Acoustic modeling

In the training stage, an acoustic model is trained to represent the relationship $\boldsymbol{y} = f(\boldsymbol{x})$ between the input features $\boldsymbol{x}$ (contextual labels of input text for TTS and speech parameters of input speech for VC) and the output speech parameters $\boldsymbol{y}$. The statistical models need to appropriately model the segmental (such as spectral parameters) and suprasegmental (such as F_0 parameters) speech parameters.

HMMs and GMMs (see Section 2.4 and Section 2.5). As we described in **Section 2.1**, HMM-based TTS and GMM-based VC have various advantages over the other approaches. The frameworks based on the HMMs and GMMs

can be extended to a variety of formulation, such as the Continuous-State (CS) HMM [78], Multi-Regressive (MR) model [39, 79], Factor-Analyzed (FA) model [80], and Eigen Voice (EV) [37, 81] model. In particular, there are several training and modeling methods for GMM-based VC [82, 83] that efficiently transmit the input speech information to output speech[8].

F0 modeling (see Section 2.4 and Section A.2.) As shown in Fig. 9, the observed F_0 contour is a multiple-dimensional feature. Multi-Space probability Density (MSD)-HMM and GMM [84, 85] have been proposed to efficiently model the F_0 contour with the mixture of the probabilities of the 1-dimensional space for voicing and the 0-dimensional space for unvoicing. Continuous F_0 modeling has proposed [86, 87][9], as a way to alleviate the weakness of the F_0 modeling with MSD-HMM/GMM; it outperforms deep neural nets-based methods in terms of quality of synthetic speech [88].

Temporal dependency (see Section 2.8.) The basic HMM/GMM frameworks don't appropriately capture the temporal dependency between speech parameters. Trajectory training [23, 89] have proposed trajectory modeling that trains the acoustic models under the temporal-delta constraint. The Auto-Regressive (AR) HMM [90] also considers the temporal dependency by assuming the speech parameter sequence conforms to an AR process.

Multiple acoustic models (see Section 2.9.) Combining a number of acoustic models improve prediction accuracy. Product of Experts (PoE) [91] is applied as a constraint of speech production, and the Global Variance (GV) in **Section 2.9** is used one of the constraints. In addition, additive [92] or hierarchical [93] acoustic models can be trained to capture the additive or suprasegmental features.

Deep Neural Nets (discussed in Section 6.) Inspired by the success of Deep Neural nets (DNNs) in machine learning and automatic speech recognition [94], researchers have attempted to use them in speech synthesis [95, 14]. The recurrent structures of DNNs [96, 97] can capture temporal dependencies of speech, and various systems can be written in a unified network [98, 99, 69].

Classification And Regression Trees (used in Section 4.6.) Whereas HMM-based TTS ties the probability density functions over multiple frames with the HMM-state-level probability density function, which is usually determined

[8] **Section A.8** discusses the implementation of the GMM-based VC with spectral differentials [82] combining the methods proposed in this thesis.

[9] **Section A.2** describes our implementation of continuous F_0 modeling for this thesis.

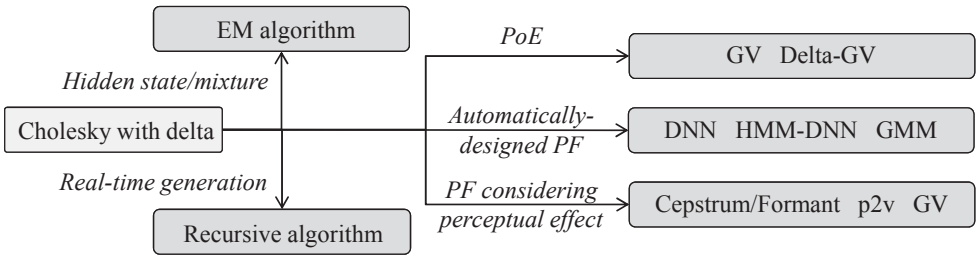


Figure 11. Speech parameter generation and their developments in statistical parametric speech synthesis. PF indicates Post-Filter.

by decision tree clustering, CLUSTERGEN [32] predicts the probability density functions frame by frame in the Classification And Regression Trees (CART) framework. The random forests algorithm [100] can be applied to this tree-based synthesis method [101].

Hybrid (see Section 2.7.) DNN-based, non-parametric (such as histogram-based method [102] and kernel regression [33, 103]), and unit selection synthesis can benefit HMM-based TTS and GMM-based VC. For example, the initial clustering of the HMMs and GMMs can be used not only for reducing the computational costs of non-parametric speech synthesis [33, 103], but also for robustly training model parameters for DNN-based speech synthesis [104, 105]. As far as improving speech quality, hybrid methods combining HMM-based TTS and unit selection synthesis have excellent capabilities [17, 18, 106, 107].

2.3.4 Speech parameter generation

The synthetic speech parameters are generated from the input parameters by using the statistics corresponding to the input features. Post-filtering processes are also used to improve the quality of the synthetic speech. The speech waveform is synthesized through a synthesis filter, such as the Mel-Log Spectrum Approximation (MLSA) filter [108].

ML-based generation using Cholesky decomposition (see Section 2.6.) The basic algorithm for generating speech parameters from HMMs was proposed in [109, 7]. By considering the temporal delta feature features, this method generates temporally-smoothed speech parameters from HMMs under the determined HMM-state sequence. Toda *et al.* [9] utilized it in GMM-based VC. Nowadays, most speech synthesizers use this method or improvements on it described below. One of the successes of these algorithms is their ability to be extended into

recursive forms [22, 7]. Using the Kalman filter, the speech parameters can be generated frame by frame and used for real-time speech synthesis [22, 110].

ML-based generation using EM algorithm (see Section 3.4.) Tokuda *et al.* and Toda *et al.* [109, 9] extended the basic algorithm for HMMs/GMMs with hidden HMM states and GMM mixture components.

Product of Experts (PoE) (see Section 2.9.) The concept of PoE can be used to address the over-smoothing effect in speech parameter generation. A promising approach to alleviating the over-smoothing effect is to extract a specific feature to quantify the effect and to generate speech parameters so that their corresponding features become more similar to those of natural speech parameters. One widely known example of such a feature is the Global Variance (GV) [19, 9], and speech parameter generation taking account of the GV is widely used [111, 112].

Post-filter (see Section 4.5.) A post-filtering process is a very simple but very effective way of alleviating the over-smoothing effect. Typically, it is done between the speech parameter generation and waveform synthesis. The post-filtering takes into account perceptually-effective features, such as GV [22, 113], cepstrum emphasis [49], and peak-to-valley (p2v) [114]. An alternative approach is to use automatically-constructed powerful statistical models to map the generate speech parameters [115, 116].

2.4 Acoustic modeling in HMM-based TTS

A Hidden Markov Model (HMM) is a statistical time series model that is widely used in various fields. Here, several refinements to the HMM idea have been used to great success by speech recognition systems. Similarly, speech synthesis has made substantial progress by using the excellent framework of HMMs. In training, the speech parameters that have been extracted from the output speech waveform in the training corpus are modeled with context-dependent HMMs. Note that the corresponding variables, e.g., the contextual labels of the input text in HMM-based TTS (in this section) and speech features of input speech in GMM-based VC (see the next section), are shared between TTS and VC.

2.4.1 Hidden Markov Model (HMM) definition

An HMM is a finite state machine that generates a sequence of discrete time observations. At each frame, the HMM changes states in accordance with a state transition probability that satisfies the Markov property, and generates

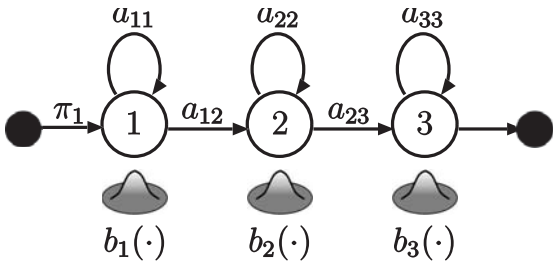


Figure 12. A three-state left-to-right HMM. The q-th HMM state ($q \in \{1, 2, 3\}$) has an individual output probability density function $b_q(\cdot)$ and transition matrix $a_{q,q+1}$.

the observed data in accordance with the output probability distribution of the current state. In TTS, HMMs that depend on a contextual label sequence $\boldsymbol{X}$ is used to model a speech parameter sequence $\boldsymbol{Y}$.

A Q-state HMM is defined by the state transition probability $\boldsymbol{A} = \{a_{qp}\}_{q,p=1}^{Q}$, the output probability distribution $\boldsymbol{B} = \{b_q(\cdot)\}_{q=1}^{Q}$, and the initial state probability $\boldsymbol{\Pi} = \{\pi_q\}_{q=1}^{Q}$. For notational simplicity, we denote the HMM parameter set $\boldsymbol{\lambda}$ as follows:

$$\boldsymbol{\lambda} = \{\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{\Pi}\}, \tag{2}$$

where q and p are HMM state indexes. A standard left-to-right HMM is shown in Fig. 12. The state index simply increases or stays equal as time processes, and this property is often used to model speech parameter sequences since they can appropriately model signals whose properties successively changes.

In HMM-based TTS, the spectral parameters and excitation parameters are jointly modeled using continuous HMMs and MSD-HMMs.

Continuous HMM: Spectral parameters (also, aperiodic parameters extracted by the STRAIGHT system) are modeled with a continuous HMM, in which its state output probability is given by

$$b_q(\boldsymbol{Y}_t) = \mathcal{N}\left(\boldsymbol{Y}_t; \boldsymbol{\mu}_q^{(Y|X)}, \boldsymbol{\Sigma}_q^{(Y|X)}\right), \tag{3}$$

where $\boldsymbol{\mu}_q^{(Y|X)}$ and $\boldsymbol{\Sigma}_q^{(Y|X)}$ are the mean vector and covariance matrix for state q. A Gaussian distribution with a mean vector $\boldsymbol{\mu}_q^{(Y|X)}$ and covariance matrix $\boldsymbol{\Sigma}_q^{(Y|X)}$ is denoted as $\mathcal{N}\left(\cdot; \boldsymbol{\mu}_q^{(Y|X)}, \boldsymbol{\Sigma}_q^{(Y|X)}\right)$, and given by

$$\begin{aligned}\mathcal{N}\left(\boldsymbol{Y}_t; \boldsymbol{\mu}_q^{(Y|X)}, \boldsymbol{\Sigma}_q^{(Y|X)}\right) &= \frac{1}{\sqrt{(2\pi)^{N_{\mathrm{w}}D}\left|\boldsymbol{\Sigma}_q^{(Y|X)}\right|}} \\ &\quad \exp\left(-\frac{1}{2}\left(\boldsymbol{Y}_t - \boldsymbol{\mu}_q^{(Y|X)}\right)^{\top} \boldsymbol{\Sigma}_q^{(Y|X)^{-1}}\left(\boldsymbol{Y}_t - \boldsymbol{\mu}_q^{(Y|X)}\right)\right),\end{aligned} \tag{4}$$

where $N_{\mathrm{w}}D$ is the number of dimensions of $\boldsymbol{Y}_t$. The dimensions of $\boldsymbol{\mu}_q^{(Y|X)}$ and $\boldsymbol{\Sigma}_q^{(Y|X)}$ are $N_{\mathrm{w}}D$ and $N_{\mathrm{w}}D$-by-$N_{\mathrm{w}}D$, respectively. In HMM-based TTS, a feature vector is defined as $\boldsymbol{Y}_t = \left[\boldsymbol{y}_t^\top, \Delta\boldsymbol{y}_t^\top, \Delta\Delta\boldsymbol{y}_t^\top\right]^\top$, which includes the D-dimensional static feature, $\boldsymbol{y}_t = \left[y_t(1), \cdots, y_t(d), \cdots, y_t(D)\right]^\top$, and dynamic features, $\Delta\boldsymbol{y}_t$, and $\Delta\Delta\boldsymbol{y}_t$. These dynamic features are computed from $\boldsymbol{y}_t$ by using

$$\Delta\boldsymbol{y}_t = \sum_{\tau=-L_-^{(1)}}^{L_+^{(1)}} \omega_\tau^{(1)} \boldsymbol{y}_{t+\tau}, \tag{5}$$

$$\Delta\Delta\boldsymbol{y}_t = \sum_{\tau=-L_-^{(2)}}^{L_+^{(2)}} \omega_\tau^{(2)} \boldsymbol{y}_{t+\tau}, \tag{6}$$

where $\omega_\tau^{(n)}$, $L_-^{(n)}$, and $L_+^{(n)}$ $(1 \leq n < N_{\mathrm{w}})$ are n-th order weight coefficients used to calculate the dynamic features. HMM-based TTS often sets $N_{\mathrm{w}} = 3$.

MSD-HMM: As we described in **Section 2.3**, an F_0 contour is modeled with MSD-HMMs [84][10]. Its state output probability is given by

$$b_q(\boldsymbol{Y}_t) = \begin{cases} w_q^{(Y|X)} \mathcal{N}\left(\boldsymbol{Y}_t; \boldsymbol{\mu}_q^{(Y|X)}, \boldsymbol{\Sigma}_q^{(Y|X)}\right), & l_t = \mathrm{V} \\ 1 - w_q^{(Y|X)}, & l_t = \mathrm{U} \end{cases}, \tag{7}$$

where l_t is a discrete voicing label that is either voiced V or unvoiced U at frame t, and $w_q^{(Y|X)}$ is the weight of the voiced space in state q, respectively. Note that l_t is observable together with $\boldsymbol{Y}_t$. In HMM-based TTS, $\boldsymbol{y}_t$, $\Delta\boldsymbol{y}_t$ and $\Delta\Delta\boldsymbol{y}_t$ are modeled with each corresponding MSD-HMM.

When a T-frame state sequence, $\boldsymbol{q} = [q_1, \cdots, q_t, \cdots, q_T]$, is determined for the input context label sequence $\boldsymbol{X}$, the probability of outputting the feature vector sequence $\boldsymbol{Y} = \left[\boldsymbol{Y}_1^\top, \cdots, \boldsymbol{Y}_t, \cdots, \boldsymbol{Y}_T^\top\right]^\top$ given the HMM parameter set $\boldsymbol{\lambda}$ is calculated by multiplying the output probabilities for each HMM-state, which is given as:

$$P(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{q}, \boldsymbol{\lambda}) = \prod_{t=1}^{T} b_{q_t}(\boldsymbol{Y}_t), \tag{8}$$

where q_t is the state index at frame t. The probability of such a state sequence $\boldsymbol{q}$ can be calculated by multiplying the state transition probabilities,

$$P(\boldsymbol{q}|\boldsymbol{X}, \boldsymbol{\lambda}) = \prod_{t=1}^{T} a_{q_{t-1}q_t}, \tag{9}$$

[10] When the continuous F_0 modeling [86] is used, the continuous F_0 contour is modeled with a continuous HMM.

where $a_{q_0 q_1}$ is given by the initial state probability π_{q_0}. Hence, the probability of the observation $\boldsymbol{Y}$ given $\boldsymbol{\lambda}$ is calculated by marginalizing over $\boldsymbol{q}$, i.e., by summing $P(\boldsymbol{Y}, \boldsymbol{q}|\boldsymbol{X}, \boldsymbol{\lambda})$ over all state sequences $\boldsymbol{q}$,

$$P(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{\lambda}) = \sum_{\text{all } \boldsymbol{q}} P(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{q}, \boldsymbol{\lambda}) P(\boldsymbol{q}|\boldsymbol{X}, \boldsymbol{\lambda}) \tag{10}$$

$$= \sum_{\text{all } \boldsymbol{q}} \prod_{t=1}^{T} a_{q_{t-1} q_t} b_{q_t}(\boldsymbol{Y}_t). \tag{11}$$

Considering that the state sequences have a trellis structure, the probability of the observation sequence can be transformed as follows:

$$P(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{\lambda}) = \sum_{q=1}^{Q} P(\boldsymbol{Y}_1, \cdots, \boldsymbol{Y}_t, q_t = q|\boldsymbol{X}, \boldsymbol{\lambda}) P(\boldsymbol{Y}_{t+1}, \cdots, \boldsymbol{Y}_T|\boldsymbol{X}, q_t = q, \boldsymbol{\lambda}). \tag{12}$$

Therefore, we can efficiently calculate the likelihood of the observation sequence by using a forward probability $\alpha_t(q)$ and a backward probability $\beta_t(q)$ defined as

$$\alpha_t(q) = P(\boldsymbol{Y}_1, \cdots, \boldsymbol{Y}_t, q_t = q|\boldsymbol{X}, \boldsymbol{\lambda}) \tag{13}$$

$$= \left[\sum_{p=1}^{Q} \alpha_{t-1}(p) a_{pq} \right] b_q(\boldsymbol{Y}_t), \tag{14}$$

$$\beta_t(q) = P(\boldsymbol{Y}_t, \cdots, \boldsymbol{Y}_T|\boldsymbol{X}, q_t = q, \boldsymbol{\lambda}) \tag{15}$$

$$= \sum_{p=1}^{Q} a_{qp} b_p(\boldsymbol{Y}_{t+1}) \beta_{t+1}(p). \tag{16}$$

This algorithm to calculate the probabilities is called the forward-backward algorithm.

2.4.2 HMM training

In the training stage, the HMM parameter set $\boldsymbol{\lambda}$ including $\boldsymbol{\mu}_q^{(Y|X)}$, $\boldsymbol{\Sigma}_q^{(Y|X)}$, and $w_q^{(Y|X)}$ is optimized with an optimization criterion such as Maximum Likelihood (ML) as follows:

$$\boldsymbol{\lambda} = \operatorname*{argmax}_{\boldsymbol{\lambda}} L_{\text{basic}}^{(\text{trn})}, \tag{17}$$

$$L_{\text{basic}}^{(\text{trn})} = P(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{\lambda}) = \sum_{\text{all } \boldsymbol{q}} P(\boldsymbol{Y}, \boldsymbol{q}|\boldsymbol{X}, \boldsymbol{\lambda}). \tag{18}$$

Since this problem is an optimization from incomplete data including a hidden variable $\boldsymbol{q}$, it is difficult to determine a $\boldsymbol{\lambda}$ which globally maximizes the likelihood $P(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{\lambda})$ for the input context label sequence $\boldsymbol{X}$ and the speech feature vector sequence $\boldsymbol{Y}$ in a closed form. However, the HMM parameter set $\boldsymbol{\lambda}$ that locally maximizes $P(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{\lambda})$ can be obtained using an iterative procedure such as the

Expectation-Maximization (EM) algorithm which conducts the optimization on an incomplete dataset [117]. This optimization algorithm is often referred to as the Baum-Welch algorithm. The auxiliary function $Q(\cdot)$ is maximized by iteratively updating the posterior probabilities of hidden variables given a current estimate $\boldsymbol{\lambda}^{(i)}$ in the E-step, and estimating the new $\boldsymbol{\lambda}^{(i+1)}$ while fixing the posterior constant in the M-step.

Continuous HMM: The auxiliary function $Q(\cdot)$ for continuous HMMs is given by:

$$Q\left(\boldsymbol{\lambda}^{(i)}, \boldsymbol{\lambda}^{(i+1)}\right) = \sum_{\text{all } \boldsymbol{q}} P\left(\boldsymbol{q}|\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{\lambda}^{(i)}\right) \log P\left(\boldsymbol{Y}, \boldsymbol{q}|\boldsymbol{X}, \boldsymbol{\lambda}^{(i+1)}\right), \tag{19}$$

where i is an iteration index. The mean vector $\boldsymbol{\mu}_q^{(Y|X)}$ and the covariance matrix $\boldsymbol{\Sigma}_q^{(Y|X)}$ of the q-th HMM-state are estimated to maximize $Q(\cdot)$, and are given by

$$\boldsymbol{\mu}_q^{(Y|X)} = \frac{\sum_{t=1}^{T} \gamma_t(q) \cdot \boldsymbol{Y}_t}{\sum_{t=1}^{T} \gamma_t(q)}, \tag{20}$$

$$\boldsymbol{\Sigma}_q^{(Y|X)} = \frac{\sum_{t=1}^{T} \gamma_t(q) \cdot \left(\boldsymbol{Y}_t - \boldsymbol{\mu}_q^{(Y|X)}\right)\left(\boldsymbol{Y}_t - \boldsymbol{\mu}_q^{(Y|X)}\right)^{\top}}{\sum_{t=1}^{T} \gamma_t(q)}, \tag{21}$$

$$\gamma_t(q) = \frac{\alpha_t(q)\,\beta_t(q)}{\sum_{q=1}^{Q} \alpha_t(q)\,\beta_t(q)}, \tag{22}$$

where $\gamma_t(q)$ is the state occupancy probability of being in the q-th HMM-state at frame t.

MSD-HMM: The auxiliary function $Q(\cdot)$ for MSD-HMMs is the same as in the case of the continuous HMM. The mean vector $\boldsymbol{\mu}_q^{(Y|X)}$, covariance matrix $\boldsymbol{\Sigma}_q^{(Y|X)}$ and weight of voiced space $w_q^{(Y|X)}$ are estimated by maximizing $Q(\cdot)$ as follows:

$$\boldsymbol{\mu}_q^{(Y|X)} = \frac{\sum_{t=1}^{T} \gamma_t(q, \mathrm{V}) \cdot \boldsymbol{Y}_t}{\sum_{t=1}^{T} \gamma_t(q, \mathrm{V})}, \tag{23}$$

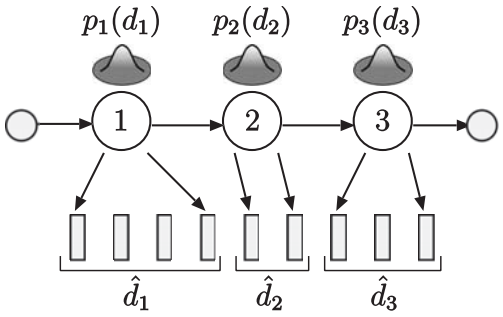


Figure 13. A three-state left-to-right HSMM. Compared with the standard HMM shown in Fig. 12, each HMM-state has an individual duration model instead of a state transition probability.

$$\boldsymbol{\Sigma}_q^{(Y|X)} = \frac{\sum_{t=1}^{T} \gamma_t(q,\mathrm{V}) \cdot \left(\boldsymbol{Y}_t - \boldsymbol{\mu}_q^{(Y|X)}\right)\left(\boldsymbol{Y}_t - \boldsymbol{\mu}_q^{(Y|X)}\right)^\top}{\sum_{t=1}^{T} \gamma_t(q,\mathrm{V})}, \tag{24}$$

$$w_q^{(Y|X)} = \frac{\sum_{t=1}^{T} \gamma_t(q,\mathrm{V})}{\sum_{t=1}^{T} \left(\gamma_t(q,\mathrm{V}) + \gamma_t(q,\mathrm{U})\right)}, \tag{25}$$

$$\gamma_t(q,\mathrm{V}) = \begin{cases} \frac{\alpha_t(q)\beta_t(q)}{\sum_{q=1}^{Q} \alpha_t(q)\beta_t(q)}, & l_t = \mathrm{V} \\ 0, & l_t = \mathrm{U} \end{cases}, \tag{26}$$

$$\gamma_t(q,\mathrm{U}) = \begin{cases} 0 & l_t = \mathrm{V} \\ \frac{\alpha_t(q)\beta_t(q)}{\sum_{q=1}^{Q} \alpha_t(q)\beta_t(q)}, & l_t = \mathrm{U} \end{cases}, \tag{27}$$

where $\gamma_t(q,\mathrm{V})$ and $\gamma_t(q,\mathrm{U})$ are the state occupancy probability of being q-th state at frame t in the voiced space and that in the unvoiced space, respectively.

In order to perform the ML-based generation described in **Section 2.6**, HMM-based TTS uses an explicit duration model. The HMM state duration distributions can be modeled using parametric probability density functions such as the following Gaussian distributions:

$$P_q(d) = \frac{1}{\sqrt{2\pi\sigma_n^2}} \exp\left(-\frac{(d-m_q)^2}{2\sigma_q^2}\right), \tag{28}$$

where m_q and σ_q are the mean and variance of the duration model of state q. The HMM including the output probability and state duration probability, which is shown in Fig. 13, is called a Hidden Semi-Markov Model (HSMM) [23].

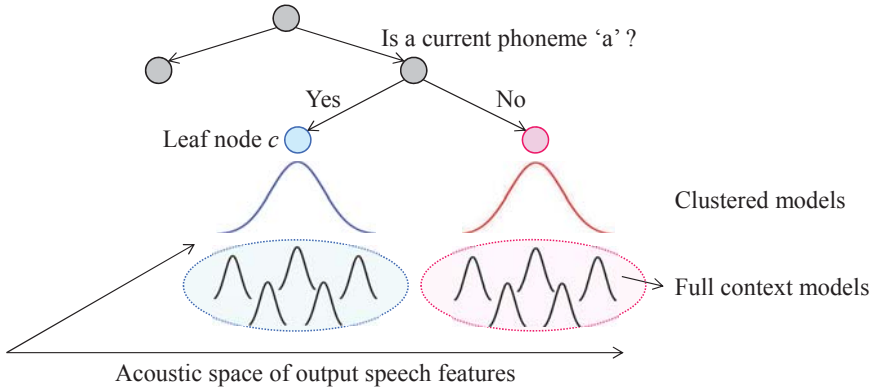


Figure 14. A decision tree for HMM-based TTS. Basically, the variance of each full context model is almost 0.

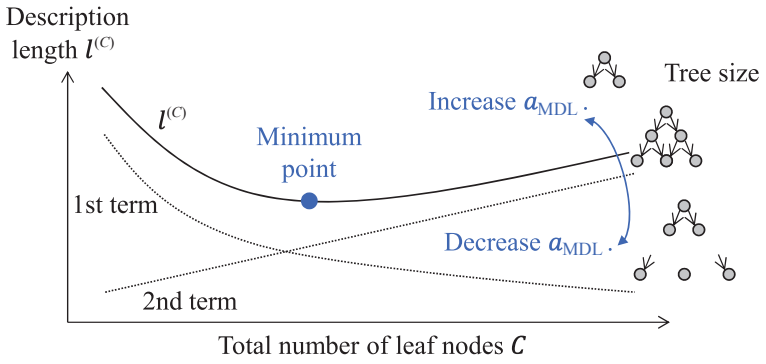


Figure 15. Description length used in tree-building with the MDL criterion. The tree size varies as the MDL parameter a_{MDL} changes.

2.4.3 Tree-based context clustering

Various contextual factors need to be considered to model speech parameters in TTS. Because the combinations of contextual factors increase exponentially and the number of them is enormous, one context label usually corresponds to only one acoustic segment in the training data. Therefore, it is impossible to prepare training data that covers all possible context-dependent HMMs. An HMM corresponding to the individual context label is called a full context model. To robustly train context-dependent HMMs, different context labels are tied together in a binary decision tree constructed from and by answering context related questions [49]. Each node (except for leaf nodes) has one context related question, such as "L-silence?" ("is the previous phoneme a silence ?"), and two child nodes representing "yes" and "no" answers to the question. Fig. 14 shows an example of the decision tree. The acoustic space is divided into sub-regions, and the

full context models are clustered in each sub-region. Generally, a decision tree for context clustering is constructed in each HMM state and is based on the Minimum Description Length (MDL) criterion [118], given by

$$l^{(C)} = \frac{1}{2}\sum_{c=1}^{C}\Gamma\left(c\right)\log\left|\boldsymbol{\Sigma}_{q_c}^{(Y|X)}\right| + a_{\mathrm{MDL}}CD\log\Gamma\left(0\right), \tag{29}$$

where c is the leaf node index, C is the total number of leaf nodes, a_{MDL} is a parameter to control C, D is the number of feature dimensions, $\boldsymbol{\Sigma}_{q_c}^{(Y|X)}$ is the covariance matrix of the c-th leaf node of the q-th HMM state, and $\Gamma\left(c\right)$ and $\Gamma\left(0\right)$ are state occupancy counts of the leaf node c and root node, respectively. The value of the first term decreases and that of the second term increases as the total number of leaf nodes C increases as shown in Fig. 15. The decision tree is constructed according to the following process.

1. Define the root node.
2. Find the node and question that maximize the difference in the description length before and after splitting.
3. If the difference is less than 0, stop splitting the nodes.
4. Split the node by using the question discovered in step 2.
5. Go to step 2.

After the tree-based context clustering, the output probability density function b_{q_c} (clustered model) of the c-th leaf node of the q-th HMM-state and its parameters (i.e., the mean vector $\boldsymbol{\mu}_{q_c}^{(Y|X)}$, covariance matrix $\boldsymbol{\Sigma}_{q_c}^{(Y|X)}$, and a weight of the voice space $w_{q_c}^{(Y|X)}$) are calculated for each leaf node.

2.5 Acoustic modeling in GMM-based VC

GMMs have been widely used to solve many classification problems. In training, the speech parameters extracted from the input and output speech waveforms are modeled with the GMM as joint probability density functions. In synthesis, a speech parameter sequence is generated from the GMMs by computing the conditional probability given the input speech parameters. As we described, some variables are shared with HMM-based TTS.

2.5.1 Gaussian Mixture Model (GMM) definition

A GMM is a mixture model of Gaussian distribution as shown in Fig. 16. The Q-mixture GMM is defined by the mixture weight $\boldsymbol{A} = \{w_q^{(Z)}\}_{q=1}^{Q}$, and the mixture

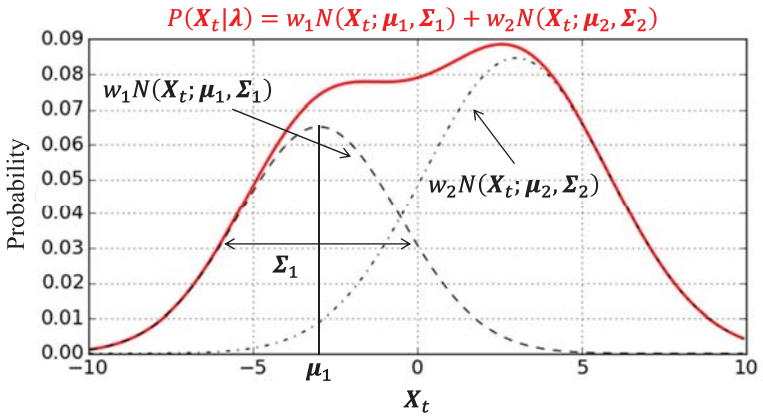


Figure 16. A 2-mixture GMM. The q-th GMM mixture component ($q \in \{1, 2\}$) has an individual output $\mathcal{N}\left(\boldsymbol{X}_t, \boldsymbol{\mu}_q, \boldsymbol{\Sigma}_q\right)$ and mixture weight w_q. Note that the variables $\boldsymbol{X}_t$, $\boldsymbol{\mu}_q$, and $\boldsymbol{\Sigma}_q$ should ideally be shown as scalar values here, but are shown as vectors for the sake of generality in the description.

component $\boldsymbol{B} = \{b_q(\cdot)\}_{q=1}^{Q}$. We denote the GMM parameter set $\boldsymbol{\lambda}$ as follows:

$$\boldsymbol{\lambda} = \{\boldsymbol{A}, \boldsymbol{B}\}. \tag{30}$$

In GMM-based VC, the spectrum is modeled with a multivariate GMM, as shown in Fig. 17. F_0 is typically modeled with a single Gaussian model. The joint probability density is modeled with:

$$P(\boldsymbol{Z}_t|\boldsymbol{\lambda}) = \sum_{q=1}^{Q} P(q|\boldsymbol{\lambda})\, P(\boldsymbol{Z}_t|q, \boldsymbol{\lambda}), \tag{31}$$

$$P(q|\boldsymbol{\lambda}) = w_q^{(Z)}, \tag{32}$$

$$P(\boldsymbol{Z}_t|q, \boldsymbol{\lambda}) = b_q(\boldsymbol{Z}_t) = \mathcal{N}\left(\boldsymbol{Z}_t; \boldsymbol{\mu}_q^{(Z)}, \boldsymbol{\Sigma}_q^{(Z)}\right), \tag{33}$$

where, $\boldsymbol{Z}_t = \left[\boldsymbol{X}_t^\top, \boldsymbol{Y}_t^\top\right]^\top$ is the joint vector of the input spectral features $\boldsymbol{X}_t$ and the output spectral features $\boldsymbol{Y}_t$ at frame t, and

$$\boldsymbol{\mu}_q^{(Z)} = \begin{bmatrix} \boldsymbol{\mu}_q^{(X)} \\ \boldsymbol{\mu}_q^{(Y)} \end{bmatrix}, \quad \boldsymbol{\Sigma}_q^{(Z)} = \begin{bmatrix} \boldsymbol{\Sigma}_q^{(XX)} & \boldsymbol{\Sigma}_q^{(XY)} \\ \boldsymbol{\Sigma}_q^{(YX)} & \boldsymbol{\Sigma}_q^{(YY)} \end{bmatrix}. \tag{34}$$

$\boldsymbol{Y}_t$ is given by $N_{\mathrm{w}}D$-dimensional joint static and dynamic feature vectors, $\left[\boldsymbol{y}_t^\top, \Delta\boldsymbol{y}_t^\top\right]^\top$, where $\boldsymbol{y}_t$ is represented as a D-dimensional static feature vector. N_{w} is often set to 2 in GMM-based VC. The source feature vector $\boldsymbol{X}_t$ is also given the same form in this thesis. $\boldsymbol{\mu}_q^{(Z)}$ consists of the input and output mean vectors, $\boldsymbol{\mu}_q^{(X)}$ and $\boldsymbol{\mu}_q^{(Y)}$. $\boldsymbol{\Sigma}_q^{(Z)}$ consists of the source and target covariance matrices, $\boldsymbol{\Sigma}_q^{(XX)}$ and $\boldsymbol{\Sigma}_q^{(YY)}$ and

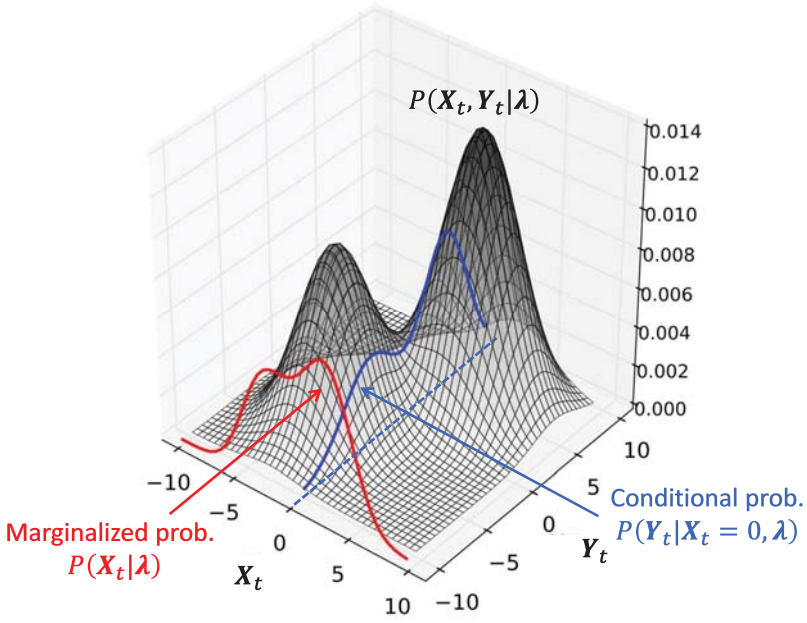


Figure 17. A 2-mixture multivariate GMM and its conditional probability and marginalized probability density function. Note that the conditional and marginalized probability density function are scaled to illustrate them clearly.

cross-covariance matrices, $\boldsymbol{\Sigma}_q^{(YX)}$ and $\boldsymbol{\Sigma}_q^{(XY)}$.

2.5.2 GMM training

The GMM parameter set $\boldsymbol{\lambda}$ is optimized with an optimization criterion as follows:

$$\boldsymbol{\lambda} = \underset{\boldsymbol{\lambda}}{\operatorname{argmax}}\, L_{\text{basic}}^{(\text{trn})}, \tag{35}$$

$$L_{\text{basic}}^{(\text{trn})} = \prod_{t=1}^{T} P\left(\boldsymbol{Z}_t|\boldsymbol{\lambda}\right) = \prod_{t=1}^{T} \sum_{q=1}^{Q} P\left(\boldsymbol{Z}_t, q|\boldsymbol{\lambda}\right). \tag{36}$$

This optimization problem can be solved using EM algorithm, the same as with HMMs. The auxiliary function $Q\left(\cdot\right)$ for the spectral component is given by:

$$Q\left(\boldsymbol{\lambda}^{(i)}, \boldsymbol{\lambda}^{(i+1)}\right) = \sum_{t=1}^{T} \sum_{q=1}^{Q} P\left(q|\boldsymbol{Z}_t, \boldsymbol{\lambda}^{(i)}\right) \log P\left(\boldsymbol{Z}_t, q|\boldsymbol{\lambda}^{(i+1)}\right), \tag{37}$$

where i is the iteration index. The mean vector $\boldsymbol{\mu}_q^{(Z)}$, covariance matrix $\boldsymbol{\Sigma}_q^{(Z)}$, and the mixture weight $w_q^{(Z)}$ of the q-th GMM mixture component are estimated

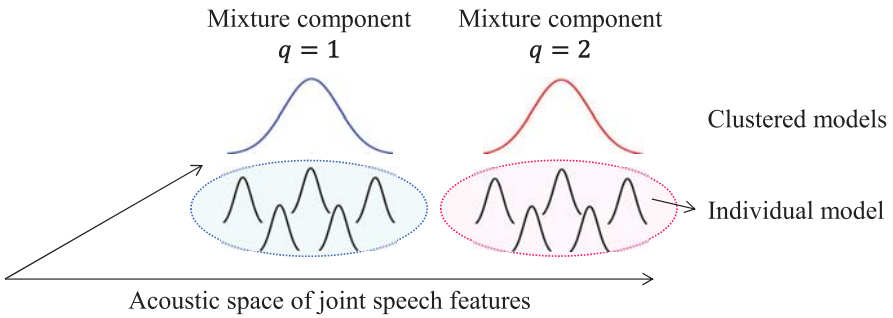


Figure 18. Hard clustering for GMM-based VC. Basically, the variance of each individual model (Gaussian distribution for the individual speech feature) is almost 0. The structure is very similar to that of the decision tree clustering in HMM-based TTS (Fig. 14).

in order to maximize $Q(\cdot)$, and are given by

$$\boldsymbol{\mu}_q^{(Z)} = \frac{\sum_{t=1}^{T} \gamma_t(q) \cdot \boldsymbol{Z}_t}{\sum_{t=1}^{T} \gamma_t(q)}, \tag{38}$$

$$\boldsymbol{\Sigma}_q^{(Z)} = \frac{\sum_{t=1}^{T} \gamma_t(q) \cdot \left(\boldsymbol{Z}_t - \boldsymbol{\mu}_q^{(Z)}\right)\left(\boldsymbol{Z}_t - \boldsymbol{\mu}_q^{(Z)}\right)^{\top}}{\sum_{t=1}^{T} \gamma_t(q)}, \tag{39}$$

$$w_q^{(Z)} = \frac{\sum_{t=1}^{T} \gamma_t(q)}{T} \tag{40}$$

$$\gamma_t(q) = \frac{\mathcal{N}\left(\boldsymbol{Z}_t; \boldsymbol{\mu}_q^{(Z)}, \boldsymbol{\Sigma}_q^{(Z)}\right)}{\sum_{q=1}^{Q} w_q^{(Z)} \mathcal{N}\left(\boldsymbol{Z}_t; \boldsymbol{\mu}_q^{(Z)}, \boldsymbol{\Sigma}_q^{(Z)}\right)}, \tag{41}$$

where $\gamma_t(q)$ is the mixture occupancy probability of being in mixture q at frame t.

Using the optimal mixture having the biggest $\gamma_t(q)$, the acoustic space can be divided into sub-regions by using Eq. (41), and each sub-region can be modeled with a GMM mixture component as shown in Fig. 18.

2.5.3 Conditional probability and marginalized probability

The conditional probability $P(\boldsymbol{Y}_t|\boldsymbol{X}_t, \boldsymbol{\lambda})$ and the marginalized probability $P(\boldsymbol{X}_t|\boldsymbol{\lambda})$ shown in Fig. 17 are analytically derived from Eq. (31), and they are used in the

speech parameter generation stage.

As derived in **Section A.4**, the conditional probability of the q-th mixture component, $P\left(\boldsymbol{Y}_t|\boldsymbol{X}_t, q, \boldsymbol{\lambda}\right)$, is given as a Gaussian distribution:

$$P\left(\boldsymbol{Y}_t|\boldsymbol{X}_t, q, \boldsymbol{\lambda}\right) = \mathcal{N}\left(\boldsymbol{Y}_t; \boldsymbol{\mu}_q^{(Y|X)}, \boldsymbol{\Sigma}_q^{(Y|X)}\right), \tag{42}$$

where,

$$\boldsymbol{\Sigma}_q^{(Y|X)} = \boldsymbol{\Sigma}_q^{(YY)} - \boldsymbol{\Sigma}_q^{(YX)}\boldsymbol{\Sigma}_q^{(XX)^{-1}}\boldsymbol{\Sigma}_q^{(XY)}, \tag{43}$$

$$\boldsymbol{\mu}_q^{(Y|X)} = \boldsymbol{\mu}_q^{(Y)} - \boldsymbol{\Sigma}_q^{(YX)}\boldsymbol{\Sigma}_q^{(XX)^{-1}}\left(\boldsymbol{X}_t - \boldsymbol{\mu}_q^{(X)}\right). \tag{44}$$

The conditional probability $P\left(\boldsymbol{Y}_t|\boldsymbol{X}_t, \boldsymbol{\lambda}\right)$ is given as a GMM mixing Eq. (42):

$$P\left(\boldsymbol{Y}_t|\boldsymbol{X}_t, \boldsymbol{\lambda}\right) = \sum_{q=1}^{Q} P\left(q|\boldsymbol{X}_t, \boldsymbol{\lambda}\right) P\left(\boldsymbol{Y}_t|q, \boldsymbol{X}_t, \boldsymbol{\lambda}\right). \tag{45}$$

$$P\left(q|\boldsymbol{X}_t, \boldsymbol{\lambda}\right) = \frac{\mathcal{N}\left(\boldsymbol{X}_t; \boldsymbol{\mu}_q^{(X)}, \boldsymbol{\Sigma}_q^{(XX)}\right)}{\sum_{q=1}^{Q} \mathcal{N}\left(\boldsymbol{X}_t; \boldsymbol{\mu}_q^{(X)}, \boldsymbol{\Sigma}_q^{(XX)}\right)}. \tag{46}$$

Next, we derive the marginalized probability. $P\left(\boldsymbol{X}_t|\boldsymbol{\lambda}\right)$ is calculated by marginalizing over all $\boldsymbol{Y}_t$:

$$P\left(\boldsymbol{X}_t|\boldsymbol{\lambda}\right) = \int P\left(\boldsymbol{X}_t, \boldsymbol{Y}_t|\boldsymbol{\lambda}\right) d\boldsymbol{Y}_t \tag{47}$$

$$= \sum_{q=1}^{Q} P\left(q|\boldsymbol{\lambda}\right) \int P\left(\boldsymbol{X}_t, \boldsymbol{Y}_t|q, \boldsymbol{\lambda}\right) d\boldsymbol{Y}_t \tag{48}$$

$$= \sum_{q=1}^{Q} w_q^{(Z)} P\left(\boldsymbol{X}_t|q, \boldsymbol{\lambda}\right). \tag{49}$$

We omit the derivation, but the q-th GMM mixture component can be derived as intuition would lead us to expect:

$$P\left(\boldsymbol{X}_t|q, \boldsymbol{\lambda}\right) = \mathcal{N}\left(\boldsymbol{X}_t; \boldsymbol{\mu}_q^{(X)}, \boldsymbol{\Sigma}_q^{(XX)}\right). \tag{50}$$

Eq. (46) is the posterior probability of this marginalized probability.

2.6 Speech parameter generation

We generate synthetic speech parameters from the input parameters $\boldsymbol{X}$. $\boldsymbol{X}$ denotes the contextual label sequence of the input text for HMM-based TTS, or the speech feature vector sequence of the input speech for GMM-based VC. After determining the optimal HMM state and GMM mixture sequence $\hat{\boldsymbol{q}} = [\hat{q}_1, \cdots, \hat{q}_t, \cdots, \hat{q}_T]$, the output speech parameter sequence $\hat{\boldsymbol{y}}_{\hat{\boldsymbol{q}}} = [\hat{\boldsymbol{y}}_1, \cdots, \hat{\boldsymbol{y}}_t, \cdots, \hat{\boldsymbol{y}}_T]^\top$ is done by maximizing the likelihood, where $\hat{q}_t$ and $\hat{\boldsymbol{y}}_t$ are the optimal HMM state

or GMM mixture and output speech parameters at frame t.

2.6.1 optimal HMM state and GMM mixture sequence

In HMM-based TTS, state duration models (Eq. (28)) corresponding to $\boldsymbol{X}$ are determined through the use of the decision tree. The optimal state duration of the q-th HMM state is determined by roughly maximizing the duration probability as follows:

$$\hat{d}_q = \underset{d}{\operatorname{argmax}}\, P_q\left(d|q, \boldsymbol{\lambda}\right) = m_q, \tag{51}$$

The optimal HMM state sequence is determined by the state duration.

In GMM-based VC, the optimal mixture component of frame t is determined by maximizing the posterior probability of the marginalized GMM (Eq. (46)):

$$\hat{q}_t = \underset{q}{\operatorname{argmax}} \frac{w_q^{(Z)} \mathcal{N}\left(\boldsymbol{X}_t; q, \boldsymbol{\lambda}\right)}{\sum_{q=1}^{Q} w_q^{(Z)} \mathcal{N}\left(\boldsymbol{X}_t; q, \boldsymbol{\lambda}\right)}, \tag{52}$$

where $\boldsymbol{X}_t$ is the input speech feature vector at frame t.

2.6.2 Maximum likelihood-based generation

Given the optimal HMM state and GMM mixture component sequence $\hat{\boldsymbol{q}}$, the output speech parameter sequence $\hat{\boldsymbol{y}}_{\hat{\boldsymbol{q}}}$ is generated by maximizing the objective function $L_{\text{basic}}^{(\text{syn})}$ using the HMM likelihood or GMM likelihood, as follows:

$$\hat{\boldsymbol{y}}_{\hat{\boldsymbol{q}}} = \operatorname{argmax} L_{\text{basic}}^{(\text{syn})}, \tag{53}$$

$$L_{\text{basic}}^{(\text{syn})} = P\left(\boldsymbol{Y}|\boldsymbol{X}, \hat{\boldsymbol{q}}, \boldsymbol{\lambda}\right) = P\left(\boldsymbol{W}\boldsymbol{y}|\boldsymbol{X}, \hat{\boldsymbol{q}}, \boldsymbol{\lambda}\right) = \mathcal{N}\left(\boldsymbol{W}\boldsymbol{y}; \boldsymbol{E}_{\hat{\boldsymbol{q}}}; \boldsymbol{D}_{\hat{\boldsymbol{q}}}\right), \tag{54}$$

where $\boldsymbol{E}_{\hat{\boldsymbol{q}}} = \left[\boldsymbol{\mu}_{\hat{q}_1,1}^\top, \cdots, \boldsymbol{\mu}_{\hat{q}_t,t}^\top, \cdots, \boldsymbol{\mu}_{\hat{q}_T,T}^\top\right]^\top$ and $\boldsymbol{D}_{\hat{\boldsymbol{q}}} = \operatorname{diag}_{N_{\text{w}}D}\left[\boldsymbol{\Sigma}_{\hat{q}_1}, \cdots, \boldsymbol{\Sigma}_{\hat{q}_t}, \cdots, \boldsymbol{\Sigma}_{\hat{q}_T}\right]$ are an $N_{\text{w}}DT$-dimensional mean vector and $N_{\text{w}}DT$-by-$N_{\text{w}}DT$ covariance matrix, respectively, and

$$\boldsymbol{\mu}_{q,t} = \begin{cases} \boldsymbol{\mu}_q^{(Y|X)} & \text{(HMM)} \\ \boldsymbol{A}_q \boldsymbol{X}_t + \boldsymbol{b}_q & \text{(GMM)} \end{cases}, \tag{55}$$

$$\boldsymbol{\Sigma}_q = \begin{cases} \boldsymbol{\Sigma}_q^{(Y|X)} & \text{(HMM)} \\ \boldsymbol{\Sigma}_q^{(YY)} - \boldsymbol{A}_q \boldsymbol{\Sigma}_q^{(XX)} \boldsymbol{A}_q^\top & \text{(GMM)} \end{cases}, \tag{56}$$

$$\boldsymbol{A}_q = \boldsymbol{\Sigma}_q^{(YX)} {\boldsymbol{\Sigma}_q^{(XX)}}^{-1}, \tag{57}$$

$$\boldsymbol{b}_q = \boldsymbol{\mu}_q^{(Y)} - \boldsymbol{A}_q \boldsymbol{\mu}_q^{(X)}, \tag{58}$$

where the notation $\operatorname{diag}_{N_{\text{w}}D}$ denotes the construction of a block diagonal matrix that has $N_{\text{w}}D$-by-$N_{\text{w}}D$ diagonal elements. This derivation is illustrated in Fig.

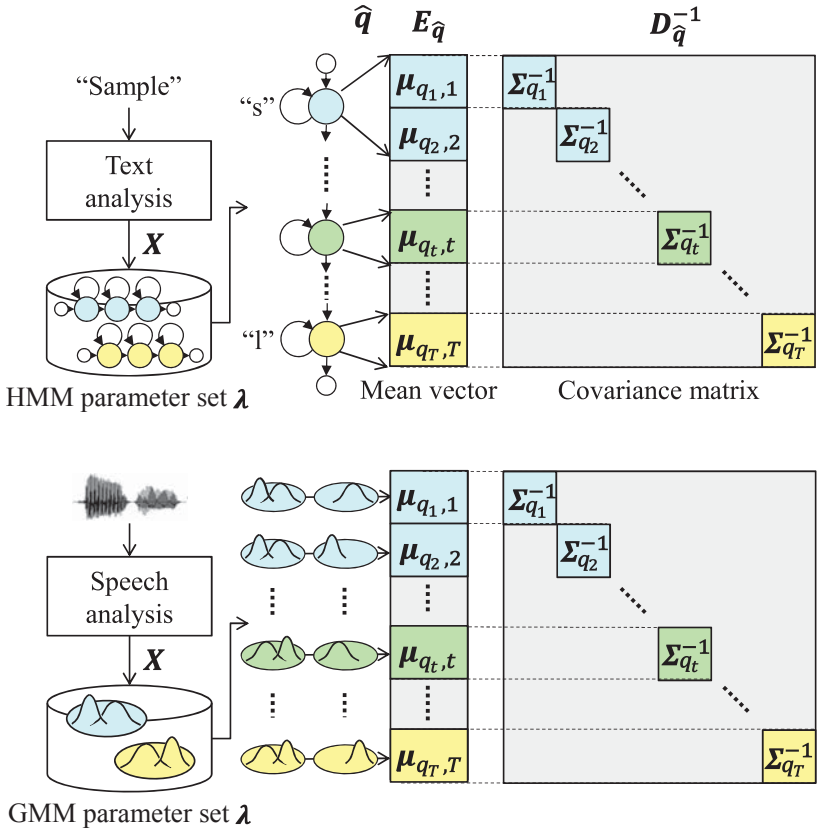


Figure 19. Output probabilities used to generate the speech parameters. The components of the mean vector and the covariance matrix are derived from the HMM state and GMM mixture.

19. $\boldsymbol{W}$ in Fig. 20 is the weighting matrix for calculating the dynamic features [109]; it is given by

$$\boldsymbol{W} = [\boldsymbol{w}_1, \cdots, \boldsymbol{w}_T]^\top, \tag{59}$$

$$\boldsymbol{w}_t = \left[\boldsymbol{w}_t^{(0)}, \ldots \boldsymbol{w}_t^{(N_\mathrm{w}-1)}\right], \tag{60}$$

$$\boldsymbol{w}_t^{(n)} = \left[\underset{1\mathrm{st}}{\boldsymbol{0}}, \cdots, \boldsymbol{0}, \underset{(t-L_-^{(n)})-\mathrm{th}}{w_{-L_-^{(n)}}\boldsymbol{I}}, \cdots, \underset{(t-L_+^{(n)})-\mathrm{th}}{w_{-L_+^{(n)}}\boldsymbol{I}}, \boldsymbol{0}, \cdots, \underset{T-\mathrm{th}}{\boldsymbol{0}}\right], \tag{61}$$

where $-L_-^{(0)}, -L_+^{(0)}$ and $w^{(0)}(0) = 1$.

The logarithm of Eq. (54) can be transformed as:

$$\ln P(\boldsymbol{W}\boldsymbol{y}; \hat{\boldsymbol{q}}, \boldsymbol{\lambda}) = -\frac{1}{2}\left(\boldsymbol{W}\boldsymbol{y} - \boldsymbol{E}_{\hat{\boldsymbol{q}}}\right)^\top \boldsymbol{D}_{\hat{\boldsymbol{q}}}^{-1}\left(\boldsymbol{W}\boldsymbol{y} - \boldsymbol{E}_{\hat{\boldsymbol{q}}}\right) + \mathrm{Const.}, \tag{62}$$

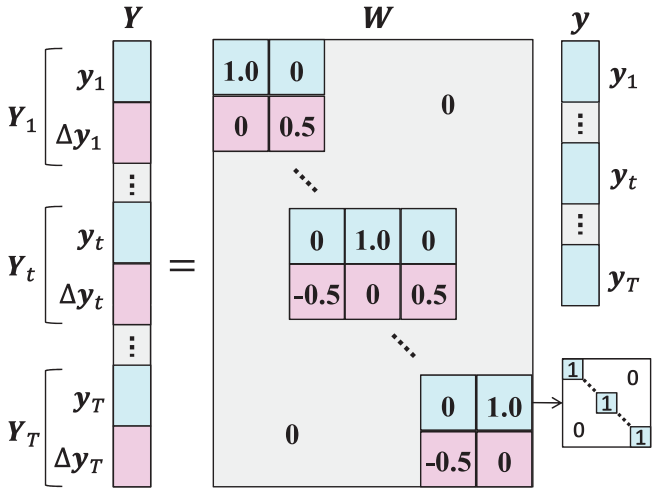


Figure 20. Delta matrix used to calculate the static and delta feature vector sequence. In this figure, $N_{\mathrm{w}} = 2$, $L_{-}^{(n)} = -0.5$, $L_{+}^{(n)} = 0.5$.

where Const. indicates a value constant to $\boldsymbol{y}$. Thus, by setting

$$\frac{\partial \ln P\left(\boldsymbol{W}\boldsymbol{y}|\hat{\boldsymbol{q}}, \boldsymbol{\lambda}\right)}{\partial \boldsymbol{y}} = \boldsymbol{0}, \tag{63}$$

the following equations are obtained.

$$\boldsymbol{W}^{\top}\boldsymbol{D}_{\hat{\boldsymbol{q}}}^{-1}\boldsymbol{W}\hat{\boldsymbol{y}}_{\hat{\boldsymbol{q}}} = \boldsymbol{W}^{\top}\boldsymbol{D}_{\hat{\boldsymbol{q}}}^{-1}\boldsymbol{E}_{\hat{\boldsymbol{q}}}, \tag{64}$$

$$\hat{\boldsymbol{y}}_{\hat{\boldsymbol{q}}} = \boldsymbol{R}_{\hat{\boldsymbol{q}}}^{-1}\boldsymbol{r}_{\hat{\boldsymbol{q}}}, \tag{65}$$

$$\boldsymbol{R}_{\hat{\boldsymbol{q}}} = \boldsymbol{W}^{\top}\boldsymbol{D}_{\hat{\boldsymbol{q}}}^{-1}\boldsymbol{W}, \tag{66}$$

$$\boldsymbol{r}_{\hat{\boldsymbol{q}}} = \boldsymbol{W}^{\top}\boldsymbol{D}_{\hat{\boldsymbol{q}}}^{-1}\boldsymbol{E}_{\hat{\boldsymbol{q}}}. \tag{67}$$

In the example shown in Fig. 21, the speech parameters are generated probabilities considering dynamic features.

The generated speech parameter sequence $\hat{\boldsymbol{y}}_{\hat{\boldsymbol{q}}}$ can be efficiently calculated sequence by sequence using Cholesky decomposition [109], but the result tends to be over-smoothed.

2.7 Hybrid synthesis

Here, we describe the hybrid methods that combine the idea of unit selection synthesis and HMM-based TTS. They include (1) hybrid synthesis with waveform concatenation that uses HMMs to guide waveform segments [17], and (2) hybrid synthesis with speech parameter generation that models the individual waveform

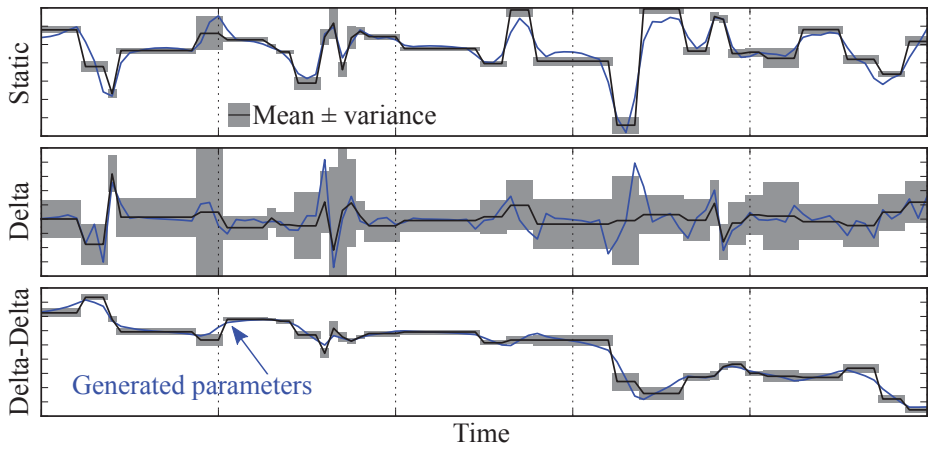


Figure 21. Example of probability distributions and speech parameters generated from the distributions in HMM-based TTS. Note that frames having the same statistics correspond to the same HMM state.

segments with Gaussian distributions [18].

Whereas HMM-based TTS causes a quality degradation as a result of using averaged information (= statistical models) of the speech parameters, these methods have the capability of improving synthetic speech quality by incorporating the ideas of unit selection synthesis. However, the flexibility of the original HMM-based TTS is lost because their mathematical formulation is different from that of HMM-based TTS.

2.7.1 Hybrid synthesis with waveform concatenation

ML-based unit selection synthesis [17] uses HMMs to guide speech segments. In the synthesis stage after training the HMMs, the optimal set of waveform segments is selected from a speech database to maximize the cost function combining the HMM likelihoods, as shown in Fig. 22. The cost function $C^{(\mathrm{ml})}$ of this approach is represented in the same form as that of unit selection, which is

$$C^{(\mathrm{ml})} = \sum_{n=1}^{N} C_{\mathrm{t}}^{(\mathrm{ml})}\left(u_n\right) + \sum_{n=2}^{N} C_{\mathrm{c}}^{(\mathrm{ml})}\left(u_{n-1}, u_n\right), \tag{68}$$

where $C_{\mathrm{t}}^{(\mathrm{ml})}$ and $C_{\mathrm{c}}^{(\mathrm{ml})}$ are the target cost and the concatenation cost, and they are weighted sums of the HMM likelihoods.

The use of waveform segments dramatically improves speech quality[11]. How-

[11] Although unit selection synthesis makes it possible to generate high-quality speech waveform, the weights of the cost functions are difficult to set and it is necessary to manually tune them. ML-based unit selection synthesis makes it possible not only to improve speech quality

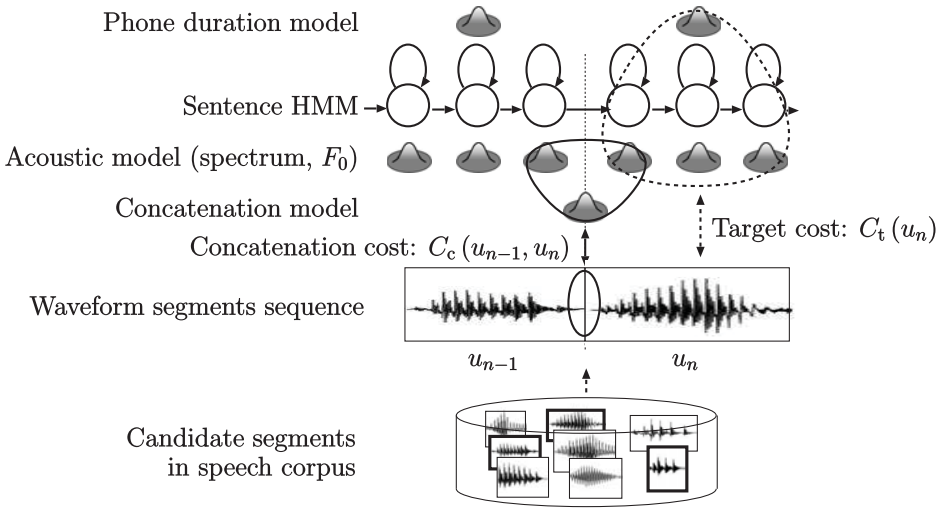


Figure 22. ML-based unit selection using HMMs. In this example, the spectrum, F_0, and phoneme-duration statistics of the HMMs are used to guide the waveform segments.

ever, the waveform generation process with waveform concatenation loses the advantage of HMM-based TTS of being able to control the voice characteristics.

2.7.2 Hybrid synthesis with parameter generation

A hybrid approaches that has more flexibility than the standard unit selection synthesis or ML-based unit selection synthesis is the use of rich context models to represent each waveform segment with probability distributions of individual speech component parameters, such as the spectrum and F_0 [18]. In the synthesis stage, the probability distributions of all components corresponding to one waveform segment are selected in each HMM state on the basis of the Kullback-Leibler Divergence (KLD) and speech parameters are generated from these distributions.

Training of rich context models In the basic HMM-based TTS, a single Gaussian distribution is used to model multiple acoustic segments belonging to the same leaf nodes in the decision tree. Consequently, its mean vector is excessively smoothed and becomes one of causes of the over-smoothing effect. On the other hand, the use of multiple acoustic segments is essential for robustly estimating the model parameters, in particular its covariance matrix. To alleviate the over-smoothing effect while preserving the robustness of the parameter estimation, in rich context model, the mean vector is trained for each full context label

but also to automatically tune the weights by using the HMMs.

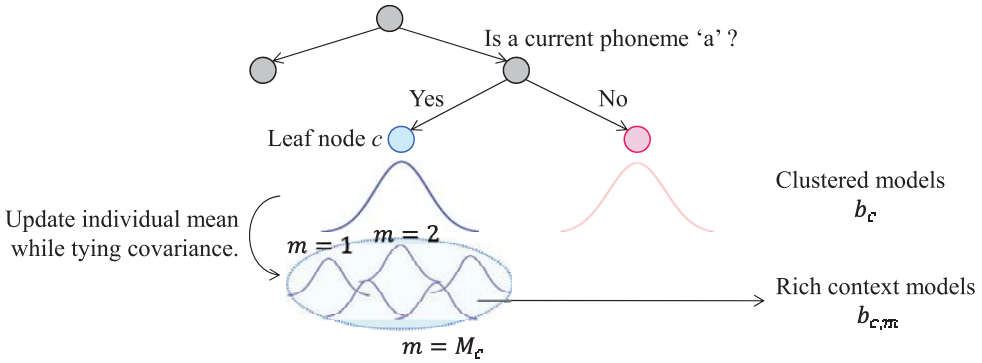


Figure 23. Training of rich context models using the clustered models of HMM-based TTS. The mean vector corresponding to the individual speech segments are updated while tying the covariance matrix of the clustered model. M_{q_c} is the number of full context labels in the c-th leaf node of the q-th HMM state. Compared with Fig. 14, the variances of the individual models are wider.

and the covariance matrix is tied among different full context labels belonging to each leaf node of the decision tree [18], as shown in Fig. 23.

For continuous HMMs, the output probability density function $b_{q_c,m}$ of the rich context model for the m-th full context label in the c-th leaf node of the q-th HMM-state is given by

$$b_{q_c,m}\left(\boldsymbol{Y}_t\right) = \mathcal{N}\left(\boldsymbol{Y}_t; \boldsymbol{\mu}_{q_c,m}^{(Y|X)}, \boldsymbol{\Sigma}_{q_c}^{(Y|X)}\right). \tag{69}$$

For MSD-HMMs, the mean vector of the Gaussian distribution in the voiced space is updated as follows:

$$b_{q_c,m}\left(\boldsymbol{Y}_t\right) = \begin{cases} w_{q_c}^{(Y|X)} \mathcal{N}\left(\boldsymbol{Y}_t; \boldsymbol{\mu}_{q_c,m}^{(Y|X)}, \boldsymbol{\Sigma}_{q_c}^{(Y|X)}\right) & l_t = \mathrm{V} \\ 1 - w_{q_c}^{(Y|X)}, & l_t = \mathrm{U} \end{cases}. \tag{70}$$

The total number of different mean vectors in a tree is equivalent to the number of full context labels in the training data. The total number of different covariance matrices is equivalent to the number of leaf nodes in the decision tree.

In the training stage, the context-clustered probability density parameters are estimated in the standard manner. Then, they are untied and only their mean vectors are updated in each full context label by using the Baum-Welch algorithm while tying the covariance matrices among full context labels in each leaf node.

Synthesis In the synthesis stage, the full context labels to be synthesized are clustered with the decision trees and the clustered models at the corresponding leaf nodes are determined to be the target $\boldsymbol{g} = \{g_1, \cdots, g_N\}$ where g_n represents the clustered model in the n-th state. Then, a sequence of rich context models

$\boldsymbol{r}$ is selected that minimizes the following KLD, considering the spectral and F_0 components, where r_n represents the rich context model in the n-th state.

$$\mathrm{KLD}\left(\boldsymbol{g}, \boldsymbol{r}\right) = \sum_{n=1}^{N} \mathrm{KLD}\left(g_n, r_n\right) T_n, \tag{71}$$

$$\mathrm{KLD}\left(g_n, r_n\right) = \mathcal{D}_{\mathrm{KL}}^{(\mathrm{SP})}\left(g_n, r_n\right) + \mathcal{D}_{\mathrm{KL}}^{(\mathrm{F0})}\left(g_n, r_n\right), \tag{72}$$

where $\mathrm{KLD}\left(\cdot\right)$ is the total KLD, $\mathcal{D}_{\mathrm{KL}}^{(\mathrm{SP})}\left(g_n, r_n\right)$ and $\mathcal{D}_{\mathrm{KL}}^{(\mathrm{F0})}\left(g_n, r_n\right)$ are KLDs for the spectral and F_0 components, respectively. This process is similar to unit selection, but each acoustic segment is represented by probability density functions in individual components. Finally, speech parameter sequences are generated from the selected probability density functions in the same manner as the original HMM-based TTS.

In this method, rich context models for the spectral and F_0 components are simultaneously selected using a constraint among different components (spectrum and F_0). This approach also yields significant improvements in speech quality. However, flexibility of the original HMM-based TTS gets lost by the use of the strong constraint in the model selection.

2.8 Trajectory modeling

The weakness of the basic HMM-based TTS is the inconsistency between the training criterion $L_{\mathrm{basic}}^{(\mathrm{trn})}$ (Eq. (18)) and the synthesis criterion $L_{\mathrm{basic}}^{(\mathrm{syn})}$ (Eq. (54)), i.e., the likelihoods for the joint static and dynamic feature vectors in the training stage compared with those for only the static feature vectors in the synthesis stage, as shown in Fig. 24. Trajectory HMM modeling [23] is a method that models static feature vector sequences under static and dynamic feature constraints.

2.8.1 Trajectory HMM definition

Here, we derive the probability density function of the trajectory HMM, $P\left(\boldsymbol{y}|\boldsymbol{W}, \boldsymbol{X}, \hat{\boldsymbol{q}}, \boldsymbol{\lambda}\right)$, by transforming the basic synthesis criterion $L_{\mathrm{basic}}^{(\mathrm{syn})}$. The basic synthesis criterion $L_{\mathrm{basic}}^{(\mathrm{syn})}$ can be written as the probability of $\boldsymbol{y}$:

$$L_{\mathrm{basic}}^{(\mathrm{syn})} = \mathcal{N}\left(\boldsymbol{Y}; \boldsymbol{E}_{\hat{\boldsymbol{q}}}, \boldsymbol{D}_{\hat{\boldsymbol{q}}}\right) = \frac{1}{Z}\mathcal{N}\left(\boldsymbol{y}; \hat{\boldsymbol{y}}_{\hat{\boldsymbol{q}}}, \boldsymbol{R}^{-1}\right), \tag{73}$$

where Z is a normalization term, and

$$\mathcal{N}\left(\boldsymbol{Y}; \boldsymbol{E}_{\hat{\boldsymbol{q}}}, \boldsymbol{D}_{\hat{\boldsymbol{q}}}\right) = \frac{1}{\sqrt{\left(2\pi\right)^{N_{\mathrm{w}}DT}\left|\boldsymbol{D}_{\hat{\boldsymbol{q}}}\right|}} \exp\left(-\frac{1}{2}\left(\boldsymbol{Y} - \boldsymbol{E}_{\hat{\boldsymbol{q}}}\right)^{\top} \boldsymbol{D}_{\hat{\boldsymbol{q}}}^{-1}\left(\boldsymbol{Y} - \boldsymbol{E}_{\hat{\boldsymbol{q}}}\right)\right), \tag{74}$$

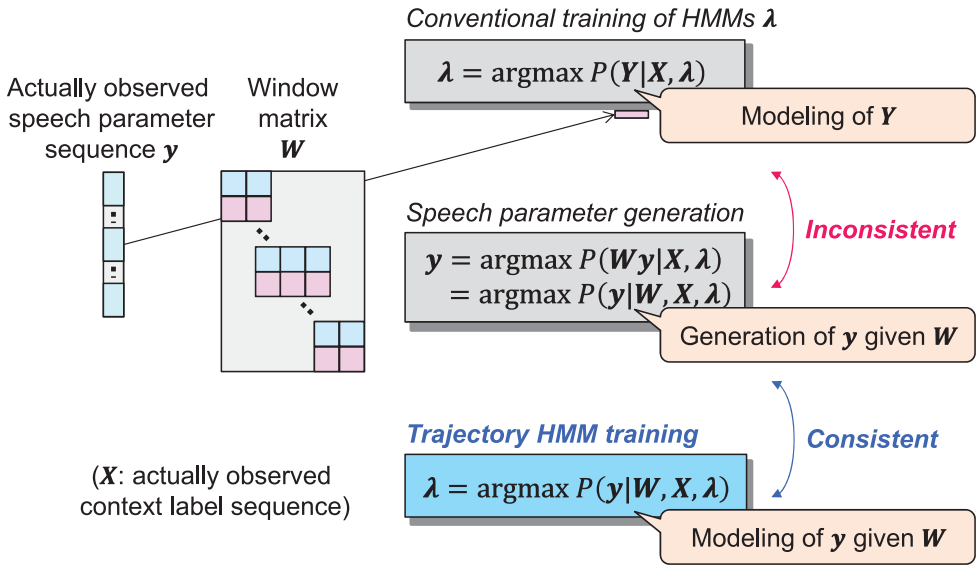


Figure 24. Consistency of training and generation. Trajectory modeling can remove the inconsistencies between the conventional HMM training and speech parameter generation. Note that we have omitted the optimal HMM state or GMM mixture sequence, $\hat{\boldsymbol{q}}$, for the sake of notational simplicity.

where $\boldsymbol{Y} = \boldsymbol{W}\boldsymbol{y}$. The exponential part of this equation can be expanded as follows:

$$-\frac{1}{2}\left(\boldsymbol{Y} - \boldsymbol{E}_{\hat{\boldsymbol{q}}}\right)^{\top} \boldsymbol{D}_{\hat{\boldsymbol{q}}}^{-1}\left(\boldsymbol{Y} - \boldsymbol{E}_{\hat{\boldsymbol{q}}}\right) \tag{75}$$

$$= -\frac{1}{2}\left((\boldsymbol{W}\boldsymbol{y})^{\top} \boldsymbol{D}_{\hat{\boldsymbol{q}}}^{-1}\boldsymbol{W}\boldsymbol{y} - (\boldsymbol{W}\boldsymbol{y})^{\top} \boldsymbol{D}_{\hat{\boldsymbol{q}}}^{-1}\boldsymbol{E}_{\hat{\boldsymbol{q}}} - \boldsymbol{E}_{\hat{\boldsymbol{q}}}\boldsymbol{D}_{\hat{\boldsymbol{q}}}^{-1}\boldsymbol{W}\boldsymbol{y} + \boldsymbol{E}_{\hat{\boldsymbol{q}}}^{\top}\boldsymbol{D}_{\hat{\boldsymbol{q}}}^{-1}\boldsymbol{E}_{\hat{\boldsymbol{q}}}\right) \tag{76}$$

$$= -\frac{1}{2}\left(\boldsymbol{y}^{\top}\boldsymbol{R}_{\hat{\boldsymbol{q}}}\boldsymbol{y} - 2\boldsymbol{r}_{\hat{\boldsymbol{q}}}\boldsymbol{y} + \boldsymbol{E}_{\hat{\boldsymbol{q}}}^{\top}\boldsymbol{D}_{\hat{\boldsymbol{q}}}^{-1}\boldsymbol{E}_{\hat{\boldsymbol{q}}}\right) \tag{77}$$

$$= -\frac{1}{2}\left(\left(\boldsymbol{y} - \boldsymbol{R}_{\hat{\boldsymbol{q}}}^{-1}\boldsymbol{r}_{\hat{\boldsymbol{q}}}\right)^{\top} \boldsymbol{R}_{\hat{\boldsymbol{q}}}\left(\boldsymbol{y} - \boldsymbol{R}_{\hat{\boldsymbol{q}}}^{-1}\boldsymbol{r}_{\hat{\boldsymbol{q}}}\right) - \boldsymbol{r}_{\hat{\boldsymbol{q}}}^{\top}\boldsymbol{R}_{\hat{\boldsymbol{q}}}^{-1}\boldsymbol{r}_{\hat{\boldsymbol{q}}} + \boldsymbol{E}_{\hat{\boldsymbol{q}}}^{\top}\boldsymbol{D}_{\hat{\boldsymbol{q}}}^{-1}\boldsymbol{E}_{\hat{\boldsymbol{q}}}\right) \tag{78}$$

$$= -\frac{1}{2}\left(\boldsymbol{y} - \hat{\boldsymbol{y}}_{\hat{\boldsymbol{q}}}\right)^{\top} \boldsymbol{R}_{\hat{\boldsymbol{q}}}\left(\boldsymbol{y} - \hat{\boldsymbol{y}}_{\hat{\boldsymbol{q}}}\right) - \frac{1}{2}\left(\boldsymbol{r}_{\hat{\boldsymbol{q}}}^{\top}\boldsymbol{R}_{\hat{\boldsymbol{q}}}^{-1}\boldsymbol{r}_{\hat{\boldsymbol{q}}} - \boldsymbol{E}_{\hat{\boldsymbol{q}}}^{\top}\boldsymbol{D}_{\hat{\boldsymbol{q}}}^{-1}\boldsymbol{E}_{\hat{\boldsymbol{q}}}\right). \tag{79}$$

Therefore, the basic criteria can be transformed as:

$$\mathcal{N}\left(\boldsymbol{Y}; \boldsymbol{E}_{\hat{\boldsymbol{q}}}, \boldsymbol{D}_{\hat{\boldsymbol{q}}}\right) = \frac{1}{\sqrt{(2\pi)^{DT}\,|\boldsymbol{R}_{\hat{\boldsymbol{q}}}^{-1}|}} \exp\left(-\frac{1}{2}\left(\boldsymbol{y} - \hat{\boldsymbol{y}}_{\hat{\boldsymbol{q}}}\right)^{\top} \boldsymbol{R}_{\hat{\boldsymbol{q}}}\left(\boldsymbol{y} - \hat{\boldsymbol{y}}_{\hat{\boldsymbol{q}}}\right)\right)$$

$$\cdot \frac{\sqrt{(2\pi)^{DT}\,|\boldsymbol{R}_{\hat{\boldsymbol{q}}}^{-1}|}}{\sqrt{(2\pi)^{N_{\mathrm{w}}DT}\,|\boldsymbol{D}_{\hat{\boldsymbol{q}}}|}} \exp\left(-\frac{1}{2}\left(\boldsymbol{E}_{\hat{\boldsymbol{q}}}^{\top}\boldsymbol{D}_{\hat{\boldsymbol{q}}}^{-1}\boldsymbol{E}_{\hat{\boldsymbol{q}}} - \boldsymbol{r}_{\hat{\boldsymbol{q}}}^{\top}\boldsymbol{R}_{\hat{\boldsymbol{q}}}^{-1}\boldsymbol{r}_{\hat{\boldsymbol{q}}}\right)\right) \quad (80)$$

$$= \frac{\sqrt{(2\pi)^{DT}\,|\boldsymbol{R}_{\hat{\boldsymbol{q}}}|}}{\sqrt{(2\pi)^{N_{\mathrm{w}}DT}\,|\boldsymbol{D}_{\hat{\boldsymbol{q}}}^{-1}|}} \exp\left(-\frac{1}{2}\left(\boldsymbol{E}_{\hat{\boldsymbol{q}}}^{\top}\boldsymbol{D}_{\hat{\boldsymbol{q}}}^{-1}\boldsymbol{E}_{\hat{\boldsymbol{q}}} - \boldsymbol{r}_{\hat{\boldsymbol{q}}}^{\top}\boldsymbol{R}_{\hat{\boldsymbol{q}}}^{-1}\boldsymbol{r}_{\hat{\boldsymbol{q}}}\right)\right)$$
$$\cdot \mathcal{N}\left(\boldsymbol{y}; \hat{\boldsymbol{y}}_{\hat{\boldsymbol{q}}}; \boldsymbol{R}_{\hat{\boldsymbol{q}}}^{-1}\right) \quad (81)$$

$$= \frac{1}{Z}\mathcal{N}\left(\boldsymbol{y}; \hat{\boldsymbol{y}}_{\hat{\boldsymbol{q}}}; \boldsymbol{R}_{\hat{\boldsymbol{q}}}^{-1}\right), \quad (82)$$

where

$$\frac{1}{Z} = \frac{\sqrt{(2\pi)^{DT}\,|\boldsymbol{R}_{\hat{\boldsymbol{q}}}^{-1}|}}{\sqrt{(2\pi)^{N_{\mathrm{w}}DT}\,|\boldsymbol{D}_{\hat{\boldsymbol{q}}}|}} \exp\left(-\frac{1}{2}\left(\boldsymbol{E}_{\hat{\boldsymbol{q}}}^{\top}\boldsymbol{D}_{\hat{\boldsymbol{q}}}^{-1}\boldsymbol{E}_{\hat{\boldsymbol{q}}} - \boldsymbol{r}_{\hat{\boldsymbol{q}}}^{\top}\boldsymbol{R}_{\hat{\boldsymbol{q}}}\boldsymbol{r}_{\hat{\boldsymbol{q}}}\right)\right) \quad (83)$$

The objective function for the trajectory training, $L_{\mathrm{trj}}^{(\mathrm{trn})}$, is written as:

$$L_{\mathrm{trj}}^{(\mathrm{trn})} = P\left(\boldsymbol{y}|\boldsymbol{W}, \boldsymbol{X}, \hat{\boldsymbol{q}}, \boldsymbol{\lambda}\right) = \mathcal{N}\left(\boldsymbol{y}; \hat{\boldsymbol{y}}_{\hat{\boldsymbol{q}}}, \boldsymbol{R}_{\hat{\boldsymbol{q}}}^{-1}\right). \quad (84)$$

The mean vector $\hat{\boldsymbol{y}}_{\hat{\boldsymbol{q}}}$ is given by Eq. (65) and the inter-frame correlation is effectively modeled with the temporal covariance matrix $\boldsymbol{R}_{\hat{\boldsymbol{q}}}^{-1}$ as shown in Fig. 25. In training, the HMM parameters are updated by maximizing $L_{\mathrm{trj}}^{(\mathrm{trn})}$. Note that the mean vector $\hat{\boldsymbol{y}}_{\hat{\boldsymbol{q}}}$ is equivalent to the generated parameter sequence in the traditional generation process. Therefore, $L_{\mathrm{trj}}^{(\mathrm{trn})}$ can be regarded as the objective function for not only the training stage but also the synthesis stage.

2.8.2 Trajectory HMM training

The HMM parameter set $\boldsymbol{\lambda}$ is updated to maximize the objective function as follows:

$$\boldsymbol{\lambda} = \underset{\boldsymbol{\lambda}}{\operatorname{argmax}}\, L_{\mathrm{trj}}^{(\mathrm{trn})}, \quad (85)$$

Here, we let $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}^{-1}$ be the joint parameters of $\boldsymbol{\mu}_q^{(Y|X)}$ and $\boldsymbol{\Sigma}_q^{(Y|X)^{-1}}$ over all HMM states:

$$\boldsymbol{\mu} = \left[\boldsymbol{\mu}_1^{(Y|X)^{\top}}, \cdots, \boldsymbol{\mu}_q^{(Y|X)^{\top}}, \cdots, \boldsymbol{\mu}_Q^{(Y|X)^{\top}}\right]^{\top}, \quad (86)$$

$$\boldsymbol{\Sigma}^{-1} = \left[\boldsymbol{\Sigma}_1^{(Y|X)^{-1}}, \cdots, \boldsymbol{\Sigma}_q^{(Y|X)^{-1}}, \cdots, \boldsymbol{\Sigma}_Q^{(Y|X)^{-1}}\right]^{\top}, \quad (87)$$

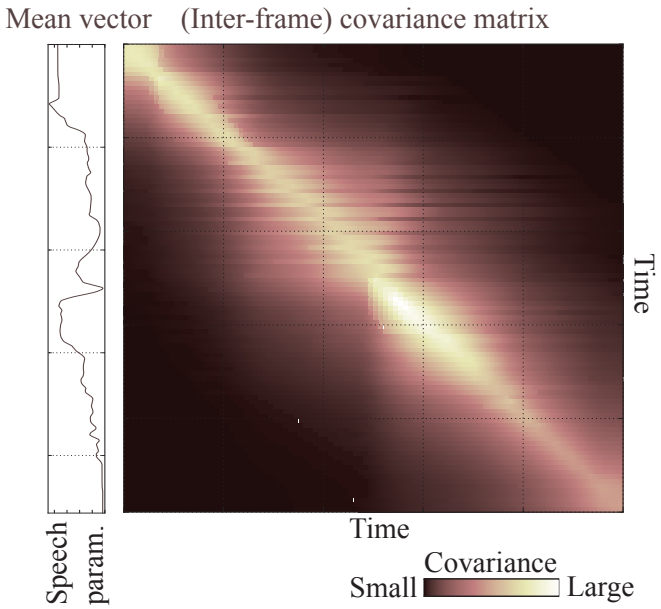


Figure 25. Example of the mean vector $\hat{\boldsymbol{y}}_{\hat{\boldsymbol{q}}}$ and covariance matrix $\boldsymbol{R}_{\hat{\boldsymbol{q}}}^{-1}$ of the trajectory HMM. The mean vector is equal to the generated speech parameters. The covariance matrix represents the temporal dependency and is generally the full covariance.

The mean vector $\boldsymbol{E}_{\hat{\boldsymbol{q}}}$ and precision matrix $\boldsymbol{D}_{\hat{\boldsymbol{q}}}^{-1}$ are represented as:

$$\boldsymbol{E}_{\hat{\boldsymbol{q}}} = \boldsymbol{S}_{\hat{\boldsymbol{q}}}\boldsymbol{\mu}, \tag{88}$$

$$\boldsymbol{D}_{\hat{\boldsymbol{q}}}^{-1} = \mathrm{diag}_{N_{\mathrm{w}}D}\left[\boldsymbol{S}_{\hat{\boldsymbol{q}}}\boldsymbol{\Sigma}^{-1}\right], \tag{89}$$

where $\boldsymbol{S}_{\hat{\boldsymbol{q}}} = \left[\boldsymbol{S}_{\hat{q}_1}, \cdots, \boldsymbol{S}_{\hat{q}_T}\right]^\top \otimes \boldsymbol{I}_{N_{\mathrm{w}}D}$ is a $N_{\mathrm{w}}DT$-by-$N_{\mathrm{w}}DQ$ matrix, $\boldsymbol{S}_{\hat{q}_t}$ is a Q-dimensional vector whose q-th component is 1 when $q = \hat{q}_t$ and 0 otherwise, as shown in Fig. 26, and $\boldsymbol{I}_{N_{\mathrm{w}}D}$ indicates the $N_{\mathrm{w}}D$-by-$N_{\mathrm{w}}D$ identity matrix.

To optimize these model parameters for the trajectory HMM training, we use the steepest descent algorithm, as follows:

$$\boldsymbol{\mu}^{(i+1)} = \boldsymbol{\mu}^{(i)} + \alpha \left.\frac{\partial \log L_{\mathrm{trj}}^{(\mathrm{trn})}}{\partial \boldsymbol{\mu}}\right|_{\boldsymbol{\mu}=\boldsymbol{\mu}^{(i)}}, \tag{90}$$

where α is the learning rate, and i is the iteration index. The $\boldsymbol{\Sigma}^{-1}$ are optimized in the same manner. The derivatives are:

$$\frac{\partial \log L_{\mathrm{trj}}^{(\mathrm{trn})}}{\partial \boldsymbol{\mu}} = \boldsymbol{S}_{\hat{\boldsymbol{q}}}^\top \boldsymbol{D}_{\hat{\boldsymbol{q}}}^{-1} \boldsymbol{W}\left(\boldsymbol{y} - \hat{\boldsymbol{y}}_{\hat{\boldsymbol{q}}}\right), \tag{91}$$

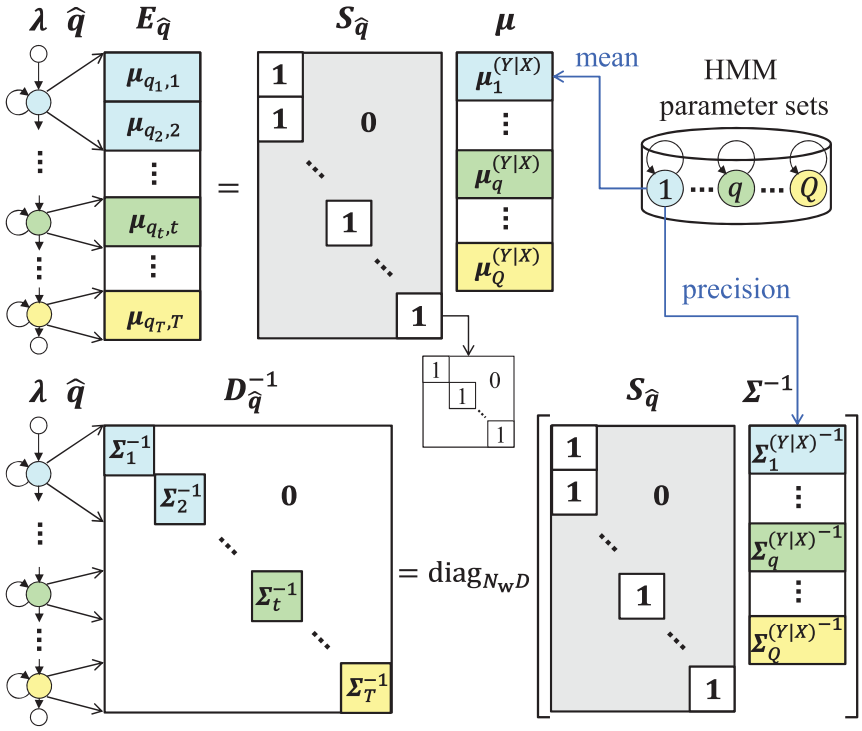


Figure 26. Relationship between variables used in trajectory HMM training. The activation matrix is determined by the tree-based clustering in HMM-based TTS.

$$\frac{\partial \log L_{\mathrm{trj}}^{(\mathrm{trn})}}{\partial \boldsymbol{\Sigma}^{-1}} = \frac{1}{2} \boldsymbol{S}_{\hat{\boldsymbol{q}}}^{\top} \mathrm{diag}_{N_{\mathrm{w}} D}^{-1} \Big[\boldsymbol{W} \left(\boldsymbol{R}_{\hat{\boldsymbol{q}}}^{-1} + \hat{\boldsymbol{y}}_{\hat{\boldsymbol{q}}} \hat{\boldsymbol{y}}_{\hat{\boldsymbol{q}}}^{\top} - \boldsymbol{y}\boldsymbol{y}^{\top} \right) - \boldsymbol{E}_{\hat{\boldsymbol{q}}} \left(\hat{\boldsymbol{y}}_{\hat{\boldsymbol{q}}} - \boldsymbol{y} \right)^{\top} \boldsymbol{W}^{\top} - \boldsymbol{W} \left(\hat{\boldsymbol{y}}_{\hat{\boldsymbol{q}}} - \boldsymbol{y} \right) \boldsymbol{E}_{\hat{\boldsymbol{q}}}^{\top} \Big], \tag{92}$$

2.9 Speech synthesis considering the global variance

The speech parameter generation described in **Section 2.6** tends to generate an over-smoothed speech parameter sequence. An intuitive way to alleviate the over-smoothing effect is to consider the features that can capture the effect. The Global Variance (GV) is a well-known examples of such a features, and it is used as part of the PoE of HMM-based TTS and GMM-based VC.

2.9.1 Global Variance (GV) definition

The GV [19, 9] is defined as the second moment of the parameter trajectory:

$$\boldsymbol{v}(\boldsymbol{y}) = \left[v(1), \cdots, v(d), \cdots, v(D) \right]^{\top}, \tag{93}$$

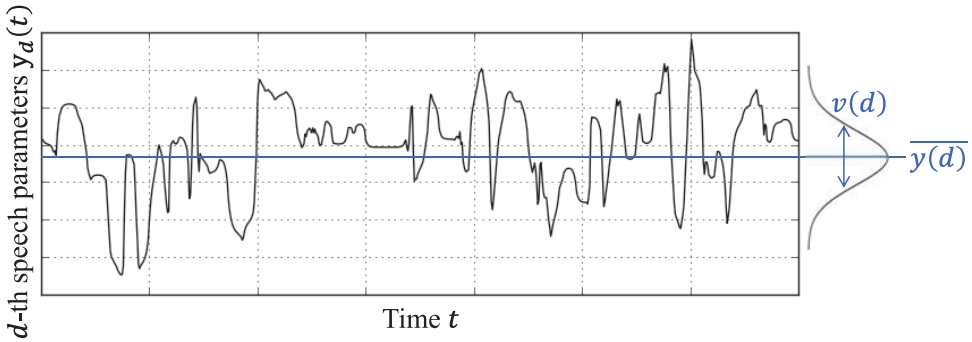


Figure 27. How to derive the Global Variance (GV). The scaling of the temporal sequence are given as the scalar value.

$$v\left(d\right) = \frac{1}{T}\sum_{t=1}^{T}\left(y_t\left(d\right) - \bar{y}\left(d\right)\right)^2, \tag{94}$$

$$\bar{y}\left(d\right) = \frac{1}{T}\sum_{\tau=1}^{T} y_\tau\left(d\right), \tag{95}$$

where $\boldsymbol{v}\left(\boldsymbol{y}\right)$ is a D-dimensional GV vector of $\boldsymbol{y}$. The GV of the generated speech parameter trajectory tends to be smaller than that of a natural speech parameter trajectory, as we will describe in **Section 5.5**. The probability density function of the GV and the GV parameter set are $\mathcal{N}\left(\boldsymbol{v}\left(\boldsymbol{y}\right); \boldsymbol{\mu}_{\mathrm{v}}, \boldsymbol{\Sigma}_{\mathrm{v}}\right)$ and $\boldsymbol{\lambda}_{\mathrm{v}} = \{\boldsymbol{\mu}_{\mathrm{v}}, \boldsymbol{\Sigma}_{\mathrm{v}}\}$, respectively. The statistics of the GV are trained from the natural speech parameters. The following synthesis and training criteria can improve the quality of the synthetic speech because the GV model alleviates the over-smoothing effect.

2.9.2 Speech parameter generation considering GV

The objective function of this generation algorithm combines the traditional generation criterion and the GV likelihood [19, 9] as follows:

$$\hat{\boldsymbol{y}}_{\hat{\boldsymbol{q}}} = \underset{\boldsymbol{y}}{\operatorname{argmax}}\, L_{\mathrm{gv}}^{(\mathrm{gen})} \tag{96}$$

$$L_{\mathrm{gv}}^{(\mathrm{gen})} = P\left(\boldsymbol{W}\boldsymbol{y}|\, \boldsymbol{q}, \boldsymbol{X}, \boldsymbol{\lambda}\right) P\left(\boldsymbol{v}\left(\boldsymbol{y}\right)|\, \boldsymbol{\lambda}_{\mathrm{v}}\right)^{w_{\mathrm{v}} N_{\mathrm{w}} T}, \tag{97}$$

w_{v} is the weight of the GV likelihoods. Because this function does not solved in a closed form, an iterative generation algorithm is used to calculate it.

2.9.3 GV-constrained HMM/GMM training

Because the speech parameter generation taking accounts of the GV requires iterations, the computational efficiency of the traditional generation algorithm using $L_{\mathrm{basic}}^{(\mathrm{syn})}$ is lost. Another approach [3, 2] defines the training criterion with

the GV in order to get a quality improvement while preserving computational efficiency. The HMM/GMM parameter sets $\boldsymbol{\lambda}$ are trained as:

$$\boldsymbol{\lambda} = \underset{\boldsymbol{\lambda}}{\operatorname{argmax}}\, L_{\mathrm{gv}}^{(\mathrm{trn})} \tag{98}$$

GV-constraint trajectory HMM training [3] has defined $L_{\mathrm{gv}}^{(\mathrm{trn})}$ that integrates the GV term into the trajectory HMM training as follows:

$$L_{\mathrm{gv}}^{(\mathrm{trn})} = P\left(\boldsymbol{y}|\boldsymbol{W}, \boldsymbol{X}, \hat{\boldsymbol{q}}, \boldsymbol{\lambda}\right) P\left(\boldsymbol{v}\left(\boldsymbol{y}\right)|\boldsymbol{X}, \hat{\boldsymbol{q}}, \boldsymbol{\lambda}, \boldsymbol{\lambda}_{\mathrm{v}}\right)^{\omega_{\mathrm{v}} T}. \tag{99}$$

We can utilize the basic generation criterion $L_{\mathrm{basic}}^{(\mathrm{syn})}$ without the GV in the synthesis stage because the generated GV is compensated in the training stage. As Toda *et al.* [3] pointed out, this approach uses a unified criteria between training and synthesis because the ML estimates of $L_{\mathrm{gv}}^{(\mathrm{trn})}$ and $L_{\mathrm{basic}}^{(\mathrm{syn})}$ are the same. This approach also makes it possible to model a GV depending on the input parameters. Similarly, GV-constraint GMM training [2] defines the training criterion in GMM-based VC as:

$$L_{\mathrm{gv}}^{(\mathrm{trn})} = P\left(\boldsymbol{W}\boldsymbol{y}|\boldsymbol{X}, \hat{\boldsymbol{q}}, \boldsymbol{\lambda}\right) P\left(\boldsymbol{v}\left(\boldsymbol{y}\right)|\boldsymbol{X}, \hat{\boldsymbol{q}}, \boldsymbol{\lambda}, \boldsymbol{\lambda}_{\mathrm{v}}\right)^{\omega_{\mathrm{v}} N_{\mathrm{w}} T}, \tag{100}$$

Note that this approach inconsistently performs training because it ignores the relationship $\boldsymbol{W}$ between the static and dynamic features.

2.10 Summary of this chapter

This chapter described the various speech synthesis techniques as follows.

Section 2.2: We reviewed unit selection synthesis, which directly uses acoustic inventories selected from a speech corpus for synthesizing speech waveforms. Unit selection synthesis can produce high-quality speech thanks to directly using speech segments. However, the voice characteristics are fully dependent on the original speech stored in the acoustic inventories.

Section 2.3: We reviewed statistical parametric speech synthesis. Here, the speech waveforms of a speech corpus are first parameterized with text or speech analyzers; then, instead of selecting a speech waveform, statistical models are trained to represent the relationship between input and output features. HMM-based TTS and GMM-based VC are examples of this kind of speech synthesis method.

Section 2.4: We described how to model the speech parameters with the HMMs for HMM-based TTS. The output probability density function models the static and dynamic speech features, and the state duration density function explicitly models the state duration. To avoid the sparsity problem with the full context

labels of the input text, tree-based clustering divides the acoustic space into sub-regions and calculates the statistics for each leaf node. The clustering loses the information of the individual context labels.

Section 2.5: We described how to model the speech parameters with the GMM for GMM-based VC. The static and dynamic features of the input and output speech are jointly modeled with the GMM. Each GMM mixture component models the sub-region of the joint acoustic space. Similarly to HMM-based TTS, this modeling loses individual information.

Section 2.6: The speech parameters of synthetic speech are generated from HMMs and GMMs on the basis of the ML criterion under the constraints on the static and dynamic features. This process is computationally efficient because it can be analytically solved. However, the synthetic speech parameters tend to be over-smoothed, and the synthetic speech sounds muffled.

Section 2.7: In order to alleviate the averaging effect in the modeling process described in **Section 2.4**, we presented two approaches that incorporate the ideas of unit selection synthesis into HMM-based TTS. The quality of the synthetic speech dramatically improves, but the flexibility of the original HMM-based TTS is lost.

Section 2.8: The training and synthesis stages of the basic HMM-based TTS are inconsistent with each other; i.e., the training stage uses the likelihoods for the joint static and dynamic feature vectors, whereas the synthesis stage uses only the static feature vectors. The trajectory HMM models the static feature vector sequence in the same way as in the synthesis stage.

Section 2.9: The Global Variance (GV) can be used to quantify the over-smoothing effect observed in **Section 2.6**. The GV is the second moment of the speech parameter sequence, and is well integrated into the speech parameter generation and training stage.

Chapter

3

Statistical sample-based speech synthesis

3.1 Introduction

Section 2.7 described two hybrid methods combining unit selection synthesis and HMM-based TTS, i.e., ML-based unit selection and KLD-based rich context model selection. Although they exploit the ideas of unit selection synthesis to make the modeling more accurate, they lose the flexibility of the original HMM-based TTS.

In this chapter, we propose statistical sample-based approaches using rich context models for speech synthesis that are both high quality and flexible (see Fig. 28). First, we apply rich context modeling to GMM-based VC. Then, we devise ML-based parameter generation methods using rich context models that preserve the flexibility of the HMM-based TTS and GMM-based VC. The trained rich context models are reformulated as a *Rich context Gaussian Mixture Model (R-GMM)* in each sub-region corresponding to one leaf node in HMM-based TTS and one GMM mixture component in GMM-based VC. The speech parameter sequence in each speech parameter component is separately and iteratively generated from the selected rich context models of R-GMMs by using the ML criterion. The methods presented here make it possible to use the probability distributions of individual components from different waveform segments as in the original HMM-based TTS and GMM-based VC. Therefore, compared with the other hybrid methods, they have more flexibility when it comes to modeling speech features. An iterative algorithm is used to generate speech parameters using the rich context models selected from the R-GMMs. As for the initialization of the iteration, we build over-trained acoustic models to generate a less-averaged initial parameter sequence. Discontinuous transitions are observed in the initial parameters, but they can be alleviated by using HMM/GMM likelihoods in the iterative parameter generation.

This chapter is organized as follows (see Fig. 29). In **Section 3.2**, we describe the method of training the rich context models[12] for GMM-based VC. In **Section 3.2**, we reformulate the rich context models belonging to one sub-region as an R-GMM in HMM-based TTS and GMM-based VC. In **Section 3.4**, we propose two ML-based speech parameter generation methods; one uses the EM algorithm, and the other one is an approximation with single Gaussian distributions. The methods described in **Section 3.5** are used to initialize these proposed generation methods. In **Section 3.6**, we compare these methods with the conventional hybrid methods in terms of the flexibility of their speech synthesis frameworks, and compare them with the basic HMM-based TTS and GMM-based VC in terms of the quality of the speech that they provide. We describe an experimental evaluation of HMM-based TTS in **Section 3.7** and GMM-based VC in **Section**

[12] The defined name "rich context model" is not strictly accurate for GMM-based VC because no context labels are used. However, we use this name to maintain consistency with the method proposed for HMM-based TTS.

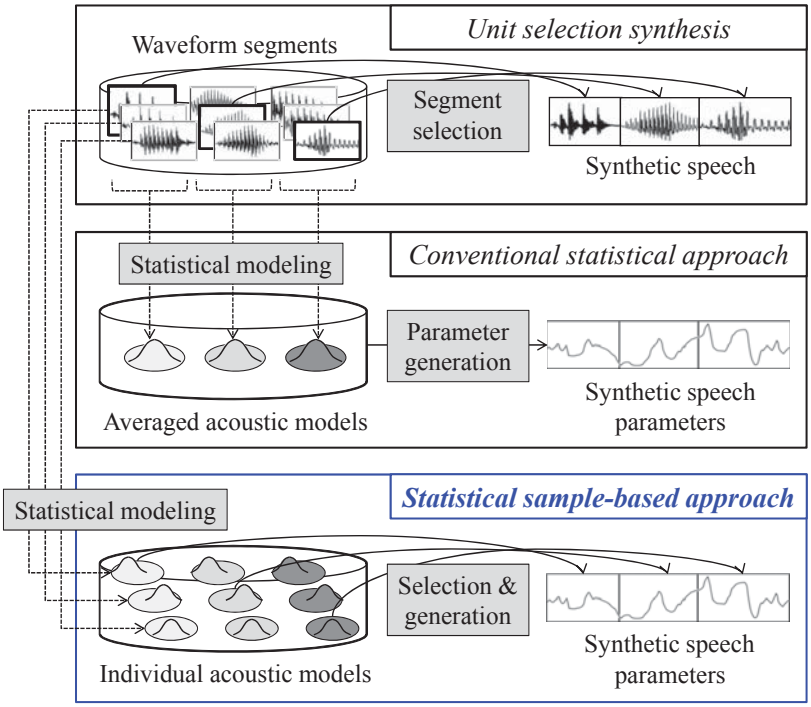


Figure 28. Comparison of unit selection synthesis, conventional statistical approaches (HMM-based TTS and GMM-based VC), and the proposed statistical sample-based approach. Whereas the acoustic model corresponds to a number of speech segments in the conventional statistical approaches, it corresponds to just one speech segment in the statistical sample-based approach. Note that the individual acoustic models are calculated using individual speech segments, but their covariance matrices are the same to those of the averaged acoustic models.

3.8, and summarize the chapter in **Section 3.9**.

3.2 Rich context modeling for GMM-based VC

As shown in Fig. 23, the rich context model of each sub-region in HMM-based TTS is constructed by estimating the mean vector while tying the covariance matrix of the clustered model. After the conventional training of GMM-based VC, rich context models are trained for individual joint speech features, $\boldsymbol{Z}_t$, by updating the mean vector of the GMM mixture components while tying its covariance matrix. The m-th rich context model of the q-th GMM mixture component is

$$P\left(\boldsymbol{Z}_t|q,m,\boldsymbol{\lambda}\right) = \mathcal{N}\left(\boldsymbol{Z}_t;\boldsymbol{\mu}_{q,m}^{(Z)},\boldsymbol{\Sigma}_q^{(Z)}\right), \tag{101}$$

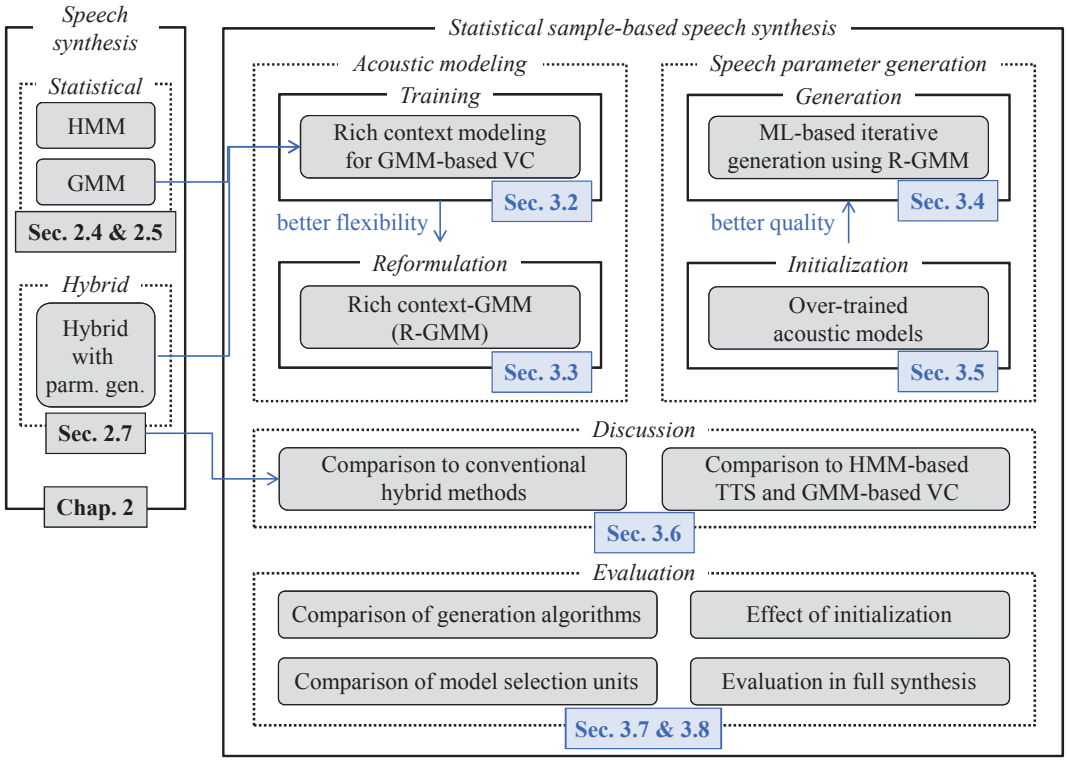


Figure 29. The rest of Chapter 3.

$$\boldsymbol{\mu}_{q,m}^{(Z)} = \begin{bmatrix} \boldsymbol{\mu}_{q,m}^{(X)} \\ \boldsymbol{\mu}_{q,m}^{(Y)} \end{bmatrix}, \tag{102}$$

where the mean vector $\boldsymbol{\mu}_{q,m}^{(Z)}$ consists of the individual input and output mean vectors, $\boldsymbol{\mu}_{q,m}^{(X)}$ and $\boldsymbol{\mu}_{q,m}^{(Y)}$. The individual mean vectors are estimated based on the ML criterion, and each of them is equal to one joint feature vector. The mixture component that $\boldsymbol{Z}_t$ belongs to is determined as follows:

$$\hat{q}_t = \underset{q}{\operatorname{argmax}}\, P\left(q|\boldsymbol{Z}_t, \boldsymbol{\lambda}\right). \tag{103}$$

This thesis performs discriminative GMM training [119] between the conventional training and rich context model training in order to alleviate the mismatch between Eq. (103) and Eq. (52)[13]. As described in the following section, the rich context models are selected from the mixture components determined with Eq. (52) in the speech parameter generation. Therefore, we expect that the discriminative training can select better rich context models.

[13] Whereas $P\left(q|\boldsymbol{Z}_t, \boldsymbol{\lambda}\right)$ is used in the training stage, $P\left(q|\boldsymbol{X}_t, \boldsymbol{\lambda}\right)$ is used in the conversion stage. The discriminative training algorithm [119] trains the GMM parameters to alleviate this inconsistency.

3.3 Reformulation of Rich context GMM (R-GMM)

After training the rich context models, the output probability density in each sub-region is modeled using an R-GMM developed with all rich context models belonging in the same sub-region. For continuous HMMs in HMM-based TTS, the output probability density of the c-th leaf node of the q-th HMM state is:

$$b_{q_c}\left(\boldsymbol{Y}_t\right) = \sum_{m=1}^{M_{q_c}} w_{q_c,m}^{(Y|X)} \mathcal{N}\left(\boldsymbol{Y}_t; \boldsymbol{\mu}_{q_c,m}^{(Y|X)}, \boldsymbol{\Sigma}_{q_c}^{(Y|X)}\right), \tag{104}$$

where $w_{q_c,m}^{(Y|X)}$ is the mixture component weight of the m-th rich context model, and the total number of mixture components is M_{q_c}. Similarly, the R-GMM for MSD-HMMs is given as:

$$b_{q_c}\left(\boldsymbol{Y}_t\right) = \begin{cases} \sum_{m=1}^{M_{q_c}} w_{q_c,m}^{(Y|X)} \mathcal{N}\left(\boldsymbol{Y}_t; \boldsymbol{\mu}_{q_c,m}^{(Y|X)}, \boldsymbol{\Sigma}_{q_c}^{(Y|X)}\right) & l_t = \mathrm{V} \\ 1 - \sum_{m=1}^{M_{q_c}} w_{q_c,m}^{(Y|X)}, & l_t = \mathrm{U} \end{cases}, \tag{105}$$

The R-GMM of the q-th GMM mixture component in GMM-based VC is:

$$b_q\left(\boldsymbol{Z}_t\right) = \sum_{m=1}^{M_q} w_{q,m}^{(Z)} \mathcal{N}\left(\boldsymbol{Z}_t; \boldsymbol{\mu}_{q,m}^{(Z)}, \boldsymbol{\Sigma}_q^{(Z)}\right), \tag{106}$$

where $w_{q_c,m}^{(Z)}$ is the weight of the m-th rich context model of the q-th GMM mixture component (not R-GMM), and M_q is the total number of mixture components. We can calculate the ML estimates of $w_{q_c,m}^{(Y|X)}$ and $w_{q_c,m}^{(Z)}$ from the occupancy counts given by the EM algorithm, but in this thesis, we set them to equivalent values as follows[14] :

$$\begin{array}{ll} w_{q_c,m}^{(Y|X)} = 1/M_{q_c} & \text{(continuous HMMs)} \\ w_{q_c,m}^{(Y|X)} = w_{q_c}^{(Y|X)}/M_{q_c} & \text{(MSD-HMMs)} \\ w_{q,m}^{(Z)} = 1/M_q & \text{(GMMs)} \end{array}, \tag{107}$$

where $w_{q_c}^{(Y|X)}$ is the weight of the voiced space. We have found this weight setting yields small quality improvements in the synthetic speech. This point is described in **Section A.5**.

3.4 Parameter generation methods

A speech parameter sequence is generated from rich context models selected from the R-GMMs. As in the same as HMM-based TTS and GMM-based VC, it is

[14] There are no duplicated joint speech features in GMM-based VC, but such features are included in the training data because we use Dynamic Time Warping (DTW) to make the joint feature vectors.

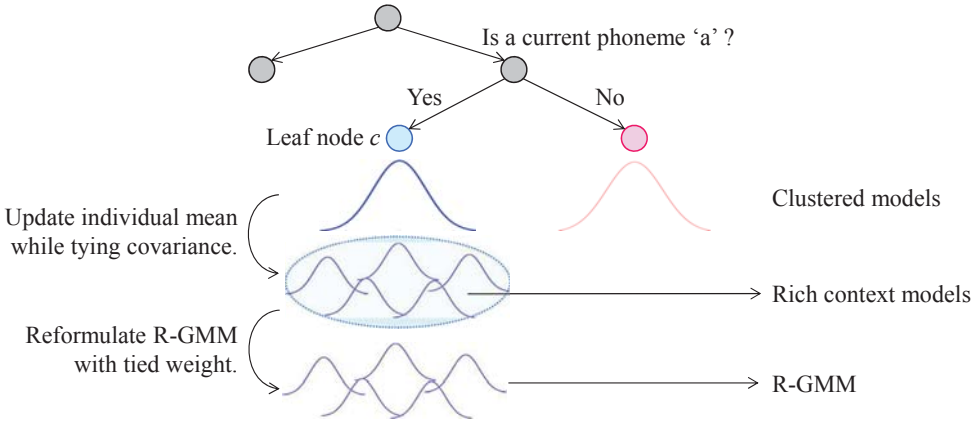


Figure 30. How to construct a R-GMM using M_{q_c} rich context models belonging to the q-th leaf node of the q-th HMM state in HMM-based TTS. Comparing Fig. 14 and Fig. 18, we can see that this construction is similar to that of GMM-based VC.

generated by maximizing the probability density function as follows[15]:

$$\hat{\boldsymbol{y}}_{\hat{\boldsymbol{q}}} = \underset{\boldsymbol{y}}{\operatorname{argmax}}\, P\left(\boldsymbol{Y}|\hat{\boldsymbol{q}}, \boldsymbol{X}, \boldsymbol{\lambda}\right) \tag{108}$$

$$= \underset{\boldsymbol{y}}{\operatorname{argmax}}\, P\left(\boldsymbol{W}\boldsymbol{y}|\hat{\boldsymbol{q}}, \boldsymbol{X}, \boldsymbol{\lambda}\right) \tag{109}$$

$$= \underset{\boldsymbol{y}}{\operatorname{argmax}} \sum_{\text{all } \boldsymbol{m}} P\left(\boldsymbol{W}\boldsymbol{y}|\hat{\boldsymbol{q}}, \boldsymbol{m}, \boldsymbol{X}, \boldsymbol{\lambda}\right) P\left(\boldsymbol{m}|\hat{\boldsymbol{q}}, \boldsymbol{X}, \boldsymbol{\lambda}\right) \tag{110}$$

where $\boldsymbol{m} = [m_1, \cdots, m_t, \cdots, m_T]$ is a mixture component sequence of R-GMM. $P\left(\boldsymbol{W}\boldsymbol{y}|\hat{\boldsymbol{q}}, \boldsymbol{m}, \boldsymbol{X}, \boldsymbol{\lambda}\right)$ is:

$$P\left(\boldsymbol{W}\boldsymbol{y}|\hat{\boldsymbol{q}}, \boldsymbol{m}, \boldsymbol{X}, \boldsymbol{\lambda}\right) = \mathcal{N}\left(\boldsymbol{W}\boldsymbol{y}; \boldsymbol{E}_{\hat{\boldsymbol{q}},\boldsymbol{m}}; \boldsymbol{D}_{\hat{\boldsymbol{q}}}\right) \tag{111}$$

$$\boldsymbol{E}_{\hat{\boldsymbol{q}},\boldsymbol{m}} = \left[\boldsymbol{\mu}^{\top}_{\hat{q}_1,m_1,1}, \cdots, \boldsymbol{\mu}^{\top}_{\hat{q}_t,m_t,t}, \cdots, \boldsymbol{\mu}^{\top}_{\hat{q}_T,m_T,T}\right] \tag{112}$$

$$\boldsymbol{D}_{\hat{\boldsymbol{q}}} = \operatorname{diag}_{N_{\mathrm{w}}D}\left[\boldsymbol{\Sigma}_{\hat{q}_1}, \cdots, \boldsymbol{\Sigma}_{\hat{q}_t}, \cdots, \boldsymbol{\Sigma}_{\hat{q}_T}\right] \tag{113}$$

$$\boldsymbol{\mu}_{q,m,t} = \begin{cases} \boldsymbol{\mu}^{(Y|X)}_{q,m} & \text{(HMM)} \\ \boldsymbol{A}_q \boldsymbol{X}_t + \boldsymbol{b}_{q,m} & \text{(GMM)} \end{cases} \tag{114}$$

$$\boldsymbol{\Sigma}_q = \begin{cases} \boldsymbol{\Sigma}^{(Y|X)}_q & \text{(HMM)} \\ \boldsymbol{\Sigma}^{(YY)}_q - \boldsymbol{A}_q \boldsymbol{\Sigma}^{(XX)}_q \boldsymbol{A}^{\top}_q & \text{(GMM)} \end{cases} \tag{115}$$

$$\boldsymbol{A}_q = \boldsymbol{\Sigma}^{(YX)}_q \boldsymbol{\Sigma}^{(XX)-1}_q, \tag{116}$$

$$\boldsymbol{b}_{q,m} = \boldsymbol{\mu}^{(Y)}_{q,m} - \boldsymbol{A}_q \boldsymbol{\mu}^{(X)}_q \tag{117}$$

[15] The optimal HMM state and GMM mixture component sequence are determined in the standard manner (Eq. (51) for HMM-based TTS and Eq. (52) for GMM-based VC).

$P\left(\boldsymbol{m}|\hat{\boldsymbol{q}}, \boldsymbol{X}, \boldsymbol{\lambda}\right)$ is given as:

$$P\left(\boldsymbol{m}|\hat{\boldsymbol{q}}, \boldsymbol{X}, \boldsymbol{\lambda}\right) = \prod_{t=1}^{T} P\left(m|\boldsymbol{X}_t, \hat{q}_t, \boldsymbol{\lambda}\right) \tag{118}$$

$$P\left(m|\boldsymbol{X}_t, q, \boldsymbol{\lambda}\right) = \begin{cases} w_{q,m}^{(Y|X)} & \text{(HMM)} \\ w_{q,m}^{(Z)} & \text{(GMM)} \end{cases} \tag{119}$$

The posterior probability $P\left(m|\boldsymbol{X}_t, q, \boldsymbol{\lambda}\right)$ of GMM-based VC is normally described as:

$$P\left(m|\boldsymbol{X}_t, q, \boldsymbol{\lambda}\right) = w_{q,m}^{(Y|X)} = \frac{w_{q,m}^{(Z)} \mathcal{N}\left(\boldsymbol{X}_t; \boldsymbol{\mu}_{q,m}^{(X)}; \boldsymbol{\Sigma}_q^{(XX)}\right)}{\sum_{m=1}^{M_q} w_{q,m}^{(Z)} \mathcal{N}\left(\boldsymbol{X}_t; \boldsymbol{\mu}_{q,m}^{(X)}; \boldsymbol{\Sigma}_q^{(XX)}\right)} \tag{120}$$

However, we set the weight $w_{\hat{q},m}^{(Y|X)}$ to constant to m in each component, like in the case of HMM-based TTS. In practice, there are enormous numbers of candidates for the rich context models in GMM-based VC[16]. Therefore, we calculate $P\left(m_t|q, \boldsymbol{X}_t, \boldsymbol{\lambda}\right)$ in a similar fashion to Eq. (52), and set $P\left(m_t|q, \boldsymbol{X}_t, \boldsymbol{\lambda}\right) = 1/M_{q,t}$ for the rich context models having the $M_{q,t}$-best posterior probabilities, and $P\left(m_t|q, \boldsymbol{X}_t, \boldsymbol{\lambda}\right) = 0$ otherwise, where $M_{q,t}$ $(1 \leq M_{q,t} \leq M_q)$ is the number of candidates at frame t.

3.4.1 EM algorithm

The synthetic speech parameter sequence $\hat{\boldsymbol{y}}_{\hat{\boldsymbol{q}}}$ is determined with the EM algorithm. First, an initial static feature vector sequence $\boldsymbol{y}_{\hat{\boldsymbol{q}}}^{(0)}$ is determined. Then, the following auxiliary function is maximized by iteratively updating the posterior probability $P\left(\boldsymbol{m}|\boldsymbol{W}\boldsymbol{y}_{\hat{\boldsymbol{q}}}^{(i)}, \hat{\boldsymbol{q}}, \boldsymbol{\lambda}\right)$ given the current estimate $\boldsymbol{y}_{\hat{\boldsymbol{q}}}^{(i)}$ in the E-step and the new estimate $\hat{\boldsymbol{y}}_{\hat{\boldsymbol{q}}}^{(i+1)}$, while keeping it constant in the M-step:

$$Q\left(\boldsymbol{y}_{\hat{\boldsymbol{q}}}^{(i)}, \boldsymbol{y}_{\hat{\boldsymbol{q}}}^{(i+1)}\right) = \sum_{\text{all } \boldsymbol{m}} P\left(\boldsymbol{m}|\boldsymbol{W}\boldsymbol{y}_{\hat{\boldsymbol{q}}}^{(i)}, \hat{\boldsymbol{q}}, \boldsymbol{X}, \boldsymbol{\lambda}\right) \ln P\left(\boldsymbol{W}\boldsymbol{y}_{\hat{\boldsymbol{q}}}^{(i+1)}, \boldsymbol{m}|\hat{\boldsymbol{q}}, \boldsymbol{X}, \boldsymbol{\lambda}\right) \tag{121}$$

The parameter sequence is given by:

$$\hat{\boldsymbol{y}}_{\hat{\boldsymbol{q}}} = \left(\boldsymbol{W}^\top \overline{\boldsymbol{D}_{\hat{\boldsymbol{q}}}^{-1}} \boldsymbol{W}\right)^{-1} \boldsymbol{W}^\top \overline{\boldsymbol{D}_{\hat{\boldsymbol{q}}}^{-1} \boldsymbol{E}_{\hat{\boldsymbol{q}}}}, \tag{122}$$

[16] Even if the training data is the same size as that of HMM-based TTS, the number of the candidates will be bigger. This is because one rich context model corresponds to one speech feature vector in GMM-based VC, whereas it corresponds to one speech segment in HMM-based TTS.

where

$$\overline{\boldsymbol{D}_{\hat{\boldsymbol{q}}}^{-1}} = \mathrm{diag}_{N_{\mathrm{w}}D}\left[\overline{\boldsymbol{\Sigma}_{\hat{q}_1}^{-1}}, \cdots, \overline{\boldsymbol{\Sigma}_{q_t}^{-1}}, \cdots, \overline{\boldsymbol{\Sigma}_{q_T}^{-1}}\right], \tag{123}$$

$$\overline{\boldsymbol{\Sigma}_{\hat{q}_t}^{-1}} = \sum_m \gamma_t(m)\,\boldsymbol{\Sigma}_{q_t}^{-1}, \tag{124}$$

$$\overline{\boldsymbol{D}_{\hat{\boldsymbol{q}}}^{-1}\boldsymbol{E}_{\hat{\boldsymbol{q}}}} = \left[\overline{\boldsymbol{\Sigma}_{\hat{q}_1}^{-1}\boldsymbol{\mu}_{\hat{q}_1,1}}^{\top}, \cdots, \overline{\boldsymbol{\Sigma}_{\hat{q}_t}^{-1}\boldsymbol{\mu}_{\hat{q}_t,t}}^{\top}, \cdots, \overline{\boldsymbol{\Sigma}_{\hat{q}_T}^{-1}\boldsymbol{\mu}_{\hat{q}_T,T}}^{\top}\right]^{\top}, \tag{125}$$

$$\overline{\boldsymbol{\Sigma}_{\hat{q}_t}^{-1}\boldsymbol{\mu}_{\hat{q}_t,t}} = \sum_m^{M_{q_t}} \gamma_t(m)\,\boldsymbol{\Sigma}_{q_t}^{-1}\boldsymbol{\mu}_{\hat{q}_t,m,t}, \tag{126}$$

$$\gamma_t(m) = P(m|\boldsymbol{Y}_t, \boldsymbol{X}, \hat{\boldsymbol{q}}, \boldsymbol{\lambda}). \tag{127}$$

3.4.2 Approximation with single Gaussian

We approximate the likelihood in Eq. (110) with a single mixture component sequence as follows:

$$P(\boldsymbol{Y}|\hat{\boldsymbol{q}}, \boldsymbol{X}, \boldsymbol{\lambda}) = \sum_{\text{all } \boldsymbol{m}} P(\boldsymbol{W}\boldsymbol{y}|\hat{\boldsymbol{q}}, \boldsymbol{m}, \boldsymbol{X}, \boldsymbol{\lambda})\, P(\boldsymbol{m}|\hat{\boldsymbol{q}}, \boldsymbol{X}, \boldsymbol{\lambda}) \tag{128}$$

$$\simeq P(\boldsymbol{W}\boldsymbol{y}|\hat{\boldsymbol{q}}, \hat{\boldsymbol{m}}, \boldsymbol{X}, \boldsymbol{\lambda})\, P(\hat{\boldsymbol{m}}|\hat{\boldsymbol{q}}, \boldsymbol{X}, \boldsymbol{\lambda}) \tag{129}$$

After determining the initial static feature vector sequence $\boldsymbol{y}_{\hat{\boldsymbol{q}}}^{(0)}$, the single mixture component sequence and the static feature vector sequence are iteratively updated as follows:

$$\hat{\boldsymbol{m}}^{(i+1)} = \underset{\boldsymbol{m}}{\mathrm{argmax}}\, P\left(\boldsymbol{m}|\boldsymbol{W}\boldsymbol{y}_{\hat{\boldsymbol{q}}}^{(i)}, \hat{\boldsymbol{q}}, \boldsymbol{X}, \boldsymbol{\lambda}\right), \tag{130}$$

$$\hat{\boldsymbol{y}}_{\hat{\boldsymbol{q}}}^{(i+1)} = \underset{\boldsymbol{y}}{\mathrm{argmax}}\, P\left(\boldsymbol{W}\boldsymbol{y}|\boldsymbol{m}^{(i+1)}, \hat{\boldsymbol{q}}, \boldsymbol{X}, \boldsymbol{\lambda}\right). \tag{131}$$

Eq. (131) is solved in the same manner as the basic generation algorithm [109]. Fig. 31 shows the procedure. Eq. (130) and Eq. (131) correspond to model selection and parameter generation, respectively.

3.5 Initialization method using over-trained acoustic models

We need to determine the initial parameter sequence. A reasonable way is to use the parameter sequence generated by the clustered models in the manner of HMM-based TTS and GMM-based VC. Although the transitions of this initial parameter sequence are continuous, the parameter sequence is over-smoothed as described in **Section 2.6**. The generated parameters are strongly influenced by this over-smoothing effect, and the improvement in speech quality is slight [120].

To generate a less-smoothed initial parameter sequence, we propose an ini-

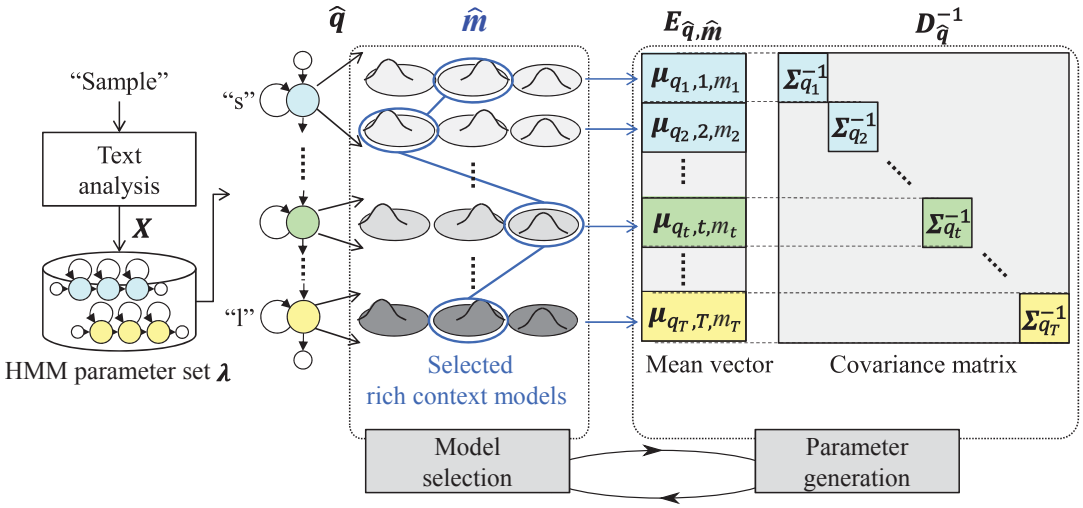


Figure 31. Speech parameter generation using rich context models. The selection stage and generation stage in this figure correspond to Eq. (130) and Eq. (131), respectively. In the case of the EM algorithm, these stages correspond to the E-step and M-step of the algorithm.

tialization with the over-trained acoustic models shown in Fig. 32. The original sub-regions are further divided up and statistics calculated. Over-trained acoustic models are calculated from only a few acoustic inventories in the sub-region. The resulting initial speech parameter sequence is less-smoothed than one generated by the conventional clustered model. It is expected that such an initial parameter sequence will help the parameter generation process with rich context models to select a less-smoothed model sequence. On the other hand, this approach causes over-training problems. In particular, the initial parameter sequence suffers from discontinuous transitions. This discontinuity problem can be addressed by the use of tied covariance matrices in the rich context models and model selection based on the likelihoods for both static and dynamic features.

In HMM-based TTS, we grow another larger decision tree by decreasing the MDL parameter a_{MDL}. Note that the sufficient statistics to build this tree are the same as those used in calculating rich context models. Therefore, its tree structure is slightly different from that of original decision tree for the conventional clustered models. Examples of initial and generated parameter sequences are shown in Fig. 33. We can see that discontinuous transitions in the initial parameter sequence are alleviated in the generated parameter sequence. For the F_0 contour generation, the voiced and unvoiced intervals are determined in initialization using the larger decision tree.

We train the over-trained acoustic models for each sub-region as shown in Fig. 32. In GMM-based VC, the acoustic space is first divided into Q sub-regions by

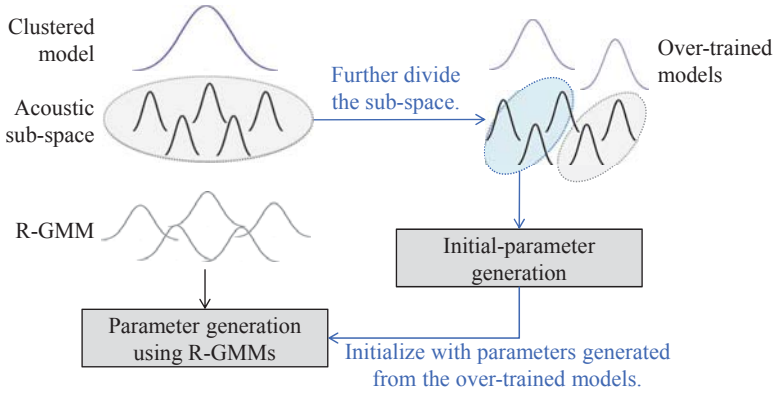


Figure 32. Initialization for iterative generation using rich context models.

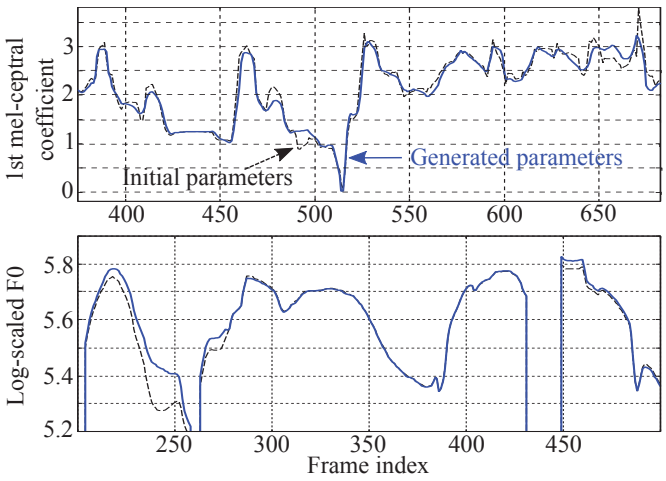


Figure 33. Example of initial and generated mel-cepstral coefficient sequences and F_0 contours in HMM-based TTS.

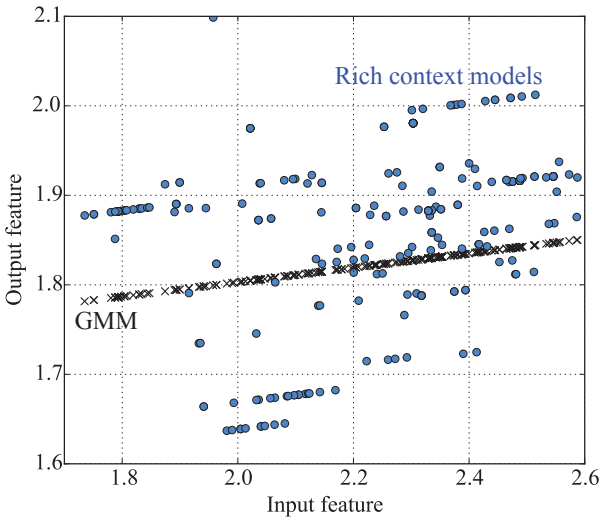


Figure 34. Example of the conversion function within one GMM mixture component. Whereas the basic function of GMM-based VC is linear, those yielded by the rich context models are piece-wise linear.

using Eq. (103), then the acoustic models are trained to fit the training data of each sub-region. This over-trained acoustic model is given as a GMM for each sub-region, and is trained in the standard manner. The total number of over-trained models is the sum of the number of mixture components of the GMMs. The MDL criterion [118] can be utilized to determine the number of over-trained models, but we determine it by using the Linde-Buzo-Gray (LBG) algorithm [121]. After determining $\hat{\boldsymbol{q}}$, the over-trained models are selected in a manner similar to that of Eq. (52), and the initial parameter sequence is generated in the standard manner using the over-trained models.

3.6 Discussion

One rich context model usually corresponds to one speech segment or speech feature vector. Therefore, the proposed processes are strongly related to unit selection synthesis. In the parameter generation methods described above, the HMM/GMM likelihood for the static features and that for the dynamic features are regarded as a target cost and a concatenation cost, respectively [122, 123].

The synthesis process using the EM algorithm is similar to the process of selecting multiple acoustic segments and mixing them to generate speech parameters [124]. On the other hand, the synthesis process with a single mixture component sequence is similar to the process of selecting a single acoustic segment sequence to generate the speech parameters [28]. Also, the model selection step

can be applied to unit selection synthesis by replacing the selected rich context models with the corresponding original speech parameters.

From the perspective of utilizing information on individual speech features, the above synthesis process is the similar to that of kernel-based speech synthesis [33, 103] and the use of a full-sized tree [120]. One of the advantages is that the individual acoustic models can be re-selected in the iterative generation process using the HMM/GMM likelihoods.

The parameter generation methods don't have to use the constraint used in the conventional selection method of rich context models as described in **Section 2.7**. Therefore, they preserve the flexibility of acoustic modeling provided by the basic HMM-based TTS (and also GMM-based VC). For instance, it is possible to separately search for the best rich context model sequences for different speech component parameters to more widely cover a joint acoustic space in HMM-based TTS.

The above parameter generation methods for HMM-based TTS selects the rich context models frame by frame. We can also select them state by state by using an additional constraint that the same rich context model must be selected within the same state in HMM-based TTS. Also, whereas the conventional GMM performs linear conversion within one mixture component, the method described above can perform piece-wise linear conversion, as shown in Fig. 34.

Even if rich context modeling is used, we cannot avoid temporal smoothing as a result of quantizing to the HMM state level because a decision tree is used to tie the HMM states in HMM-based TTS[17]. To address this problem, we can use a micro-level context such as a frame [120] and state level [125].

3.7 Experimental evaluation in HMM-based TTS

3.7.1 Experimental conditions

In the experiments on HMM-based TTS, we trained a context-dependent phoneme HSMM [23] for a Japanese female speaker. We used 450 sentences for training and 53 sentences for evaluation from 503 phonetically balanced sentences including in the ATR Japanese speech database [24]. The speech signals were sampled at 16 kHz. The shift length was set to 5 ms. The 0th-through-24th mel-cepstral coefficients were extracted as spectral parameters and log-scaled F_0 and five band-aperiodicity [74] were extracted as excitation parameters by using the STRAIGHT analysis system [10]. The feature vector consisted of spectral and excitation parameters and their delta and delta-delta features. Five-state left-to-right HSMMs were used. In the synthesis stage, GV [19] was not considered, unless otherwise described.

[17] In GMM-based VC, there is no such phenomenon because each rich context model corresponds to one speech feature vector.

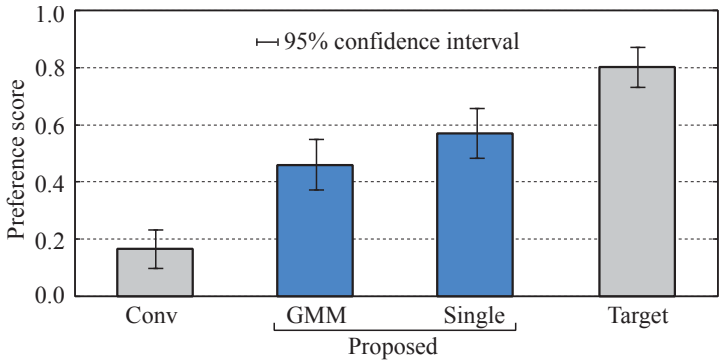


Figure 35. Preference scores on speech quality for comparing two proposed generation methods. The use of a single Gaussian ("Single") produces higher-quality synthetic speech compared with the use of a GMM ("GMM").

We conducted five different experimental evaluations. In the first evaluation, we compared the two parameter generation methods described in **Section 3.4**. In the second evaluation, we compared frame-level and state-level model selection to investigate the model selection unit. In the third evaluation, we investigated the effects of the initial parameter sequence on the generated parameter sequence, and we investigated the effectiveness of the proposed initialization method in the fourth evaluation. In the last evaluation, we applied these methods to both spectral and F_0 components in order to evaluate the effectiveness of our methods with rich context models. Conventional clustered models were used for the duration and aperiodic components in all the evaluations. The parameter generation method using the approximation with a single mixture component sequence is used in all experiments except the first evaluation. To clarify the effectiveness of the methods in a better setting, natural state duration determined by the state-level forced alignment with conventional context-clustered models was used in the first and second evaluations.

3.7.2 Comparison of parameter generation methods

We compared the synthetic speech generated by the conventional clustered model (Conv), our generation method with the EM algorithm (Proposed (GMM)), our generation method using a single mixture component sequence (Proposed (Single)), and a single mixture component sequence selected by the natural speech parameters as a reference (Target). "Conv." was used to generate the initial parameter sequence in our generation methods. Note that our generation methods were applied to only the spectral component, and the clustered model was used for the F_0 component. A preference test (AB test) on speech quality was conducted. Pairs of these four types of synthetic speech were presented to seven

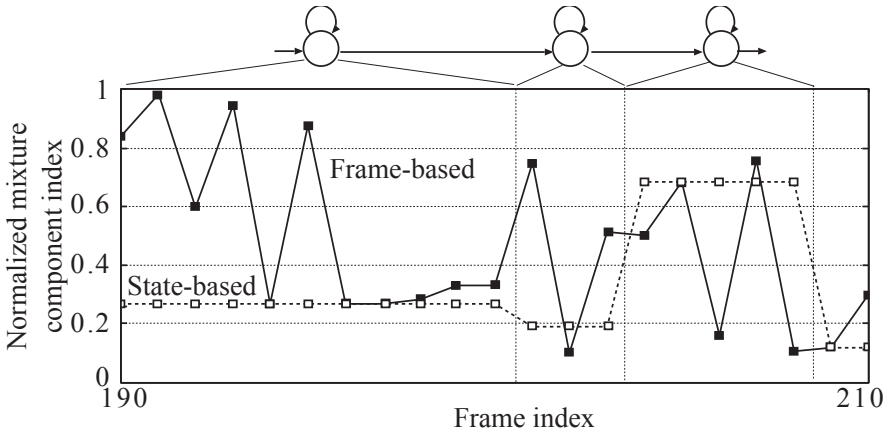


Figure 36. Example of the mixture component sequence selected by frame-based and state-based model selection. Whereas the mixture component varies frame by frame in the frame-based selection, it is tied during one HMM-state in the state-based selection.

listeners in random order. The listeners were asked which sample sounded better in terms of speech quality.

The results are shown in Fig. 35. Our methods significantly improve speech quality. Moreover, the use of a single mixture component sequence yields better-quality speech in comparison with the use of the EM algorithm. We can also see that the likelihood measure has trouble selecting the best rich context model sequence, which is approximated with "Target".

3.7.3 Comparison of model selection unit

We investigate the effect of using different model selection units by comparing synthetic speech generated by our method with a single mixture component sequence selected frame by frame (Proposed (Frame)) or state by state (Proposed (State)) with the conventional clustered model (Conv). Our generation method was applied to each spectral and F_0 component, and the natural speech parameter sequence was used as an initial parameter. In the F_0 component, this initial parameter sequence had both voiced parameters and unvoiced symbol in same HMM-state. In the state-level model selection, we unified the Unvoiced/Voiced (U/V) intervals in the HMM-state by comparing the number of frames of voiced parameter and that of unvoiced symbols. A test involving seven listeners was conducted to compare the spectral and F_0 components of the different selection methods in the same manner as **Section 3.7.2**. We found that the mixture component sequences selected by the two methods were different, as shown in Fig. 36. Note that the selected mixture component index is normalized by the total

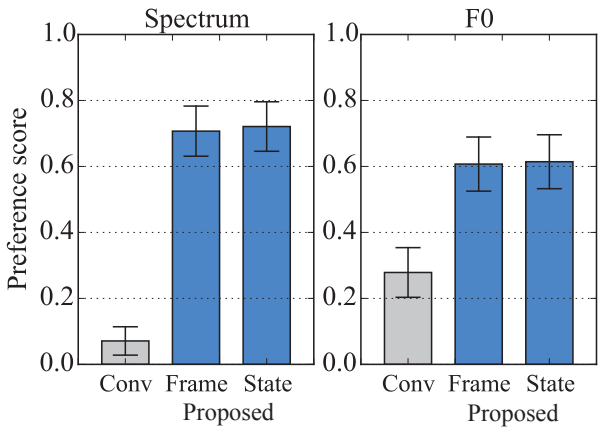


Figure 37. Preference scores on speech quality with 95% confidence interval for comparing the selection unit for spectrum and F_0 in HMM-based TTS. We can see that the frame-level and state-level have the same quality.

number of mixtures.

The results for the spectral and F_0 components are shown in Fig. 37. We can see that the spectral component shows no significant difference between frame-based selection and state-based selection. Moreover, F_0 components of different U/V intervals show a similar tendency. These results indicate that state-based selection is as effective as frame-based selection at improving the quality of the synthetic speech in terms of in both spectral and F_0 components. We can also see that the difference between "Conv" and "Proposed (Frame)" is larger in the spectral component than in the F_0 component. This means that the improvement in the spectral component is more significant than that in the F_0 component.

3.7.4 Objective evaluation for investigating dependency on initial parameter sequence

To investigate the dependency on the setting of the initial parameter sequence on the finally generated speech parameter sequence, we evaluated three settings of the initial parameter sequence: 1) Rand: generated from rich context models randomly selected in individual leaf nodes, 2) Clus: generated from the conventional context-clustered models, and 3) Target: natural target speech parameters. The initial and a final rich context model sequences were evaluated in terms of the likelihoods for the generated speech parameters and natural speech parameters. These evaluation is conducted for each spectral and F_0 components under the natural state duration.

The results of the HMM likelihood for the generated parameters in the spectral

components are shown in Fig. 38(a); those for the natural parameters in the spectral components are shown in Fig. 38(b), and those for the F_0 component are shown in Fig. 39(a) and Fig. 39(b). It is reasonable that the likelihood for the generated speech parameters increases through iteration in both components, as shown in Fig. 38(a) and Fig. 39(a). On the other hand, the likelihood for the natural speech parameters does not always increase through the iteration and its value strongly depends on the initial parameter sequence as shown in Fig. 38(b) and Fig. 39(b). We can also see that the likelihood differences in Fig. 38(b) and Fig. 39(b) are much larger than those in Fig. 38(a) and Fig. 39(a). These results suggest that the HMM likelihood for the generated parameters increases through iteration in every initial parameter setting. In contrast, the HMM likelihood for the natural parameters does not change much. Therefore, these results show that our generation method strongly depends on the choice of the the initial parameters.

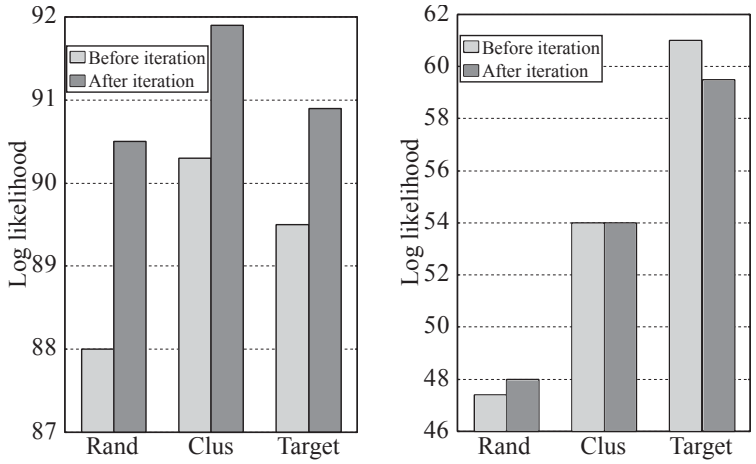
3.7.5 Subjective evaluation for investigating dependency on initial parameter sequence

To confirm the results of objective evaluations, we compared synthetic speech under the generated duration: 1) Conv: generated from conventional clustered models, 2) Proposed (Clus): generated using the parameter sequence of "Conv" as the initial parameters in our generation method, 3) Target: generated using natural target speech parameters as the initial parameters in our generation method. A preference test (AB test) by seven listeners on speech quality was conducted in the same manner as in **Section 3.7.2**. Note that the proposed method was applied to only spectral parameters.

The results of the preference test are shown in Fig. 40. The proposed generation method yields only slight improvements in synthetic speech. On the other hand, we can find that the difference between "Proposed (Clus)" and "Target" is large. Hence, there is a strong dependency on the initial parameters; an appropriate setting is essential. Although we did not do a comparison using the F_0 component, we believe that would have shown similar results.

3.7.6 Alleviating discontinuous transitions arising in initialization

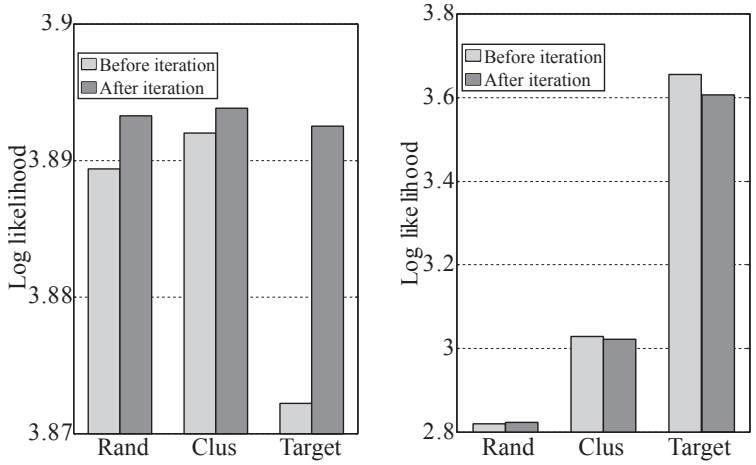
Before investigating the effectiveness of the initialization method, we conducted a preliminary experiment to confirm whether or not the iterative parameter generation method effectively alleviates discontinuous transitions in the initial parameter sequence. We evaluated three settings: 1) Clus: initial parameters generated from the conventional clustered models, 2) $a_{\mathrm{MDL}} = 0.1$: initial parameters generated with a large decision tree ($a_{\mathrm{MDL}} = 0.1$), and 3) Target: natural target speech parameter sequence as a target reference. The difference in the HMM likelihoods



(a) For the generated parameters.

(b) For the natural speech parameters.

Figure 38. HMM likelihoods using rich context models for the spectrum parameter sequences.



(a) For the generated parameters.

(b) For the natural speech parameters.

Figure 39. HMM likelihoods using rich context models for the F_0 contours.

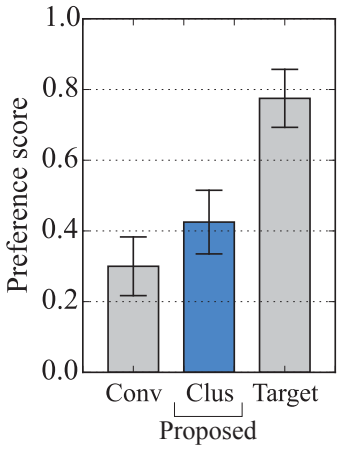


Figure 40. Preference scores on speech quality with 95% confidence interval for determining the dependency on the initial parameter sequence. We can see that the speech quality of "Proposed (Clus)" is heavily degraded compared with "Target."

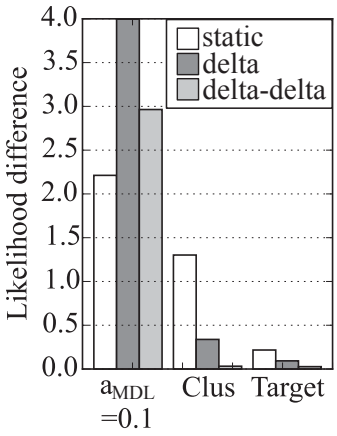


Figure 41. HMM Likelihood differences between before and after iteration. The initialization ("$a_{\mathrm{MDL}} = 0.1$") dramatically increases the likelihood of the temporal delta feature.

for the generated parameters between the initially selected rich context model sequence and the finally selected rich context model sequence was calculated for each static and dynamic features in the spectral parameter.

Fig. 41 shows the result of the likelihood differences yielded by the iterative parameter generation. Here, the HMM likelihood for dynamic features of "$a_{\mathrm{MDL}} = 0.1$" increases more than that in other initial parameter sequences. This means that iterative parameter generation alleviated the discontinuous transitions in the initial parameter sequence. We can also see that the difference in HMM likelihoods for the static feature of "$a_{\mathrm{MDL}} = 0.1$" is the smaller than that of the dynamic feature,s whereas it is the largest for the other methods. These results show that "$a_{\mathrm{MDL}} = 0.1$" better alleviates discontinuous transitions than other methods do.

3.7.7 Objective evaluation of initialization method

To investigate the tree size used to generate the initial parameter sequence, we evaluated three settings of the initial parameters: 1) Clus: initial parameters generated from the conventional clustered models, 2) Proposed: initial parameters generated with a large decision tree ($a_{\mathrm{MDL}} = 0.1, 0.2, \cdots, 1.0$), and 3) Target: natural target speech parameter sequence as a target reference for each spectral

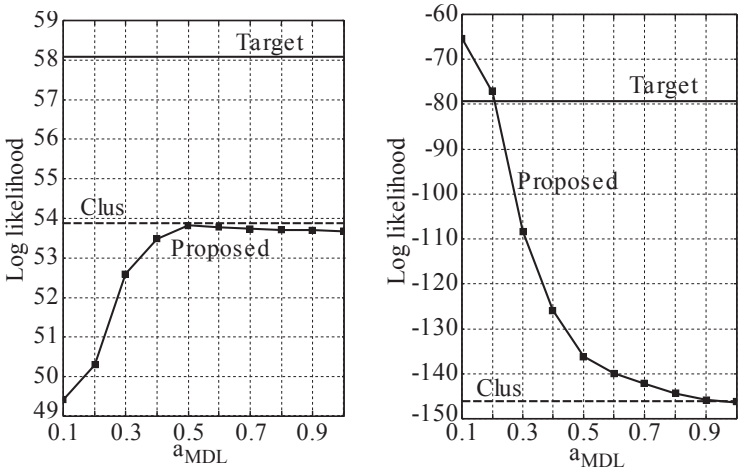
and F_0 component. The rich context model sequences selected by the parameter generation method were evaluated with the HMM likelihood for the natural speech parameters. Moreover, the parameter sequences generated by the selected rich context models (i.e., those generated by the proposed parameter generation method) were evaluated with both the GV likelihood [19] and U/V error rate. The U/V error rate for the F_0 component was calculated as the percentage of U/V mismatched frames in the generated parameter sequence compared with the natural parameter sequence.

Fig. 42(a) shows the results of the HMM likelihood for the spectral component, while Fig. 42(b) shows those of the GV likelihood for the spectral component. Fig. 43(a) and Fig. 43(b) show the results for the F_0 component. Moreover, Fig. 44 shows the size of the decision trees used in our initialization method, and Fig. 45 shows the resulting U/V error rate. From Fig. 42(a), we can see that the HMM likelihood of "Proposed" very slightly increases as the parameter a_{MDL} decreases from 1.0 to 0.5, and it rapidly decreases as the parameter a_{MDL} decreases in the spectral components. We can see that the HMM likelihood at $a_{\mathrm{MDL}} = 0.5$ is almost the same as that of "Clus" but it is significantly lower than that of "Target." The result for the F_0 component shown in Fig. 43(a) are similar, except that no peaks appear as the parameter a_{MDL} decreases. On the other hand, Fig. 42(b) indicates the GV likelihood of "Proposed" rapidly increases as the parameter a_{MDL} decreases, and its value at $a_{\mathrm{MDL}} = 0.1$ is higher than that of "Target" in the spectral component. Regarding the F_0 component, the GV likelihood of "Proposed" rapidly increases as the parameter a_{MDL} decreases from 1.0 to 0.6, and it rapidly decreases beyond 0.6. Moreover from Fig. 45, we can see that the U/V error rate increases as the parameter a_{MDL} decreases. From these results, it is cleared that the HMM likelihood and GV likelihood change as a result of having a tree whose is size controlled by the parameter a_{MDL}.

3.7.8 Subjective evaluation of initialization method

Two preference tests (AB test) by seven listeners were conducted in the same manner as in **Section 3.7.2**. The synthetic speech was generated from the rich context models by using 1) "Clus", 2) "Proposed ($a_{\mathrm{MDL}} = 0.1$)", 3) "Proposed ($a_{\mathrm{MDL}} = 0.5$)", and 4) "Target" as the initial parameter for the spectral parameter. For the F_0 component, they were 1) Conv: speech generated from conventional clustered model and generated from the rich context models with using 2) Clus, 3) Proposed ($a_{\mathrm{MDL}} = 0.6$), and 4) Target as the initial parameter.

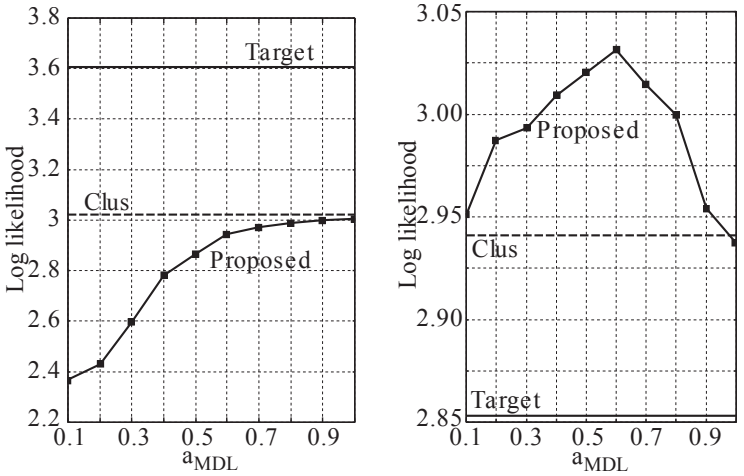
The results of the preference test as to the spectral component are shown in Fig. 46(a), and those for the F_0 component are shown in Fig. 46(b). From Fig. 46(a), we can see that our initialization method significantly improves speech quality over that of the conventional initialization method "Clus." We can also see that the score of "Proposed ($a_{\mathrm{MDL}} = 0.1$)" is higher than that of "Proposed



(a) HMM likelihood for natural parameters

(b) GV likelihood for generated parameters

Figure 42. Likelihoods used to tune the tree size for initialization of spectral parameters.



(a) HMM likelihood for natural parameters

(b) GV likelihood for generated parameters

Figure 43. Likelihoods used to tune the tree size for initialization pf F_0 contours.

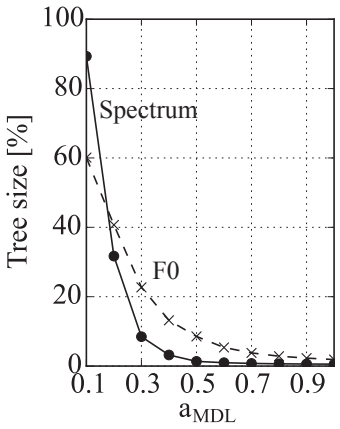


Figure 44. Size of the decision trees for initialization.

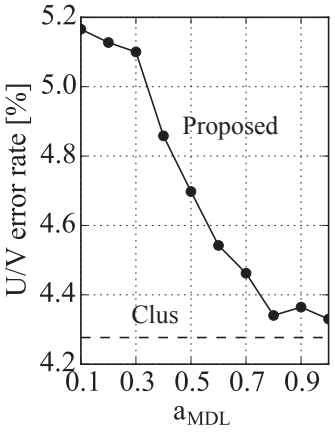


Figure 45. Error rates of the unvoiced/voiced decision using various tree sizes for initialization.

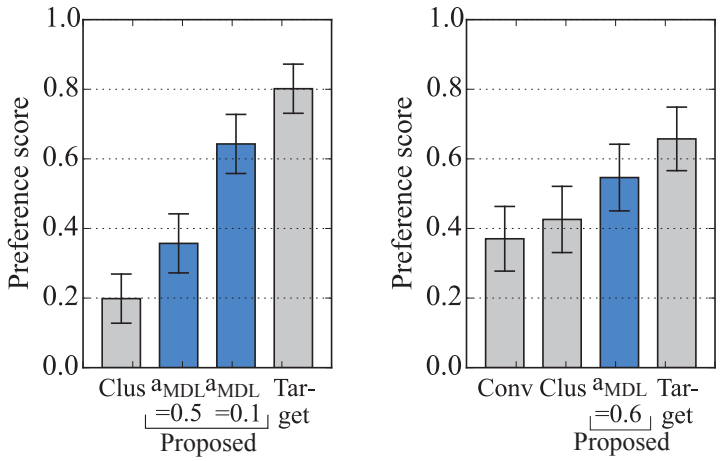
($a_{\mathrm{MDL}} = 0.5$)." This tendency is the same as the one for the GV likelihood shown in Fig. 42(b). From Fig. 45 and Fig. 46(b)(b), although the setting of the parameter a_{MDL} to maximize the GV likelihood slightly increases the U/V error rate, it still improves speech quality, even in terms of the F_0 component.

3.7.9 Evaluation in full synthesis

To investigate the effectiveness of all of proposed methods, we evaluated five kinds of synthetic speech listed in Table 1. A preference test (AB test) on speech quality was conducted by eight listeners in the same manner as in **Section 3.7.2**. Note that "Target" was generated by parameter generation with rich context models using the natural speech parameter sequence as the initial parameters.

Next, we investigated the effectiveness of methods considering the GV. Here, the spectral and F_0 sequences were generated considering the GV. A preference test (AB test) on speech quality was conducted by seven listeners in the same manner as described in **Section 3.7.2**.

Fig. 47(a) shows the result of the preference test in full synthesis, and Fig. 48 and Fig. 49 show the spectrograms and the F_0 contours of "Conventional," "Proposed," and natural speech. It is clear that applying the proposed method yields a larger improvement in the spectral component than in the F_0 component. Moreover, applying it to both the spectral and F_0 components ("PP") improves the speech quality to the point that it is close to the target ("TT"). From this result, we can see that the proposed parameter generation with rich context models for the spectral and F_0 components improves the quality of synthetic



(a) Spectral parameters (b) F_0

Figure 46. Preference scores on speech quality with 95% confidence interval for investigating the effectiveness of the initialization method. Our initialization method improves the quality the most for both the spectral and F_0 components.

speech.

Fig. 47(b) shows the results of the preference test in full synthesis considering the GV. Here, our method improves speech quality for the F_0 component even when considering the GV. We can also see that the score of "PC (GV)" is lower than that of "CC (GV)." Thus, although most of the discontinuous transitions of the initial parameter are alleviated, some of them are slightly emphasized as a result of considering the GV, and this causes a quality degradation in the synthetic speech. Since the tree for the spectral component is larger than that for the F_0 component, the over-emphasis effect affects the spectral parameter.

3.8 Experimental evaluation in GMM-based VC

3.8.1 Experimental conditions

We selected 450 parallel sentences of subsets A-through-I from the 503 phonetically balanced sentences included in the ATR Japanese speech database [126] for training, and the 53 sentences of subset J for evaluation. We trained female-to-male GMMs. Speech signals were sampled at 16 kHz. The shift length was set to 5 ms. The 0th-through-24th mel-cepstral coefficients were extracted as spectral parameters and log-scaled F_0 and five-band aperiodicity [74, 127] were extracted as excitation parameters. The STRAIGHT analysis-synthesis system [10] was

Table 1. Synthetic speech samples used for full synthesis evaluation using rich context models in HMM-based TTS.

Method	Spectrum	F_0
CC	Conventional	Conventional
CP	Conventional	Proposed ($a_{\mathrm{MDL}} = 0.6$)
PC	Proposed ($a_{\mathrm{MDL}} = 0.1$)	Conventional
PP	Proposed ($a_{\mathrm{MDL}} = 0.1$)	Proposed ($a_{\mathrm{MDL}} = 0.6$)
TT	Target	Target

used for parameter extraction and waveform generation. The feature vector consisted of spectral and excitation parameters and their delta and features. We built a 128-mixture GMM for spectral parameter conversion and a 16-mixture GMM for band-aperiodicity conversion. Our method was applied to the spectral parameters. The log-scaled F_0 was linearly converted. The band-aperiodicity was converted using the conventional GMM. The total number of rich context models was 590,745. In the parameter generation, we selected the 128-best candidates for each frame. GV [9] was not considered in speech parameter generation.

We compared the following approaches:

Cnv: conventional GMM-based VC[18]

Pro: our approach using rich context models

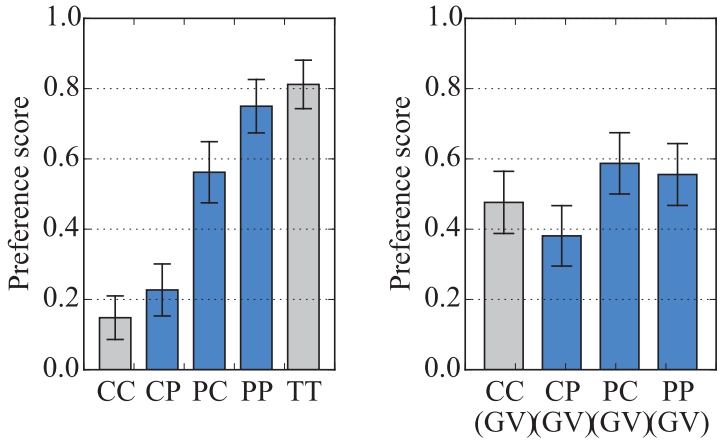
Tar: rich context models selected by reference data

In the initialization for "Tar," the best rich context models were selected by using the target reference speech parameters. We first calculated the misclassification rate for the training data to see the effect of discriminative training [119]. Then, after determining the number of over-trained models for initialization, subjective evaluations were conducted to confirm effectiveness of our method.

3.8.2 Effect of discriminative training

We evaluated the effect of the discriminative training done after the conventional joint density model training. The misclassification error rates were calculated for the training algorithms. The error rate was calculated as the number of misclassified training data divided by the number of the training data. Here "misclassified data" indicates the joint speech feature that the mixture component determined with Eq. (103) is different from the one determined with Eq. (52).

[18] The discriminative training [119] was performed.



(a) w/o GV (b) w/ GV

Figure 47. Preference scores on speech quality with 95% confidence intervals for full synthesis in HMM-based TTS. When the GV is not considered, the proposed method for the spectral and F_0 components matches the target in quality. However, the results deteriorate when the GV is considered.

The error rates are shown in Fig. 50. The 1.4% reduction in error rates means that discriminative training [119] makes it possible to select better rich context models.

3.8.3 The number of the over-trained models

We calculated the GV likelihoods[19] for the generated speech parameters in order to determine the number of the over-trained models. In each sub-region, we increased the number with the LBG algorithm until we could not estimate the model parameters. Although we can change the number sub-region by sub-region, the number was the same among the sub-regions[20].

The GV likelihood is shown in Fig. 51. We can see that the GV likelihood of "Pro" is the biggest around the compression ratio of 0.6 (3616 over-trained models). Therefore, we determine the number of over-trained models to be 3616.

[19] We did not calculate the GMM likelihood because we have demonstrated that the number of over-trained models is determined by the GV likelihood rather than by the HMM likelihood in HMM-based TTS.

[20] Except we cannot estimate the GMM parameters of the sub-region.

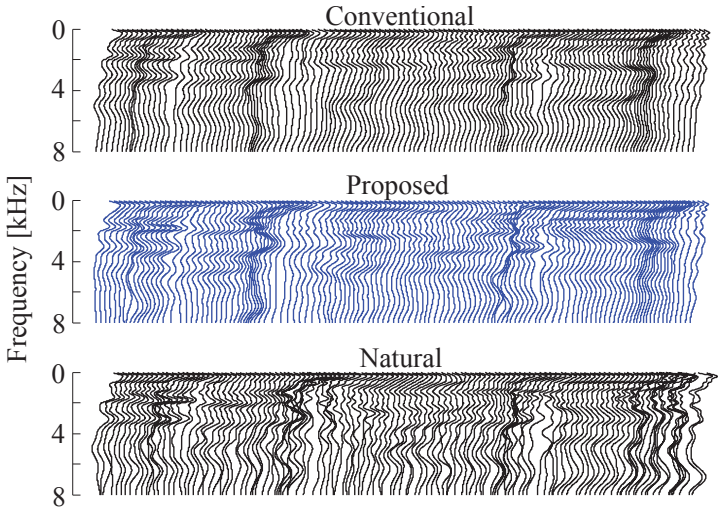


Figure 48. Example of spectrograms of synthetic speech. "Natural" represents the spectrograms of natural speech.

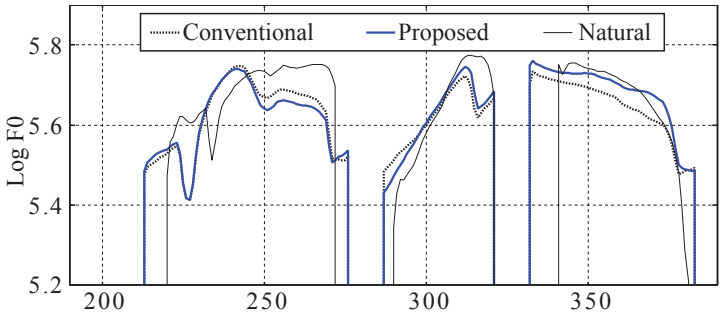


Figure 49. Example of F_0 contours of synthetic speech for the sentence fragment "sorewa taitei." "Natural" represents the spectrograms of natural speech.

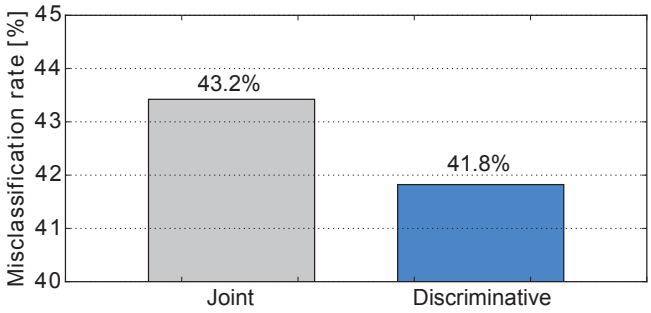


Figure 50. Misclassification rate for the training data to confirm the effect of the discriminative GMM training.

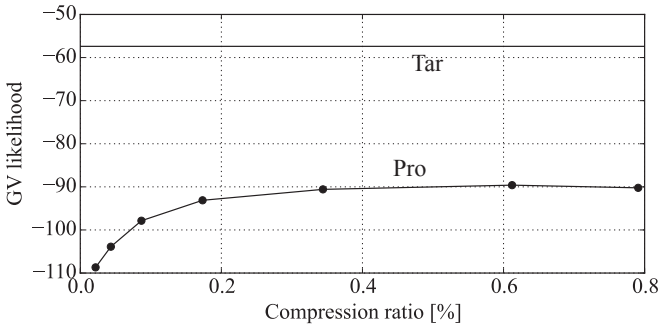
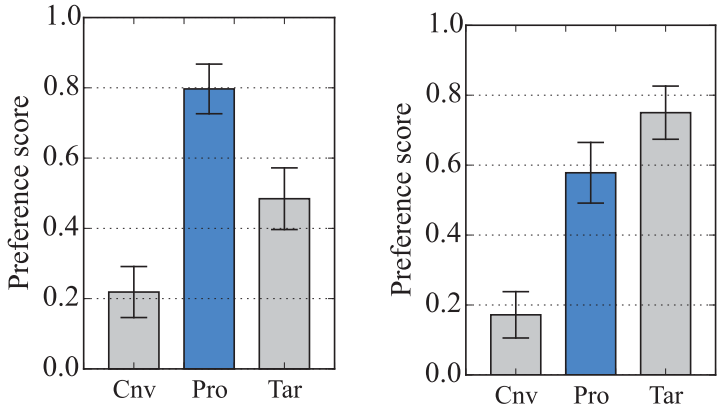


Figure 51. GV likelihoods for the finally generated speech parameter sequence. The compression ratio (x-axis) is the number of over-trained models divided by the number of training data.

3.8.4 Evaluation in speech quality and speaker individuality

A preference test (AB test) was conducted in the perceptual evaluation. We presented every pair of generated speech of the three algorithms in random order, and we made the listeners to select the better-quality speech sample. Similarly, an XAB test on speaker individuality was conducted using the analysis-synthesized speech as reference "X." Eight listeners participated in each evaluation.

The results of the preference tests are shown in Fig. 52(a) and Fig. 52(b). We can see that our method achieves higher scores in both speech quality and speaker individuality compared with the conventional GMM-based VC ("Cnv"). This demonstrates the effectiveness of our method. The score of "Tar" is lower than that of "Pro" in speech quality. We found some speech samples of "Tar" sound discontinuous, and it is expected that small training data size caused this phenomenon. Whereas "Pro" can alleviate the discontinuity by using slightly averaged initial speech parameters, "Target" uses non-averaged initial speech



(a) Speech quality (b) Speaker individuality

Figure 52. Preference scores with 95% confidence intervals for examining the effectiveness of rich context modeling for GMM-based VC.

parameters[21]. We expect that this degradation using non-averaged parameters can be avoided by increasing the size of the training data[22]. An alternative solution is to perform the iterative generation after the "Tar" initialization, but this is not an aim of this study.

3.9 Summary of this chapter

This chapter described statistical sample-based speech synthesis using rich context models to address the problem of inaccurate modeling causing quality degradation.

Section 3.2: We applied the rich context modeling originally proposed in HMM-based TTS [18] to GMM-based VC. The rich context models were trained for each joint speech feature vector belonging to each GMM mixture component.

Section 3.3: The rich context models belonging to one sub-region were gathered to construct the R-GMM in both HMM-based TTS and GMM-based VC. The mixing weights were tied instead of using the ML estimates.

[21] **Section 3.7.1** and this section used the same amount of training data, but **Section 3.7.1** did not show such a result. We think that this is because the rich context models used in **Section 3.7.1** are temporally averaged, but those of this work are not.

[22] This solution is known in unit selection synthesis.

Section 3.4: We proposed two ML-based speech parameter generation methods; the use of EM algorithm and approximation with single Gaussian distributions. The speech parameter generation with the hidden variable [109] have been utilized in the EM-based generation, and the likelihood is approximated with the single Gaussian distributions having the highest posterior probability.

Section 3.5: The iterative generation algorithms were initialized using a less-smoothed parameter sequence. Each subspace, which corresponds to one leaf node in HMM-based TTS and one mixture component in GMM-based VC, was further divided, and over-trained acoustic models were built to fit the training data of each sub-region. The initial speech parameter sequence in synthesis is generated from the over-trained acoustic models.

Section 3.6: We compared our methods with the conventional approaches. Compared with basic HMM-based TTS and GMM-based VC, our methods produce higher-quality synthetic speech by modeling individual speech features. Moreover, compared with the conventional hybrid methods combining unit selection synthesis, our methods retain the flexibility of the basic HMM-based TTS and GMM-based VC because it doesn't have the constraints used in the conventional hybrid methods.

Section 3.7: We conducted several experiments to confirm the effectiveness of our methods in HMM-based TTS. The results demonstrated: (1) the use of an approximation with a single Gaussian component sequence yields synthetic speech higher in quality than that produced by the EM algorithm, (2) the state-based model selection yields quality improvements at the same level as the frame-based model selection, (3) the use of the initial parameters generated from the over-trained speech probability distributions is very effective at improving speech quality, and (4) our methods for spectral and F_0 components yield significant improvements in quality compared with the use of basic HMM-based TTS.

Section 3.8: We conducted experiments proving the effectiveness of our methods in GMM-based VC. In particular, our methods achieved better scores in speech quality and speaker individuality in comparison with basic GMM-based VC.

Chapter

4

Modulation spectrum-based post-filter

4.1 Introduction

The Global Variance (GV) [19, 9] in **Section 2.9** is a well-known example to capture the over-smoothing effect. However, the use of this metric in the parameter generation tends to additionally generate artificial sounds [128, 21] and the quality gap between natural and synthetic speech is still large.

In this chapter, we first propose a new feature more sensitively correlated to the over-smoothing effect than the GV, the Modulation Spectrum (MS). The MS of a speech parameter sequence is given as the power spectrum of the sequence. The linear-scaled MS is a second order moments of the parameter sequence as the same as the GV, and can be regarded as a mathematical extension of the GV. The effectiveness of the MS in capturing speech properties has been noted in other research areas, such as spectral cues of speech perception [129], the use as acoustic features in HMM-based speech recognition [130] and acoustic signal classification [131], and as a counter-measure to discriminate synthetic speech from natural speech in speaker verification [132]. Related to the perceptual effect, [133, 134] investigated the effect of the MS (especially, lower modulation frequency band) on the perceptual intelligibility. Because generated speech parameter sequences tend to be temporally smoothed by the statistical generation process, the MS of synthetic speech tends to be degraded compared to that of natural speech. This MS degradation is still observed even when GV is used in parameter generation.

Furthermore, we also propose the post-filtering processes based on the MS. As we described in **Section 2.3**, the post-filtering process is very simple, portable, and effective approach to alleviate the over-smoothing effect, and it is done between speech parameter generation and waveform generation, as shown in Fig. 53. The post-filtering approach proposed in this chapter remedies the over-smoothing problem by modifying the generated speech parameter sequence so that its MS becomes more similar to that of natural speech. The proposed post-filter modifies the MS utterance by utterance and can be automatically constructed using natural speech and synthetic speech as training data. This utterance-level post-filter is further extended to a segment-level post-filter to modify the MS segment by segment in order to achieve low-delay parameter generation [135, 96].

In the experimental evaluation, we first evaluate the proposed post-filters in HMM-based TTS [8] from various perspectives. Then, we evaluate them in other speech synthesizers to confirm the high portability of the MS-based post-filters: the utterance-level post-filter in GMM-based VC [9] and the segment-level post-filter in CART-based TTS (a.k.a., CLUSTERGEN) [32].

The rest of this chapter is organized as follows, and shown in Fig. 54. In **Section 4.2**, We define the MS in this section, and we analyze the difference between natural and generated speech parameters, including spectral parameters, F_0 contours, and HMM-state duration sequences. The MS-based post-filters are proposed in **Section 4.3** and **Section 4.4**. We first describe the basic process

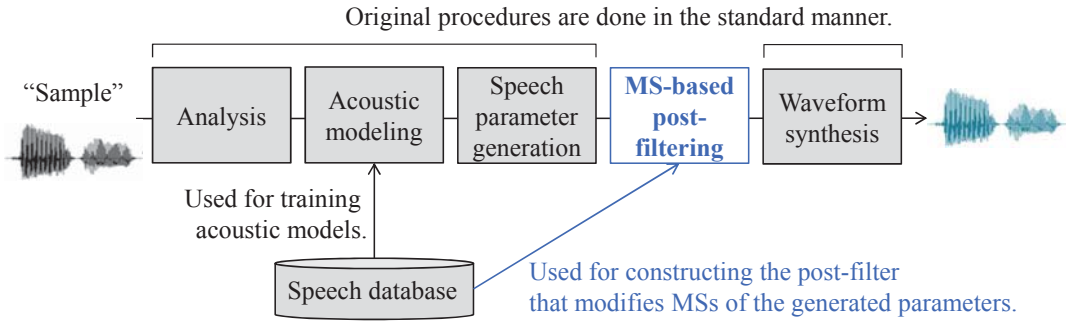


Figure 53. The proposed MS-based post-filter added in statistical parametric speech synthesis procedures. The post-filter is automatically constructed using speech database, and its process is independent on the original speech synthesis procedure.

of the MS-based post-filter, which is performed utterance by utterance. Then, the utterance-level post-filtering process is further modified into the segment-level process. In **Section 4.5**, We discuss about several terms, such as (1) the relationship between the conventional GV and the proposed MS, and (2) intuitive understandings of the effect of the MS-based post-filter. **Section 4.6** and **Section 4.7** are the experiments and summary.

4.2 Modulation Spectrum (MS) analysis

4.2.1 MS definition

Roughly speaking, in the HMM-based TTS framework, the context-dependent HMM averages the corresponding natural speech parameters in the training stage, and then outputs the averaged parameters in the synthesis stage. In practice, this averaging has a similar effect to low-pass filtering applied to the speech parameter sequence. Therefore, we expect that frequency characteristics of the speech parameters can measure the difference between natural and generated speech parameter sequences. In this chapter, we focus on the MS as a such quantitative measure of these frequency characteristics.

Though the MS is traditionally defined as a value calculated using the Fourier transform of the parameter sequence [136], this chapter defines the MS as its log-scaled power spectrum. The temporal fluctuation of the parameter sequence is modeled as power values of individual modulation frequency components of the parameter sequence. The MS $\boldsymbol{s}(\boldsymbol{y})$ of the parameter sequence $\boldsymbol{y}$ is calculated as:

$$\boldsymbol{s}(\boldsymbol{y}) = \left[\boldsymbol{s}(1)^\top, \cdots, \boldsymbol{s}(d)^\top, \cdots, \boldsymbol{s}(D)^\top\right]^\top, \tag{132}$$

$$\boldsymbol{s}(d) = \left[s_d(0), \cdots, s_d(f), \cdots, s_d(D_{\mathrm{s}})\right]^\top, \tag{133}$$

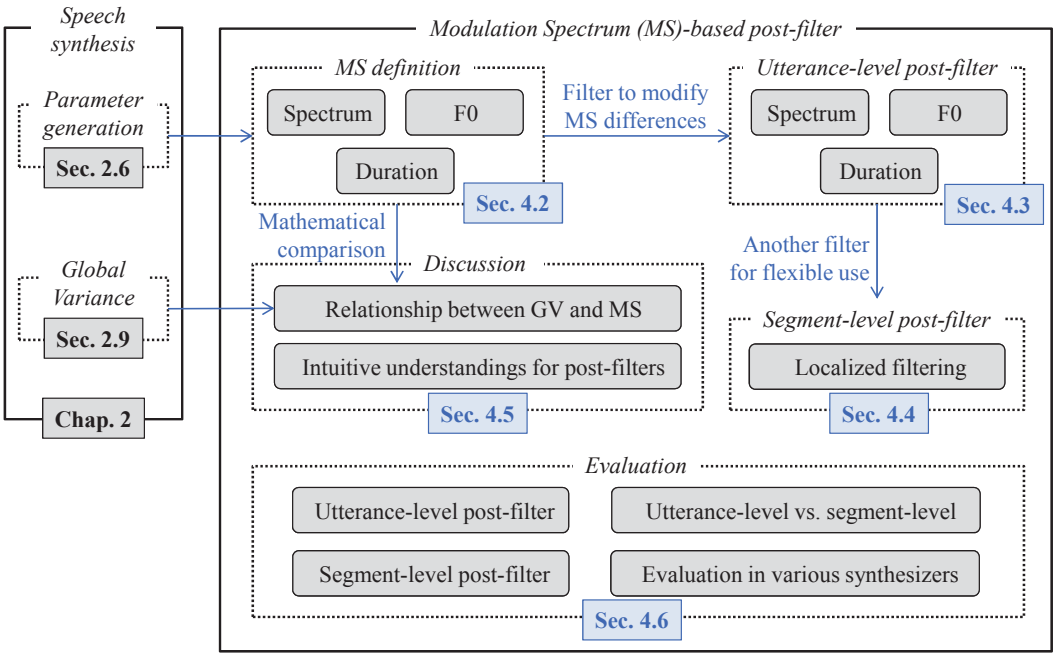


Figure 54. The rest of Chapter 4.

$$s_d(f) = \log\left(\left(\sum_{t=1}^{T} y_t(d)\cos mt\right)^2 + \left(\sum_{t=1}^{T} y_t(d)\sin mt\right)^2\right), \qquad (134)$$

where f is a modulation frequency index, $m = -\pi f / D_{\mathrm{s}}$ is a modulation frequency, and D_{s} is one half of the Discrete Fourier Transform (DFT) length. The MS is calculated from zero-padded parameter sequences so its length is $2D_{\mathrm{s}}$. As shown in Fig. 55, $\boldsymbol{s}(\boldsymbol{y})$ is given as a super vector consisting of the MSs corresponding to individual feature dimensions.

4.2.2 Over-smoothing effect quantified by MS

To demonstrate how the MS allows us to capture relevant frequency characteristics, we first demonstrate some characteristics of the MS of natural and synthetic speech. Figure 56 shows the MS mean of the mel-cepstral coefficient sequences generated using Eq. (54) ("HMM") and Eq. (97) ("HMM+GV") in HMM-based TTS. Additionally, the MS mean of a natural speech parameter sequence ("natural") is shown in the same figure for comparison. It can be observed that the MS of "HMM" is markedly degraded compared to that of "natural." This is because temporal fluctuation observed in the natural speech parameter sequences is lost in the HMM framework. We can also find that the MS of "HMM+GV" is closer to natural speech in lower modulation frequency bands but there is still a large

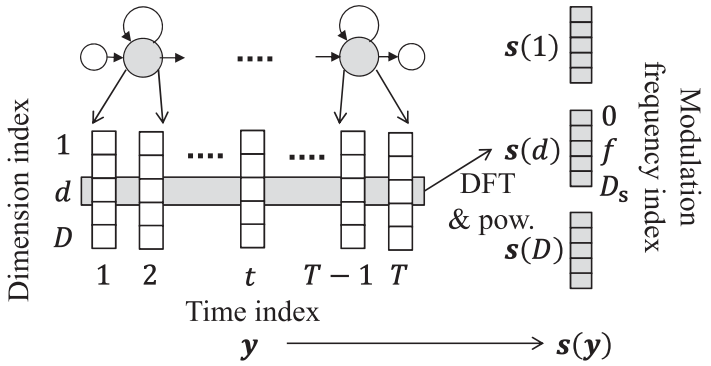


Figure 55. Graphic representation of how to derive the MS $\boldsymbol{s}(\boldsymbol{y})$ from the speech parameter sequence $\boldsymbol{y}$. Note that a zero-padding process is skipped in this figure.

gap between the MSs of "HMM+GV" and "natural speech" in higher modulation frequency bands (more than 10 Hz). From these results, we can expect that further quality improvements will be yielded by compensating for these differences in the MS.

In addition, we consider the spectral tilt of the MS (defined as "MS tilt") which indicates the power difference between the lower and the higher modulation frequency components in Fig. 56. We can observe that the MS tilt of the natural mel-cepstrum tends to increase in the higher order mel-cepstral coefficients. On the other hand, the MS tilt of the generated mel-cepstrum is similar among different order mel-cepstral coefficients. Even when using the GV in the parameter generation "HMM+GV," the MS is just shifted and the MS tilt is not changed. These results show that the parameter generation process shown by Eq. (54) or Eq. (97) tends to constrain the MS tilt of the generated speech parameter sequence to be almost constant.

In addition to the cepstral coefficients, we can also calculate the MSs of the other features. as described in the following section. The MS of the F_0 contour shown in Fig. 57 is also degraded by the statistical process. Higher modulation frequency components of the generated MS are almost the same as those of natural speech, but lower components are slightly different. HMM-state duration determined by Eq. (51) is also affected by the over-smoothing effect due to the statistical averaging process implicit in conventional parameter generation, as in the spectrum and F_0 components [49, 112]. Figure 58 shows the MS mean of phoneme duration sequences. We can see that the generated MS is generally smaller than that of natural speech.

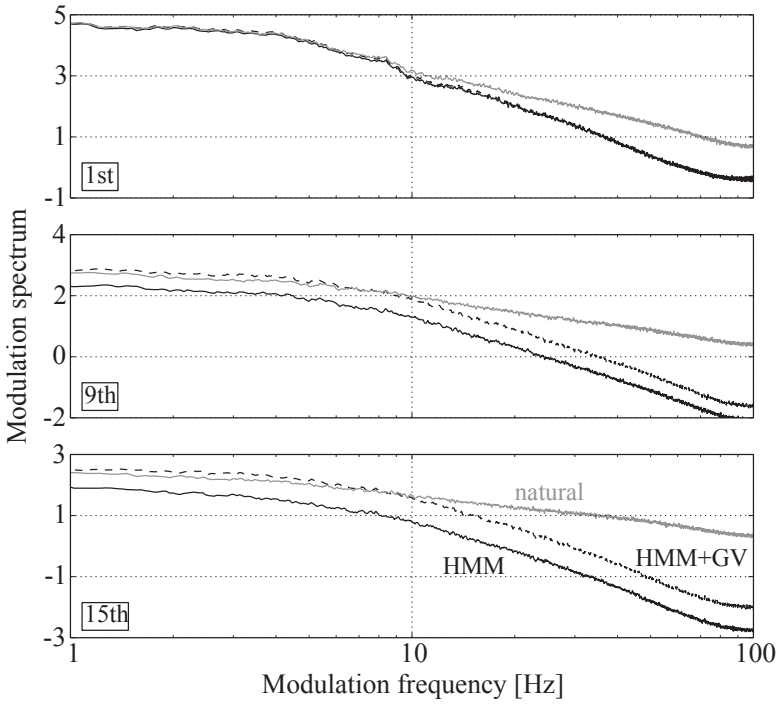


Figure 56. Averaged log-scaled MSs of the 1st, 9th and 15th mel-cepstral coefficient sequences from above in HMM-based TTS. Note that the modulation frequency (vertical axis) is in a log-scale. We didn't draw the MSs generated using the rich context models proposed in Chapter 3, but they are plotted in the middle between "HMM" and "natural."

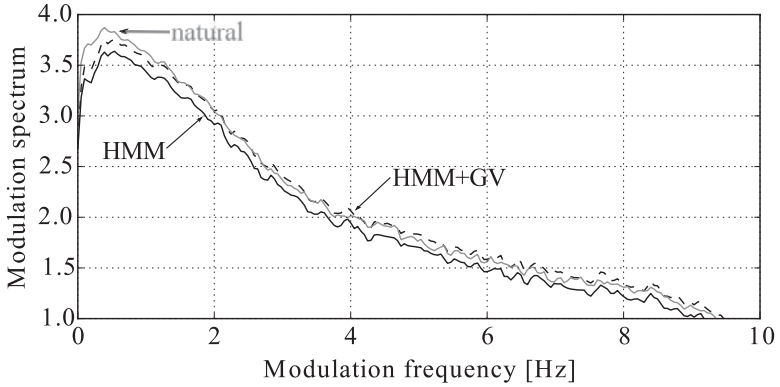


Figure 57. Averaged log-scaled MSs of the log-scaled F0 contours in HMM-based TTS. Note that the Nyquist frequency is 100 Hz similarly to the spectral parameters, but only < 10 Hz components are shown.

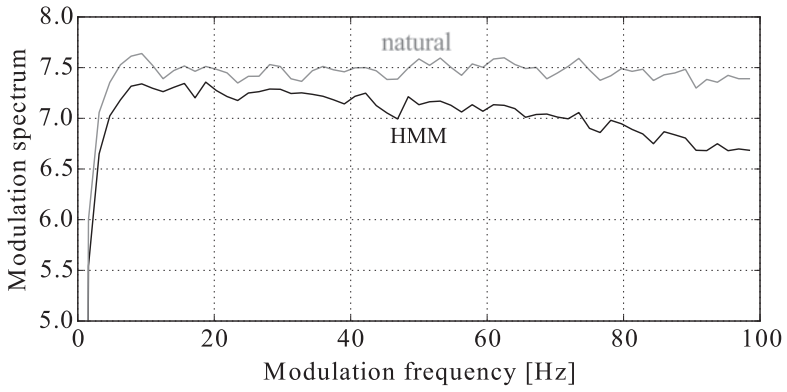


Figure 58. Averaged log-scaled MSs of the phoneme-level duration in HMM-based TTS. Note that the pseudo Nyquist frequency is set to 100 Hz because we cannot define the Nyquist frequency for duration.

4.3 Utterance-level post-filter

This section proposes post-filters to modify the MS of the generated parameter sequence. Figure 59 shows a schematic diagram of the proposed method. Parameters of the proposed post-filter are automatically trained using natural and generated speech parameter sequences in the training data. The speech parameters are generated by an individual speech synthesizer. First, the utterance-level MS-based post-filter is described for spectrum, F_0, and HMM-state duration. Then, the segment-level MS-based post-filter is derived by localizing the utterance-level post-filtering process.

4.3.1 Basic processes

The MS is calculated from a parameter sequence that is zero-padded to set its sequence length to $2D_{\mathrm{s}}$.

Training process The following probability distribution function is estimated from natural speech parameter sequences:

$$P\left(\boldsymbol{s}\left(\boldsymbol{y}\right)|\boldsymbol{\lambda}_s\right) = \mathcal{N}\left(\boldsymbol{s}\left(\boldsymbol{y}\right); \boldsymbol{\mu}_{\mathrm{s}}^{(\mathrm{N})}, \boldsymbol{\Sigma}_{\mathrm{s}}^{(\mathrm{N})}\right), \tag{135}$$

where $\boldsymbol{\mu}_{\mathrm{s}}^{(\mathrm{N})}$ and $\boldsymbol{\Sigma}_{\mathrm{s}}^{(\mathrm{N})}$ are the mean vector and the diagonal covariance matrix of the MS,

$$\boldsymbol{\mu}_{\mathrm{s}}^{(\mathrm{N})} = \left[{\boldsymbol{\mu}_1^{(\mathrm{N})}}^{\top}, \cdots, {\boldsymbol{\mu}_d^{(\mathrm{N})}}^{\top}, \cdots, {\boldsymbol{\mu}_D^{(\mathrm{N})}}^{\top}\right]^{\top}, \tag{136}$$

$$\boldsymbol{\Sigma}_{\mathrm{s}}^{(\mathrm{N})} = \mathrm{diag}\left[\boldsymbol{\Sigma}_1^{(\mathrm{N})}, \cdots, \boldsymbol{\Sigma}_d^{(\mathrm{N})}, \cdots, \boldsymbol{\Sigma}_D^{(\mathrm{N})}\right], \tag{137}$$

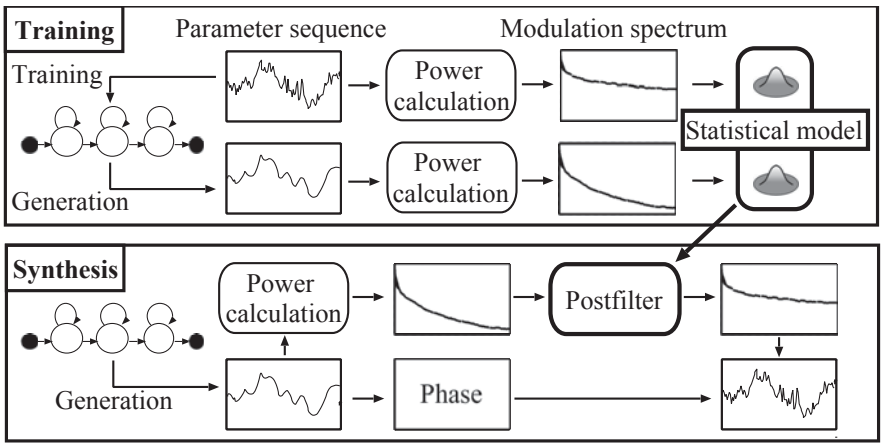


Figure 59. A schematic diagram of the proposed MS-based post-filter to modify the MS of the generated parameter sequence in the case of HMM-based TTS. When the post-filter is applied to GMM-based VC, the statistics of the generated MS are calculated using the speech parameters generated through the GMM-based conversion process.

$$\boldsymbol{\mu}_d^{(\mathrm{N})} = \left[\mu_{d,0}^{(\mathrm{N})}, \cdots, \mu_{d,f}^{(\mathrm{N})}, \cdots, \mu_{d,D_\mathrm{s}}^{(\mathrm{N})}\right]^\top, \tag{138}$$

$$\boldsymbol{\Sigma}_d^{(\mathrm{N})} = \mathrm{diag}\left[{\sigma_{d,0}^{(\mathrm{N})}}^2, \cdots, {\sigma_{d,f}^{(\mathrm{N})}}^2, \cdots, {\sigma_{d,D_\mathrm{s}}^{(\mathrm{N})}}^2\right], \tag{139}$$

where $\mu_{d,f}^{(\mathrm{N})}$ and ${\sigma_{d,f}^{(\mathrm{N})}}^2$ are the mean and the variance of $s_d(f)$, respectively. $\boldsymbol{\lambda}_s$ is the parameter set of the MS. $\mathcal{N}\left(\cdot; \boldsymbol{\mu}_\mathrm{s}^{(\mathrm{G})}, \boldsymbol{\Sigma}_\mathrm{s}^{(\mathrm{G})}\right)$ is also estimated in the same manner using the speech parameter sequences generated as described in Chapter 2. To avoid the effect of the duration difference between natural and generated speech parameter sequences in HMM-based TTS, the parameter sequence is generated using the natural speech duration. In the case of GMM-based VC, temporally-aligned input speech parameter sequence $\boldsymbol{X}$ is used to generated the speech parameter sequence $\hat{\boldsymbol{y}}_{\hat{\boldsymbol{q}}}$.

Synthesis process The following filter is applied to the generated speech parameter sequence $\hat{\boldsymbol{y}}_{\hat{\boldsymbol{q}}}$ (see Fig. 60.):

$$s'_d(f) = (1-k)s_d(f) + k\left[\frac{\sigma_{d,f}^{(\mathrm{N})}}{\sigma_{d,f}^{(\mathrm{G})}}\left(s_d(f) - \mu_{d,f}^{(\mathrm{G})}\right) + \mu_{d,f}^{(\mathrm{N})}\right], \tag{140}$$

where k is a post-filter emphasis coefficient between 0 and 1. If $k = 1$, the MS will be modified to be close to the MS of natural speech parameter sequences. On the other hand, if $k = 0$, the filtered sequence will be the same as the non-filtered

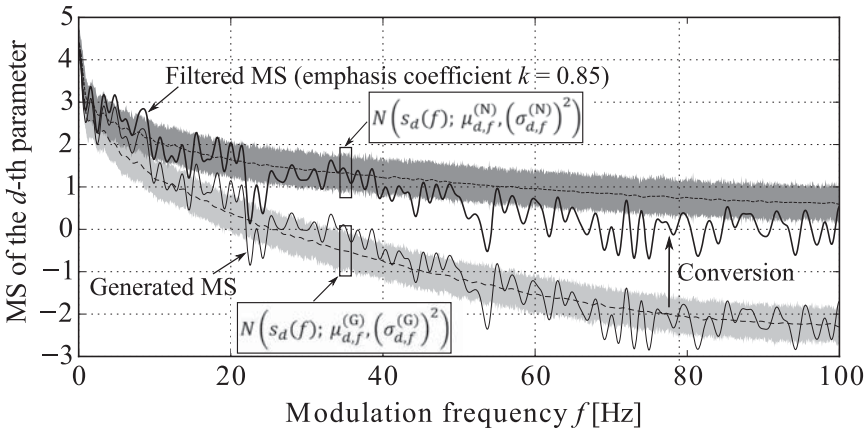


Figure 60. An example of the MS conversion in the synthesis stage. Note that the MS envelope ("Generated MS" and "Filtered MS") is drawn instead of the MS itself for clear illustration. The MS envelope is calculated by liftering the cepstrum of the MS.

sequence. The filtered parameter sequence is calculated from the modified MS and original phase characteristics of the parameter sequence before filtering. The detailed procedure is listed below,

1. Zero-pad the original parameter sequence.
2. Take the DFT and store the phase characteristics.
3. Calculate the log-scaled power spectrum (= MS).
4. Apply the post-filter to the MS.
5. Compute the power and add the original phase.
6. Take the inverse DFT.
7. Truncate the resulting signal to have an appropriate length.

4.3.2 Application to F0 contour

While the proposed post-filter can be directly applied to the spectral component, additional processing is required for its application to the F_0 component because observed F_0 contours are not a continuous sequence. To solve this problem, we use continuous F_0 modeling [86] which also estimates F_0 values at the unvoiced frames. Following [137], F_0 values of the unvoiced frames are estimated with spline-based interpolation. Because the effect of micro prosody on speech quality is small [53] but the effect on the MS is not negligible, we remove it with a Low Pass Filter (LPF). Moreover, the utterance-level F_0 mean is subtracted from original F_0 values before estimating continuous F_0 contours to avoid discontinuous transitions in the zero-padding process. These procedures are shown in Fig. 61.

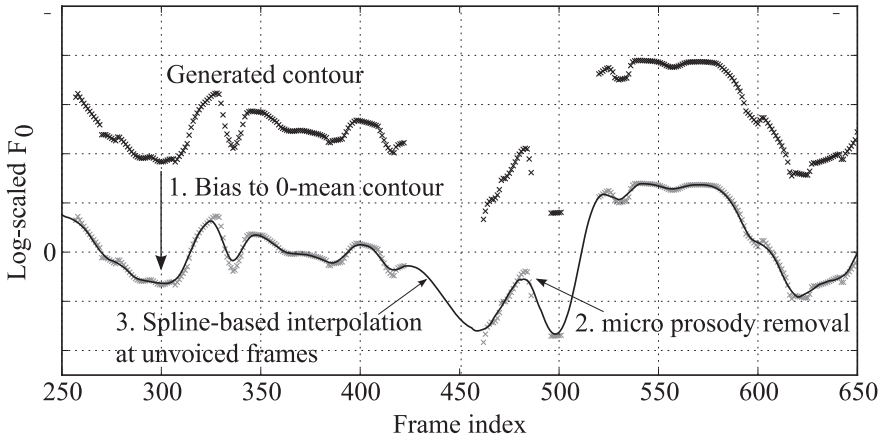


Figure 61. An illustration of the pre-processing procedures to calculate the continuous F_0 contour from the original F_0 contour. A low pass filter is used for removing the micro prosody.

Because spline-based methods are inappropriate for extrapolation, i.e., silence frames, we calculate the MS from the non-silence frames[23].

In the synthesis stage, the utterance-level mean and unvoiced/voiced regions of the generated F_0 contour are extracted before applying the proposed post-filter. First, the filtered continuous F_0 contour is calculated in the same manner as the spectral component. Then, the filtered F_0 contour is calculated by adding the mean to the filtered continuous F_0 contour and restoring the unvoiced/voiced regions.

4.3.3 Application to HMM-state duration

The proposed utterance-level post-filter modifies the MS of the phoneme-level duration calculated from the state-level duration determined by Eq. (51). The phoneme-level duration sequence is filtered after excluding silence and its mean value is normalized as with the F_0 parameters. After restoring the utterance-level mean, the phoneme-level duration is revised if it is smaller than the number of states of the phoneme HMM. Finally, the HMM-state duration is updated by maximizing the state duration while fixing the phoneme duration to the filtered values.

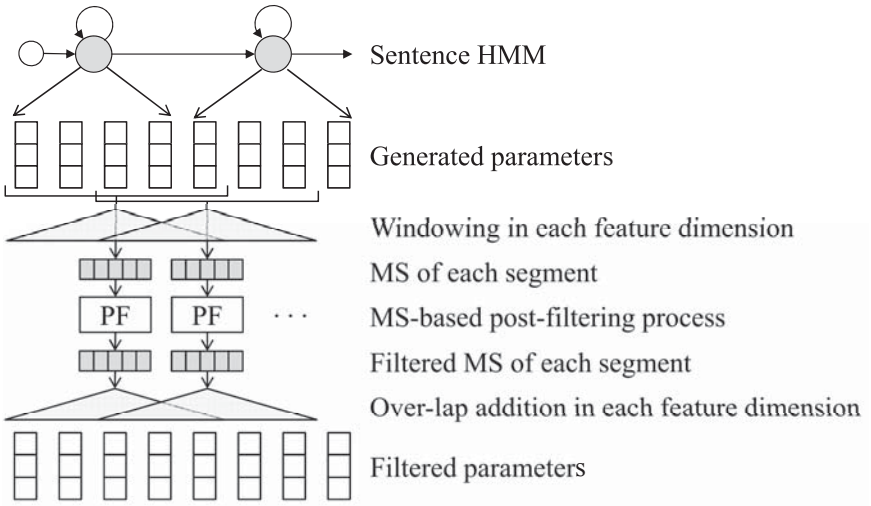


Figure 62. Procedures of the segment-level MS-based post-filter in HMM-based TTS. The window length and DFT length must be determined in this filtering process. The shift length is a half of the window length.

4.4 Segment-level post-filter

Because the proposed utterance-level MS-based post-filter calculates the MS utterance by utterance, the DFT length needs to be set large enough to cover various lengths of utterances. This MS calculation causes some problems: if the length of an utterance to be synthesized is longer than the previously determined DFT length, the MS can not be calculated accurately, and thus it is difficult to apply the utterance-level filtering process to a low-latency speech synthesis framework [135, 96] and a incremental speech synthesis framework [59, 60, 61] where frame-level or segment-level processing based on the recursive parameter generation is essential [7].

In order to handle these cases, we propose a segment-level post-filter that is effective on shorter segments. The segment-level post-filter is derived by localizing the post-filtering process as illustrated in Fig. 62. A part of the speech parameter sequence that is windowed by a triangular window with constant length is used as a segment to calculate the MS and its statistics. The window shift length is set to a half of the window length. The MS-based post-filtering process is performed segment by segment in the same manner as the trajectory-level post-filtering process. The filtered speech parameter sequence is generated by overlapping and adding the filtered segments. The Hanning window may also be used instead of

[23]We also considered simple approaches to estimate F_0 of silence such as the use of the utterance-level mean of F_0 or the use of the F_0 value in the nearest voiced frame. However, we have confirmed that the current method is better to model the MS.

the triangular window. Note that for the spectrum parameters, silence frames are removed in calculating the MS statistics to alleviate the over-fitting problem [14]. The segment-level post-filtering can be applicable to low-delay speech waveform generation. Moreover, it is possible to further implement context-dependent post-filtering.

4.5 Discussions

The proposed post-filters can be automatically constructed in a data-driven manner. Whereas conventional post-filtering processes [49, 114, 115, 116] requires the rule-based design [49], or manual tuning [114], the proposed post-filters enable automatic design and tuning.

Another data-driven approach is the post-filtering process to maintain the GV of the generated parameter sequence [22]. The generated speech parameters are linearly converted as follows:

$$\hat{y}_t(d) = \sqrt{\frac{\mu_d^{(\mathrm{GV,N})}}{\mu_d^{(\mathrm{GV,G})}}}\{y_t(d) - \langle y_t(d)\rangle\} + \langle y_t(d)\rangle, \tag{141}$$

where $\mu_d^{(\mathrm{GV,N})}$ and $\mu_d^{(\mathrm{GV,G})}$ are the GV mean of the d-th dimension of the natural and synthetic speech parameters in the training data, respectively, and $\langle y_t(d)\rangle$ is the mean of the d-th dimension of the synthetic speech parameters. In this method, since only the variance of the sequence is considered, the MS degradation is not completely recovered. Thus, temporal fluctuation of the generated speech parameters after filtering is still very different from that of natural speech. On the other hand, the proposed post-filters can recover this fluctuation because we directly consider the MS itself.

According to the Parseval's theorem, the power of a temporal sequence is preserved during a DFT. The GV defined in Eq. (94) represents the power of the sequence except the bias component. Because the utterance-level MS is defined as the power spectrum of the sequence, the sum of the MS over all modulation spectra except the bias component (frequency zero) is equal to the GV[24]. As the another interpretation, MS can be regarded as the frequency-domain GV as shown in Fig. 63. The temporal sequence are decomposed into the frequency components, and GV of one frequency component is given as one of the MS component.

In the GV-based post-filtering process, MSs of all modulation frequencies other than the bias are converted in the same way. Namely, the GV-based post-filtering process is a special case of the proposed MS-based post-filtering process

[24]Properly described, the sum of linear-scaled MS except the bias is equivalent to GV.

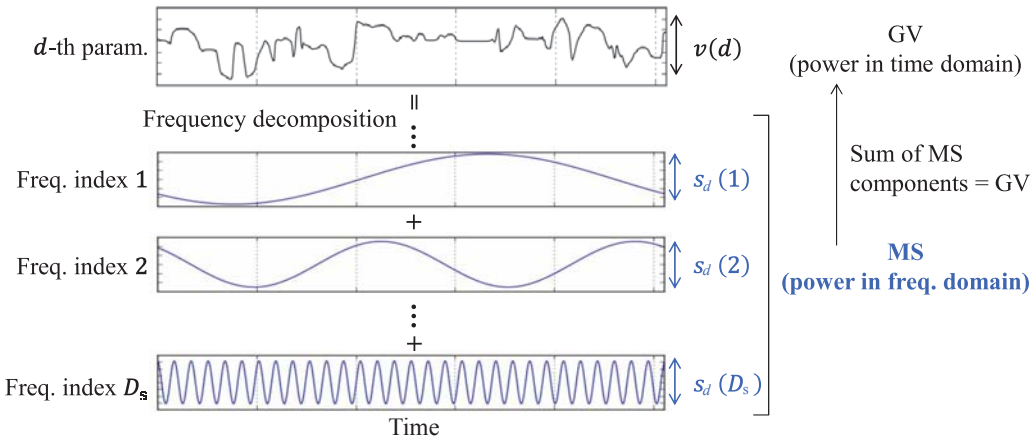


Figure 63. The relationship between Global Variance (GV) and Modulation Spectrum (MS). The MS can be regarded as the frequency-dependent GV.

under the following conditions:

$$\mu_{d,f}^{(\cdot)} = \log \mu^{(\mathrm{GV},\cdot)} \quad (f > 0), \tag{142}$$

$$\mu_{d,f}^{(\mathrm{N})} = \mu_{d,f}^{(\mathrm{G})} \quad (f = 0), \tag{143}$$

$$\sigma_{d,f}^{(\mathrm{N})} = \sigma_{d,f}^{(\mathrm{G})}, \tag{144}$$

in which the post-filter emphasis coefficient is set to 1. Namely, the GV-based post-filtering process only causes the unnatural MS shift as shown in Fig. 56[25]. On the other hand, the proposed methods can directly convert the MS components at individual modulation frequencies.

Figure 64 draws an example of the filtered/non-filtered mel-cepstral coefficient sequences. It is observed that the proposed post-filter generates the fluctuated parameter sequence, and the effect is larger in the higher order of the mel-cepstral coefficients. This is because the MS difference between natural and generated parameter sequences is larger in higher-order mel-cepstral coefficients as shown in Fig. 56. Similarly, Fig. 65 and Fig. 66 show the F_0 contour and duration. We can also find the fluctuated parameter sequences are generated by the proposed post-filter.

As the another implementation of the MS-based post-filters, we can also consider the use of frequency-delta MS as used in the GV [111] and the non-parametric MS modeling such as [140].

[25] In Fig. 56, the parameter generation algorithm considering the GV rather than the GV-based post-filter is used. Although it tends to make the GV of the generated speech parameter sequence almost equal to the GV mean $\boldsymbol{\mu}_{\mathrm{v}}$ [138, 139], it still causes only a MS shift in practical effect, although the amount of the MS shift changes utterance by utterance.

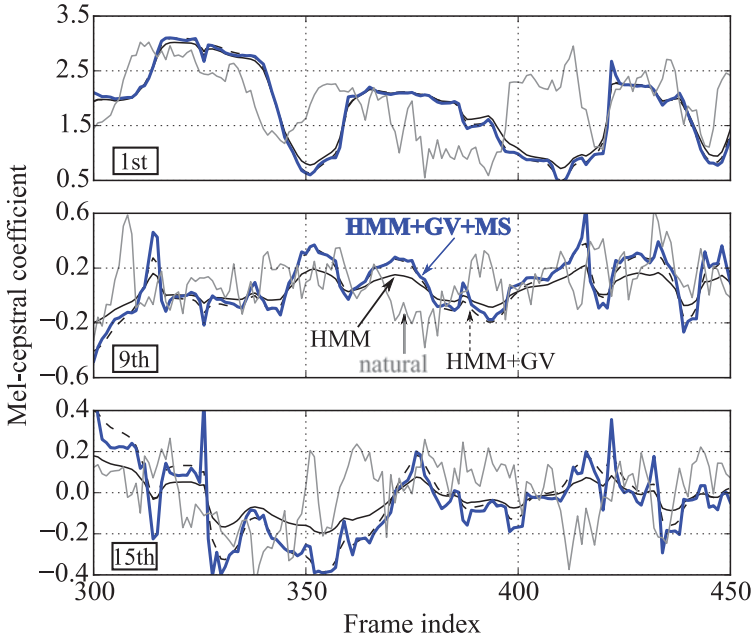


Figure 64. An example of the filtered and non-filtered 1st, 9th, and 15th mel-cepstral coefficient sequences from above in HMM-based TTS. We can see that the effect of the post-filter is larger in the higher order of the mel-cepstral coefficients.

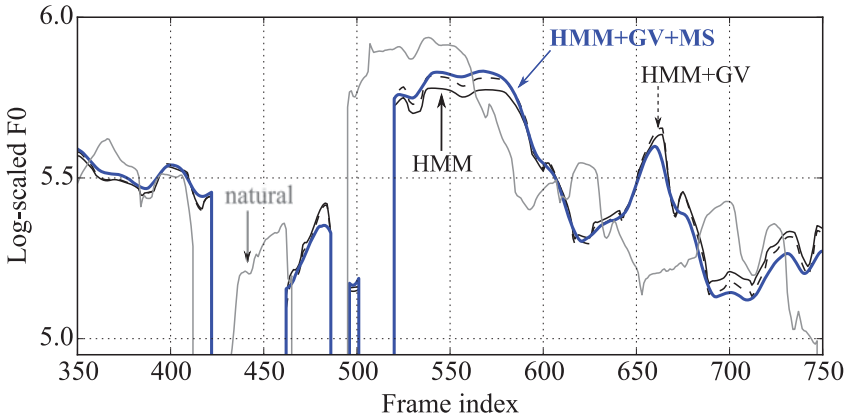


Figure 65. An example of the filtered and non-filtered F_0 contours in HMM-based TTS.

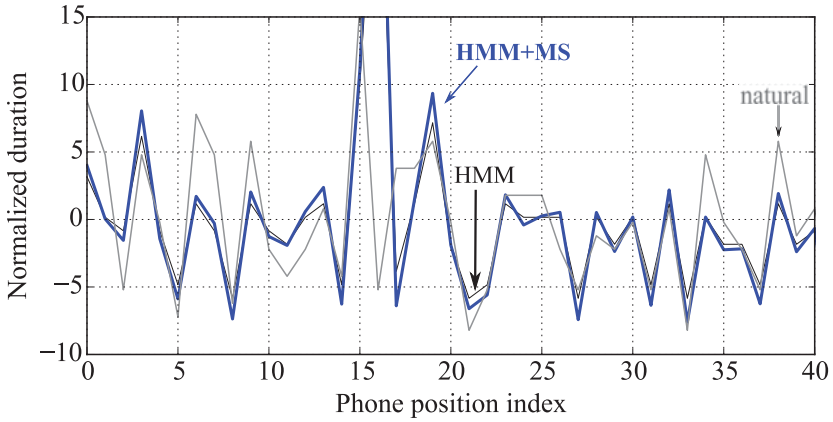


Figure 66. An example of the filtered and non-filtered phoneme-level duration in HMM-based TTS.

Note that although these fluctuated parameter sequences are effective for improving naturalness of synthetic speech, they sometimes make synthetic speech warbling. It is expected that this problem will be addressed by incorporating the MS metric into the metric for the parameter generation as done in the GV-based parameter generation [19, 9]. We will confirm it in the next chapter.

4.6 Experimental evaluation

First, we investigate the effects of the proposed utterance-level and segment-level post-filters from various perspectives in HMM-based TTS. Then, we evaluate them in other statistical parametric speech synthesis frameworks: the effect of the utterance-level post-filter in GMM-based VC and the effect of the segment-level post-filter in CLUSTERGEN.

4.6.1 Experimental conditions for evaluation in HMM-Based TTS

We trained a context-dependent phoneme Hidden Semi-Markov Model (HSMM) [141] for a Japanese female speaker for evaluation in HMM-based TTS. We used 450 sentences for training and 53 sentences for evaluation from the 503 phonetically balanced sentences included in the ATR Japanese speech database [24]. Speech signals were sampled at 16 kHz. The shift length was set to 5 ms. The 0th-through-24th mel-cepstral coefficients were extracted as spectral parameters and log-scaled F_0 and 5 band-aperiodicity [74, 127] were extracted as excitation parameters. The STRAIGHT analysis-synthesis system [10] was employed for parameter extraction and waveform generation. The feature vector consisted of spectral and excitation parameters and their delta and delta-delta features. Five-

state left-to-right HSMMs were used. The proposed post-filter was trained in a context-independent manner. A 10 Hz-cutoff LPF was used to remove the micro prosody from the continuous F_0 contours[26].

We conducted evaluation with the following systems:

HMM: The spectrum and F_0 are generated with Eq. (54), and the HMM-state duration is determined with Eq. (51).
HMM+MS: The proposed post-filter is applied to "HMM."
HMM+GV: The spectrum and F_0 are generated with Eq. (97).
HMM+GV+MS: The proposed post-filter is applied to "HMM+GV."

Note that the post-filter of "HMM+GV+MS" was trained using parameter sequences generated with the GV. The "HMM" system was used for the components that the proposed methods were not applied to. The post-filters were not applied to the aperiodicity component because there is no quality gain achieved by the post-filters[27].

4.6.2 Coefficient tuning for utterance-level post-filter

We investigate the effectiveness of the proposed utterance-level post-filter in HMM-based TTS. The filter emphasis coefficients for spectrum, F_0 and duration are first tuned by the likelihoods. The synthetic speech quality is then evaluated using the tuned emphasis coefficients. The DFT length to calculate MS ($= 2D_{\mathrm{s}}$) was set to 4096, which is over the maximum frame length in training and evaluation data.

Here, in order to determine the filter emphasis coefficients, we calculated the HMM likelihood, GV likelihood, and MS likelihood for filtered spectrum, F_0, and HMM-state duration for settings of the emphasis coefficient from 0 to 1. The duration likelihood was calculated instead of the HMM likelihood when tuning the coefficient for duration. For comparison, the likelihood for natural speech parameter sequences was calculated, which was labeled as "natural." Note that the HMM likelihood and the MS likelihood were normalized by the total number of frames T and one half of the DFT length D_{s}, respectively.

Figure 67 shows the likelihoods for the filtered spectral parameters. It is observed that the HMM likelihoods of "HMM+MS" and "HMM+GV+MS" decrease as the emphasis coefficient increases. Nevertheless, their values are always higher than that of "natural." In the GV likelihood, we can see that these likelihoods cross that of "natural speech" at $k = 0.85$. On the other hand, MS

[26] We evaluated training accuracy of MS likelihood for various cutoff frequencies, and confirmed that this setting was the best.

[27] The same tendency is reported in the parameter generation algorithm considering the GV [127].

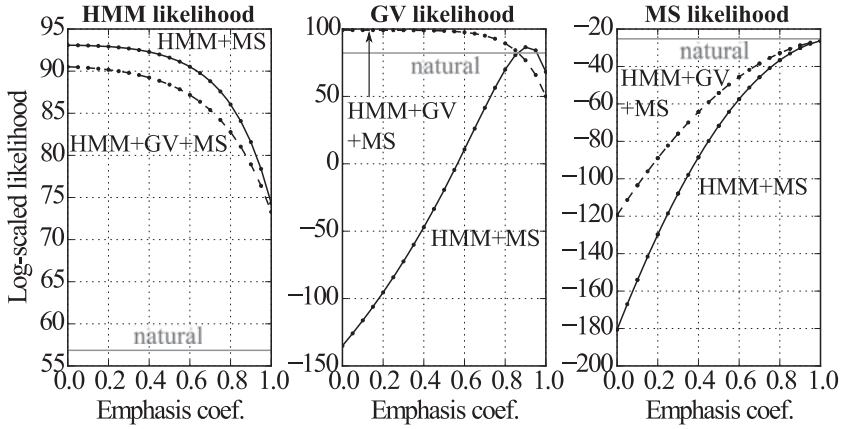


Figure 67. HMM, GV, and MS likelihoods for the spectral parameter sequences filtered by the proposed utterance-level post-filter in HMM-based TTS.

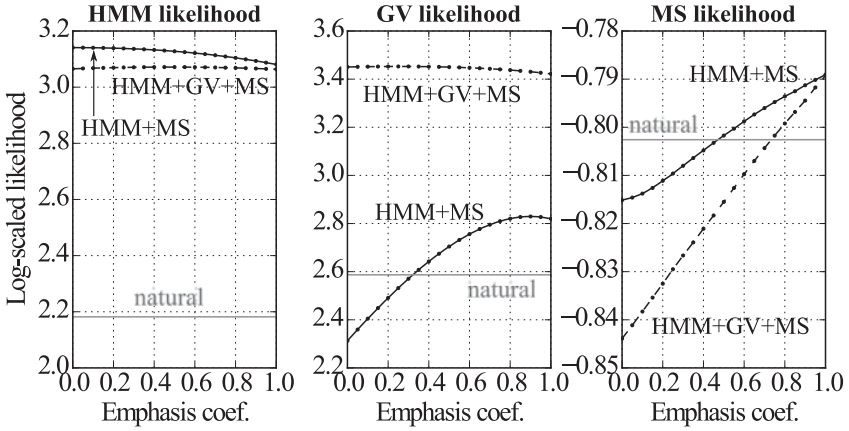


Figure 68. HMM, GV, and MS likelihoods for the F_0 contours filtered by the proposed utterance-level post-filter in HMM-based TTS.

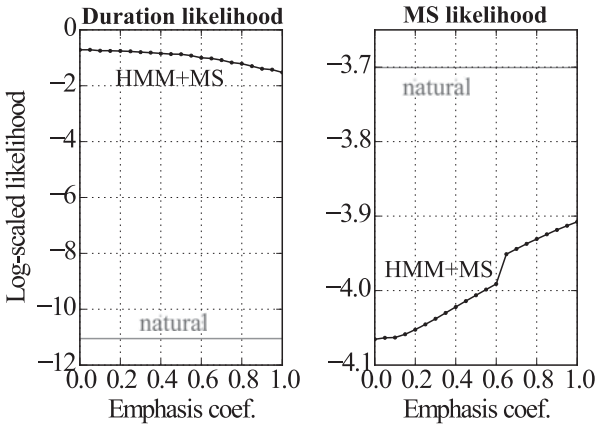


Figure 69. Duration and MS likelihoods for the phoneme-level duration sequences filtered by the proposed utterance-level post-filter in HMM-based TTS.

likelihoods increase as the coefficient increases but their values always lower than "natural speech." Considering these results, we determined the filter emphasis coefficient for spectral component to be 0.85.

Figure 68 shows the likelihoods for the filtered F_0 contour. The change of these likelihoods as the coefficient varies show the same tendency as those for the spectral components except the relation with the likelihoods of "natural speech." We can find that all likelihoods of "HMM+MS" and "HMM+GV+MS" are higher than "natural speech" when setting the emphasis coefficient over $k = 0.75$, and we can also find that the coefficient $k = 1.0$ is the highest point of MS likelihood. From these results, we set the coefficient to 1.0.

Figure 69 shows the likelihoods for the filtered phoneme-level duration. The tendency of the likelihood change is similar to those of the spectrum and F_0, and the MS likelihood is the highest at $k = 1.0$. Therefore, we set the coefficient $k = 1.0$. We can also see discontinuous transitions of the MS likelihood. We expect that this was caused by the effect of rounding the filtered duration values into integer values after filtering.

4.6.3 Subjective evaluation for utterance-level post-filter

To investigate whether or not quality improvements are yielded by applying the proposed post-filter to the spectrum, F_0, and duration components, we conducted a preference AB test on speech quality. Every pair of these types of synthetic speech was presented to listeners in random order. Listeners were asked which sample sounded better in terms of speech quality. Evaluation for spectrum, F_0, and duration was conducted by 8, 8, and 6 listeners, respectively.

Figure 70 shows the preference test for the spectrum, F_0, and duration. For

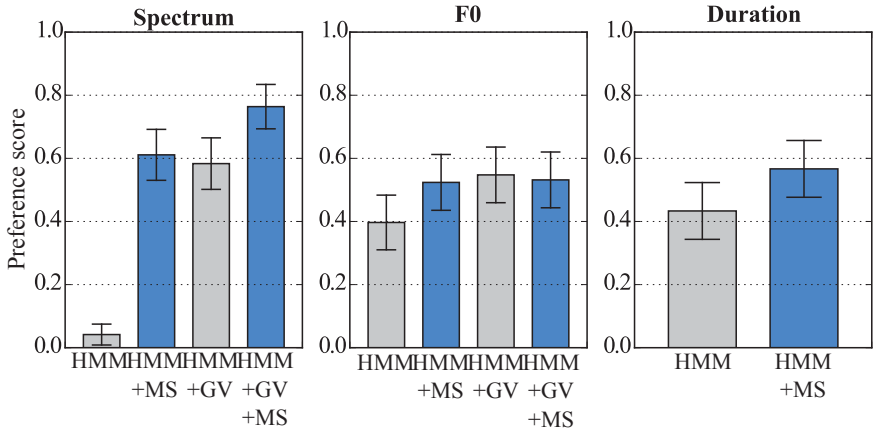


Figure 70. Preference scores on speech quality with 95% confidence interval for confirming the quality gain by the proposed utterance-level post-filter in HMM-based TTS.

spectrum, we can see that the score of the "HMM+MS" system dramatically increases over the "HMM" system, and achieves a similar score to the "HMM+GV" system. Additionally, further improvement can be observed by applying the proposed method to "HMM+GV." From these results, the effectiveness of the proposed method for the spectral component is confirmed. For F_0, "HMM+MS" and "HMM+GV+MS" achieve a better score than "HMM," but there are not additional gains over when GV is considered. The reason why the score differences among conventional and proposed methods are smaller than those in the spectral components is that the MS of the generated F_0 contours is quite close to that of the natural F_0 contours, as shown in Fig. 57, even if not applying the proposed post-filter. Finally, we can also see a slight improvement in quality for duration. These results demonstrate a quality gains by the proposed utterance-level post-filter for spectrum, F_0 and duration.

4.6.4 Coefficient tuning for segment-level post-filter

We evaluate the effectiveness of the segment-level post-filter in HMM-based TTS. The window length and window shift length were set to 125 ms (25 samples) and 60 ms (12 samples) [142]. A 64-taps DFT was used to calculate the MS. The tuning step and evaluation step were conducted in the same way as the evaluation of the proposed utterance-level post-filter. Note that the post-filter was not applied to the duration because we could not observe a large difference between filtered and non-filtered sequences.

The HMM likelihood, GV likelihood, and MS likelihood for the filtered spectral parameters and F_0 contours were calculated. The results are shown in Fig.

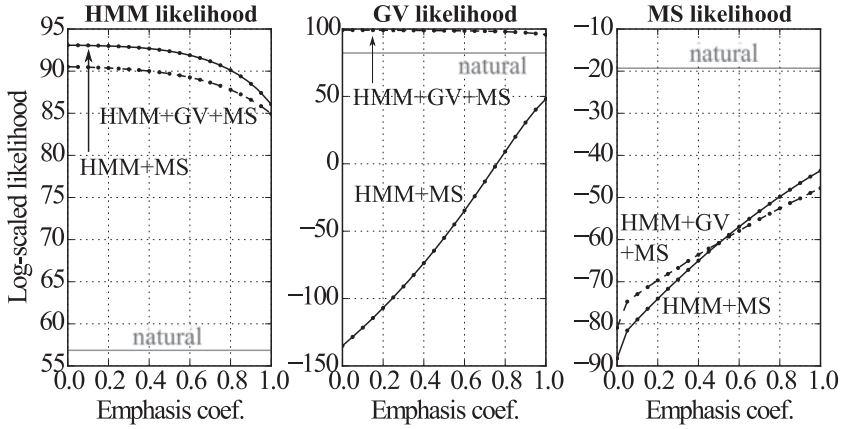


Figure 71. HMM, GV, and MS likelihoods for the spectral parameter sequences filtered by the proposed segment-level post-filter in HMM-based TTS.

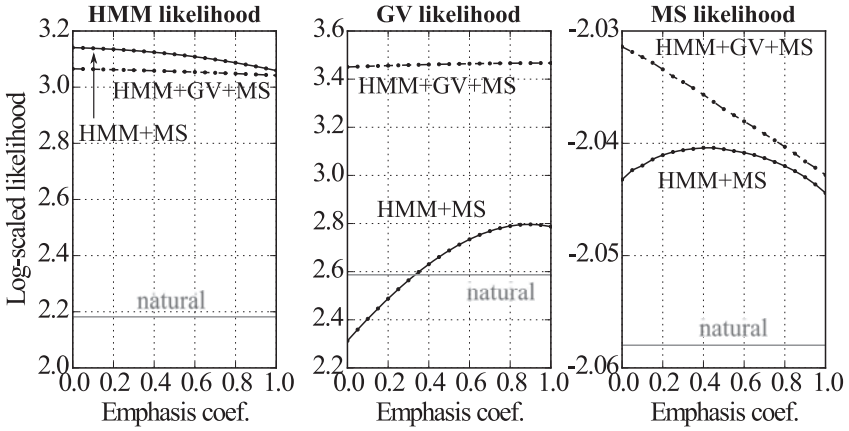


Figure 72. HMM, GV, and MS likelihoods for the F_0 contours filtered by the proposed segment-level post-filter in HMM-based TTS.

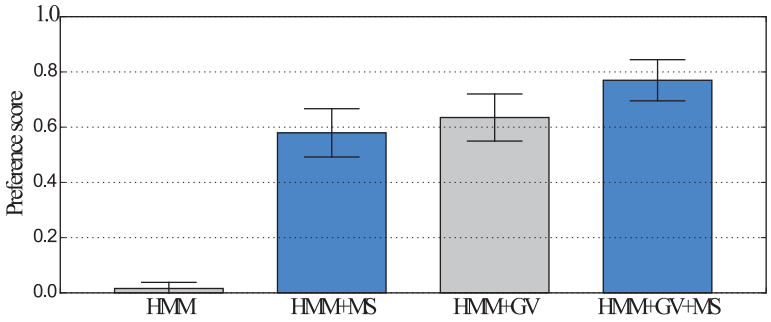


Figure 73. Preference scores on speech quality with 95% confidence interval for confirming the quality gain by the proposed segment-level post-filter in HMM-based TTS.

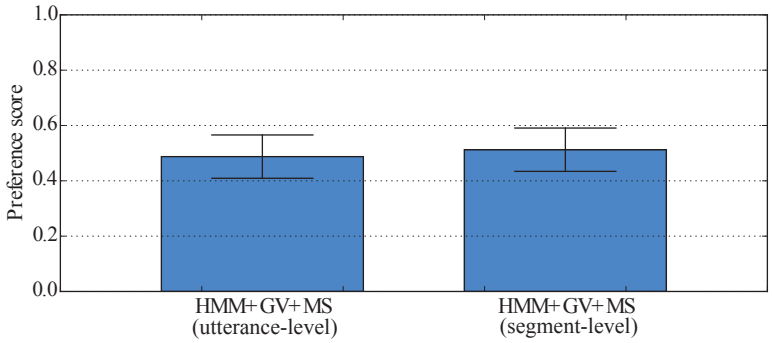


Figure 74. Preference scores on speech quality with 95% confidence interval for comparing the proposed utterance-level and segment-level post-filters in HMM-based TTS.

71 and Fig. 72. Their tendencies are similar to those of the utterance-level post-filter. Although the segment-level post-filtering process causes a degradation of the HMM likelihoods, they are still greater than those of natural parameters. Almost all likelihoods tend to increase as the filter coefficient approaches 1. We observed a degradation of the MS likelihood for F_0, but it is always greater than that of natural parameters. From these results, we tuned the emphasis coefficient to 1.0 for both spectrum and F_0. As the general tendency, the change of the MS likelihoods is smaller than that in the utterance-level post-filter.

4.6.5 Subjective evaluation for segment-level post-filter

The preference AB test on speech quality by 7 listeners was conducted in the same manner as in the previous section. The post-filtering was applied to both

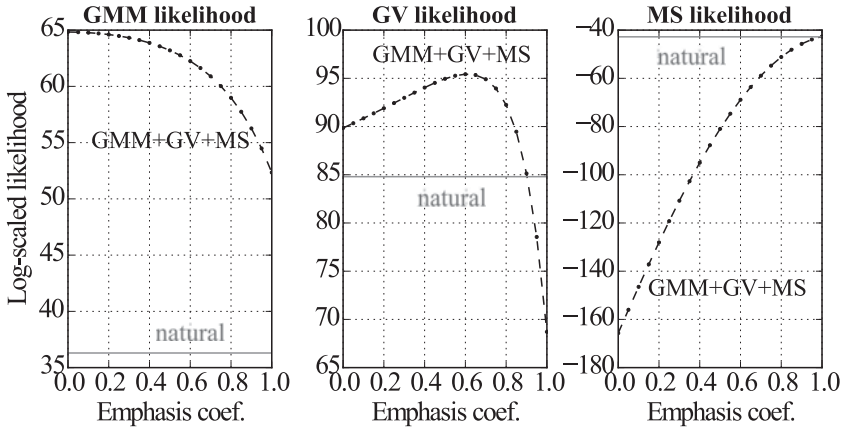


Figure 75. GMM, GV, and MS likelihoods for the spectral parameters filtered by the proposed utterance-level post-filter in GMM-based VC.

spectrum and F_0.

The preference score is shown in Fig. 73. It is observed that a significant quality gain is yielded by "HMM+MS" compared to "HMM," and it is comparable to that yielded by "HMM+GV." Furthermore, we can see that an additional gain is yielded by "HMM+GV+MS" compared to "HMM+GV." This tendency is similar to that observed in the utterance-level post-filter. Note that the segment-level post-filter is applicable to speech parameter sequences of various lengths but the utterance-level post-filter is not.

4.6.6 Comparison of utterance-level and segment-level post-filters

We compare the proposed utterance-level and segment-level post-filters that are applied to "HMM+GV" for spectrum and F_0. We used the emphasis coefficients tuned in this and the previous section. The preference AB test on speech quality by 8 listeners was conducted.

Fig. 74 shows the result. Because there is no significant difference between two post-filters, we can find that the proposed post-filters have the same capability in the speech quality improvement.

4.6.7 Evaluation in GMM-Based VC

The proposed utterance-level post-filter was applied to GMM-based VC. the tuning step and evaluation step are conducted in the same manner as the evaluation for HMM-based TTS. Here, "HMM+GV" and "HMM+GV+MS" were relabeled as "GMM+GV" and "GMM+GV+MS," respectively. The systems corresponding to "HMM" and "HMM+MS" were not used in the evaluation.

We prepared speech from two Japanese male and female speakers[28]. We selected 50 parallel sentences of subset A from the 503 phonetically balanced sentences included in the ATR Japanese speech database [126] for training, and 50 sentences of subset B for evaluation. We trained female-to-male GMMs. The speech features were the same as in the evaluations for HMM-based TTS. The spectral parameters and aperiodic components were converted with a 64-mixture GMM and a 16-mixture GMM, respectively. The log-scaled F_0 was linearly converted. The DFT length to calculate MS was set to 2048, which is over the maximum frame length in the training and evaluation data. The proposed utterance-level post-filter was applied to the spectral parameters.

The GMM likelihood, GV likelihood, and MS likelihood for the filtered spectral parameters were shown in Fig. 75. From this result, we can see that the tendency of the likelihood changes is almost the same as that in Fig. 67, but the GV likelihood of "GMM+GV+MS" starts to fall below "natural" at the emphasis coefficient $k = 0.90$. Therefore, the emphasis coefficient is set to 0.90.

We conducted a preference AB test on speech quality, and a preference XAB test on speaker individuality. We first presented an analysis-synthesized reference speech as "X", then we presented random-ordered synthesized speech. 7 listeners participated in each evaluation. Fig. 76 shows the results. In term of speech quality, a significant quality gain is observed. However, there is no significant difference in the preference score on speaker individuality. We expect that no cues for individuality are at higher modulation frequencies that are recovered by the MS-based post-filter.

4.6.8 Evaluation in CLUSTERGEN

The proposed segment-level post-filter was also applied to CLUSTERGEN. We also tuned the emphasis coefficient as in the previous experiments. We observed that the likelihoods didn't vary very much as shown in Figs. 71 and 72. We also confirmed that a quality gain was yielded by setting k to 1.0. Here, the methods corresponding to "HMM" and "HMM+MS" were relabeled as "CNV" and "CNV+MS," respectively.

We prepared an English female speaker. 418 and 46 sentences of news reader speech were used for training and evaluation, respectively. The speech features were the same as those in the evaluation for HMM-based TTS, but they were extracted by Speech signal Processing ToolKit (SPTK) [143] and the aperiodicity component was not used. The window length and window shift length of the segment-level post-filter were set to 125 ms (25 samples) and 60 ms (12 samples). A 64-taps DFT was used to calculate the MS. The segment-level post-filter was applied to both spectrum and F_0. parameters.

[28]The female speaker here is a different person from the speaker we used in the evaluation for HMM-based TTS.

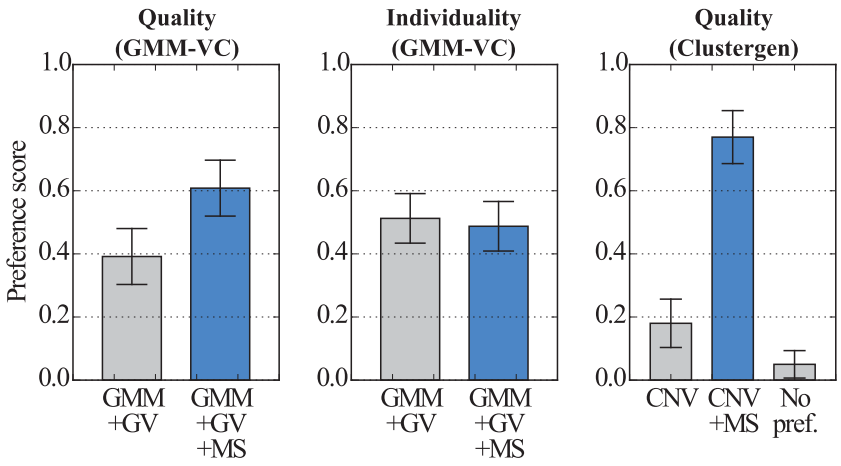


Figure 76. Preference scores on speech quality with 95% confidence interval in GMM-based VC and CLUSTERGEN

A preference AB test on speech quality was conducted by 6 listeners on the Amazon Mechanical Turk service [144]. Because many listening environments are expected, a no preference option was prepared. The right side of Fig. 76 shows the result. We can see that large improvements are yielded by the segment-level post-filter.

The results presented in this section suggest that the proposed MS-based post-filters are effective for a variety of statistical parametric speech synthesis frameworks.

4.7 Summary of this chapter

This chapter have introduced the Modulation Spectrum (MS) of speech parameter trajectory as a new feature to effectively quantify the over-smoothing effect, which is cause of the synthetic speech quality degradation. We have further proposed the MS-based post-filters for high-quality speech synthesis.

Section 4.2: We have defined the MS, and compared the natural speech parameters and synthetic speech parameters that are over-smoothed by the statistical processing. We have found the MS degradation caused by the over-smoothing effect in not only spectral parameters but also other speech parameters.

Section 4.3: We have proposed the MS-based post-filter. The post-filter is automatically trained using the natural and synthetic speech parameters included in the training data. In the synthesis stage, the generated speech parameters have been filtered utterance by utterance to make the MS close to natural MS.

Section 4.4: We have extended the filtering processes to the segment level to support the low latency speech synthesis. The generated speech parameters are windowed and filtered by the proposed segment-level MS-based post-filter.

Section 4.5: We have discussed about the MS and the MS-based post-filter and have described: (1) the MS is the mathematical extension of the GV, and (2) the MS-based post-filter generates the fluctuating speech parameter sequence.

Section 4.6: We have conducted experimental evaluation to confirm the effectiveness of the proposed post-filters, and have demonstrated: (1) the proposed utterance-level post-filter achieves better quality for spectrum, F_0, and HMM-state duration in HMM-based TTS, (2) the proposed segment-level post-filter capable of achieving low-delay synthesis also yields significant improvements in synthetic speech quality, (3) the proposed utterance-level and segment-level post-filters have the capability in the speech quality improvement, and (4) the proposed post-filters are also effective in not only HMM-based TTS but also GMM-based VC and CLUSTERGEN.

Chapter

5

Speech synthesis integrating modulation spectrum

5.1 Introduction

In Chapter 4, we have introduced the Modulation Spectrum (MS) as the features that can quantify the over-smoothing effect. Because generated speech parameter sequences tend to be temporally smoothed by the statistical generation process, we could find that the MS of synthetic speech tends to be degraded compared to that of natural speech. We have also proposed the MS-based post-filter in Chapter 4, which modifies the generated speech parameter sequences so that its MS gets closer to that of natural speech. Although the post-filtering approaches can improve the synthetic speech quality, this framework based on the post-filtering possibly causes adverse effects due to completely ignoring the basic criteria. Moreover, it is expected that the use of the MS model as one of the acoustic models is straightforward to apply various useful techniques of the original HMM-based TTS and GMM-based VC.

In this chapter, we integrate the MS into the speech synthesis criteria as similar as in **Section 2.9**. Integrating into the speech parameter generation is a straightforward way to alleviate the over-smoothing effect observed in the synthesis stage. The speech parameters of synthetic speech is generated to consider both the basic criterion and the additional criterion. However, we should avoid the use of such a generation algorithm for speech-based systems that require the computationally-efficient speech synthesis when the generation algorithm loses the basic computationally-efficient generation ability as described in **Section 2.6**. Yet another way avoiding the high computational cost is to integrating into the acoustic model training. The acoustic model are trained to generate the speech parameters that satisfy the additional criterion.

In this chapter, we first propose a speech parameter generation algorithm considering the MS. The proposed algorithm generates the parameter trajectories by maximizing a novel objective function consisting of the traditional criterion and the MS likelihoods. The MS likelihood works as a penalty term to make the MS of the generated parameters close to that of natural ones. Furthermore, we proposes a training algorithm considering the MS as yet another approach with the MS to improve the speech quality while preserving the traditional computationally-efficient generation. After implementing the trajectory GMM training for GMM-based VC as the same as the trajectory HMM in **Section 2.8**, we integrate the MS into the trajectory training for both HMM-based TTS and GMM-based VC. The HMM or GMM are trained to recover the MS of the generated speech parameters, and the proposed training algorithm gives a unified framework for both training and generation which provides both a consistent optimization criterion and a closed form solution for parameter generation considering the MS. Also, the proposed training algorithm makes it possible to perform the MS modeling depending on the input parameters. The objective functions listed in Table 2 are compared in this chapter. The experimental results demonstrate the pro-

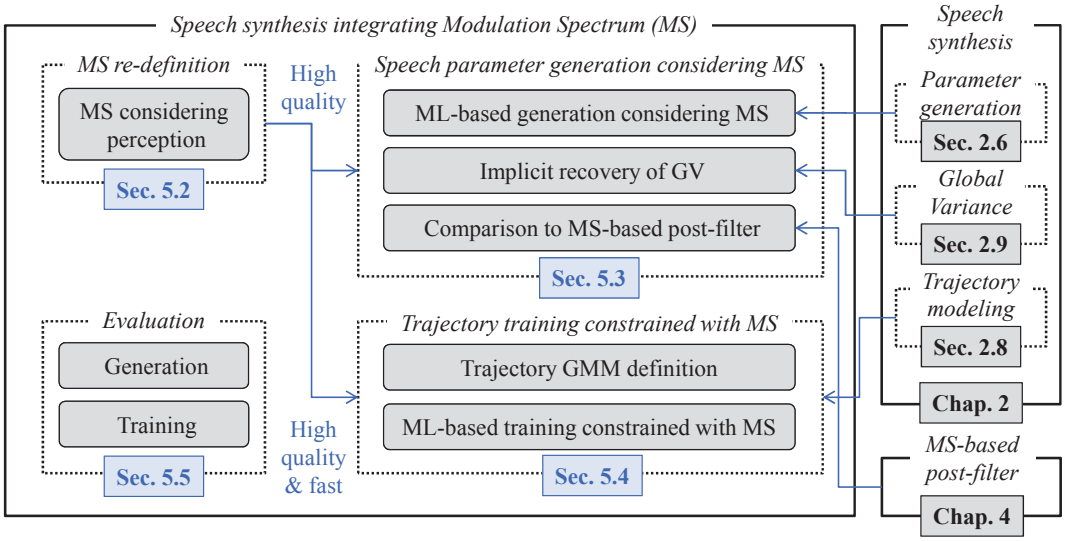


Figure 77. The rest of Chapter 5.

posed approaches achieved the best in synthetic speech quality. Summarizing the proposed methods based on the MS, the post-filter, parameter generation, and trajectory training have the following advantages.

Post-filter (Chapter 4) has the high portability meaning it can be easily used in the various speech synthesis system.

Parameter generation (This chapter) can generate the most high-quality speech parameters by directly alleviating the over-smoothing effect observed in the parameter generation stage.

Trajectory training (This chapter) performs the computationally-efficient and high-quality speech synthesis.

We further discuss this in **Section 5.4**.

The rest of this chapter is organized as follows and shown in Fig. 77. In **Section 5.2**, we slightly fix the MS definition to consider the perceptual effect. In **Section 5.3** and **Section 5.4**, We integrate the MS into the speech parameter generation algorithm and the training algorithm.

Section 5.5 and Section 5.6 are the experimental evaluation and summary of this chapter.

Table 2. Objective functions $L_{\cdot}^{(\cdot)}$ compared in Chapter 5. The training criterion $L_{\cdot}^{(\mathrm{trn})}$ is maximized to estimate the HMM/GMM parameter set $\boldsymbol{\lambda}$, and the synthesis criterion $L_{\cdot}^{(\cdot)}$ is maximized to generate a synthetic speech parameter sequence $\hat{\boldsymbol{y}}_{\hat{\boldsymbol{q}}}$. Note that the objective function $L_{\mathrm{gv}}^{(\mathrm{trn})}$ of [2] and [3] are obviously different as described in **Section 2.9**, but we use the same notation for simplicity.

	Training	Synthesis
Basic	$L_{\mathrm{basic}}^{(\mathrm{trn})}$ (**Section 2.4** and **Section 2.5**)	$L_{\mathrm{basic}}^{(\mathrm{syn})}$ (**Section 2.6**)
GV	$L_{\mathrm{gv}}^{(\mathrm{trn})}$ (**Section 2.9**)	$L_{\mathrm{gv}}^{(\mathrm{syn})}$ (**Section 2.9**)
Trajectory	$L_{\mathrm{trj}}^{(\mathrm{trn})}$ (**Section 2.8**)	-
MS	$L_{\mathrm{ms}}^{(\mathrm{trn})}$	$L_{\mathrm{ms}}^{(\mathrm{syn})}$

5.2 Modulation spectrum re-definition

As [133, 134] reported[29], the lower modulation frequency components are dominant in speech perception. Therefore, we re-define the MS $\boldsymbol{s}(\boldsymbol{y})$ of the parameter sequence $\boldsymbol{y}$ as the following form considering speech perception,

$$\boldsymbol{s}(\boldsymbol{y}) = \left[\boldsymbol{s}(1)^{\top}, \cdots, \boldsymbol{s}(d)^{\top}, \cdots, \boldsymbol{s}(D)^{\top}\right]^{\top}, \tag{145}$$

$$\boldsymbol{s}(d) = \left[s_d(0), \cdots, s_d(f), \cdots, s_d(D'_{\mathrm{s}} - 1)\right]^{\top}, \tag{146}$$

$$s_d(f) = \left(\sum_{t=1}^{T} y_t(d) \cos mt\right)^2 + \left(\sum_{t=1}^{T} y_t(d) \sin mt\right)^2, \tag{147}$$

where f is the modulation frequency index, $m = -\pi f / D_{\mathrm{s}}$ is a modulation frequency, and D_{s} is one half of the DFT length. The MS is calculated from zero-padded parameter sequences so its length is $2D_{\mathrm{s}}$. D'_{s} is the fixed number of dimension of MS. As shown in Fig. 78, the re-defined MS consists of only lower modulation frequency components where the originally-defined MS in Chapter 4 have consisted of all the components. Also, we calculate the linear-scaled MS in this chapter because we find that there is no significant difference between the linear- and log-scaled MS in synthetic speech quality.

5.3 Parameter generation algorithm considering MS

This section describes the speech parameter generation algorithm that maximizes a function $L_{\mathrm{ms}}^{(\mathrm{syn})}$ combining the basic criteria $L_{\mathrm{basic}}^{(\mathrm{syn})}$ and the MS likelihood.

[29] Whereas [133, 134] have investigated the effect of the MS on intelligibility, **Section A.10** have investigated it on the speech quality. We have found that there was no significant quality difference between analysis-synthesized speech samples with/without the MS components over 50 Hz for mel-cepstrum.

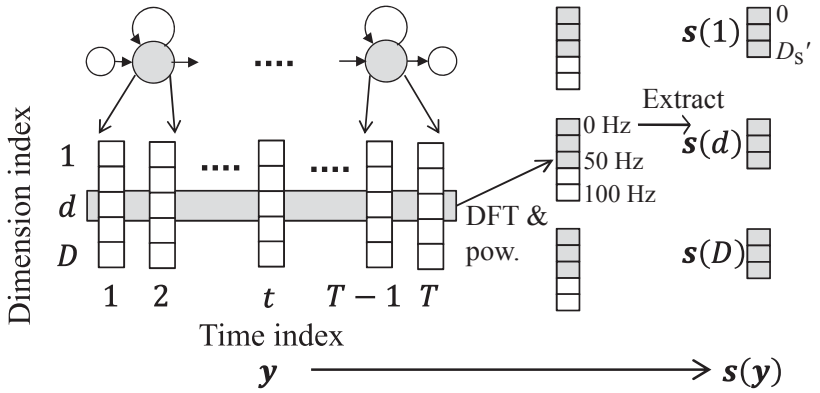


Figure 78. Re-defined Modulation Spectrum (MS) $\boldsymbol{s}(\boldsymbol{y})$ of the speech parameter sequence $\boldsymbol{y}$. Compared to the original definition in **Section 4.2**, only the lower modulation frequency components are used. In this figure, we assume that the shift length of speech parameter sequence is 5 msec (Nyquist frequency is 100 Hz.) and the modulation frequency components lower than 50 Hz are used.

5.3.1 Objective function

Let the MS likelihood be $\mathcal{N}(\boldsymbol{s}(\boldsymbol{y}); \boldsymbol{\mu}_{\rm s}, \boldsymbol{\Sigma}_{\rm s})$ where $\boldsymbol{\mu}_{\rm s}$ and $\boldsymbol{\Sigma}_{\rm s}$ are a $DD'_{\rm s}$-by-1 mean vector and a $DD'_{\rm s}$-by-$DD'_{\rm s}$ covariance matrix, respectively. $\boldsymbol{\Sigma}_{\rm s}^{-1}$ is represented as $\left[\boldsymbol{p}_{\rm s}^{(1)}, \cdots, \boldsymbol{p}_{\rm s}^{(d)}, \cdots, \boldsymbol{p}_{\rm s}^{(D)}\right]$ where $\boldsymbol{p}_{\rm s}^{(d)}$ is $DD'_{\rm s}$-by-$D'_{\rm s}$ matrix whose columns correspond to $\boldsymbol{s}(d)$. The MS is calculated utterance by utterance and its mean vector and covariance matrix are calculated from the whole utterances of the training data.

The objective function is as follows:

$$
\begin{aligned}
L_{\rm ms}^{(\rm syn)} &= P(\boldsymbol{W}\boldsymbol{y}|\boldsymbol{X}, \hat{\boldsymbol{q}}, \boldsymbol{\lambda}) P(\boldsymbol{s}(\boldsymbol{y})|\boldsymbol{\lambda}_{\rm s})^{\omega_{\rm s}\frac{N_{\rm w}T}{D'_{\rm s}}} && (148)\\
&= \mathcal{N}\left(\boldsymbol{W}\boldsymbol{y}; \boldsymbol{\mu}_{\hat{\boldsymbol{q}}}, \boldsymbol{\Sigma}_{\hat{\boldsymbol{q}}}\right) \mathcal{N}(\boldsymbol{s}(\boldsymbol{y}); \boldsymbol{\mu}_{\rm s}, \boldsymbol{\Sigma}_{\rm s})^{\omega_{\rm s}\frac{N_{\rm w}T}{D'_{\rm s}}} && (149)\\
&= L_{\rm basic}^{(\rm syn)} \left(L_{\rm s}^{(\rm syn)}\right)^{\omega_{\rm s}\frac{N_{\rm w}T}{D'_{\rm s}}}, && (150)
\end{aligned}
$$

where $\omega_{\rm s}$ denotes the MS weight for controlling the balance between the traditional and MS likelihoods, and

$$
L_{\rm s}^{(\rm syn)} = \mathcal{N}(\boldsymbol{s}(\boldsymbol{y}); \boldsymbol{\mu}_{\rm s}, \boldsymbol{\Sigma}_{\rm s}) \tag{151}
$$

is the MS likelihood. The basic likelihood and the MS likelihood are normalized when $\omega_{\rm s} = 1$.

5.3.2 Parameter generation

The speech parameter sequence $\hat{\boldsymbol{y}}_{\hat{\boldsymbol{q}}}$ is generated by maximizing the proposed objective function $L_{\rm ms}^{(\rm syn)}$ as follows:

$$\hat{\boldsymbol{y}}_{\hat{\boldsymbol{q}}} = \mathop{\rm argmax}_{\boldsymbol{y}} L_{\rm ms}^{(\rm syn)}. \tag{152}$$

Because the proposed objective function has a 4-order form like that in the algorithm considering the GV, it is hard to analytically solve its maximization problem. Instead, we use the steepest descent algorithm to iteratively update the generated parameter trajectory as follows:

$$\hat{\boldsymbol{y}}_{\hat{\boldsymbol{q}}}^{(i+1)} = \hat{\boldsymbol{y}}_{\hat{\boldsymbol{q}}}^{(i)} + \alpha \left. \frac{\partial \log L_{\rm ms}^{(\rm syn)}}{\partial \boldsymbol{y}} \right|_{\boldsymbol{y}=\hat{\boldsymbol{y}}_{\hat{\boldsymbol{q}}}^{(i)}}. \tag{153}$$

The logarithm function of $L_{\rm ms}^{(\rm syn)}$ is given by:

$$\log L_{\rm ms}^{(\rm syn)} = \log L_{\rm basic}^{(\rm syn)} + \omega_{\rm s} \frac{N_{\rm w} T}{D_{\rm s}'} \log L_{\rm s}^{(\rm syn)}, \tag{154}$$

where α and i are the learning rate and the iteration index, respectively. Referring Eq. (65), the first derivative of $\log L_{\rm basic}^{(\rm syn)}$ is $\boldsymbol{R}_{\hat{\boldsymbol{q}}}\boldsymbol{y} - \boldsymbol{r}_{\hat{\boldsymbol{q}}}$, and the first derivative of $\log L_{\rm s}^{(\rm syn)}$ is calculated as:

$$\frac{\partial L_{\rm s}^{(\rm syn)}}{\partial \boldsymbol{y}} = \left[\boldsymbol{s}_1'^{\top}, \cdots, \boldsymbol{s}_t'^{\top}, \cdots, \boldsymbol{s}_T'^{\top} \right]^{\top}, \tag{155}$$

$$\boldsymbol{s}_t' = \left[s_t(1), \cdots, s_t(d), \cdots, s_t(D) \right]^{\top}, \tag{156}$$

$$s_t(d) = \left(\boldsymbol{s}(\boldsymbol{y}) - \boldsymbol{\mu}_{\rm s} \right)^{\top} \boldsymbol{p}_{\rm s}^{(d)} \boldsymbol{f}_t(d), \tag{157}$$

$$\boldsymbol{f}_t(d) = \left[f_{t,d}(0), \cdots, f_{t,d}(f), \cdots, f_{t,d}(D_{\rm s}' - 1) \right]^{\top}, \tag{158}$$

$$f_{t,d}(f) = -2 \left(R_{d,f} \cos mt + I_{d,f} \sin mt \right). \tag{159}$$

This derivation is graphically shown in Fig. 79. In this chapter, $D_{\rm s}'/D_{\rm s}$ is set to 1.0.[30] Instead of controlling this ratio, we apply 50 Hz-cutoff low pass filter (LPF)[31] to the generated parameter trajectories in after iteration in order to avoid slightly artificial sounds caused by enhancing the high modulation frequency components.

5.3.3 Initialization

For initialization, we basically use the same idea in the conventional algorithm considering the GV, i.e., first generating the parameter trajectory by maximizing

[30] We set it to 0.5 for spectrum in **Section A.1**. Also, we remove in advance the < 50 Hz MS components of the spectral parameters in the training data.

[31] This cutoff frequency corresponds to $D_{\rm s}'/D_{\rm s} = 0.5$.

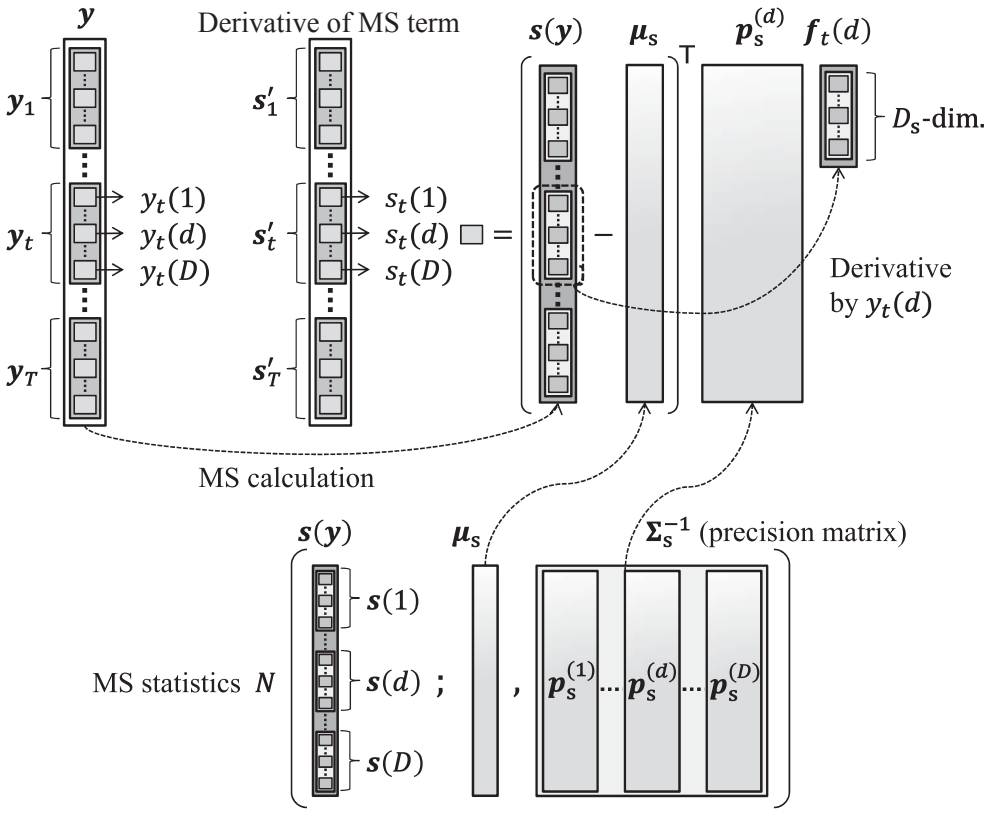


Figure 79. Graphical representation of how to derive the first derivative used in the proposed speech parameter generation considering the MS. We can find that all modulation frequency components are considered to calculate the derivative of one speech parameter.

only the traditional criterion $L^{(\mathrm{syn})}_{\mathrm{basic}}$ and then transforming it further by maximizing the other likelihood. To transform the parameter trajectory so that the MS likelihood increases, we use the MS-based post-filter proposed in Chapter 4, which is given by

$$s_d(f)' = \frac{\sigma'_{d,f}}{\sigma_{d,f}}\left(s_d(f) - \mu_{d,f}\right) + \mu'_{d,f}, \tag{160}$$

where $\mu_{d,f}$ and $\sigma_{d,f}$ are the mean and standard deviation of $s_d(f)$, respectively. $\mu'_{d,f}$ and $\sigma'_{d,f}$ are those of the MS of the generated trajectories. We estimate $\mu'_{d,f}$ and $\sigma'_{d,f}$ using the generated trajectories included in the training data. Finally, the initial parameter trajectory $\boldsymbol{y}^{(0)}_{\hat{\boldsymbol{q}}}$ is determined using the filtered MS and the original phase components of the parameter trajectory before the filtering.

5.3.4 Application to F0 component

The proposed parameter generation is also applied to the F_0 components modeled with MSD-HMM [84]. In this case, after unvoiced/voiced determination, F_0 values at only voiced frames are generated from the corresponding probability density functions, while the precision matrices (inverse matrix of the covariance matrix) at the unvoiced/voiced boundaries are set to zero matrices to allow discontinuous transitions[32]. Therefore, the MS is calculated from the concatenated voiced frames in this chapter. Moreover, we reform $y_t(d)$ of Eq. (147) as $y_t(d) - \bar{y}(d)$ as pointed out in [145]. The MS is directly affected by the discontinuous transitions at the unvoiced/voiced boundaries. This causes some adverse effects in the post-filtering process. To avoid them, we use the initialization method of the conventional GV-based algorithm rather than the MS-based post-filtering.

5.3.5 Discussions

Although we can also integrate the GV term into the proposed objective function, i.e., a product of the HMM/GMM, GV, and MS likelihoods, the proposed objective function effectively recovers the GV likelihood without it because the MS involves the GV, as we described. Figures 80 and 81 illustrate examples of the GV and the MS of the generated parameter trajectories. "HMM," "GV," and "MS" indicate the results of the generated parameter trajectories of the traditional generation algorithm using $L^{(\mathrm{syn})}_{\mathrm{basic}}$, conventional generation algorithm with the GV using $L^{(\mathrm{syn})}_{\mathrm{gv}}$, and the proposed algorithm with the MS using $L^{(\mathrm{syn})}_{\mathrm{ms}}$ in HMM-based TTS, respectively. "nat" indicates those of natural speech parameter trajectories. We can see that the proposed generation algorithm well recovers not only the MS but also the GV. On the other hand, "GV" cannot recover the MS appropriately. Although it makes the MS slightly larger, the resulting MS is still very different from the natural one. This is because the GV models only average values of the MS components over the modulation frequencies.

The footprint of the synthesis system using the proposed algorithm is slightly larger than that of the one using the algorithm with the GV because the MS is DD'_{s}-dimensional vector, whereas the GV is D-dimensional vector. We may reduce the footprint by considering only low modulation frequency components which have a larger effect on speech perception [129].

We can localize the MS constraint that captures the segment-level fluctuation, but we can't find any difference in synthetic speech quality between the proposed method considering utterance-level MS and segment-level MS.

[32] When we apply the proposed parameter generation to the continuous F_0 modeling [86], it is applied as the same as the spectral parameters.

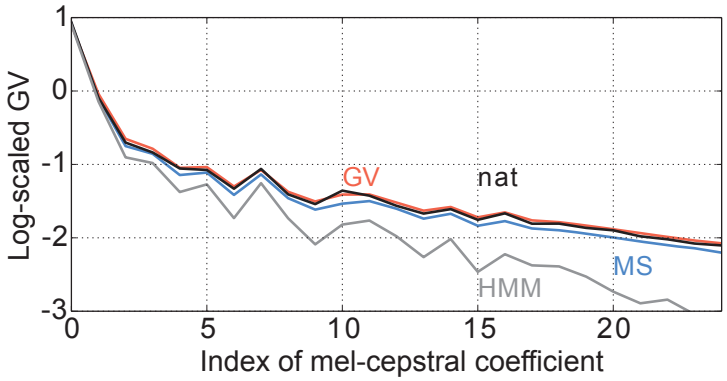


Figure 80. An example of the GV of the generated mel-cepstral coefficients. We can find that not only "GV" (conventional generation considering GV) but also "MS" (proposed generation algorithm) are close to "nat" (natural speech). This is because the MS involves the GV, and the proposed generation algorithm considering the MS implicitly recover the GV.

Finally, the MS-based post-filter proposed in Chapter 4 tends to generate over-transformed trajectories and synthesize over-emphasized speech because it completely ignores the traditional criterion, e.g., the HMM/GMM likelihood. On the other hand, the proposed algorithm effectively generates naturally fluctuated parameter trajectories by jointly maximizing the HMM/GMM and MS likelihoods. Fig. 82 shows an example of the final speech parameter trajectory ("After iteration") and initial speech parameter trajectory ("Before iteration") determined by applying the MS-based post-filter to the trajectory generated from HMMs. We can see that over-fluctuated transition is alleviated by iterating with the HMM and MS likelihood.

5.4 MS-constrained trajectory training

This section proposes a novel trajectory training algorithm that maximizes the novel function $L_{\text{ms}}^{(\text{trn})}$ combining the traditional criterion and the MS likelihood. Before defining $L_{\text{ms}}^{(\text{trn})}$, we reform the basic GMM in **Section 2.5** as the trajectory GMM as the similar as the trajectory HMM in **Section 2.8**, and define the trajectory training criterion $L_{\text{trj}}^{(\text{trn})}$ for GMM-based VC.

5.4.1 Trajectory GMM training

The trajectory GMM training has been implemented for the joint probability density modeling [6] in GMM-based VC. In this section, we present another implementation by reformulating the conditional probability density function by

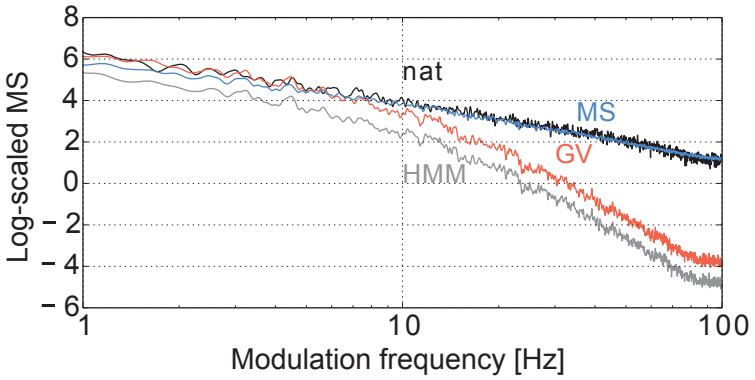


Figure 81. An examples of the MS of the generated 9-th mel-cepstral coefficient. As we described, conventional parameter generation algorithm considering the GV performs bias-like effect in the MS domain, but the proposed generation algorithm efficiently recovers the MS.

imposing the explicit relationship between the static and dynamic features.

The objective function $L_{\mathrm{trj}}^{(\mathrm{trn})}$ is the same as Eq. (84). Similarly to Eq. (87), the joint parameters of $\boldsymbol{A}_q$ and $\boldsymbol{b}_q$ over all GMM-mixture components are defined as:

$$\boldsymbol{\xi}_{\mathrm{A}} = \left[\boldsymbol{A}_1^\top, \cdots, \boldsymbol{A}_q^\top, \cdots, \boldsymbol{A}_Q^\top\right]^\top, \tag{161}$$

$$\boldsymbol{\xi}_{\mathrm{b}} = \left[\boldsymbol{b}_1^\top, \cdots, \boldsymbol{b}_q^\top, \cdots, \boldsymbol{b}_Q^\top\right]^\top, \tag{162}$$

and the mean vector $\boldsymbol{E}_{\hat{\boldsymbol{q}}}$ is represented as:

$$\boldsymbol{E}_{\hat{\boldsymbol{q}}} = \mathrm{diag}_{N_{\mathrm{w}}D}\left[\boldsymbol{S}_{\hat{\boldsymbol{q}}}\boldsymbol{\xi}_{\mathrm{A}}\right]\boldsymbol{X} + \boldsymbol{S}_{\hat{\boldsymbol{q}}}\boldsymbol{\xi}_{\mathrm{b}}. \tag{163}$$

$\boldsymbol{D}_{\hat{\boldsymbol{q}}}^{-1}$ is represented as the same as Eq. (89). We use the steepest descent algorithm to optimize $\boldsymbol{A}_q$, $\boldsymbol{b}_q$ and $\boldsymbol{\Sigma}_q^{(Y|X)^{-1}}$ [33], and the first derivatives with respect to $\boldsymbol{A}_q$ and $\boldsymbol{b}_q$ are

$$\frac{\partial \log L_{\mathrm{trj}}^{(\mathrm{trn})}}{\partial \boldsymbol{\xi}_{\mathrm{A}}} = \boldsymbol{S}_{\hat{\boldsymbol{q}}}^\top \mathrm{diag}_{N_{\mathrm{w}}D}^{-1}\left[\boldsymbol{D}_{\hat{\boldsymbol{q}}}^{-1}\boldsymbol{W}\left(\boldsymbol{y} - \hat{\boldsymbol{y}}_{\hat{\boldsymbol{q}}}\right)\boldsymbol{X}^\top\right], \tag{164}$$

$$\frac{\partial \log L_{\mathrm{trj}}^{(\mathrm{trn})}}{\partial \boldsymbol{\xi}_{\mathrm{b}}} = \boldsymbol{S}_{\hat{\boldsymbol{q}}}^\top \boldsymbol{D}_{\hat{\boldsymbol{q}}}^{-1}\boldsymbol{W}\left(\boldsymbol{y} - \hat{\boldsymbol{y}}_{\hat{\boldsymbol{q}}}\right), \tag{165}$$

The traditional joint density training using $L_{\mathrm{basic}}^{(\mathrm{trn})}$ is performed first in order to estimate $\boldsymbol{\lambda}$. Then, the proposed algorithms updates $\{\boldsymbol{\xi}_{\mathrm{A}}, \boldsymbol{\xi}_{\mathrm{b}}, \boldsymbol{\Sigma}_q^{(Y|X)^{-1}}\}$ while

[33] Closed form solutions also exist for $\boldsymbol{\xi}_{\mathrm{A}}$ and $\boldsymbol{\xi}_{\mathrm{b}}$.

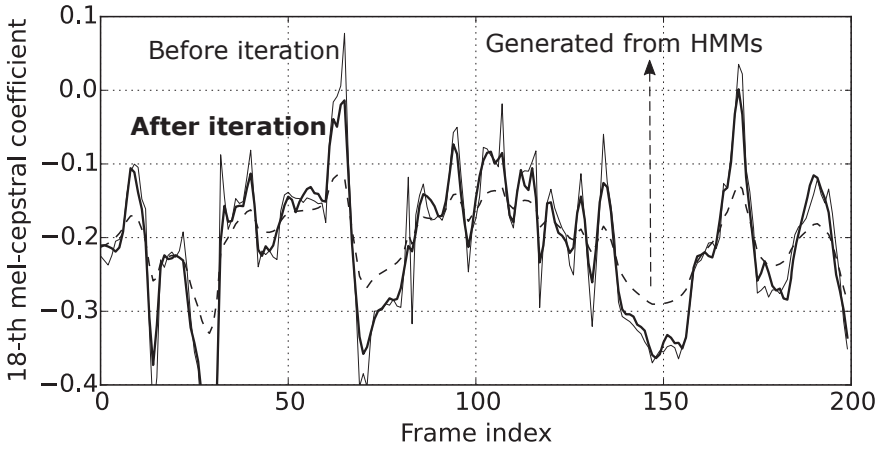


Figure 82. Examples of mel-cepstral coefficient sequences before and after iteration of the proposed speech parameter generation algorithm. Whereas the initial parameters generated using MS-based post-filter causes unnatural changes of the sequence, we can see that it is alleviated by the iteration.

keeping $\{w_q^{(Z)}, \boldsymbol{\mu}_q^{(X)}, \boldsymbol{\Sigma}_q^{(XX)}\}$ constant. Note that the sub-optimum GMM-mixture component sequence $\hat{\boldsymbol{q}}$ never changes.

5.4.2 Objective function

Because the lower modulation frequency components mainly affect speech perception [129], it is better to train the HMM/GMM parameters with only the lower modulation frequency components. Therefore, we redefine $\boldsymbol{s}(d)$ as $[s_d(0), \cdots, s_d(f), \cdots, s_d(D'_{\rm s}-1)]$. $D'_{\rm s}$ is the fixed number of MS dimensions in each feature dimension, where $D'_{\rm s} \leq D_{\rm s}$. Note that the numbers of dimension of $\boldsymbol{s}(\boldsymbol{y})$, $\boldsymbol{\mu}_{\rm s}$, $\boldsymbol{\Sigma}_{\rm s}$, and $\boldsymbol{p}_{\rm s}^{(d)}$ are fixed to $DD'_{\rm s}$, $D'_{\rm s}D$-by-1, $D'_{\rm s}D$-by-$D'_{\rm s}D$, and $D'_{\rm s}D$-by-D, respectively.

We integrate the MS likelihood into the trajectory training as follows:

$$L_{\rm ms}^{(\rm trn)} = P(\boldsymbol{y}|\boldsymbol{X}, \hat{\boldsymbol{q}}, \boldsymbol{\lambda}) P(\boldsymbol{s}(\boldsymbol{y})|\boldsymbol{X}, \hat{\boldsymbol{q}}, \boldsymbol{\lambda}, \boldsymbol{\lambda}_{\rm s})^{\omega_{\rm s}\frac{T}{D'_{\rm s}}}, \tag{166}$$

$$P(\boldsymbol{s}(\boldsymbol{y})|\boldsymbol{X}, \hat{\boldsymbol{q}}, \boldsymbol{\lambda}, \boldsymbol{\lambda}_{\rm s}) = \mathcal{N}\left(\boldsymbol{s}(\boldsymbol{y}); \boldsymbol{s}\left(\hat{\boldsymbol{y}}_{\hat{\boldsymbol{q}}}\right), \boldsymbol{\Sigma}_{\rm s}\right), \tag{167}$$

The MS likelihood works as a penalty term to reduce the temporal fluctuation of the generated parameter sequence.

5.4.3 Model parameter estimation

The HMM/GMM parameter sets $\boldsymbol{\lambda}$ are estimated in the same way as in the trajectory training. Let $L_{\rm s}^{(\rm trn)}$ be the MS likelihood $\mathcal{N}\left(\boldsymbol{s}(\boldsymbol{y}); \boldsymbol{s}\left(\hat{\boldsymbol{y}}_{\hat{\boldsymbol{q}}}\right), \boldsymbol{\Sigma}_{\rm s}\right)$. The

logarithm of $L_{\mathrm{ms}}^{(\mathrm{trn})}$ is

$$\log L_{\mathrm{ms}}^{(\mathrm{trn})} = \log L_{\mathrm{trj}}^{(\mathrm{trn})} + \omega_{\mathrm{s}} \frac{T}{D_{\mathrm{s}}'} \log L_{\mathrm{s}}^{(\mathrm{trn})}, \tag{168}$$

and the gradients of $\log L_{\mathrm{s}}^{(\mathrm{trn})}$ are given as

$$\frac{\partial \log L_{\mathrm{s}}^{(\mathrm{trn})}}{\partial \boldsymbol{\xi}_{\mathrm{A}}} = \boldsymbol{S}_{\hat{\boldsymbol{q}}}^{\top} \mathrm{diag}_{N_{\mathrm{w}}D}^{-1} \left[\boldsymbol{D}_{\hat{\boldsymbol{q}}}^{-1} \boldsymbol{W} \boldsymbol{R}_{\hat{\boldsymbol{q}}}^{-1} \boldsymbol{s}_{\hat{\boldsymbol{q}}} \boldsymbol{X}^{\top} \right], \tag{169}$$

$$\frac{\partial \log L_{\mathrm{s}}^{(\mathrm{trn})}}{\partial \boldsymbol{\xi}_{\mathrm{b}}} = \frac{\partial \log L_{\mathrm{s}}^{(\mathrm{trn})}}{\partial \boldsymbol{\mu}} = \boldsymbol{S}_{\hat{\boldsymbol{q}}}^{\top} \boldsymbol{D}_{\hat{\boldsymbol{q}}}^{-1} \boldsymbol{W} \boldsymbol{R}_{\hat{\boldsymbol{q}}}^{-1} \boldsymbol{s}_{\hat{\boldsymbol{q}}}, \tag{170}$$

$$\frac{\partial \log L_{\mathrm{s}}^{(\mathrm{trn})}}{\partial \boldsymbol{\Sigma}^{-1}} = \boldsymbol{S}_{\hat{\boldsymbol{q}}}^{\top} \mathrm{diag}_{N_{\mathrm{w}}D}^{-1} \left[\boldsymbol{W} \boldsymbol{R}_{\hat{\boldsymbol{q}}}^{-1} \boldsymbol{s}_{\hat{\boldsymbol{q}}} \left(\boldsymbol{E}_{\hat{\boldsymbol{q}}} - \boldsymbol{W} \hat{\boldsymbol{y}}_{\hat{\boldsymbol{q}}} \right) \right], \tag{171}$$

where

$$\boldsymbol{s}_{\hat{\boldsymbol{q}}} = \left[\boldsymbol{s}_1'^{\top}, \cdots, \boldsymbol{s}_t'^{\top}, \cdots, \boldsymbol{s}_T'^{\top} \right]^{\top}, \tag{172}$$

$$\boldsymbol{s}_t' = \left[s_t(1), \cdots, s_t(d), \cdots, s_t(D) \right]^{\top}, \tag{173}$$

$$s_t(d) = 2 \boldsymbol{f}_t(d) \boldsymbol{p}_{\mathrm{s}}^{(d)\top} \left(\boldsymbol{s}(\boldsymbol{y}) - \boldsymbol{s}(\hat{\boldsymbol{y}}) \right), \tag{174}$$

$$\boldsymbol{f}_t(d) = \left[f_{t,d}(0) \cdots, f_{t,d}(f), \cdots, f_{t,d}(D_{\mathrm{s}}' - 1) \right]^{\top}, \tag{175}$$

$$f_{t,d}(f) = \hat{R}_{d,f} \cos kt + \hat{I}_{d,f} \sin kt, \tag{176}$$

$\hat{R}_{d,f}$ and $\hat{I}_{d,f}$ are calculated using the d-th dimensional components of $\hat{\boldsymbol{y}}_{\hat{\boldsymbol{q}}}$.

We perform the basic training algorithm using $L_{\mathrm{basic}}^{(\mathrm{trn})}$ first, then the trajectory training algorithm using $L_{\mathrm{trj}}^{(\mathrm{trn})}$. Finally, we update the model parameter sets with the proposed training algorithm.

5.4.4 Application to F0 component

MSD-HMM [84] is unsuitable for MS modeling as described in Section V-D. Therefore, we decided to use continuous F_0 modeling [86]. Moreover, we reformulate $y_t(d)$ of Eq. (147) as $y_t(d) - \bar{y}(d)$ in the same way as in the proposed parameter generation.

5.4.5 Discussions

It is unnecessary to consider the MS in parameter generation because the HMM/GMM parameters are optimized to make the MS of the generated parameter sequence close to the natural one. Consequently, the basic parameter generation using the $L_{\mathrm{basic}}^{(\mathrm{syn})}$ algorithm can be straightforwardly employed. If the proposed objective function $L_{\mathrm{ms}}^{(\mathrm{trn})}$ is used in the parameter generation, the generated parameter sequence to maximize it is equivalent to $\hat{\boldsymbol{y}}_{\hat{\boldsymbol{q}}}$ which can be solved analytically.

Therefore, the proposed training algorithm can also be regarded as a unified framework of the training and generation processes. Also, the proposed training algorithm makes it possible to the context-dependent MS modeling because the mean vectors of the MS model are calculated from the input parameters. This also enables one to avoid a large footprint, as discussed above.

Fig. 83 plots the output probabilities at each frame in HMM-based TTS. We can see that the variance of the trajectory training using $L_{\text{trj}}^{(\text{trn})}$ ("TRJ") is slightly larger than that of the traditional training using $L_{\text{basic}}^{(\text{trn})}$ ("BSC"), and the mean of the GV-constrained trajectory training using $L_{\text{gv}}^{(\text{trn})}$ ("GV"), or MS-constrained trajectory training using $L_{\text{ms}}^{(\text{trn})}$ ("MS"), is significantly changed compared to "TRJ," Moreover, the mean of "MS" tends to move far from the neighboring HMM-state[34].

We didn't investigate the quality difference between the proposed generation algorithm and the proposed training algorithm, but we expect that the generation algorithm will achieve higher quality as the similar result in GV has been reported in [146]. One of the reasons is the limitation of the model structures. To explain this, we assume that there is one HMM-state having too long duration. The MS tries to fluctuate a speech parameter sequence generated from HMMs. The proposed generation allows such a transition varying frame by frame. However, the HMMs trained by the proposed training algorithm can not produce such the parameter sequence because one HMM-state have only one output probability. The same problem occurs in GMM-based VC because the conversion function within one GMM-state must be a linear function[35].

Table 3 summarizes three proposed methods using the MS. The MS-based post-filter has the best portability because the process is independent on the original speech synthesis procedures. Comparing $L_{\text{ms}}^{(\text{syn})}$ and $L_{\text{ms}}^{(\text{trn})}$, it is cleared that the proposed training algorithm is strongly constrained with the model structures compared to the generation algorithm. In term of quality, the proposed generation algorithm is the best as discussed above[36]. Finally, the proposed training algorithm makes it possible to perform real-time speech generation. It is impossible for the post-filter to perform the real-time process, but it is possible to perform low-delay process. The proposed generation algorithm needs iterations in synthesis.

[34] Note that the frames that have same statistics correspond to the same HMM-state.

[35] However, the effect in GMM-based VC is expected to be less than that in HMM-based TTS because the output probability varies frame by frame in GMM-based VC.

[36] However, as described in **Section 5.5**, the post-filtering process after parameter generation taking account of GV is the similar in quality compared to the parameter generation taking account of the MS.

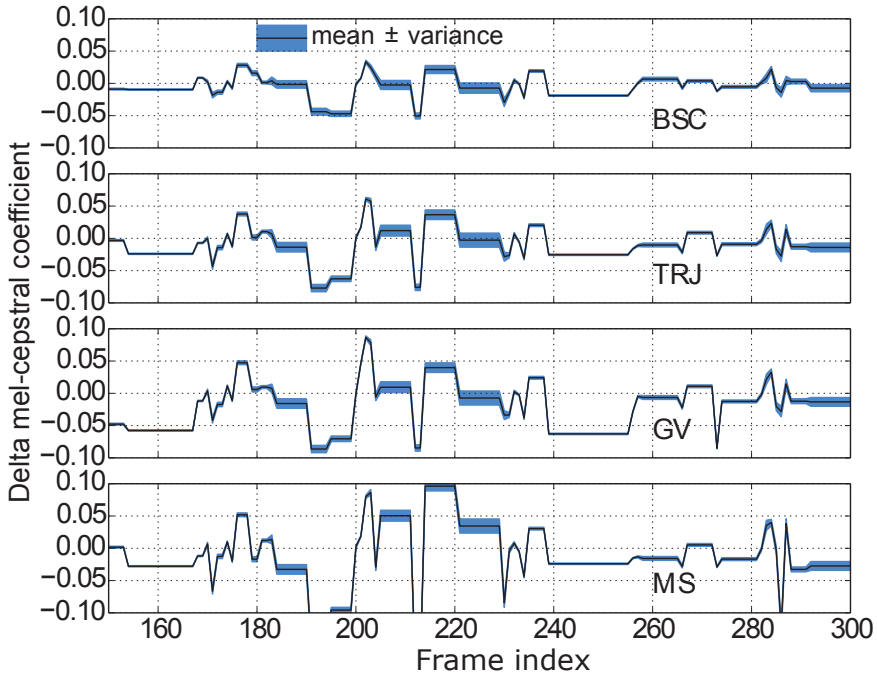


Figure 83. Example of statistics of the 10th mel-cepstral coefficient of the HMMs trained by the several training algorithms in HMM-based TTS. Note that the frames having same statistics correspond to the same HMM-state. We can see that the statistics by the proposed training algorithm ("MS") varies more than the other algorithms.

5.5 Experimental evaluation

5.5.1 Experimental conditions for speech parameter generation algorithm

We used an English male speaker "RMS" and an English female speaker "SLT" from the CMU ARCTIC database [147]. Speech signals were sampled at 16 kHz. The shift length was set to 5 ms. The 0th-through-24th mel-cepstral coefficients were extracted as spectral parameters and log-scaled F_0 and 5 band-aperiodicity [74, 127] were extracted as excitation parameters. The STRAIGHT analysis-synthesis system [10] was employed for parameter extraction and waveform generation. The DFT length used to calculate the MS was set to 8192. Diagonal covariance matrices were used in the context-independent GV models and context-independent MS models. The GV weight ω_{v} and MS weight ω_{s} were set to 1.0.

For HMM-based TTS, we trained a five-state context-dependent phoneme Hidden Semi-Markov Model (HSMM) [141] for the speakers RMS and SLT. Di-

Table 3. Comparison of three proposed methods using the MS in term of portability, speech quality, and computation time in synthesis. 120 ms is the computation time when the segment-level post-filter is used[†].

	Portability	Speech quality	Computation time
Post-filter	**Best**	Better	Better (120 ms)[†]
Parameter generation	Better	**Best**	Worse [‡]
Training	Worse	Better	**Best** (5 ms)

agonal covariance matrices were used in the HSMM. We used 593 sentences from subset A for training and 100 sentences from subset B for evaluation. The feature vector consisted of spectral and excitation parameters and their delta and delta-delta features. MSD-HMM was used for modeling F_0 contours. For GMM-based VC, we trained a 64-mixture and 16-mixture GMMs for spectrum and aperiodicity, respectively. The covariance matrices and cross-covariance matrices were diagonal matrices. F_0 was linearly converted. The GMMs were for RMS-to-SLT and SLT-to-RMS conversion. We used 50 sentences from subset A for training and 100 sentences from subset B for evaluation. The feature vector consisted of spectral and excitation parameters and their delta features.

We evaluated the following systems:

BSC: generation using $L^{(\mathrm{syn})}_{\mathrm{basic}}$
GV: generation using $L^{(\mathrm{syn})}_{\mathrm{gv}}$
MS: proposed generation using $L^{(\mathrm{syn})}_{\mathrm{ms}}$
nat: natural speech parameters

We first conducted an objective evaluation with the likelihoods used in the algorithms[37]. Then, we conducted subjective evaluations on speech quality. The traditional training using $L^{(\mathrm{trn})}_{\mathrm{basic}}$ was performed. These systems were used to generate the spectrum and F_0 of the synthetic speech. The "GV" system was used to generate the aperiodicity of the synthetic speech.

5.5.2 Objective evaluation for parameter generation algorithm

The generation algorithms were evaluated using the HMM/GMM, GV, and MS likelihoods for the generated trajectories. Additionally, we estimated the log-MS $\log s_d(f)$ probability density function and also calculated its likelihood to deeply discuss the results. Note that the HMM/GMM likelihood was normalized by the total number of frames T, and the MS and log-MS likelihoods were similarly

[37] The 50 Hz LPF was not applied to the parameters in the objective evaluation.

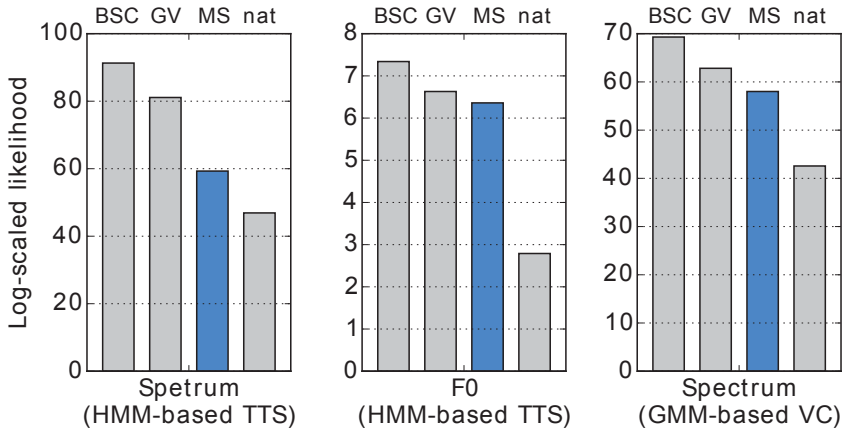


Figure 84. HMM/GMM likelihoods for parameter sequences generated by the several training algorithms.

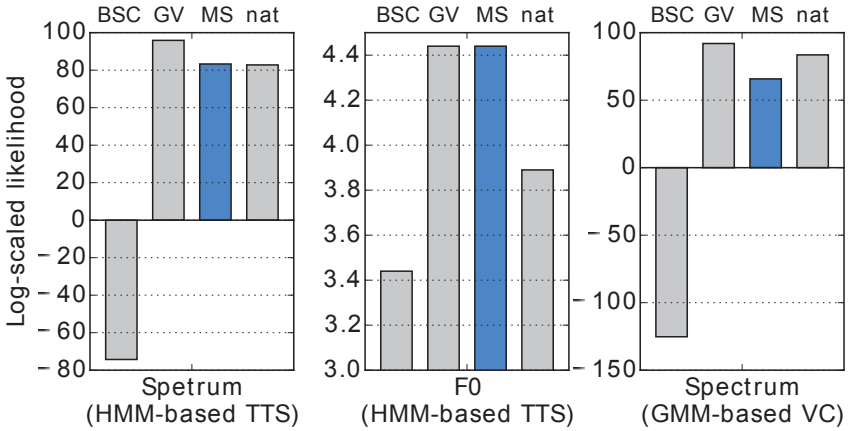


Figure 85. GV likelihoods for parameter sequences generated by the several training algorithms.

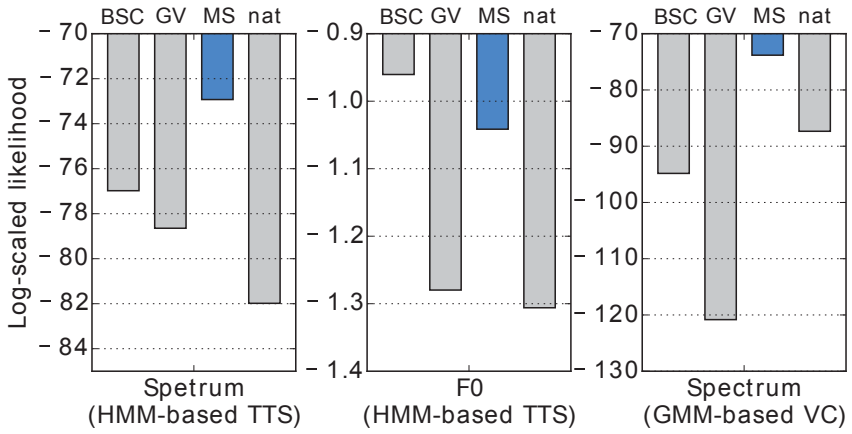


Figure 86. MS likelihood for parameter sequences generated by the several training algorithms.

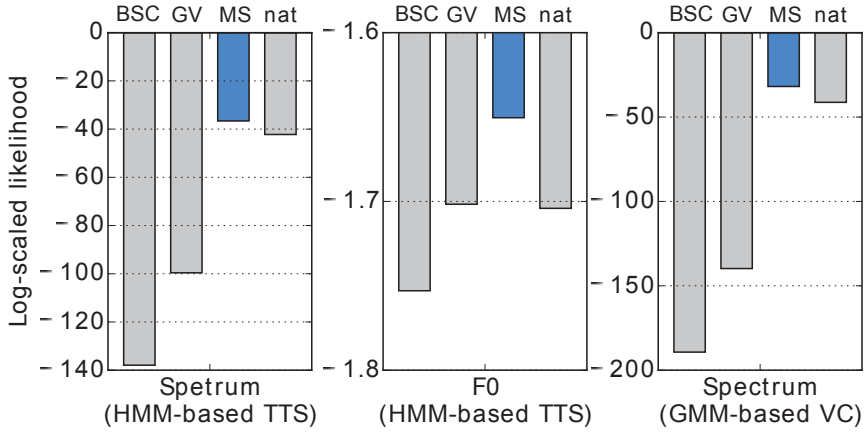


Figure 87. Log-MS likelihood for parameter sequences generated by the several training algorithms.

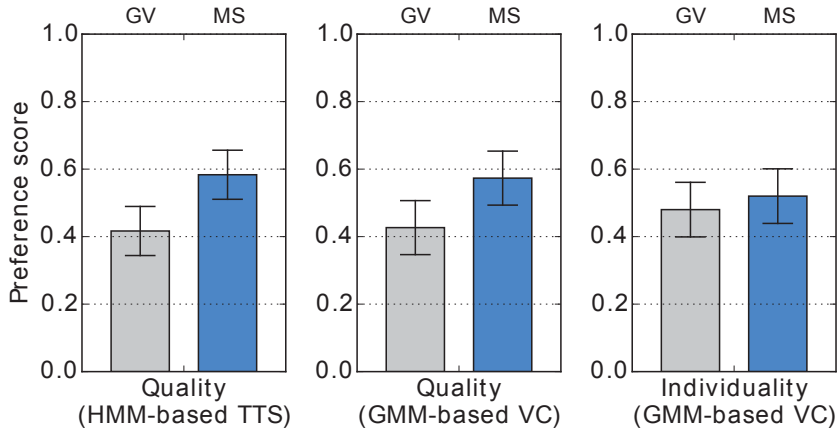


Figure 88. Results of subjective evaluation on speech quality and speaker individuality for confirming the effectiveness of the proposed speech parameter generation algorithm considering the MS ("MS"). We can find that "MS" achieved the best scores.

normalized by the number of dimension D_{s}. Figures 84–87 show these four types of likelihoods. Note that these results are the averages of the two speakers.

HMM/GMM and GV likelihoods: We can see in Fig. 84 that the HMM/GMM likelihoods for both the spectral and F_0 components in the proposed algorithm ("MS") are lower than those of the traditional algorithm ("BSC") and algorithm with the GV ("GV"), but higher than those of natural speech parameter trajectories ("nat"). For the GV likelihoods shown in Fig. 85, "MS" can effectively recovers the GV likelihood as in the "GV." These results demonstrate that the proposed generation algorithm preserves the conventional criteria.

MS and log-MS likelihoods: Fig. 86 shows that the MS likelihood of "MS" is larger than that of "BSC" and "GV" in both HMM-based TTS (left) and GMM-based VC (right). Regarding the F_0 component, the MS likelihood of "MS" is larger than that of the "GV." A comparison of the results for "MS" and "GV," which use the same initial parameter trajectories but different objective functions, reveals that the proposed objective function is effective at recovering the MS likelihood. In contrast, in both the spectral and F_0 components in HMM-based TTS, the MS likelihoods of "BSC" are higher than those of "nat"[38]. These results are hard to interpret. To analyze them, the Fig. 87 illustrates log-MS likelihoods. They show more reasonable results implying that the probability density of the MS is well modeled by the Gaussian distribution in the logarithm domain. Nevertheless, we found that there was no perceptual quality difference

[38] The MS likelihood of "BSC" is lower than that of "nat" in GMM-based VC. This is because the overall likelihood of "BSC" of GMM-based VC tends to be lower than that of HMM-based TTS.

between the MS modeling and the log-MS modeling in the proposed parameter generation algorithm.

5.5.3 Subjective evaluation for speech parameter generation algorithm

We conducted a preference test (AB test) on speech quality with eight listeners. Synthetic speech pairs of "GV" and "MS" were presented to the listeners in random order[39]. The listeners were asked which sample sounded better in terms of speech quality. Similarly, an XAB test on speaker individuality was conducted with six listeners, wherein the analysis-synthesized speech was used as the reference "X." Because the results for the two speakers were similar [148], we here show only the results for one speaker.

The results of the preference test is illustrated in Fig. 88. We can see that the score of "MS" is higher than that of "GV." This means that the proposed algorithm can generate better-quality synthetic speech than the conventional algorithm using GV can. However, unfortunately, there is no significant difference in the preference test on speaker individuality. We suppose that there were no cues for individuality at the higher modulation frequency recovered by the MS.

5.5.4 Comparison of the post-filter and speech parameter generation with the MS

We conducted a preference test (AB test) on speech quality in order to compare a MS-based post-filter (Chapter 4) and speech parameter generation considering the MS. The speaker, training/evaluation data, and speech parameters were the same to those used in **Section 4.6**. The proposed segment-level MS-based post-filter was used, and filter-related parameters (e.g., window length) were the same to those in **Section 4.6.4**.

The following systems in HMM-based TTS were evaluated:

GV+MSPF: MS-based post-filter after speech parameter generation using $L_{\mathrm{gv}}^{(\mathrm{syn})}$

MS: proposed speech parameter generation using $L_{\mathrm{ms}}^{(\mathrm{syn})}$

Note that "GV+MSPF" in this evaluation is equal to "HMM+GV+MS" labeled in **Section 4.6**. We applied 50 Hz cut-off LPF to generated and filtered speech parameter sequences. Because the effect of the MS compensation is significant for spectral parameters, we have used these methods for only spectral parameters. GV and MSs were not used in speech parameter generation and post-filtering process for other speech parameters.

[39] We didn't use "BSC" in the subjective evaluation because it is known that "GV" is better in quality than "BSC."

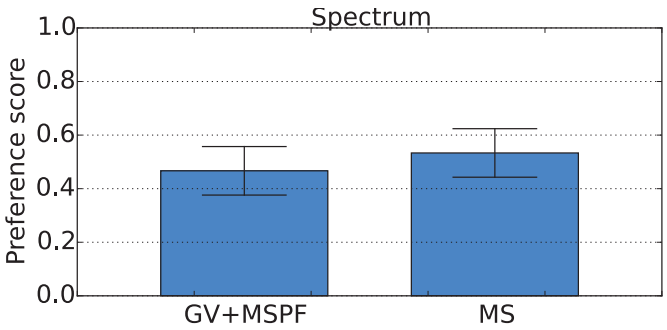


Figure 89. Results of subjective evaluation on speech quality for comparing the post-filter ("GV+MSPF") and speech parameter generation ("MS") using the MS. There is no difference between their scores.

Fig. 89 shows the result of the preference AB test by six listeners. Because there is no significant difference, these two methods have the same capability in quality improvements.

5.5.5 Experimental conditions for training algorithm

The speech features were the same as those used in the previous evaluation, but the length of the DFT used to calculate the MS was 2048. The likelihood weight ω_{v} was set to 0.5 [3] for HMM-based TTS and 1.0 [2] for GMM-based VC. ω_{s} was set to 1.0. D'_{s} for the spectrum and F_0 were set to $D_{\mathrm{s}}/2$ (= 50Hz) and $D_{\mathrm{s}}/10$ (= 10Hz) [145], respectively.

We trained HSMMs with continuous F_0 modeling for speaker RMS in HMM-based TTS, and SLT-to-RMS GMMs in GMM-based VC. The training and evaluation data were the same as in the previous experiment.

We compared the following training algorithms:

BSC: training using $L^{(\mathrm{trn})}_{\mathrm{basic}}$
TRJ: training using $L^{(\mathrm{trn})}_{\mathrm{trj}}$
GV: training using $L^{(\mathrm{trn})}_{\mathrm{gv}}$
MS: proposed training using $L^{(\mathrm{trn})}_{\mathrm{ms}}$

The evaluation was conducted in a similar way to the previous one. The systems were used to train the HMM/GMM for the spectrum and F_0. The "BSC" system was used to train for aperiodicity. The traditional generation algorithm using $L^{(\mathrm{trn})}_{\mathrm{basic}}$ was used in the synthesis stage. Note that the voiced/unvoiced regions of the F_0 contour never changed in all the training algorithms.

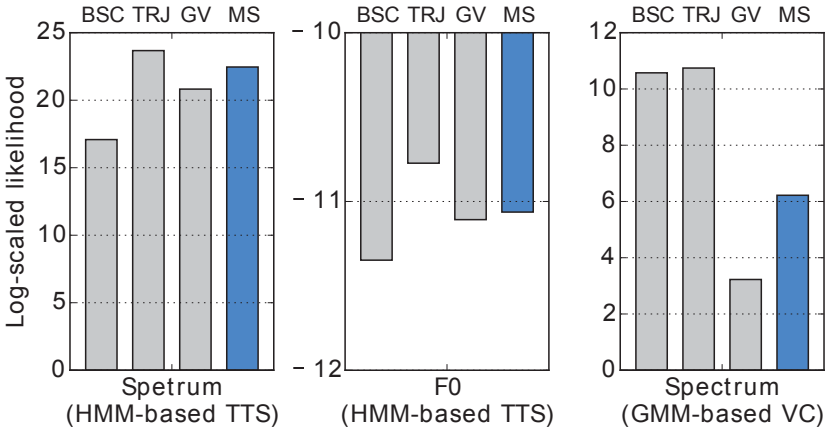


Figure 90. Trajectory likelihoods for natural speech parameters in HMM-based TTS or GMM-based VC trained using the several training algorithms. Blue bars indicate the proposed training algorithm. The trajectory training algorithm ("TRJ") for GMM-based VC is also the proposed in this thesis, but the bar is gray-colored.

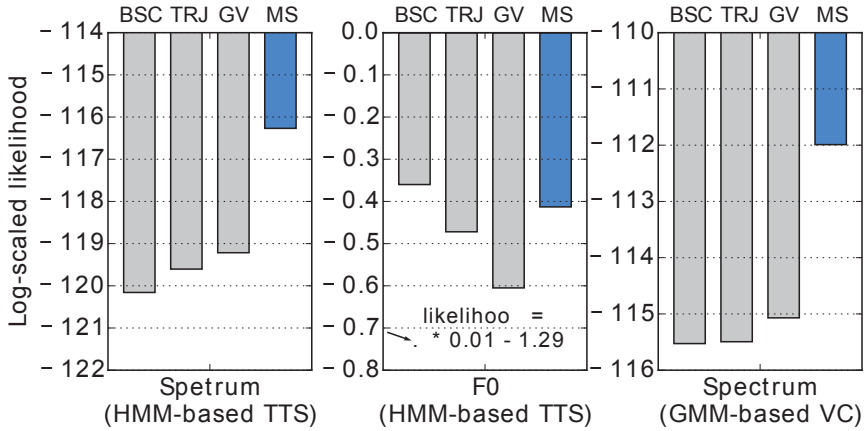


Figure 91. MS likelihoods for natural speech parameters in HMM-based TTS or GMM-based VC trained using the several training algorithms. Blue bars indicate the proposed training algorithm.

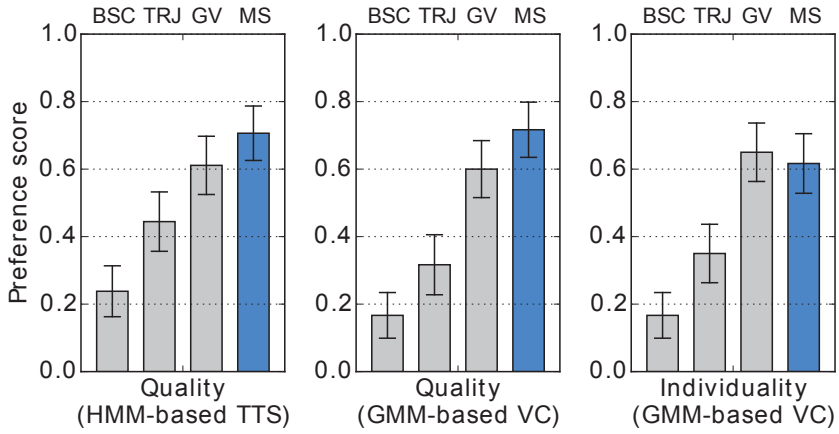


Figure 92. Results of subjective evaluation on speech quality and speaker individuality for confirming the effectiveness of the proposed training algorithm constraint with the MS ("MS"). The trajectory training algorithm ("TRJ") for GMM-based VC is also proposed in this thesis, but the bar is gray-colored. We can find that "MS" achieves the best scores.

5.5.6 Objective evaluation of training algorithms

Fig. 90 and Fig. 91 illustrate the trajectory likelihood and the MS likelihood for the natural parameter trajectories of the evaluation data. The trajectory likelihood is normalized by the total number of frames T. In GMM-based VC, the trajectory training ("TRJ") slightly improves the MS likelihood as well as the trajectory likelihood compared with the basic training ("BSC"). This result shows that the proposed trajectory training models the parameter trajectories more accurately than the traditional training.

The MS-constrained trajectory training ("MS") improves the MS likelihood more than the other algorithms do in HMM-based TTS and GMM-based VC. This result demonstrates that "MS" recovered the MS of the generated parameter trajectory. By comparison, the conventional GV-constrained training ("GV") significantly degraded the trajectory likelihood in GMM-based VC.

5.5.7 Subjective evaluation of training algorithm

We conducted the subjective evaluations in the same manner as above. Six listeners participated in each evaluation.

The results are illustrated in Fig. 92. IN GMM-based VC, it is observed that "TRJ" has higher scores than "BSC" in term of both the speech quality and the speaker individuality. Therefore, we can confirm the quality improvements by the proposed trajectory GMM training in GMM-based VC. Moreover, we can see that "MS" achieves the best scores than others in term of the speech quality in

both HMM-based TTS and GMM-based VC. This result demonstrates that the proposed MS-constrained trajectory training yields the best performance among the training methods.

5.6 Summary of this chapter

This chapter integrated the Modulation Spectrum (MS) into speech synthesis framework to jointly optimize the basic criteria and the proposed MS criterion that is effective to alleviate the over-smoothing effect.

Section 5.2: We have re-defined the MS to consider only the lower modulation frequency components that are dominant in speech perception.

Section 5.3: The MS was straightforwardly integrated into the speech parameter generation stage that causes the over-smoothing effect. The synthetic speech parameter sequences were generated by iteratively maximizing the weighted sum of the HMM/GMM likelihoods and the MS likelihood. Because the MS is extension of the GV, the proposed generation algorithm can perform not only the explicit MS compensation but also implicit GV compensation.

Section 5.4: Instead of integrating the MS into the parameter generation, we have integrated it into the training stage in order to produce high-quality speech with adopting the basic computationally-efficient generation algorithm. The trajectory GMM have been proposed in advance as the same as the trajectory HMM. The HMM/GMM parameter sets have been trained by maximizing the weighted sum of the trajectory HMM/MGM likelihoods and the MS likelihood. Because they are trained so that the synthetic speech parameter sequence generated using the basic generation algorithm has the compensated MS, we don't need to consider the MS term in the synthesis stage but the quality gain by the MS is benefited.

Section 5.5: We have conducted the several experimental evaluation. The experimental results for the proposed generation algorithm have demonstrated the quality gain overcoming the conventional parameter generation considering the GV. Also, the results for the proposed training algorithm have demonstrated that the proposed training algorithm overcomes the several training algorithms in synthetic speech quality.

Chapter

6

Conclusion

6.1 Contribution

In this thesis, we have addressed improvements of the synthetic speech quality in statistical parametric speech synthesis including Hidden Markov Model (HMM)-based Text-To-Speech (TTS) and Gaussian Mixture Model (GMM)-based Voice Conversion (VC). They have the promising techniques to control the characteristics of the synthesized speech beyond the limitations of unit selection synthesis, but the critical drawback in the statistical parametric speech synthesis is significant degradation in synthetic speech quality. The synthetic speech often sounds muffled, and we can still distinguish the synthetic speech from natural speech. There are three main reasons causing the quality degradation: parameterization errors in the analysis/synthesis stage, insufficient modeling in the training stage, and over-smoothing effect in the synthesis stage.

Chapter 2 has described the basic speech synthesis frameworks and the conventional methods for better training and synthesis. We have explained that there is a trade-off between unit selection synthesis and statistical parametric speech synthesis on the synthetic speech quality and the flexibility. In order to alleviate the averaging effect in the modeling process, we have presented 2 approaches that introduced the idea of unit selection synthesis into HMM-based TTS. The speech quality in synthetic speech are dramatically improved, but the flexibility of original HMM-based TTS is lost. A trajectory HMM has been presented to train the HMM parameters considering the temporal dependency as the similar as in the synthesis stage. Finally, Global Variance (GV) have been introduce to quantify the over-smoothing effect observed in the synthesis stage.

In Chapter 3, we have proposed statistical sample-based speech synthesis with rich context models to address the insufficiency modeling causing the quality degradation in synthetic speech. We first have applied the rich context modeling originally proposed in HMM-based TTS to GMM-based VC, then, the have reformulated the Rich context GMM (R-GMM) using the rich context models belonging to the same acoustic sub-space. The synthetic speech parameter sequences have been generated by iteratively maximizing the likelihood. The generation process have been initialized by the use of the less-smoothed speech parameters generated from the statistics of the further divided sub-space. Compared to the basic HMM-based TTS and GMM-based VC, the proposed methods can improve the quality in synthetic speech by introducing the modeling of the individual speech feature segments. Also, compared to the conventional hybrid methods combining unit selection synthesis and HMM-based TTS, the proposed methods have preserved the advantage of flexible acoustic modeling provided by the basic HMM-based TTS and GMM-based VC because the proposed methods don't have any constraint used in the conventional methods. We have conducted several experiments to confirm the effectiveness of the proposed methods in HMM-based TTS. The experimental results have demonstrated: (1) the use of approximation

with a single Gaussian component sequence yields better synthetic speech quality than the use of EM algorithm, (2) the state-based model selection yields quality improvements at the same level as the frame-based model selection, (3) the use of the initial parameters generated from the over-trained speech probability distributions is very effective to further improve speech quality, and (4) the proposed methods for spectral and F_0 components yields significant improvements in synthetic speech quality compared with the basic HMM-based TTS.

Chapter 4 has introduced the Modulation Spectrum (MS) of speech parameter trajectory as a new feature to effectively quantify the over-smoothing effect, which is cause of the synthetic speech quality degradation. We have further proposed the MS-based post-filters for high-quality speech synthesis. We have defined the MS as the log-scaled power spectrum of the speech parameter sequence, have used it to find the over-smoothing effect. 2 types of the MS-based post-filters (utterance- and segment-level post-filters) have been proposed. In the synthesis stage, the generated speech parameters have been filtered utterance by utterance to make the MS close to natural MS. In discussion, we have clarified that (1) the MS is the mathematical extension of the GV, and (2) the MS-based post-filter generates the fluctuating speech parameter sequence. We have conducted experimental evaluation to confirm the effectiveness of the proposed post-filters, and have demonstrated: (1) the proposed utterance-level post-filter achieves better quality for spectrum, F_0, and HMM-state duration in HMM-based TTS, (2) the proposed segment-level post-filter capable of achieving low-delay synthesis also yields significant improvements in synthetic speech quality, (3) the proposed utterance-level and segment-level post-filters have the capability in the speech quality improvement, and (4) the proposed post-filters are also effective in not only HMM-based TTS but also GMM-based VC and CLUSTERGEN.

Chapter 5 has integrated the Modulation Spectrum (MS) into speech synthesis framework to jointly optimize the basic criteria and the proposed MS criterion that is effective to alleviate the over-smoothing effect. Before the integration, we have re-defined the MS to consider only the lower modulation frequency components that are dominant in speech perception. The MS was first straightforwardly integrated into the speech parameter generation stage that causes the over-smoothing effect. The synthetic speech parameter sequences were generated by iteratively maximizing the weighted sum of the HMM/GMM likelihoods and the MS likelihood. Because the MS is extension of the GV, the proposed generation algorithm can perform not only the explicit MS compensation but also implicit GV compensation. Then, we have integrated it into the training stage in order to produce high-quality speech with adopting the basic computationally-efficient generation algorithm. The trajectory GMM have been proposed in advance as the same as the trajectory HMM. The HMM/GMM parameter sets have been trained by maximizing the weighted sum of the trajectory HMM/MGM likelihoods and the MS likelihood. Because they are trained so that the synthetic

speech parameter sequence generated using the basic generation algorithm has the compensated MS, we don't need to consider the MS term in the synthesis stage but the quality gain by the MS is benefited. We have conducted the several experimental evaluation to confirm the quality gain by the proposed methods.

6.2 Future work

As mentioned in **Section 1**, research is an action toward blurring such boundaries between objects. Fully developed high-quality speech synthesis can remove the boundaries between a human and a computer, or between human beings. In the future, every object (not only human beings but also computers) living in such future will not aware of differences between each others' speech production, and speech synthesis will be a black box or a magic. The final goal for high-quality speech synthesis is to realize such future.

Toward the future, this thesis addressed high-quality speech synthesis. The quality gain was confirmed by preference AB tests, and we have observed 1.0 MOS gain (by the MS-based post-filters) as shown in the experimental evaluation in **Section A.3**. However, the *real* quality of synthetic speech is still far from that of natural speech. For example, assuming the 5-point MOS scores of natural speech is 5.0, the scores of synthetic speech using all methods proposed in this thesis will be lower than 4.0 [40]. Consequently, there are many issues to be solved for high-quality speech synthesis.

What are the meaningful factors that are still different between natural/synthetic speech? In this thesis, we have found that the MSs are different between natural and synthetic speech parameters. Its effect in quality was confirmed but it is still uncleared why the MS causes such the effects. We should investigate more from the perspectives of the physical constraint of the speech production and the auditory characteristics. Also, we should investigate how the MS is modeled. In this thesis, we modeled power spectra of each modulation frequency bin. This is too much strong to constrain the speech parameter sequence,

As described in **Section 4.5**, even if the MSs are compensated, something is still different between natural speech and synthetic speech. Inefficient way to find the difference is to take account of an anti-spoofing technique. An anti-spoofing technique [149, 150] is to detect the spoofing attack by speech synthesis. As we described in Chapter 4, the modulation spectrum is originally used to distinguish natural speech from synthetic speech in the anti-spoofing, but it have became a non-meaningful feature when we consider it in the speech synthesis

[40] The score using the MS-based post-filter is around 3.5, as described in **Section A.3**.

side. Similarly, other features found in the anti-spoofing side will be effective for training the better speech synthesis system.

The use of better acoustic models As demonstrated in **Section A.6** and also reported in [120], we observe quality degradation even if we can select the best rich context models in HMM-based TTS. This is caused by the HMM state-level temporal quantization and the use of macro-level context labels. Also, HMM-based TTS and GMM-based VC frameworks are insufficient for integrating the MS as discussed in **Section 5.4**.

DNNs have a capability of solving these limitation. For example, frame-level contexts are acceptable in DNN-based acoustic modeling. Also, we can integrate the MS into the DNN training by several methods, such as minimum generation error training [151], multi-task learning [152] and trajectory modeling [153].

Adaptation using rich context models and MS models This thesis confirmed quality gain by rich context models or MS models, but we need to investigate their adaptation methods in order to benefit of statistical parametric speech synthesis.

For rich context models, it is impossible to estimate the individual adaptation rules for each rich context models. Therefore, what we need to consider is how we apply the standard adaptation techniques with the suitable constraints.

Because the MS is the higher-ordered feature strongly depending on the speaker, it is inappropriate to perform the MS-integrated algorithms using the few amounts of the speech data. We need to propose the MS model adaptation technique and the HMM/GMM (and DNN) adaptation techniques constrained with the MS.

Where is the upper bound in quality in the real situation? I have noted that the perceptual quality of synthetic speech is lower than that of natural speech. However, it is resulted in silence and sound-only environments that are not *real* environment for speech synthesis. Especially, I'm curious of the use of speech synthesis in multimedia including visual information. Visual information is dominant in human perception, and sound perception tends to be excessively affected by the visual information. We need to investigate the upper bound of speech quality in such situation, which means what extent we should improve the quality.

As the related topic, one of the reasons why we can distinguish synthetic speech from natural speech is that synthetic speech includes errors human beings never do. Using the better modeling and synthesis methods is, of course, an effective way, and the another way is to allow the errors but make it close to human-like errors.

Publication, Reference, and Appendix

Publication

Journal papers

1. Shinnosuke Takamichi, Tomoki Toda, Alan W. Black, Graham Neubig, Sakriani Sakti and Satoshi Nakamura, "Post-Filters to Modify the Modulation Spectrum for Statistical Parametric Speech Synthesis,"
   IEEE Transactions on Audio, Speech and Language Processing.
   (accepted, corresponds to Chapter 4)
2. Shinnosuke Takamichi, Tomoki Toda, Yoshinori Shiga, Sakriani Sakti, Graham Neubig, and Satoshi Nakamura,
   "Parameter Generation Methods with Rich Context Models for High-Quality and Flexible Text-To-Speech Synthesis,"
   IEEE Journal of Selected Topics of Speech Processing, 2014.
   (corresponds to Chapter 3)

International conferences

1. Shinnosuke Takamichi, Kazuhiro Kobayashi, Kou Tanaka, Tomoki Toda, and Satoshi Nakamura,
   "The NAIST Text-to-Speech System for the Blizzard Challenge 2015,"
   Proc. of Blizzard Challenge Workshop, Berlin, Germany, Sep., 2015.
   (corresponds to Appendix)
2. Shinnosuke Takamichi, Tomoki Toda, Alan W. Black, and Satoshi Nakamura,
   "Modulation Spectrum-Constrained Trajectory Training Algorithm for HMM-Based Speech Synthesis,"
   Proc. of INTERSPEECH, pp. 1206-1210, Dresden, Germany, Sep., 2015.
   (corresponds to Chapter 5)
3. Shinnosuke Takamichi, Tomoki Toda, Alan W. Black, and Satoshi Nakamura,
   "Parameter generation algorithm considering modulation spectrum for HMM-based speech synthesis,"
   Proc. of ICASSP, pp. 4210-4214, Brisbane, Australia, Apr., 2015.
   (corresponds to Chapter 5)
4. Shinnosuke Takamichi, Tomoki Toda, Alan W. Black, and Satoshi Nakamura,
   "Modulation Spectrum-Constrained Trajectory Training for GMM-Based Voice Conversion,"
   Proc. of ICASSP, pp. 4859-4863, Brisbane, Australia, Apr., 2015.
   (corresponds to Chapter 5)
5. Shinnosuke Takamichi, Tomoki Toda, Alan W. Black, and Satoshi Nakamura,
   "Modulation Spectrum-based Post-filter for GMM-based Voice Conversion,"
   Proc. of APSIPA, Siem Reap, Cambodia, Dec., 2014.
   (corresponds to Chapter 4)
6. Shinnosuke Takamichi, Tomoki Toda, Alan W. Black, and Satoshi Nakamura,
   "Modified Modulation Spectrum-based Post-filter for HMM-based Speech Synthesis,"
   Proc. of GlobalSIP, pp. 710-714, Atlanta, U.S.A., Dec., 2014.
   (corresponds to Chapter 4)
7. Shinnosuke Takamichi, Tomoki Toda, Graham Neubig, Sakriani Sakti, and Satoshi Nakamura,

"A Postfilter to Modify The Modulation Spectrum in HMM-based Speech Synthesis,"
Proc. of ICASSP, pp. 290-294, 2014.
(corresponds to Chapter 4)

8. Shinnosuke Takamichi, Tomoki Toda, Yoshinori Shiga, Graham Neubig, Sakriani Sakti, and Satoshi Nakamura,
"Improvements to HMM-based Speech Synthesis Based on Parameter Generation with Rich Context Models,"
Proc. of INTERSPEECH, pp. 364-368, 2013.
(corresponds to Chapter 3)
9. Shinnosuke Takamichi, Tomoki Toda, Yoshinori Shiga, Hisashi Kawai, Sakriani Sakti, and Satoshi Nakamura,
"An Evaluation of Parameter Generation Methods with Rich Context Models in HMM-Based Speech Synthesis,"
Proc. of INTERSPEECH, Portland, U.S.A., Sep., 2012.
(corresponds to Chapter 3)

Technical reports

1. Shinnosuke Takamichi, Tomoki Toda, Alan W. Black, Satoshi Nakamura,
"Quality Improvements Approaches Based on the Modulation Spectrum to Statistical Parametric Speech Synthesis,"
IPSJ SIG Tech. Rep., 2015-MUS-107, pp. 1-4, Mar., 2015.
(in Japanese, corresponds to Chapter 4 and 5)
2. Shinnosuke Takamichi, Tomoki Toda, Alan W. Black, Satoshi Nakamura,
"Modulation Spectrum-Constrained Trajectory Training Algorithm for Statistical Parametric Speech Synthesis,"
IEICE Tech. Rep., SP2014-140, pp. 31-36, Mar., 2015.
(in Japanese, corresponds to Chapter 5)
3. Shinnosuke Takamichi, Tomoki Toda, Alan W. Black, Satoshi Nakamura,
"Speech Parameter Generation Algorithm Considering Modulation Spectrum for Statistical Parametric Speech Synthesis,"
IPSJ SIG Tech. Rep., 2015-SLP-105, No. 1, pp. 1-6, Feb., 2015.
(in Japanese, corresponds to Chapter 5)
4. Shinnosuke Takamichi, Tomoki Toda, Graham Neubig, Sakriani Sakti, Satoshi Nakamura,
"Postfilter Based on Modulation Spectrum in HMM-Based Speech Synthesis,"
IEICE Tech. Rep., SP2013-74, pp. 19-24, Nov., 2013.
(in Japanese, corresponds to Chapter 4)
5. Shinnosuke Takamichi, Tomoki Toda, Yoshinori Shiga, Sakriani Sakti, Graham Neubig, Satoshi Nakamura,
"F0 Contour Generation Using Rich Context Models in HMM-Based Speech Synthesis,"
IEICE Tech. Rep., SP2012-104, pp. 37-42, Jap., 2013.
(in Japanese, corresponds to Chapter 3)
6. Shinnosuke Takamichi, Tomoki Toda, Yoshinori Shiga, Sakriani Sakti, Graham Neubig, Satoshi Nakamura,
"Improvements of HMM-based Speech Synthesis Using Rich Context Models,"
IEICE Tech. Rep., SP2012-78, pp.37-42, Nov., 2012.
(in Japanese, corresponds to Chapter 3)

7. Shinnosuke Takamichi, Tomoki Toda, Yoshinori Shiga, Hisashi Kawai, Sakriani Sakti, Graham Neubig, Satoshi Nakamura,
"A Study on HMM-Based Speech Synthesis Using Rich Context Models,"
IPSJ SIG Tech. Rep., SLP-10, No. 10, pp. 1-6, Jul., 2012.
(in Japanese, corresponds to Chapter 3)

Domestic conferences

1. Shinnosuke Takamichi, Tomoki Toda, Alan W. Black, Satoshi Nakamura,
"Modulation Spectrum-Constrained Trajectory Training Algorithm in Statistical Parametric Speech Synthesis,"
Proc. of Spring Meeting, Acoust. Soc. Jpn., 2-2-3, 2015.
(in Japanese, corresponds to Chapter 5)
2. Shinnosuke Takamichi, Tomoki Toda, Alan W. Black, Satoshi Nakamura,
"Parameter Generation Algorithm Considering Modulation Spectrum in Statistical Parametric Speech Synthesis,"
Proc. of Spring Meeting, Acoust. Soc. Jpn., 2-2-2, 2015.
(in Japanese, corresponds to Chapter 5)
3. Shinnosuke Takamichi, Tomoki Toda, Graham Neubig, Sakriani Sakti, Satoshi Nakamura,
"HMM-based speech synthesis considering modulation spectrum,"
Proc. of Autumn Meeting, Acoust. Soc. Jpn., 2-7-10, 2013.
(in Japanese, corresponds to Chapter 4)
4. Shinnosuke Takamichi, Tomoki Toda, Yoshinori Shiga, Sakriani Sakti, Graham Neubig, Satoshi Nakamura,
"Quality Improvements with Rich Context Models for Spectral and F0 Components in HMM-based Speech Synthesis,"
Proc. of Spring Meeting, Acoust. Soc. Jpn., 2-7-10, 2013.
(in Japanese, corresponds to Chapter 3)
5. Shinnosuke Takamichi, Tomoki Toda, Yoshinori Shiga, Sakriani Sakti, Graham Neubig, Satoshi Nakamura,
"A Study on a Selection Method of Rich Context Models in HMM-based Speech Synthesis,"
Proc. of Autumn Meeting, Acoust. Soc. Jpn., 2-2-1, 2012.
(in Japanese, corresponds to Chapter 3)
6. Shinnosuke Takamichi, Tomoki Toda, Yoshinori Shiga, Hisashi Kawai, Sakriani Sakti, Satoshi Nakamura,
"A Study on the Effectiveness of Full-context Models with Tied-covariance Matrices in HMM-based Speech Synthesis,"
Proc. of Spring Meeting, Acoust. Soc. Jpn., 1-11-4, 2012.
(in Japanese, corresponds to Chapter 3)

Awards

1. 2014 IEICE ISS Young Researcher's Award in Speech Field, Aug., 2015.
2. 30th TELECOM System Technology Award for Students from TAF, Mar., 2015.
3. IEEE Kansai Section Student Paper Award, Feb., 2015.
4. APSIPA ASC 2014 Best Paper Award, Dec., 2014.

5. The 8th IEEE Japan SPS Outstanding Student Paper Award, Nov., 2014.
6. The 35th Awaya Prize Young Researcher Award of ASJ, Mar., 2014.
7. Award of Campus Venture Grand Prix in Osaka, Jan., 2014.
8. The 7th Best Student Presentation Award of ASJ, Sep., 2013.
9. The Best Student of Nara Institute of Science and Technology, Jul., 2013.

Articles

1. Shinnosuke Takamichi, Tomoki Toda, Alan W. Black, and Satoshi Nakamura,
"Post-Filter Using Modulation Spectrum as a Metric to Quantify Over-Smoothing Effects in Statistical Parametric Speech Synthesis,"
APSIPA newsletter, No. 9, pp. 14-16, 2015. (Invited article, corresponds to Chapter 4)
2. Shinnosuke Takamichi,
"Coffee break, Q&A,"
Acoustical Science and Technology, Vol.70, No. 8, Aug., 2014.
(in Japanese)

Software

1. HMM-based Speech Synthesis System (HTS) [1]
(I provided the segment-level MS-based post-filter proposed in Chapter 4 to HTS ver. 2.3 beta)

Research talks

1. Shinnosuke Takamichi,
"High-quality Statistical Parametric Speech Synthesis Considering the Modulation Spectrum,"
ICS Research Seminar, Technical University Munich, Sep., 2015.
2. Shinnosuke Takamichi,
"Modulation Spectrum-based Approaches for High-Quality Speech Synthesis,"
IEEE MileStone Pre-Event, Kyoto, Japan, May, 2015.
3. Shinnosuke Takamichi,
"Modulation Spectrum-based Approach to High-quality Statistical Parametric Speech Synthesis,"
Techtalk, Google London, UK, Nov., 2014.
4. Shinnosuke Takamichi,
"Modulation Spectrum-based Approach to High-quality Statistical Parametric Speech Synthesis,"
CUED seminars, Univ. of Cambridge, UK, Nov., 2014.
5. Shinnosuke Takamichi,
"Modulation Spectrum-based Approach to High-quality Statistical Parametric Speech Synthesis,"
Speech! Meeting, Edinburgh Univ., UK, Nov., 2014.
6. Shinnosuke Takamichi,
"Modulation Sectrum (MS) in HMM-based speech synthesis,"
Sphinx lunch, Carnegie Mellon Univ., U.S.A., Mar., 2014.

Related publications

International conferences

1. Yuri Nishigaki, Shinnosuke Takamichi, Tomoki Toda, Graham Neubig, Sakti Sakriani, and Satoshi Nakamura,
"Prosody-Controllable HMM-Based Speech Synthesis Using Speech Input,"
Proc. of 2015 First Workshop on MLSLP, Aizu, Japan, Sep., 2015.
2. Quoc Truong Do, Sakriani Sakti, Shinnosuke Takamichi, Graham Neubig, Tomoki Toda, and Satoshi Nakamura,
"Preserving Word-level Emphasis in Speech-to-speech Translation using Linear Regression HSMMs,"
Proc. of INTERSPEECH, pp. 3665-3669, Dresden, Germany, Sep., 2015.
3. Yuji Oshima, Shinnosuke Takamichi, Tomoki Toda, Graham Neubig, Sakriani Sakti, and Satoshi Nakamura,
"Non-native Speech Synthesis Preserving Speaker Individuality Based on Partial Correction of Prosodic and Phonetic Characteristics,"
Proc. of INTERSPEECH, pp. 299-303, Dresden, Germany, Sep., 2015.
4. Nozomi Jinbo, Shinnosuke Takamichi, Tomoki Toda, Graham Neubig, Sakriani Sakti, and Satoshi Nakamura,
"A Hearing Impairment Simulation Method Using Audiogram-based Approximation of Auditory Characteristics,"
Proc. of INTERSPEECH, pp. 490-494, MAX Atria, Singapore, Sep., 2014.
5. Takatomo Kano, Shinnosuke Takamichi, Sakriani Sakti, Graham Neubig, Tomoki Toda, and Satoshi Nakamura,
"Generalizing Continuous-space Translation of Paralinguistic Information,"
Proc. of INTERSPEECH, pp. 2614-2618, 2013.
6. Takatomo Kano, Sakriani Sakti, Shinnosuke Takamichi, Graham Neubig, Tomoki Toda, and Satoshi Nakamura,
"A Method For Translation of Paralinguistic Information,"
Proc. of IWSLT, pp. 158-163, 2012.

Technical reports

1. Shinya Kura, Shinnosuke Takamichi, Tomoki Toda, Graham Neubig, Sakriani Sakti, Satoshi Nakamura,
"Evaluation and Analysis of Duration Correction for Non-Native Speech Based on Waveform Modification,"
IEICE Tech. Rep., Dec., 2015. (in Japanese)
2. Yuji Oshima, Shinnosuke Takamichi, Tomoki Toda, Sakriani Sakti, Graham Neubig, Satoshi Nakamura,
"English-Read-By-Japanese Speech Synthesis Preserving Speaker Individuality Based on Partial Correction of Prosody and Phonetic Sounds and Effects of English Proficiency Level on Its Performance,"
IPSJ SIG Tech. Rep., SLP-105, pp. 1–6, Feb., 2015. (in Japanese)
3. Shinnosuke Takamichi, Yuji Oshima, Tomoki Toda, Graham Neubig, Sakriani Sakti, Satoshi Nakamura,
"A Study on Computer Assisted Language Learning Using English-Read-By-Japanese Speech Synthesis Techniques,"
JSiSE research report, Vol. 29, No. 5, pp. 111-116, Jan., 2015. (in Japanese)

4. Yuri Nishigaki, Shinnosuke Takamichi, Tomoki Toda, Graham Neubig, Sakriani Sakti, Satoshi Nakamura,
"HMM-Based Speech Synthesis System with Prosody Modification Based on Speech Input,"
IEICE Tech. Rep., SP2014-115, pp. 81-86, Dec., 2014. (in Japanese)
5. Yuji Oshima, Shinnosuke Takamichi, Tomoki Toda, Graham Neubig, Sakriani Sakti, Satoshi Nakamura,
"Prosody Correction Preserving Speaker Individuality in English-Read-By-Japanese Speech Synthesis Based on HMM,"
IEICE Tech. Rep., SP2014-112, pp. 63-68, Dec., 2014. (in Japanese)
6. Nozomi Jinbo, Shinnosuke Takamichi, Tomoki Toda, Graham Neubig, Sakriani Sakti, Satoshi Nakamura,
"Hearing Impairment Simulation using Audiogram-based Approximation of Auditory Filter and Loudness Compensation,"
IEICE Tech. Rep., SP2013-96, pp. 1-6, Jan., 2014. (in Japanese)

Domestic conferences

1. Truong Do, Shinnosuke Takamichi, Sakriani Sakti, Graham Neubig, Tomoki Toda, and Satoshi Nakamura,
"Word-level Emphasis Transfer in Speech-to-speech Translation,"
Proc. of Spring Meeting, Acoust. Soc. Jpn., 2016. (to appear)
2. Shinya Kura, Shinnosuke Takamichi, Tomoki Toda, Graham Neubig, Sakriani Sakti, Satoshi Nakamura, "Analysis of quality degradation caused by duration correction of non-native speech using direct waveform modification,"
Proc. of Spring Meeting, Acoust. Soc. Jpn., 2016. (in Japanese, to appear)
3. Shinnosuke Takamichi, Keita Higuchi, Satoshi Nakamura,
"Identity reflection using speech synthesis into avatar,"
23-th Workshop on Interactive Systems and Software, 1-R-16, 2015. (in Japanese)
4. Shinnosuke Takamichi, Tomoki Toda, Masanori Morise, Satoshi Nakamura,
" STRAIGHT vs. WORLD, Comparison of Speech Anaylysis-Synthesis Systems in HMM-Based Speech Synthesis,"
Proc. of Autumn Meeting, Acoust. Soc. Jpn., 1-Q-27, 2015. (in Japanese)
5. Shinnosuke Takamichi, Tomoki Toda, Graham Neubig, Sakriani Sakti, Satoshi Nakamura,
"An Evaluation of HMM-Based English-Read-by-Japanese Speech Synthesis Using English Speech Read by Japanese Junior High School Students,"
Proc. of Autumn Meeting, Acoust. Soc. Jpn., 2-5-8, 2015. (in Japanese)
6. Truong Do, Shinnosuke Takamichi, Sakriani Sakti, Graham Neubig, Tomoki Toda, and Satoshi Nakamura,
"Study on Word-Level Emphasis Across English and Japanese,"
Proc. of Autumn Meeting, Acoust. Soc. Jpn., 3-1-6, 2015.
7. Yuri Nishigaki, Shinnosuke Takamichi, Tomoki Toda, Graham Neubig, Sakriani Sakti, Satoshi Nakamura,
"Improvements to HMM-Based Speech Synthesis System with Prosody Modification Based on Speech Input,"
Proc. of Spring Meeting, Acoust. Soc. Jpn., 2-2-4, 2015. (in Japanese)

8. Yuji Oshima, Shinnosuke Takamichi, Tomoki Toda, Graham Neubig, Sakriani Sakti, Satoshi Nakamura,
"Prosody Correction Preserving Speaker Individuality in English-Read-By-Japanese Speech Synthesis and Effects of English Proficiency Level,"
Proc. of Spring Meeting, Acoust. Soc. Jpn., 1-2-9, 2015. (in Japanese)
9. Shinya Kura, Shinnosuke Takamichi, Tomoki Toda, Graham Neubig, Sakriani Sakti, Satoshi Nakamura,
"An Evaluation of Duration Correction for Non-Native Speech,"
Proc. of Spring Meeting, Acoust. Soc. Jpn., 1-2-8, 2015. (in Japanese)
10. Yuji Oshima, Shinnosuke Takamichi, Tomoki Toda, Graham Neubig, Sakriani Sakti, Satoshi Nakamura,
"Prosody Correction Preserving Speaker Individuality in English-Read-By-Japanese Speech Synthesis,"
Proc. of Autumn Meeting, Acoust. Soc. Jpn., 2-7-5, 2014. (in Japanese)
11. Yuri Nishigaki, Shinnosuke Takamichi, Tomoki Toda, Graham Neubig, Sakriani Sakti, Satoshi Nakamura,
"HMM-Based Speech Synthesis System with Speech-driven Prosody Modification,"
Proc. of Spring Meeting, Acoust, Soc. Jpn., 3-6-1, 2014. (in Japanese)
12. Nozomi Jinbo, Shinnosuke Takamichi, Tomoki Toda, Graham Neubig, Sakriani Sakti, Satoshi Nakamura,
"Audiogram-based Approximation of Loudness and Auditory Filter Characteristics in Hearing Impairment Simulation,"
Proc. of Spring Meeting, Acoust, Soc. Jpn., 2-Q-7, 2014. (in Japanese)
13. Yasuhiro Hamada, Keisuke Imoto, Shinnosuke Takamichi,
"How to survey and manage research papers,"
The 4th acoustic seminar, 2013. (in Japanese)
14. Daisuke Morikawa, Yoji Ishii, Shinnosuke Takamichi, Jorge Trevino,
"History of 3D Sounds Techniques,"
The 4th acoustic seminar, 2013. (in Japanese)
15. Nozomi Jinbo, Shinnosuke Takamichi, Tomoki Toda, Graham Neubig, Sakriani Sakti, Satoshi Nakamura,
"Hearing Impairment Simulation System with Audiogram-Based Auditory filter approximation,"
16th Young Researchers' Interactive Meeting of ASJ Kansai Section, 2013. (in Japanese)

Award

1. ASJ Kansai Section Young Researchers' Interactive Meeting Encouragement Award (Awardee: Nozomi Jinbo)

Research talks

1. Shinnosuke Takamichi,
"Hearing impairment simulation to assist hearing-impaired people,"
Sphinx lunch, Carnegie Mellon Univ., U.S.A., Sep., 2014.

Master's thesis

1. Shinnosuke Takamichi,
"Hybrid Approach to High-Quality and Flexible Text-To-Speech Synthesis,"
Master's thesis, Graduate School of Information Science, Nara Institute of Science and Technology, Mar., 2013.

References

[1] "HMM-based speech synthesis system (HTS) http://hts.sp.nitech.ac.jp/."

[2] H. Hwang, Y. Tsao, H. Wang, Y. Wang, and S. Chen, "Incorporating global variance in the training phase of GMM-based voice conversion," in *Proc. APSIPA*, Kaohsiung, Taiwan, Oct. 2013, pp. 1–6.

[3] T. Toda and S. Young, "Trajectory training considering global variance for HMM-based speech synthesis," in *Proc. ICASSP*, Taipei, Taiwan, Aug. 2009, pp. 4025–4028.

[4] Y. Sagisaka, "Speech synthesis by rule using an optimal selection of non-uniform synthesis units," in *Proc. ICASSP*, New York, U.S.A., Apr. 1988, pp. 679–682.

[5] Y. Stylianou, O. Cappe, and E. Moulines, "Continuous probabilistic transform for voice conversion," *IEEE Trans. Speech and Audio Processing*, vol. 6, no. 2, pp. 131–142, Mar. 1988.

[6] H. Zen, K. Tokuda, and A. Black, "Statistical parametric speech synthesis," *Speech Commun.*, vol. 51, no. 11, pp. 1039–1064, 2009.

[7] K. Tokuda, T. Kobayashi, and S. Imai, "Speech parameter generation from HMM using dynamic features," in *Proc. ICASSP*, Detroit, U.S.A., May 1995, pp. 660–663.

[8] K. Tokuda, Y. Nankaku, T. Toda, H. Zen, J. Yamagishi, and K. Oura, "Speech synthesis based on hidden Markov models," *Proceedings of the IEEE*, vol. 101, no. 5, pp. 1234–1252, 2013.

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

[10] H. Kawahara, I. Masuda-Katsuse, and A. D. 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," *Speech Commun.*, vol. 27, no. 3–4, pp. 187–207, 1999.

[11] M. Morise, "An attempt to develop a singing synthesizer by collaborative creation," in *Proc. SMAC*, Stockholm, Aug. 2013.

[12] Y. Agiomyrgiannakis, "Vocaine the vocoder and applications in speech synthesis," in *Proc. ICASSP*, Brisbane, Australia, Apr. 2015, pp. 4230–4234.

[13] P. K. Muthukumar, A. W. Black, and H. T. Bunnell, "Optimizations and fitting procedures for the Liljencrants-Fant model for statistical parametric speech synthesis," in *Proc. ICASSP*, Vancouver, Canada, May 2013.

[14] H. Zen and A. Senior, "Deep mixture density networks for acoustic modeling in statistical parametric speech synthesis," in *Proc. ICASSP*, Florence, Italy, May 2014, pp. 3872–3876.

[15] S. King and V. Karaiskos, "The blizzard challenge 2011," in *Proc. Blizzard Challenge workshop*, Turin, Italy, Sept. 2011.

[16] Y. Stylianou, "Voice transformation: A survey," in *Proc. ICASSP*, Taipei, Taiwan, Apr. 2009, pp. 3585–3588.

[17] Z. Ling, L. Qin, H. Lu, Y. Gao, L. Dai, R. Wang, Y. Jiang, Z. Zhao, J. Yang, J. Chen, and G. Hu, "The USTC and iflytek speech synthesis systems for blizzard challenge 2007," in *Proc. Blizzard Challenge workshop*, Bonn, Germany, Aug. 2007.

[18] Z. Yan, Q. Yao, and S. K. Frank, "Rich context modeling for high quality HMM-based TTS," in *Proc. INTERSPEECH*, Brighton, U.K., Sept. 2009, pp. 1755–1758.

[19] 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.

[20] T. Nose, V. Chunwijitra, and T. Kobayashi, "A parameter generation algorithm using local variance for HMM-based speech synthesis," *IEEE Journal of Selected Topics in Signal Processing*, vol. 8, no. 2, pp. 221–228, 2014.

[21] M. Shannon and W. Byrne, "Fast, low-artifact speech synthesis considering global variance," in *Proc. ICASSP*, Vancouver, Canada, May. 2013, pp. 7869–7873.

[22] T. Toda, T. Muramatsu, and H. Banno, "Implementation of computationally efficient real-time voice conversion," in *Proc. INTERSPEECH*, Portland, U.S.A., Sept. 2012.

[23] H. Zen, K. Tokuda, and T. Kitamura, "Refomulating the HMM as a trajectory model by imposing explicit relationships between static and dynamic feature vector sequences," *Computer Speech and Language*, vol. 21, no. 1, pp. 153–173, Jan. 2007.

[24] M. Abe, Y. Sagisaka, T. Umeda, and H. Kuwabara, "ATR technical repoart," no. TR-I-0166M, 1990.

[25] D. Klatt, "Review of text-to-speech conversion for English," *J. Acoust. Soc. Am.*, vol. 82, no. 3, pp. 737–793, 1987.

[26] T. Nose and T. Kobayashi, "Speaker-independent hmm-based voice conversion using adaptive quantization of the fundamental frequency," *Speech Commun.*, vol. 53, no. 7, pp. 973—985, 2011.

[27] Y. J. Wu, Y. Nankaku, and K. Tokuda, "State mapping based method for cross-lingual speaker adaptation in HMM-based speech synthesis," in *Proc. INTERSPEECH*, Brighton, U. K., 2009, pp. 528–531.

[28] N. Iwahashi, N. Kaiki, and Y. Sagisaka, "Speech segment selection for concatenative synthesis based on spectral distortion minimization," *IEICE Trans., Fundamentals*, vol. E76-A, no. 11, pp. 1942–1948, 1993.

[29] A. J. Hunt and A. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," in *Proc. ICASSP*, Atlanta, U.S.A., May 1996, pp. 373–376.

[30] T. Dutoit, A. Holzapfel, M. Jottrand, A. Moinet, J. Perez, and Y. Stylianou, "Towards a voice conversion system based on frame selection," in *Proc. ICASSP*, Hawaii, U.S.A., Apr. 2007, pp. 513–516.

[31] A. K. Syrdal, C. W. Wightman, A. Conkie, Y. Stylianou, M. Beutnagel, J. Schroeter, V. Strom, K.-S. Lee, and M. Makashay, "Corpus-based techniques in the AT&T NextGen synthesis system," in *Proc. ICSLP*, Beijing, China, Oct 2000, pp. 410–415.

[32] A. W. Black, "CLUSTERGEN: A statistical parametric synthesizer using trajectory modeling," in *Proc. INTERSPEECH*, Pittsburgh, U.S.A., Sep. 2006.

[33] T. Koriyama, T. Nose, and T. Kobayashi, "Statistical parametric speech synthesis based on gaussian process regression," *IEEE Journal of Selected Topics in Signal Processing*, vol. 8, no. 2, pp. 173–183, Apr. 2014.

[34] E. Helander, T. V. H. Silen, and M. Gabbouj, "Voice conversion using dynamic kernel partial least squares regression," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 20, no. 3, pp. 806–817, Mar. 2012.

[35] K. Oura, H. Zen, Y. Nankaku, A. Lee, and K. Tokuda, "An investigation of implementation performance analysis of DNN based speech synthesis system," in *Proc. INTERSPEECH*, Brighton, U. K., 2014, pp. 577–582.

[36] J. Yamagishi and T. Kobayashi., "Average-voice-based speech synthesis using HSMM-based speaker adaptation and adaptive training," *IEICE Trans., Inf. and Syst.*, vol. E90-D, no. 2, pp. 533–543, 2007.

[37] T. Toda, O. Ohtani, and K. Shikano, "One-to-many and many-to-one voice conversion based on Eigenvoices," in *Proc. ICASSP*, Hawaii, U.S.A., Apr. 2007, pp. 1249–1252.

[38] T. Yoshimura, T. Masuko, K. Tokuda, T. Kobayashi, and T. Kitamura, "Speaker interpolation for HMM-based speech synthesis system," *J. Acoust. Soc. Jpn. (E)*, vol. 21, no. 4, pp. 199–206, 2000.

[39] T. Nose, J. Yamagishi, T. Masuko, and T. Kobayashi, "A style control technique for HMM-based expressive speech synthesis," *IEICE Trans., Inf. and Syst.*, vol. E90-D, no. 9, pp. 1406–1413, 2007.

[40] L. Chen, M. J. F. Gales, L. Chen, K. Chin, K. Knull, and M. Akamine, "Exploring rich expressive information from audiobook data using cluster adaptive training," in *Proc. INTERSPEECH*, Portland, U.S.A., Sep. 2012.

[41] F. Jelinek, *Statistical methods for speech recognition.* MIT press, 1997.

[42] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 788–798, 2011.

[43] T. Nakashika, R. Takashima, T. Takiguchi, and Y. Ariki, "Voice conversion in high-order Eigen space using deep belief nets," in *Proc. INTERSPEECH*, Lyon, France, Aug. 2013, pp. 369–372.

[44] H. Fujisaki and K. Hirose, "Analysis of voice fundamental frequency contours for declarative sentence of Japanense," *J. Acoust. Soc. Jpn. (E)*, vol. 5, no. 4, pp. 233–242, 1984.

[45] H. Kawai, T. Toda, J. Ni, M. Tsuzaki, and K. Tokuda., "XIMERA: a new TTS from ATR based on corpus-based technologies," in *Proc. 5th ISCA Speech Synthesis Workshop (SSW5)*, Pittsburgh, USA, June 2004, pp. 179–184.

[46] Y. Qian, J. Xu, and F. K. Soong, "A frame mapping based HMM approach to cross-lingual voice transformation," in *Proc. ICASSP*, Prague, Czech Republic, 2011, pp. 5120–5123.

[47] T. Toda, H. Kawai, and M. Tsuzaki, "Optimizing sub-cost functions for segment selection based on perceptual evaluations in concatenative speech synthesis," in *Proc. ICASSP*, Montreal, Canada, May 2004, pp. 657–660.

[48] E. Moulines and F. Charpentier, "Pitch-synchronous waveform processing techniques for Text-To-Speech synthesis using diphones," *Speech Commun.*, vol. 9, no. 5-6, pp. 453–467, 1990.

[49] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis," in *Proc. EUROSPEECH*, Budapest, Hungary, Apr. 1999, pp. 2347–2350.

[50] T. Kudo, K. Yamamoto, and Y. Matsumoto, "Applying conditional random fields to Japanese morphological analysis," in *Proc. EMNLP*, Barcelona, Spain, Jul. 2004, pp. 230–237.

[51] K. Tokuda and H. Z. adn A. W. Black, "An HMM-based speech synthesis system applied to English," in *Proc. IEEE SSW*, 2002, pp. 227–230.

[52] Y. Qian, F. Soong, Y. Chen, and M. Chu, "An HMM-based Mandarin Chinese text-to-speech system," in *Proc. ICSLP*, 2006, pp. 223–232.

[53] P. Taylor, *Text-To-Speech Synthesis*. Cambridge Univ. Press, 2009.

[54] D. Hirst and A. D. Cristo, *Intonation Systems: A Survey of Twenty Languages*. Cambridge Univ. Press, 1998.

[55] G. Esther and E. L. Low, "Durational variability in speech and the rhythm class hypothesis," *Papers in laboratory phonology 7*, pp. 515–546, 2002.

[56] H. Lu and S. King, "Factorized context modeling for Text-to-Speech synthesis," in *Proc. ICASSP*, Vancouver, Canada, May 2013.

[57] S. Yokomizo, T. Nose, and T. Kobayashi, "Evaluation of prosodic contextual factors for HMM-based speech synthesis," in *Proc. INTERSPEECH*, Chiba, Japan, Sept. 2010, pp. 430–433.

[58] S. L. Maguer, N. Barbot, and O. Boeffard, "Evaluation of contextual descriptors for HMM-based speech synthesis in French," in *Proc. SSW8*, Barcelona, Spain, Aug. 2013.

[59] F. Eyben and Y. Agiomyrgiannakis, "Decision tree usage for incremental parametric speech synthesis," in *Proc. ICASSP*, Florence, Italy, May 2014, pp. 3819—-3823.

[60] T. Baumann, "Partial representations improve the prosody of incremental speech synthesis," in *Proc. INTERSPEECH*, MAX Atria, Singapore, May 2014, pp. 2932—-2936.

[61] M. Pouget, T. Hueber, G. Bailly, and T. Baumann, "Hmm training strategy for incremental speech synthesis," in *Proc. INTERSPEECH*, Dresden, Germany, Sep. 2015, pp. 1201–1205.

[62] K. Silverman, M. E. Beckman, J. F. Pitrelli, M. Ostendorf, C. Wightman, and P. Price, "ToBI: A standard for labeling English prosody," in *Proc. ICSLP*, Banff, Alberta, Canada, Oct. 1992, pp. 867–870.

[63] M. Kikuo, K. Hideaki, and I. Yosuke, "X-JToBI: An intonation labeling scheme for spontaneous Japanese," in *Technical Report of IEICE*, vol. SP2001-106, Lyon, France, Dec. 2001, pp. 25–30.

[64] K. Morizane, K. Nakamura, T. Toda, H. Saruwatari, and K. Shikano, "Emphasized speech synthesis based on hidden Markov models," in *Proc. Oriental COCOSDA*, Urumqi, China, Aug. 2009, pp. 76–81.

[65] Q. T. Do, S. Sakti, S. Takamichi, G. Neubig, T. Toda, and S. Nakamura, "Preserving word-level emphasis in speech-to-speech translation using linear regression HSMMs," in *Proc. INTERSPEECH*, Dresden, Germany, Sep. 2015, pp. 3665–3669.

[66] S. Sitaram, G. Anumanchipalli, J. Chiu, A. Parlikar, and A. W. Black, "Text to speech in new languages without a standardized orthography," in *Proc. SSW8*, Barcelona, Spain, Aug. 2013.

[67] H. Liang, Y. Qian, F. K. Soong, and L. Gongshen, "A cross-language state mapping approach to bilingual (Mandarin-English) TTS, year = 2008,," in *Proc. ICASSP*, Las Vegas, U. S. A., Apr., pp. 4641–4644.

[68] K. Tokuda, T. Kobayashi, T. Masuko, and S. Imai, "Mel-generalized cepstral analysis - a unified approach to speech spectral estimation," in *Proc. ICSLP*, Yokohama, Japan, Sep. 1994, pp. 410–415.

[69] S. Takaki, Z. Wu, and J. Yamagishi, "A function-wise pre-training technique for constructing a deep neural network based spectral model in statistical parametric speech synthesis," in *Proc. MLSLP*, Aizu, Fukushima, Sep. 2015.

[70] N. Hojo, K. Yoshizato, H. Kameoka, D. Saito, and S. Sagayama, "Text-to-speech synthesizer based on combination of composite wavelet and hidden Markov models," in *Proc. SSW8*, Barcelona, Spain, Aug. 2013.

[71] K. Yoshizato, H. Kameoka, D. Saito, and S. Sagayama, "Statistical approach to Fujisaki-model parameter estimation from speech signals and its quantitative evaluation," *Speech Prosody*, vol. 1, pp. 175–178, 2012.

[72] Z.-H. Ling, K. Richmond, J. Yamagishi, and R.-H. Wang, "Articulatory control of HMM-based parametric speech synthesis driven by phonetic knowledge," in *Proc. INTERSPEECH*, Brisbane, Australia, Sep. 2008, pp. 573–576.

[73] P. L. Tobing, T. Toda, G. Neubig, S. Sakti, S. Nakamura, and A. Purwarianti, "Articulatory controllable speech modification based on statistical feature mapping with Gaussian mixture models," in *Proc. INTERSPEECH*, Max Atria, Singapore, Sep. 2014, pp. 2298–2302.

[74] H. Kawahara, J. Estill, and 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," in *MAVEBA 2001*, Firentze, Italy, Sept. 2001, pp. 1–6.

[75] H. Banno, H. Hata, M. Morise, T. Takahashi, T. Irino, and H. Kawahara, "Implementation of realtime STRAIGHT speech manipulation system: Report on its first implementation," *J. Acoust. Soc. Jpn. (E)*, vol. 21, no. 3, pp. 79–86, 2000.

[76] M. Morise, H. Kawahara, and H. Katayose, "Fast and reliable F0 estimation method based on the period extraction of vocal fold vibration of singing voice and speech," in *Proc. AES 35th International Conference*, London, United Kingdom, Feb. 2009.

[77] M. Morise, "CheapTrick, a spectral envelope estimator for high quality speech synthesis," *Speech Commun.*, vol. 67, pp. 1–7, 2015.

[78] G. E. Henter, M. R. Frean, and W. B. Kleijn, "Gaussian process dynamical models for nonparametric speech representation and synthesis," in *Proc. ICASSP*, Kyoto, Japan, Mar. 2012.

[79] K. Ohta, T. Toda, Y. Ohtani, H. Saruwatari, and K. Shikano, "Adaptive voice-quality control based on one-to-many Eigenvoice conversion," in *Proc. INTERSPEECH*, Chiba, Japan, Sept. 2010, pp. 2158–2161.

[80] K. Kazumi, Y. Nankaku, and K. Tokuda, "Factor analyzed voice models for HMM-based speech synthesis," in *Proc. ICASSP*, Dallas, Texas, U.S.A., Apr. 2010, pp. 4234–4237.

[81] Y. Ohtani, Y. Nasu, M. Morita, and M. Akamine, "Emotional transplant in statistical speech synthesis based on emotion additive model," in *Proc. INTERSPEECH*, Dresden, Germany, Sep. 2015, pp. 274–278.

[82] K. Kobayashi, T. Toda, G. Neubig, S. Sakti, and S. Nakamura, "Statistical singing voice conversion with direct waveform modification based on the spectrum differential," in *Proc. INTERSPEECH*, Max Atria, Singapore, Sep. 2014, pp. 2514–2518.

[83] H.-T. Hwang, Y. Tsao, H.-M. Wang, Y.-R. Wang, and S.-H. Chen, "A study of mutual information for GMM-based spectral conversion," in *Proc. INTERSPEECH*, Portland, U.S.A., Sep. 2012.

[84] K. Tokuda, T. Masuko, B. Miyazaki, and T. Kobayashi, "Multi-space probability distribution HMM," *IEICE Trans., Inf. and Syst.*, vol. E85-D, no. 3, pp. 455–464, 2002.

[85] K. Yutani, Y. Uto, Y. Nankaku, A. Lee, and K. Tokuda, "Voice conversion based on simulatenous modeling of spectrum and f0," in *Proc. INTERSPEECH*, Brighton, U. K., 2009, pp. 3897–3900.

[86] K. Yu and S. Young, "Continuous F0 modeling for HMM based statistical parametric speech synthesis," *IEEE Trans. Audio, Speech and Language*, vol. 19, no. 5, pp. 1071–1079, 2011.

[87] J. Latorre, M. J. F. Gales, S. Buchholz, K. Knill, M. Tamura, Y. Ohtani, and M. Akamine, "Continuous f0 in the source-excitation generation for HMM-based TTS: Do we need voiced/unvoiced classification?" in *Proc. ICASSP*, Prague, Czech Republic, 2011, pp. 4724–4727.

[88] Z. Chen and K. Yu, "Tying covariance matrices to reduce the footprint of HMM-based speech synthesis systems," in *Proc. ICSP*, Zhejiang, China, 2009, pp. 1759–1762.

[89] H. Zen, Y. Nankaku, and K. Tokuda, "Continuous stochastic feature mapping based on trajectory HMMs," *IEEE Trans. on Audio, Speech, and Language processing*, vol. 19, pp. 417–430, Jan. 2011.

[90] M. Shannon, H. Zen, and W. Byrne, "Autoregressive models for statistical parametric speech synthesis," *IEEE Trans. on Audio, Speech, and Language processing*, vol. 21, no. 3, pp. 587–597, 2013.

[91] G. Hinton, "Product of experts," in *Proc. ICANN*, 1999, pp. 1–6.

[92] S. Takaki, Y. Nankaku, and K. Tokuda, "Contextual additive structure for hmm-based speech synthesis," *IEEE Journal of Selected Topics in Signal Processing*, vol. 8, no. 2, pp. 229–238, 2014.

[93] Y. Wu and F. Soong, "Modeling pitch trajectory by hierarchical HMM with minimum generation error training," in *Proc. ICASSP*, Kyoto, Japan, Mar. 2012.

[94] G. Hinton, L. Deng, D. Yu, G. Dahl, A. r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *Signal Processing Magazine of IEEE*, vol. 29, no. 6, pp. 82–97, 2012.

[95] H. Zen, A. Senior, and M. Schuster, "Statistical parametric speech synthesis using deep neural networks," in *Proc. ICASSP*, Vancouver, Canada, May 2013.

[96] H. Zen and H. Sak, "Unidirectional long short-term memory recurrent neural network with recurrent output layer for low-latency speech synthesis," in *Proc. ICASSP*, Brisbane, Australia, Apr. 2015, pp. 4470–4474.

[97] S. Fan, Y. Qian, and F. Soong, "TTS synthesis with bidirectional LSTM based recurrent neural networks," in *Proc. INTERSPEECH*, Max Atria, Singapore, Sep. 2014, pp. 1964–1968.

[98] K. Tokuda and H. Zen, "Directly modeling speech waveforms by neural networks for statistical parametric speech synthesis," in *Proc. ICASSP*, Brisbane, Australia, Apr. 2015, pp. 4215–4219.

[99] O. Watts, Z. Wu, and S. King, "Sentence-level control vectors for deep neural network speech synthesis," in *Proc. INTERSPEECH*, Dresden, Germany, Sep. 2015.

[100] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[101] A. W. Black and P. K. Muthukumar, in *Proc. INTERSPEECH*, Dresden, Germany, Sep., pp. 1211–1215.

[102] Z.-Z. Wu, T. K. E.-S. Chng, and H. Li, "Text-independent F0 transformation with non-parallel data for voice conversion," in *Proc. INTERSPEECH*, Chiba, Japan, Sept. 2010, pp. 1732–1735.

[103] H. Z. N. C. V. Pilkington and M. J. F. Gales, "Gaussian process experts for voice conversion," in *Proc. INTERSPEECH*, Florence, Italy, Jul. 2011.

[104] L.-H. Chen, Z.-H. Ling, Y. Song, and L.-R. Dai, "Joint spectral distribution modeling using restricted Boltzmann machines for voice conversion," in *Proc. INTERSPEECH*, Lyon, France, Sep. 2013, pp. 3052–3056.

[105] B. Chen, Z. Chen, J. Xu, and K. Yu, "An investigation of context clustering for statistical speech synthesis with deep neural network," in *Proc. INTERSPEECH*, Dresden, Germany, Sep. 2015, pp. 2212–2216.

[106] Y. Qian, Z. Yan, Y. Wu, and F. K. Soong, "An HMM trajectory tiling (HTT) approach to high quality TTS," in *Proc. INTERSPEECH*, Chiba, Japan, Sept. 2010, pp. 422–425.

[107] T.-N. Phung, C. M. Luong, and M. Akagi, "A hybrid TTS between unit selection and HMM-based TTS under limited data conditions," in *Proc. SSW8*, Barcelona, Spain, Aug. 2013.

[108] S. Imai, K. Sumita, and C. Furuichi, "Mel log spectrum approximation (MLSA) filter for speech synthesis," *Electronics and Communications in Japan*, vol. 66, no. 2, pp. 10–18, 1983.

[109] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, "Speech parameter generation algorithms for HMM-based speech synthesis," in *Proc. ICASSP*, Istanbul, Turkey, June 2000, pp. 1315–1318.

[110] T. Moriguchi, T. Toda, M. Sano, H. Sato, G. Neubig, S. Sakti, and S. Nakamura, "A digital signal processor implementation of silent/electrolaryngeal speech enhancement based on real-time statistical voice conversion," in *Proc. INTERSPEECH*, Lyon, France, Aug. 2013, pp. 3072–3076.

[111] S. Pan, Y. Nankaku, K. Tokuda, and J.Tao, "Global variance modeling on the log power spectrum of lsps for HMM-based speech synthesis," in *Proc. ICASSP*, Prague, Czech Republic, 2011, pp. 4716–4719.

[112] S. Pan, J. Tao, and Y. Wang, "A state duration generation algorithm considering global variance for HMM-based speech synthesis," in *Proc. APSIPA ASC*, Xi'an, China, 2011.

[113] H. Silen, E. Helander, J. Nurminen, and M. Gabbouj, "Ways to implement global variance in statistical speech synthesis," in *Proc. INTERSPEECH*, Portland, U.S.A., Sep. 2012.

[114] F. Eyben and Y. Agiomyrgiannakis, "A frequency-weighted post-filtering transform for compensation of the over-smoothing effect in HMM-based speech synthesis," in *Proc. ICASSP*, Florence, Italy, May 2014, pp. 275–279.

[115] K. Hashimoto, K. Oura, Y. Nankaku, and K. Tokuda, "The effect of neural networks in statistical parametric speech synthesis," in *Proc. ICASSP*, Brisbane, Australia, Apr. 2015, pp. 4455–4459.

[116] L.-H. Chen, T. Raitio, C. V.-Botinhao, J. Yamagishi, and Z.-H. Ling, "DNN-based stochastic postfilter for HMM-based speech synthesis," in *Proc. INTERSPEECH*, MAX Atria, Singapore, May 2014, pp. 1954–1958.

[117] G. J. McLachlan and T. Krishnan, *The EM Algorithm and Extensions*. John Wiley & Sons, 1977.

[118] K. Shinoda and T. Watanabe, "MDL-based context-dependent subword modeling for speech recognition," *J. Acoust. Soc. Jpn.(E)*, vol. 28, no. 3, pp. 140–146, 2007.

[119] H.-T. Hwang, Y. Tsao, H.-M. Wang, Y.-R. Wang, and S.-H. Chen, "Alleviating the over-smoothing problem in GMM-based voice conversion with discriminative training," in *Proc. INTERSPEECH*, Lyon, France, Sep. 2013, pp. 3062–3066.

[120] T. Merritt, J. Yamagishi, Z. Wu, O. Watts, and S. King, "Deep neural network context embeddings for model selection in rich-context HMM synthesis," in *Proc. INTERSPEECH*, Dresden, Germany, Sep. 2015, pp. 2207–2211.

[121] Y. Linde, A. Buzo, and R. M. Gray, "An algorithm for vector quantizer design," *IEEE Trans. on communications*, pp. 84–95, 1980.

[122] S. Kataoka, N. Mizutani, K. Tokuda, and T. Kitamura, "Decision tree backing-off in HMM-based speech synthesis," in *Proc. INTERSPEECH*, vol. 2, Jeju, Korea, Oct. 2004, pp. 1205–1208.

[123] Z. Ling and R. Wang, "HMM-based unit selection using frame sized speech segments," in *Proc. INTERSPEECH*, Pittsburgh U.S.A., Sept. 2006.

[124] T. Mizutani and T. Kagoshima, "Concatenative speech synthesis based on the plural unit selection and fusion method," *IEICE Trans. on Inf. and Syst.*, vol. E88-D, no. 11, pp. 2565–2572, 2005.

[125] K. Oura, Y. Nankaku, and K. Tokuda, "The use of state-level contexts in HMM-based speech synthesis," in *Proc. spring meeting of ASJ 2014*, Tokyo, Japan, Mar. 2014 (In Japanese), pp. 341–342.

[126] Y. Sagisaka, K. Takeda, M. Abe, S. Katagiri, T. Umeda, and H. Kuawhara, "A large-scale Japanese speech database," in *ICSLP90*, Kobe, Japan, Nov. 1990, pp. 1089–1092.

[127] Y. Ohtani, T. Toda, H. Saruwatari, and K. Shikano, "Maximum likelihood voice conversion based on GMM with STRAIGHT mixed excitation," in *Proc. INTERSPEECH*, Pittsburgh, U.S.A., Sep. 2006, pp. 2266–2269.

[128] S. Takamichi, T. Toda, Y. Shiga, S. Sakti, G. Neubig, and S. Nakamura, "Parameter generation methods with rich context models for high-quality and flexible text-to-speech synthesis," *IEEE Journal of Selected Topics in Signal Processing*, vol. 8, no. 2, pp. 239–250, 2014.

[129] R. Drullman, J. M. Festen, and R. Plomp, "Effect of reducing slow temporal modulations on speech reception," *J. Acoust. Soc. of America*, vol. 95, pp. 2670–2680, 1994.

[130] S. Thomas, S. Ganapathy, and H. Hermansky, "Phoneme recgnition usng spectral envelop and modulation frequency features," in *Proc. ICASSP*, Taipei, Taiwan, April 2009, pp. 4453–4456.

[131] S. Gergen, A. Nagathil, and R. Martin, "Reduction of reverberation effects in the MFCC modulation spectrum for improved classification of acoustic signals," in *Proc. INTERSPEECH*, Dresden, Germany, Sep. 2015, pp. 1992–1995.

[132] Z. Wu, X. Xiao, E. S. Chng, and H. Li, "Synthetic speech detection using temporal modulation feature," in *Proc. ICASSP*, Vancouver, Canada, May. 2013, pp. 7234–7238.

[133] R. Drullman, J. M. Festen, and R. Plomp, "Effect of temporal envelope smearing on speech perception," *J. Acoust. Soc. of America*, vol. 95, pp. 1053–1064, 1994.

[134] T. Arai, M. Pavel, H. Hermansky, and C. Avendano, "Intelligibility of speech with filtered time trajectories of spectral envelopes," in *Proc. ICSLP*, vol. 4, 1996, pp. 2490–2493.

[135] T. Muramatsu, Y. Ohtani, T. Toda, H. Saruwatari, and K. Shikano, "Low-delay voice conversion based on maximum likelihood estimation of spectral parameter trajectory," in *Proc. INTERSPEECH*, Brisbane, Australia, Sep. 2008, pp. 1076–1079.

[136] L. Atlas and S. A.Shamma, "Joint acoustic and modulation frequency," *EURASIP Journal on Applied Signal Processing*, vol. 7, pp. 668–675, 2003.

[137] K. Tanaka, T. Toda, G. Neubig, S. Sakti, and S. Nakamura, "A hybrid approach to electrolaryngeal speech enhansement based on spectral subtraction and statistical voice conversion," in *Proc. INTERSPEECH*, Lyon, France, Sep. 2013, pp. 3067–3071.

[138] H. Zen, M. J. F. Gales, Y. Nankaku, and K. Tokuda, "Product of experts for statistical parametric speech synthesis," *IEEE Trans. on Audio, Speech, and Language processing*, vol. 20, no. 3, pp. 794–805, Mar. 2011.

[139] T. Nose and A. Ito, "Analysis of spectral enhancement using global variance in HMM-based speech synthesis," in *Proc. INTERSPEECH*, MAX Atria, Singapore, May 2014, pp. 2917–2921.

[140] Y. Ohtani, M. Tamura, M. Morita, T. Kagoshima, and M. Akamine, "Histogram-based spectral equalization for HMM-based speech synthesis using mel-LSP," in *Proc. INTERSPEECH*, Portland, U.S.A., Sept. 2012.

[141] H. Zen, K. Tokuda, T. K. T. Masuko, and T. Kitamura, "Hidden semi-Markov model based speech synthesis system," *IEICE Trans., Inf. and Syst., E90-D*, no. 5, pp. 825–834, 2007.

[142] V. Tyagi, I. McCowan, H. Misra, and H. Bourlard, "Mel-cepstrum modulation spectrum (MCMS) features for robust ASR," in *Proc. ASRU*, MAX Atria, Singapore, Nov. 2003, pp. 399–404.

[143] "Speech signal processing toolkit (SPTK) http://sp-tk.sourceforge.net/."

[144] "Amazon mechanical turk https://www.mturk.com/."

[145] S. Takamichi, T. Toda, G. Neubig, S. Sakti, and S. Nakamura, "A postfilter to modify modulation spectrum in HMM-based speech synthesis," in *Proc. ICASSP*, Florence, Italy, May 2014, pp. 290–294.

[146] Y. Koretake, T. Toda, Y. Kisaki, H. Saruwatari, and K. Shikano, "An evaluation of modeling methods of global variance in HMM-based speech synthesis," in *IPSJ SIG Technical Report*, vol. 2010-SLP-84, no. 29, Dec. 2010 (In Japanese), pp. 1–6.

[147] J. Kominek and A. W. Black, "The CMU ARCTIC speech databases for speech synthesis research," in *Tech. Rep. CMU-LTI-03-177*, Language Technologies Institute, Carnegie Mellon University, Pittsburgh, U.S.A., 2003.

[148] S. Takamichi, T. Toda, A. W. Black, and S. Nakamura, "Parameter generation algorithm considering modulation spectrum for HMM-based speech synthesis," in *Proc. ICASSP*, Brisbane, Australia, Apr. 2015.

[149] Z. Wu and H. Li, "Voice conversion versus speaker verification: an overview," *APSIPA Transactions on Signal and Information Processing*, vol. 3, 2014.

[150] Z. Wu, T. Kinnunen, N. Evans, J. Yamagishi, C. Hanilci, M. Sahidullah, and A. Sizov, "ASVspoof 2015: the first automatic speaker verification spoofing and countermeasures challenge," in *Proc. INTERSPEECH*, Dresden, Germany, Sep. 2015, pp. 2037–2041.

[151] Z. Wu and S. King, "Minimum trajectory error training for deep neural networks, combined with stacked bottleneck features," in *Proc. INTERSPEECH*, Dresden, Germany, Sep. 2015, pp. 309–313.

[152] Z. Wu, C. Valentini-Botinhao, O. Watts, and S. King, "Deep neural networks employing multi-task learning and stacked bottleneck features for speech synthesis," in *Proc. ICASSP*, Brisbane, Australia, Apr. 2015.

[153] K. Hashimoto, K. Oura, Y. Nankaku, and K. Tokuda, "Trajectory model training considering global variance for speech synthesis based on neural network," in *Proc. autumn meeting of ASJ 2015*, Fukushima, Japan, Sep. 2015 (In Japanese), pp. 237–238.

[154] A. W. Black and K. Tokuda, "The blizzard challenge-2005: Evaluating corpus-based speech synthesis on common datasets," in *Proc. INTERSPEECH*, Lisbon, Portugal, Sep. 2005.

[155] "Blizzard challenge http://www.synsig.org/index.php/blizzard_challenge/."

[156] H. A. Patil, T. B. Patel, N. J. Shah, H. B. Sailor, R. Krishnan, G. R. Kasthuri, T. Nagarajan, L. Christina, N. Kumar, V. Raghavendra, S. P. Kishore, S. R. M. Prasanna, N. Adiga, S. R. Singh, K. Anand, P. Kumar, B. C. Singh, S. L. B. Kumar, T. G. Bhadran, T. Sajini, A. Saha, T. Basu, K. S. Rao, N. P. Narendra, A. K. Sao, R. Kumar, P. Talukdar, P. Acharyaa, S. Chandra, S. Lata, and H. A. Murthy, "A syllable-based framework for unit selection synthesis in 13 indian languages," in *Proc. O-COCOSDA*, Gurgaon, India, Nov. 2013, pp. 1–8.

[157] K. Sawada, S. Takaki, K. Hashimoto, K. Oura, and K. Tokuda, "Overview of NITECH HMM-based text-to-speech system for Blizzard Challenge 2014," in *Proc. Blizzard Challenge*, Singapore, Sep. 2014.

[158] A. Suni, T. Raitio, D. Gowda, R. Karhila, M. Gibson, and O. Watts, "The Simple4All entry to the Blizzard Challenge 2014," in *Proc. Blizzard Challenge*, Singapore, Sep. 2014.

[159] "Festvox http://festvox.org/download.html."

[160] S. S. Nair, R. C. R, and C. S. Kumar, "Rule-based grapheme to phoneme converter for malayalam," *International Journal of Computational Linguistics and Natural Language Processing*, vol. 2, no. 7, pp. 417–420, Jul. 2013.

[161] S. Kang and H. Meng, "Statistical parametric speech synthesis using weighted multi-distribution deep belief network," in *Proc. INTERSPEECH*, Max Atria, Singapore, Sep. 2014, pp. 1959–1963.

[162] C. M. Bishop, *Pattern recognition and machine learning*. Springer, 2006.

[163] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks," in *Proc. AISTATS*, Lauderdale, U.S.A., Apr. 2011, pp. 315–323.

[164] D. Kingma and B. Jimmy, "Adam: A method for stochastic optimization," in *arXiv preprint arXiv:1412.6980*, 2014.

[165] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, Apr. 2014.

[166] "Chainer: A powerful, flexible, and intuitive framework of neural networks http://chainer.org/."

Appendix

A.1 Text-to-speech of Indian languages for Blizzard Challenge 2015

In order to better understand different speech synthesis techniques to develop a corpus-based text-to-speech (TTS) system using a common dataset, Blizzard Challenge was devised in January 2005 [154] and has been held every year since then [155]. Blizzard Challenge 2015 has two tasks, 1) a mono-lingual speech synthesis task (IH1 hub task) for 6 Indian languages consisting of Bengali, Hindi, Malayalam, Marathi, Tamil, and Telugu, and 2) a multi-lingual speech synthesis task (IH2 spoke task) for Indian language and English. The Indian datasets [156] provided in the challenge consist of speech waveform and the corresponding texts only. The size of the speech data in each Indian language is about 4 hours for Hindi, Tamil and Telugu, and 2 hours for Bengali, Malayalam, and Marathi. They are sampled at 16 kHz. The text data is provided in UTF-8 format. As only the plain text data is provided without any additional information, such as a phoneme set, syllable definition, and prosodic labels, participants need to develop a natural language processing module (front-end) as well as a speech waveform generation module (back-end) to develop their own TTS systems.

To submit a TTS system from our group to the Blizzard Challenge 2015, we have developed our own system, the NAIST TTS system based on a statistical parametric speech synthesis technique using hidden Markov model (HMM) [8]. To improve quality of synthetic speech, two techniques are newly implemented for the traditional HMM-based speech synthesis framework, 1) pre-processing for producing smooth parameter trajectories to be modeled with HMM and 2) speech parameter generation considering the modulation spectrum (MS) of speech parameters [145][148]. The developed system has been submitted to the mono-lingual task and its performance has been demonstrated from the results of large-scaled subjective evaluations.

A.1.1 HMM-based TTS for mono-lingual task

The NAIST TTS system has 4 main modules; a text processing module, a speech processing module, a training module, and a speech synthesis module, as shown in Fig. 93. Context labels used for HMM training are generated using the existing toolkit or our developed rule-based grapheme-to-phoneme converter and syllable estimator in the text processing module. Smoothly varying speech parameter sequences are extracted in the speech processing module. The context-dependent phoneme HMMs and the MS probability density functions are trained using the context labels and the speech parameters in the training module. Finally, a speech waveform is generated from these trained models corresponding to a given text

to be synthesized in the synthesis module.

Text processing module Because the provided Indian datasets do not include any linguistic information, such as a phoneme set and prosodic labels, which is usually needed to describe speech parameters corresponding to a given text, it is indispensable to predict these in- formation from the given text. In the last year ' s challenge, some participants used several techniques to cope with this issue, e.g., the use of an existing speech recognizer for a different language to extract auxiliary linguistic information [157] or the development of a fully data-driven text analyzer [158].

we used hand-crafted text analyzers. We used text analyzers developed with language-specific recipes distributed by Festvox [159] for Bengali, Hindi, Tamil, and Telugu. Additionally, we also developed a text analyzer for Marathi with the recipe for Hindi because Marathi has a certain similarity to Hindi. For Malayalam, we developed a rule-based grapheme-to-phoneme converter [160] dealing with chillus and a rule-based syllable estimator considering specific characteristics of Malayalam, such as dependent vowel signs.

In the context generation stage, the context labels are required to train the context-dependent phoneme HMMs. Our context labels were designed on the basis of the contextual factors used in HTS speaker adaptation/adaptive training demo for English [1]. An example of the contextual factors used in our context label definition is shown as follows:

- phoneme, syllable structure, and stress
- vowel/consonant, articulator position, and voicing/unvoicing
- position of phoneme, syllable, and word
- the number of phonemes, syllables, and words.

Note that stress information is not used for Malayalam because it is not extracted in our text analysis module.

Speech analysis module A high-quality speech analysis-synthesis system is required to develop a high-quality TTS synthesizer. We conducted preliminary evaluation to compare analysis-synthesized speech quality by STRAIGHT [10, 74] and WORLD [77, 76] as a high-quality analysis-synthesis system. From the result of this preliminary evaluation, we decided that spectral envelope and aperiodicity were extracted with STRAIGHT, given F 0 extracted with WORLD. They were parameterized into the 0th-through-60th mel-cepstral coefficients, band aperiodicity, and log-scaled F 0, where the band aperiodicity was calculated by averaging aperiodicity of each frequency component in 5 frequency bands [127]. The shift

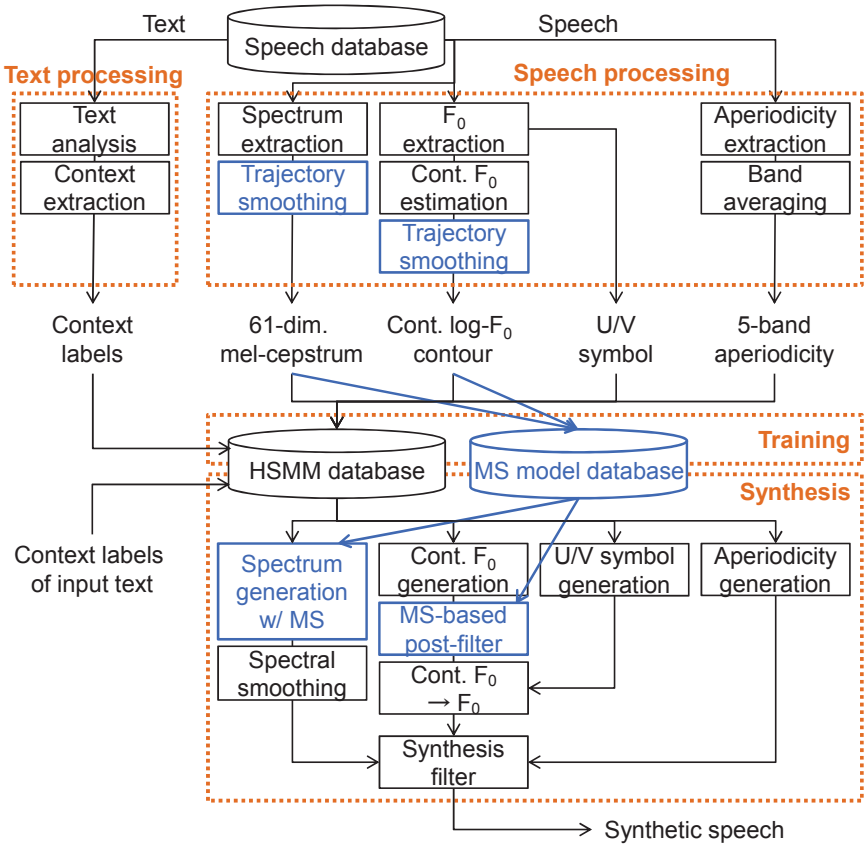


Figure 93. An overview of the NAIST TTS system for the Blizzard Challenge 2015. The orange-colored boxes indicate 4 main modules, a text processing module, a speech processing module, a training module, and a synthesis module. The blue-colored items are techniques newly implemented for the traditional HMM-based speech synthesis framework to improve synthetic speech quality, where "cont. F_0" and "MS" indicate the continuous F_0 and the modulation spectrum, respectively.

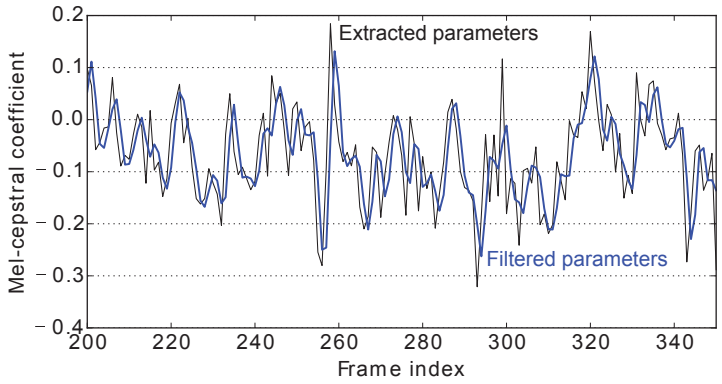


Figure 94. An example of the 20-th mel-cepstral coefficient sequences before and after the low pass filtering that removes the MS components over than 50 Hz. We can see that some fluctuation have been removed.

length was set to 5 ms. Moreover, the continuous F_0 contour [86] was additionally produced from the extracted F_0 contour. The spline-based interpolation algorithm was used to estimate F_0 values at unvoiced regions (see **Section A.2**.).

After the speech parameter extraction, we perform the speech parameter trajectory smoothing. Many fluctuations are usually observed over a time sequence of some speech parameters, such as mel-cepstral coefficients. They are represented as the MS of the temporal parameter sequence, i.e., power spectrum of the parameter sequence. As described in **Section A.10**, we have found that the effect of the MS components in high MS frequency bands on quality of analysis-synthesized speech is negligible, e.g., more than 50 Hz MS frequency components for the mel-cepstral coefficient sequence and more than 10 Hz MS frequency components for the continuous F_0 contour[41]. To make the HMMs focus on the modeling of only auditory informal components, low-pass filter (LPF) was applied to each parameter sequence. The cutoff frequency of LPF was set to 50 Hz for the mel-cepstral coefficients and 10 Hz for the continuous F 0 contour, respectively. An example of this parameter trajectory smoothing for the mel-cepstral coefficients is shown in Fig. 95.

Training module The context-dependent phoneme hidden semi-Markov models (HSMMs) were trained on the basis of a maximum likelihood criterion in a unified framework to model individual speech components [141]. Five-state left-to-right HSMMs were used for every Indian language. The feature vector consisted of mel-cepstral coefficients (61 dimensions), continuous log-scaled F_0 contour (1 dimension), band aperiodicity (5 dimensions), and their delta and

[41] Micro-prosody is captured by these components [53].

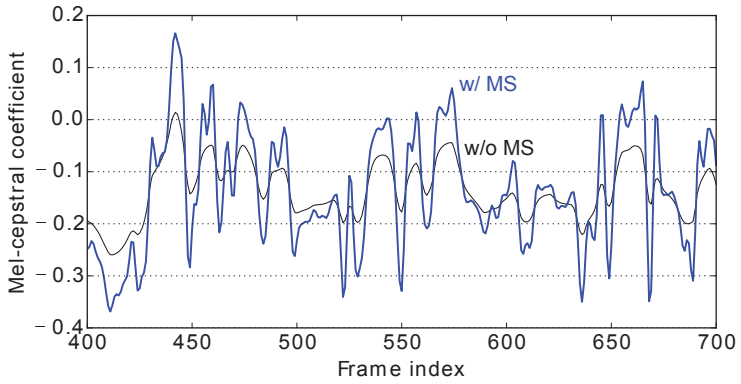


Figure 95. An example of the 20-th mel-cepstral coefficient sequence generated without considering the MS ("w/o MS") and that with considering the MS ("w/ MS").

delta-delta features, and discrete log-scaled F 0 contour (1 dimension) consisting of unvoiced symbols. The total dimensionality of the feature vector is 202. Only for Hindi, we used the 0th-through-24th mel-cepstral coefficients as we found that the spectral parameter because the 61-dimentional mel-cepstral coefficients were not well modeled in the HSMMs. The spectrum, continuous F_0, band aperiodicity components were modeled with the multi-stream continuous distributions. The discrete F_0 contour was additionally modeled with the multi-space distributions [84] to determine the voiced/unvoiced region of the continuous F_0 contour in the synthesis module. The tree-based clustering with the minimum description length (MDL) criterion [118] was employed. The stream weights were set to 1.0 (spectrum), 1.0 (continuous F_0), 1.0 (discrete F_0)[42] , and 0.0 (aperiodicity).

Gaussian distributions were also trained as the context- independent MS models for the spectrum and continuous F 0 contour. The utterance-level mean was first subtracted from the temporal parameter sequence, and then its MS was calculated. The length of discrete Fourier transform to calculate the MS was set to cover the maximum utterance length of the training data. These MS models were used in the synthesis module to reproduce the MS components, which were not well reproduced from the HSMMs only.

Synthesis module In the synthesis module, the context labels were first generated in the text processing module, and then the sentence HSMM corresponding to the text to be synthesized were constructed to generate the spectrum, continuous F 0, aperiodicity, and voiced/unvoiced regions. The spectral parameters were

[42] This stream setting is similar to the duplicated feature training [161] and the stream weights for continuous F_0 and discrete F_0 should be determined. We informally evaluated synthetic speech quality using some stream weight settings and chose this setting.

generated based on the speech parameter generation algorithm considering the MS components lower than 50 Hz. The other parameters were generated based on the ML-based parameter generation [109]. Additionally, we applied the MS-based post-filter (Chapter 4) to the generated continuous F_0 contour.[43] The MS was not considered in the aperiodicity component because there was no quality gain by the MS modification. An example of the generated mel-cepstrum sequences is illustrated in Fig. 95. We can find that more fluctuations are observed on the mel-cepstral sequence generated with the MS than that without the MS. Note that the global variance (GV) is also recovered because the MS can also represent the GV.

A.1.2 Experimental results

To submit the NAIST TTS system to the Blizzard Challenge 2015, we synthesized 50 reading texts (RD) and 50 semantically unpredictable sentences (SUS) in each language. The following 3 subjective evaluations were conducted in the challenge: (1) a mean opinion score (MOS) test on naturalness, (2) a degradation MOS (DMOS) test on similarity to the original speaker, and (3) a manual dictation test on intelligibility to calculate the word error rate (WER). Fig. 96-through-Fig. 100 show the result. Alphabets "A" and "J" indicate natural speech and our system, respectively. The other alphabets indicate the other participants' systems. We have found that our system was ranked in the highest group among the submitted systems in terms of naturalness in most of Indian languages but the gap between natural speech and synthetic speech was still large. Although our system was evaluated as the best in terms of intelligibility in Marathi (which was better than natural speech), such a result was not observed consistently over the other languages. Finally, our system was usually ranked in the middle group among the submitted systems in terms of similarity.

[43] No significant quality difference was observed between the continuous F_0 contour generated by speech parameter generation considering the MS and that filtered by the MS-based post-filter.

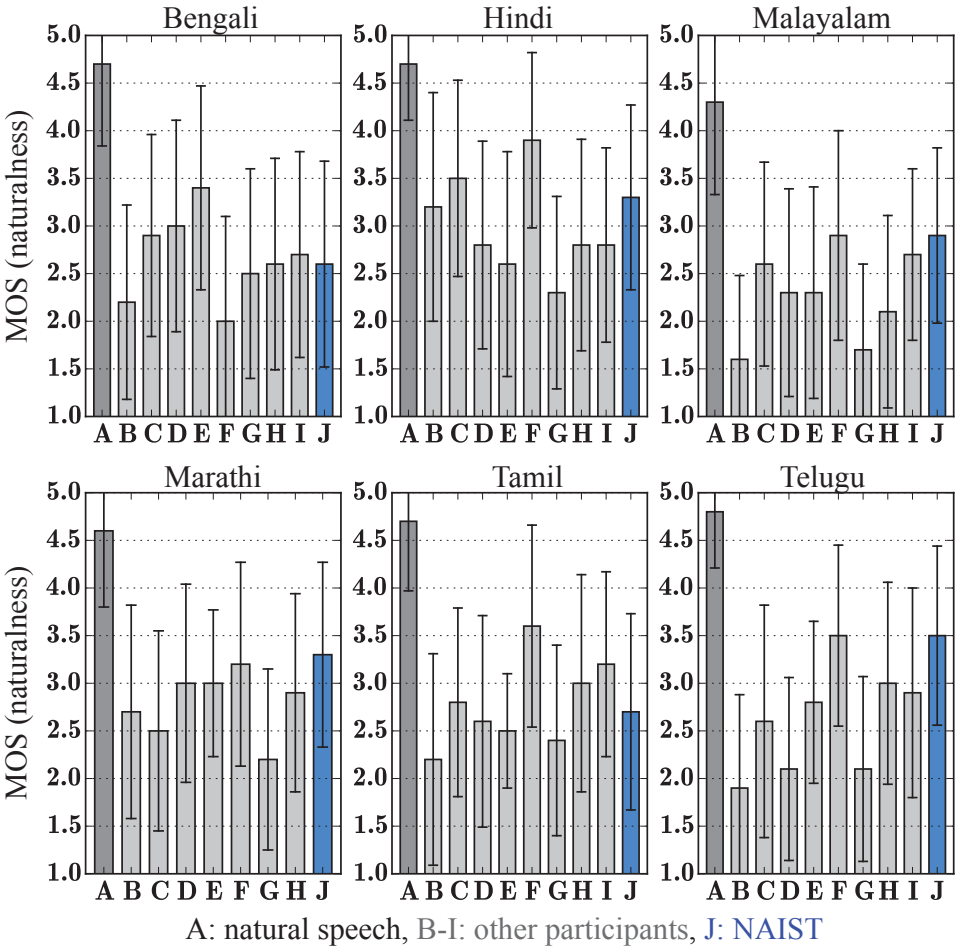


Figure 96. A result of MOS test on naturalness in the RD task.

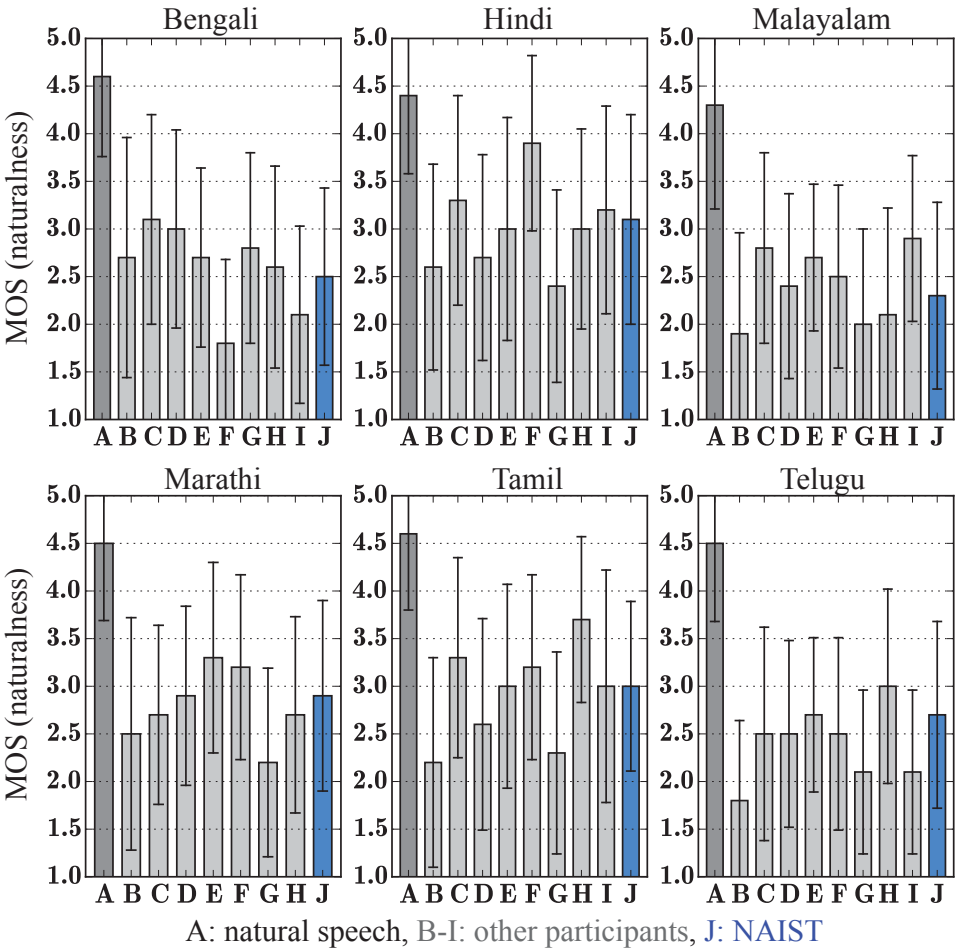


Figure 97. A result of MOS test on naturalness in the SUS task.

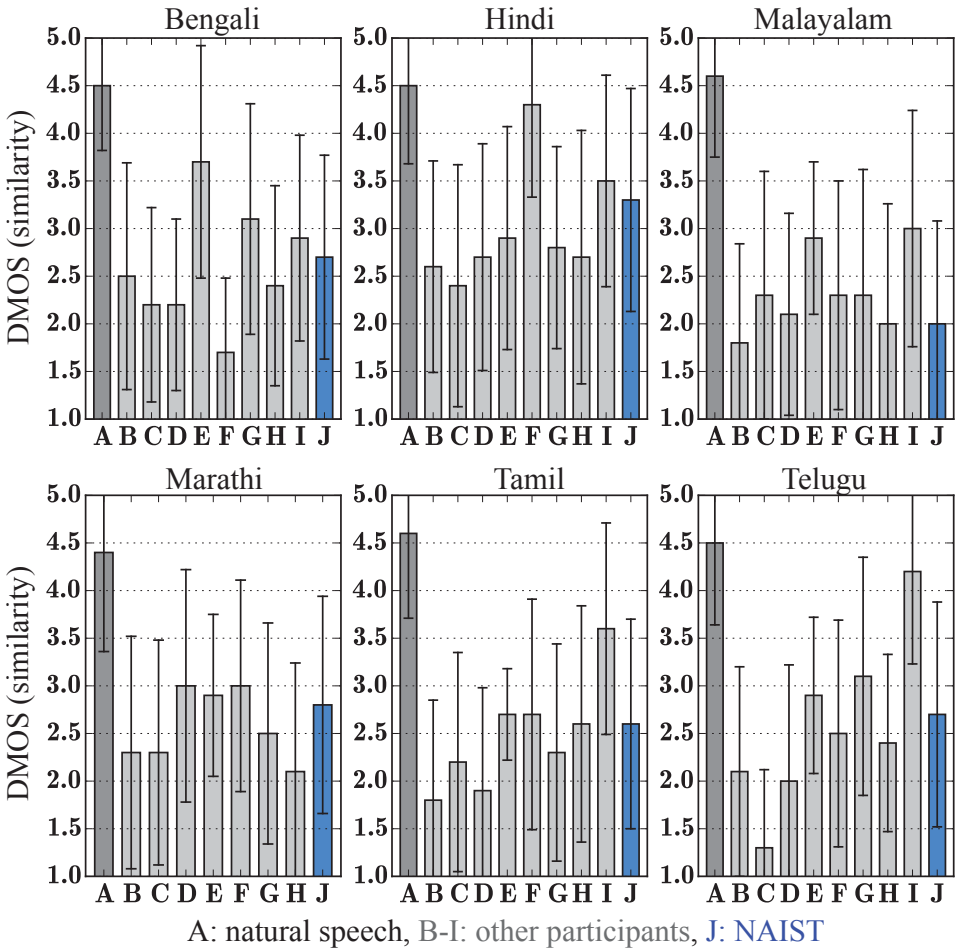


Figure 98. A result of MOS test on similarity to the original speaker in the RD task.

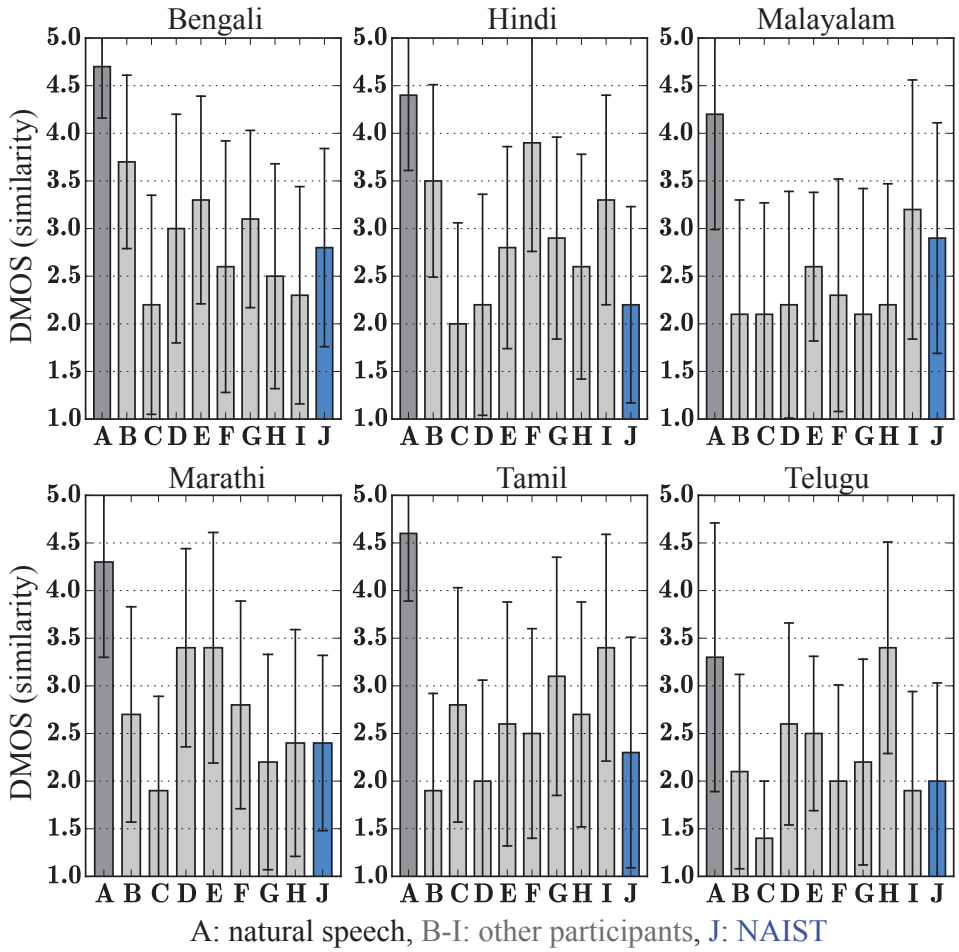


Figure 99. A result of MOS test on similarity to the original speaker in the SUS task.

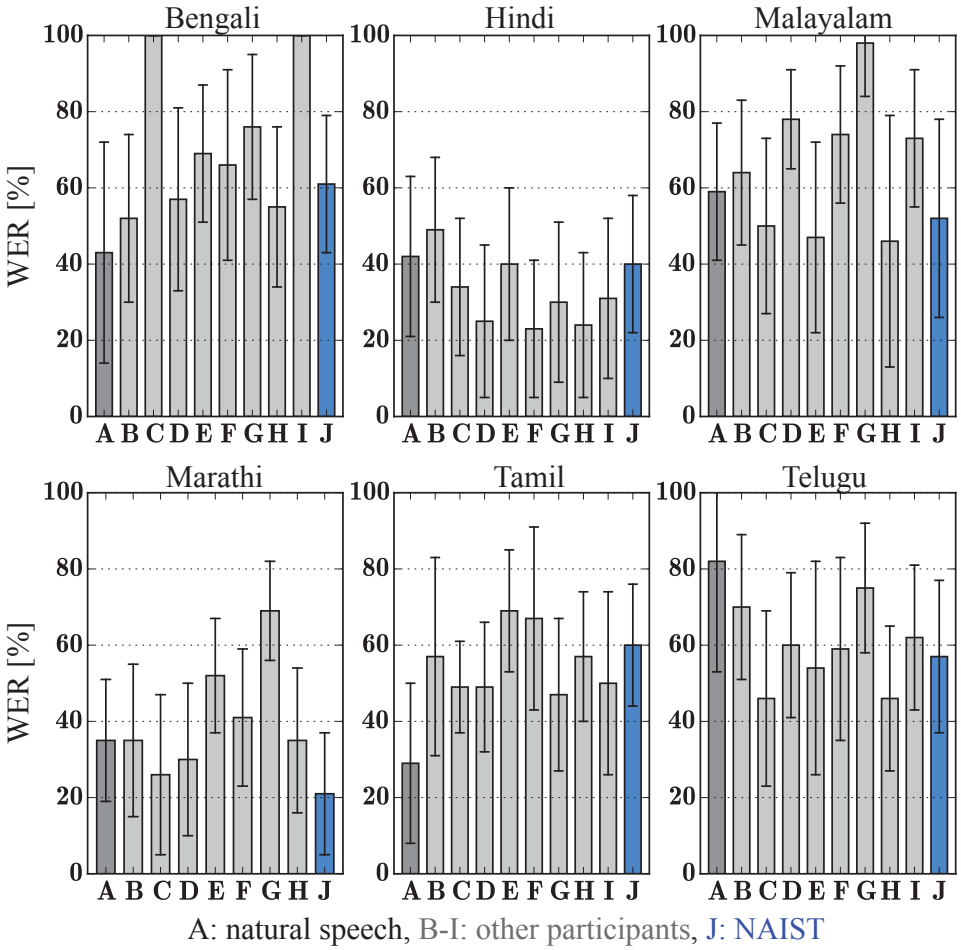


Figure 100. A result of intelligibility test.

A.2 Implementation of continuous F0 contour

The continuous F_0 contour is estimated for the continuous F_0 modeling [86]. Fig. 101 shows the example of how to calculate the continuous F_0 contour from the original discrete F_0 contour.

First, we apply the Low Pass Filter (LPF) to improve the performance of spline-based interpolation. After the spline-based interpolation, the original F_0 values are restored in the voiced region. For the silence part, we first copy the nearest F_0 value, then, reply the LPF again to remove the discontinuity.

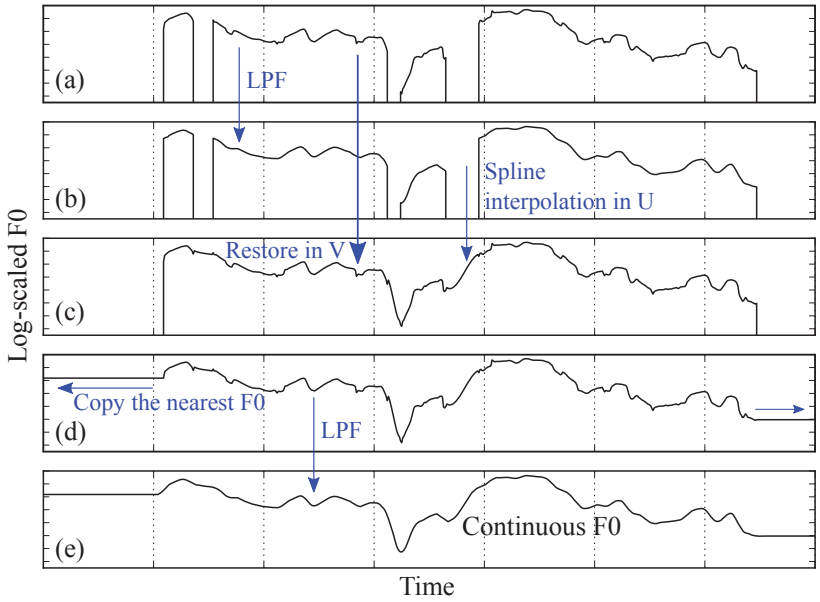


Figure 101. How to calculate (e) the continuous F_0 contour from (a) the discrete F_0 contour. V and U indicate the voiced/unvoiced regions, respectively.

A.3 Comparison of STRAIGHT and WORLD in HMM-based TTS

The STRAIGHT system, which we used in experimental evaluations, is high-quality speech analysis-synthesis system. However, deployments of speech synthesis systems with the STRAIGHT is limited because the system is patented. Recently, a novel system WORLD has been proposed and provided as a BSD-licensed system. This section investigates the effect of the WORLD in HMM-based TTS in order to accelerate the deployment.

A.3.1 Implementation of HMM-based TTS with WORLD

Speech parameters Parameter formats of spectrum and F_0 is the same between STRAIGHT and WORLD, but that of aperiodicity is different. Whereas STRAIGHT aperiodicity is extracted in each frequency bin and averaged in every frequency bands [127] (Fig. 102 (a)), the WORLD extracts it through explicit estimation of band-aperiodicity in voiced frames[44]. The band-aperiodicity by the WORLD is 0-dimensional vector in unvoiced frames. Therefore, it is appropriate to model the WORLD band-aperiodicity with MSD-HMMs [84] as shown in Fig. 102 (b).

Continuous aperiodicity Yu et al. proposed a novel method to model a "continuous" F_0 sequence with continuous HMMs in order to avoid weakness of the MSD-HMMs. To address the same problem with WORLD band-aperiodicity. we first extract a continuous F_0 sequence after the F_0 extraction. Then, we perform band-aperiodicity extraction given the continuous F_0 sequence[45]. The continuous band-aperiodicity is shown in Fig. 103, and the modeling is shown in Fig. 102(c).

A.3.2 Experimental evaluation

We trained five-state left-to-right HSMM using 4 speakers (2 male and 2 female) from CMU ARCTIC speech database [147]. The number of training and test data are 593 and 100, respectively. Speech signals were sampled at 16 kHz. The shift length was set to 5 ms. The 0th-through-24th mel-cepstral coefficients were extracted as spectral parameters and log-scaled F_0 and band-aperiodicity were extracted as excitation parameters. The stream weights are 1.0 for spectrum, 1.0 for F_0 and 0.0 for aperiodicity. In synthesis, we adopted ML-based parameter generation [109] and the MS-based post-filter described in Chapter 4. The frame

[44] The band width is 3 kHz.

[45] Note that both STRAIGHT and WORLD are F_0 adaptive system, which means that spectral parameters and aperiodicity parameters are estimated given the estimated F_0 parameters.

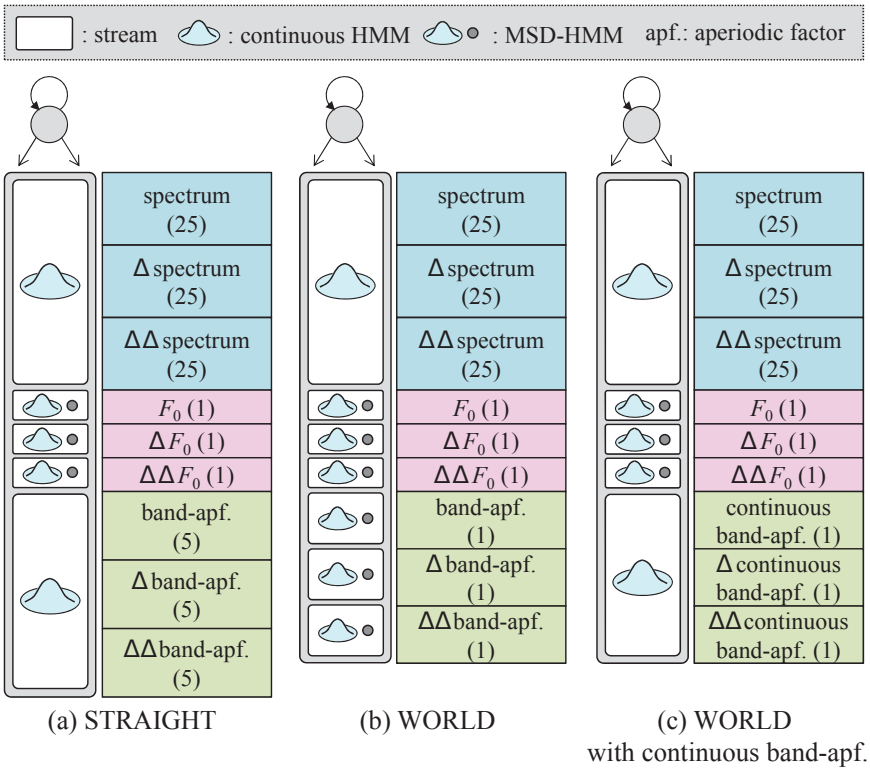


Figure 102. Stream structure for HMM-based TTS with WORLD. (·) is the number of dimensions in the evaluation.

length, shift length, and DFT length for the segment-level post-filter are 25, 12, and 64 frames, respectively.

In the preliminary evaluation, we didn't observe differences in quality between Fig. 102 (b) and (c). Therefore, we compare the synthetic speech quality of Fig. 102 (a) and (c). We conducted 5-scale MOS test on speech quality by 8 listeners.

The result is shown in Fig. 104. "*+MSPF" denotes that we applied the MS-based post-filter. "NATURAL" indicates natural speech. Although the WORLD system is worse than STRAIGHT system for the male speakers, there is no significant difference in total.

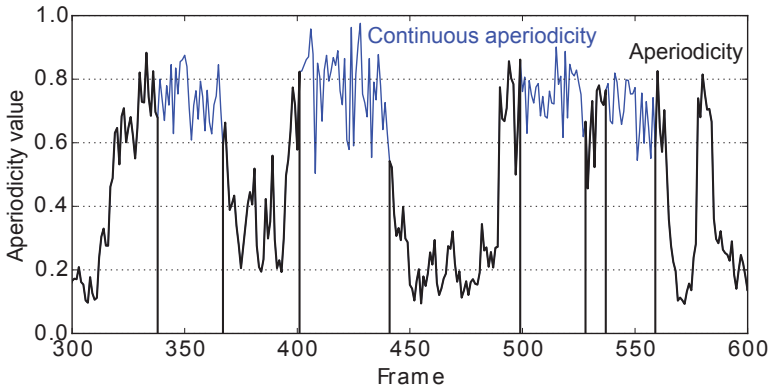


Figure 103. Continuous band-aperiodicity of the WORLD. It is extracted after the continuous F_0 estimation.

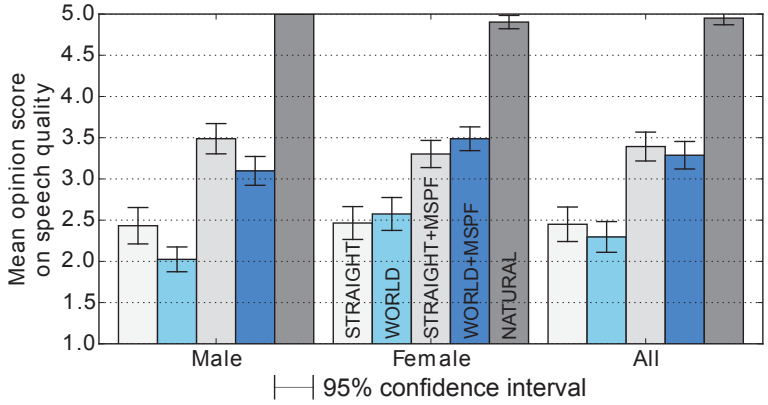


Figure 104. Subjective evaluation using STRAIGHT and WORLD. There is no significant difference in quality in total.

A.4 Derivation of conditional probability of the GMM

First, we derive the conditional probability of q-th mixture component, $P\left(\boldsymbol{Y}_t|\boldsymbol{X}_t, q, \boldsymbol{\lambda}\right)$, which is given as Gaussian distribution as follows:

$$P\left(\boldsymbol{Y}_t|\boldsymbol{X}_t, q, \boldsymbol{\lambda}\right) = \mathcal{N}\left(\boldsymbol{Y}_t; \boldsymbol{\mu}_q^{(Y|X)}, \boldsymbol{\Sigma}_q^{(Y|X)}\right), \tag{177}$$

where $\boldsymbol{\mu}_q^{(Y|X)}$ and $\boldsymbol{\Sigma}_q^{(Y|X)}$ are the $N_{\mathrm{w}}D$-dimensional mean vector and $N_{\mathrm{w}}D$-by-$N_{\mathrm{w}}D$ covariance matrix of the conditional probability of q-th GMM-mixture, respectively. The logarithmic probability is given as:

$$\begin{aligned}
\log P\left(\boldsymbol{Y}_t|\boldsymbol{X}_t, q, \boldsymbol{\lambda}\right) &= -\frac{1}{2}\left(\boldsymbol{Y}_t - \boldsymbol{\mu}_q^{(Y|X)}\right)^\top \boldsymbol{\Sigma}_q^{(Y|X)^{-1}}\left(\boldsymbol{Y}_t - \boldsymbol{\mu}_q^{(Y|X)}\right) && (178)\\
&= -\frac{1}{2}\boldsymbol{Y}_t^\top \boldsymbol{\Sigma}_q^{(Y|X)^{-1}}\boldsymbol{Y}_t + \boldsymbol{Y}_t^\top \boldsymbol{\Sigma}_q^{(Y|X)^{-1}}\boldsymbol{\mu}_q^{(Y|X)} \\
&\quad +\mathrm{Const.}, && (179)
\end{aligned}$$

where Const. is the value constant to $\boldsymbol{Y}_t$. Here, we define the precision matrix $\boldsymbol{P}_q^{(Z)}$ (inverse of the covariance matrix $\boldsymbol{\Sigma}_q^{(Z)^{-1}}$) as follows:

$$\boldsymbol{\Sigma}_q^{(Z)^{-1}} = \begin{bmatrix} \boldsymbol{\Sigma}_q^{(XX)} & \boldsymbol{\Sigma}_q^{(XY)} \\ \boldsymbol{\Sigma}_q^{(YX)} & \boldsymbol{\Sigma}_q^{(YY)} \end{bmatrix}^{-1} = \boldsymbol{P}_q^{(Z)} = \begin{bmatrix} \boldsymbol{P}_q^{(XX)} & \boldsymbol{P}_q^{(XY)} \\ \boldsymbol{P}_q^{(YX)} & \boldsymbol{P}_q^{(YY)} \end{bmatrix}, \tag{180}$$

where $\boldsymbol{P}_q^{(XY)^\top} = \boldsymbol{P}_q^{(YX)}$. The following formula holds between the previous equation [162].

$$\begin{bmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{C} & \boldsymbol{D} \end{bmatrix}^{-1} = \begin{bmatrix} \boldsymbol{M}\boldsymbol{A}^{-1} + \boldsymbol{A}^{-1}\boldsymbol{B}\boldsymbol{M}\boldsymbol{C}\boldsymbol{A}^{-1} & -\boldsymbol{A}^{-1}\boldsymbol{B}\boldsymbol{M} \\ -\boldsymbol{M}\boldsymbol{C}\boldsymbol{A}^{-1} & \boldsymbol{M} \end{bmatrix}, \tag{181}$$

where

$$\boldsymbol{M} = \left(\boldsymbol{D} - \boldsymbol{C}\boldsymbol{A}^{-1}\boldsymbol{B}\right)^{-1}. \tag{182}$$

Logarithmic probability $\ln P\left(\boldsymbol{Z}_t|q, \boldsymbol{\lambda}\right)$ are decomposed as:

$$\begin{aligned}
\ln P\left(\boldsymbol{Z}_t|q, \boldsymbol{\lambda}\right) &= -\frac{1}{2}\left(\boldsymbol{Z}_t - \boldsymbol{\mu}_q^{(Z)}\right)^\top \boldsymbol{P}_q\left(\boldsymbol{Z}_t - \boldsymbol{\mu}_q^{(Z)}\right) + \mathrm{Const.} && (183)\\
&= -\frac{1}{2}\Big(\left(\boldsymbol{X}_t - \boldsymbol{\mu}_q^{(X)}\right)^\top \boldsymbol{P}_q^{(XX)}\left(\boldsymbol{X}_t - \boldsymbol{\mu}_q^{(X)}\right) \\
&\quad + \left(\boldsymbol{X}_t - \boldsymbol{\mu}_q^{(X)}\right)^\top \boldsymbol{P}_q^{(XY)}\left(\boldsymbol{Y}_t - \boldsymbol{\mu}_q^{(Y)}\right) \\
&\quad + \left(\boldsymbol{Y}_t - \boldsymbol{\mu}_q^{(Y)}\right)^\top \boldsymbol{P}_q^{(YX)}\left(\boldsymbol{X}_t - \boldsymbol{\mu}_q^{(X)}\right) \\
&\quad + \left(\boldsymbol{Y}_t - \boldsymbol{\mu}_q^{(Y)}\right)^\top \boldsymbol{P}_q^{(YY)}\left(\boldsymbol{Y}_t - \boldsymbol{\mu}_q^{(Y)}\right)\Big) + \mathrm{Const.} && (184)
\end{aligned}$$

Given the input speech feature $\boldsymbol{X}_t$, it was expressed as:

$$\log P\left(\boldsymbol{Y}_t|\boldsymbol{X}_t, q, \boldsymbol{\lambda}\right) = -\frac{1}{2}\Big(\left(\boldsymbol{Y}_t - \boldsymbol{\mu}_q^{(Y)}\right)^\top \boldsymbol{P}_q^{(YY)}\left(\boldsymbol{Y}_t - \boldsymbol{\mu}_q^{(Y)}\right)$$

$$
\begin{aligned}
&+\left(\boldsymbol{Y}_t-\boldsymbol{\mu}_q^{(Y)}\right)^\top \boldsymbol{P}_q^{(YX)}\left(\boldsymbol{X}_t-\boldsymbol{\mu}_q^{(X)}\right)\\
&+\left(\boldsymbol{X}_t-\boldsymbol{\mu}_q^{(X)}\right)^\top \boldsymbol{P}_q^{(XY)}\left(\boldsymbol{Y}_t-\boldsymbol{\mu}_q^{(Y)}\right)\Big) + \text{Const.} \quad (185)\\
= \; & -\frac{1}{2}\boldsymbol{Y}_t^\top \boldsymbol{P}_q^{(YY)}\boldsymbol{Y}_t\\
&+\boldsymbol{Y}_t^\top\left(\boldsymbol{P}_q^{(YY)}\boldsymbol{\mu}_q^{(Y)}-\boldsymbol{P}_q^{(YX)}\left(\boldsymbol{X}_t-\boldsymbol{\mu}_q^{(X)}\right)\right) + \text{Const.}, \quad (186)
\end{aligned}
$$

Compared Eq. (42), Eq. (180) and Eq. (181), we can derive $\boldsymbol{\mu}_q^{(Y|X)}$ and $\boldsymbol{\Sigma}_q^{(Y|X)}$ as follows:

$$
\begin{aligned}
\boldsymbol{\Sigma}_q^{(Y|X)} &= \boldsymbol{P}_q^{(YY)^{-1}} = \boldsymbol{\Sigma}_q^{(YY)} - \boldsymbol{\Sigma}_q^{(YX)}\boldsymbol{\Sigma}_q^{(XX)^{-1}}\boldsymbol{\Sigma}_q^{(XY)}, && (187)\\
\boldsymbol{\mu}_q^{(Y|X)} &= \boldsymbol{\Sigma}_q^{(Y|X)}\left(\boldsymbol{P}_q^{(YY)}\boldsymbol{\mu}_q^{(Y)}-\boldsymbol{P}_q^{(YX)}\left(\boldsymbol{X}_t-\boldsymbol{\mu}_q^{(X)}\right)\right) && (188)\\
&= \boldsymbol{\mu}_q^{(Y)} - \boldsymbol{P}_q^{(YY)^{-1}}\boldsymbol{P}_q^{(YX)}\left(\boldsymbol{X}_t-\boldsymbol{\mu}_q^{(X)}\right) && (189)\\
&= \boldsymbol{\mu}_q^{(Y)} - \boldsymbol{P}_q^{(YY)^{-1}}\boldsymbol{P}_q^{(YY)}\boldsymbol{\Sigma}_q^{(YX)}\boldsymbol{\Sigma}_q^{(XX)^{-1}}\left(\boldsymbol{X}_t-\boldsymbol{\mu}_q^{(X)}\right) && (190)\\
&= \boldsymbol{\mu}_q^{(Y)} - \boldsymbol{\Sigma}_q^{(YX)}\boldsymbol{\Sigma}_q^{(XX)^{-1}}\left(\boldsymbol{X}_t-\boldsymbol{\mu}_q^{(X)}\right). && (191)
\end{aligned}
$$

A conditional probability $P\left(\boldsymbol{Y}_t|\boldsymbol{X}_t,\boldsymbol{\lambda}\right)$ are given as the following GMM mixing Eq. (42):

$$
\begin{aligned}
P\left(\boldsymbol{Y}_t|\boldsymbol{X}_t,\boldsymbol{\lambda}\right) &= \sum_{q=1}^{Q} P\left(q|\boldsymbol{X}_t,\boldsymbol{\lambda}\right) P\left(\boldsymbol{Y}_t|q,\boldsymbol{X}_t,\boldsymbol{\lambda}\right). && (192)\\
P\left(q|\boldsymbol{X}_t,\boldsymbol{\lambda}\right) &= \frac{\mathcal{N}\left(\boldsymbol{X}_t;\boldsymbol{\mu}_q^{(X)},\boldsymbol{\Sigma}_q^{(XX)}\right)}{\sum\limits_{q=1}^{Q}\mathcal{N}\left(\boldsymbol{X}_t;\boldsymbol{\mu}_q^{(X)},\boldsymbol{\Sigma}_q^{(XX)}\right)}. && (193)
\end{aligned}
$$

A.5 Comparison of mixture component weight for statistical sample-based speech synthesis

In general, The mixture component weight ω_m of the GMM using rich context models is calculated with the ML estimate, which is given by $\omega_m = \Gamma(c,m) \,/\, \sum_{m=1}^{Mc} \Gamma(c,m)$, where $\Gamma(c,m)$ is the occupancy count of the m-th mixture component in leaf node c. However, we set the weight to $\omega_m = 1/M_c$. The spectrogram is in Fig. 105. "Conventional", "Proposed (Occ)", "Proposed (Same)", and "Natural" represent spectrograms of generated parameters from conventional clustered model, occupancy-weighted GMM, the GMM with identical weights, and parameters of natural speech. Natural state duration was used. Parameters of "Proposed (Occ)" and "Proposed (Same)" is generated with proposed parameter generation with single Gaussian approximation, which is set initial parameter sequence to natural speech parameter. We can find that further improvement is realized in the generated parameter with GMMs of the same weight compared to that with conventional clustered models and occupancy-weighted GMMs.

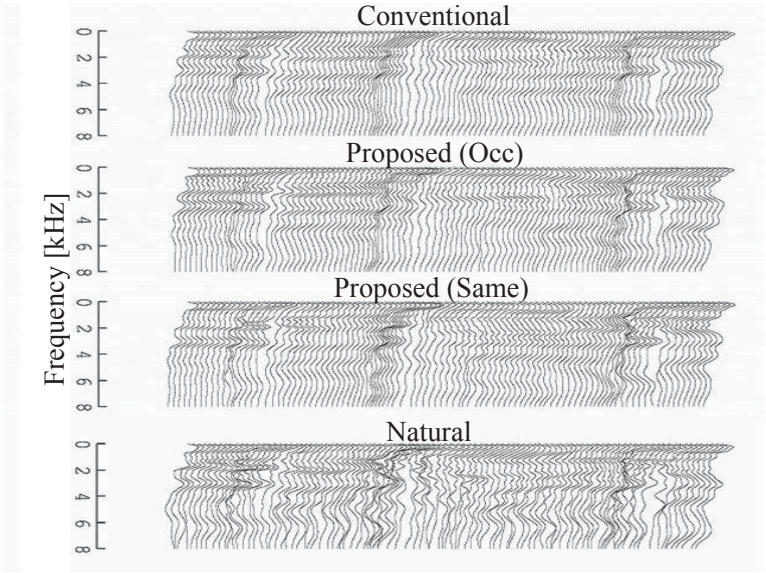


Figure 105. An example of spectrograms using Rich context-GMM (R-GMM) with different settings of the mixture weight. We can see that the tied weight has the structure similar to the natural speech parameters.

A.6 Investigation of quality degradation caused by rich context modeling

To confirm the degradation caused by the use of rich context models, we compared 4 kinds of synthetic speech shown in Table 4. "Target" is generated by rich context models using natural speech parameter as a initial parameter, and "Natural" is natural speech parameter. A opinion test on speech quality was conducted by 6 listeners. Natural state duration is used.

The result of mean opinion score in shown in Fig. 106. We can see that the degradation caused in the F_0 component is slightly where as that caused in spectral components is larger. This is because the spectral feature changes dynamically in the time domain.

Table 4. Synthetic speech samples used for investigating the quality degradation by the rich context modeling in HMM-based TTS.

Method	Spectrum	F_0
TT	Target	Target
TN	Target	Natural
NT	Natural	Target
NN	Natural	Natural

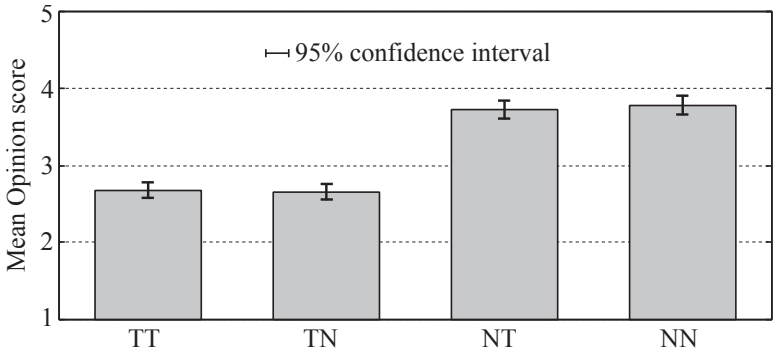


Figure 106. Mean opinion scores on speech quality to confirm degradation. We can find the degradation by the rich context modeling for the spectral component.

A.7 Time-invariant MS-based post-filter

As the yet another approach to the utterance-level MS-based post-filter, a time-invariant post-filter is derived by assuming that $\sigma_{d,f}^{(\mathrm{N})}$ is equal to $\sigma_{d,f}^{(\mathrm{G})}$ in Eq. (140) as follows:

$$\begin{aligned} s'_d(f) &= (1-k)s_d(f) + k\left[s_d(f) - \mu_{d,f}^{(\mathrm{G})} + \mu_{d,f}^{(\mathrm{N})}\right] \\ &= s_d(f) + k\left[\mu_{d,f}^{(\mathrm{N})} - \mu_{d,f}^{(\mathrm{G})}\right]. \end{aligned} \tag{194}$$

Because the second term in R.H.S. is independent of $s_d(f)$, this conversion process can be represented as a filtering process for the generated speech parameter sequence with a time-invariant FIR filter.

The result of the preference test on speech quality by 6 listeners are shown in Fig. 107. The experimental settings are the same to **Section 4.6.3**. "*+MS" indicates that we applied the utterance-level post-filter to the generated speech parameters in HMM-based TTS. "*+MS(ti)" indicate the time-invariant post-filter. We can see that a quality improvement is yielded by applying the time-invariant post-filter to the generated speech parameters. Although the improved quality is not comparable to that yielded by the utterance-level post-filter, the time-invariant post-filter is applicable to various lengths of speech parameter sequences.

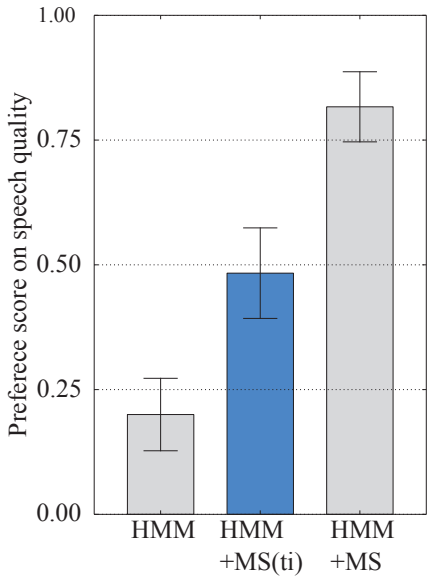


Figure 107. Preference scores on speech quality using time-invariant MS-based post-filter.

A.8 Modulation spectrum-based post-filter for GMM-based VC with spectral differentials

GMM-based VC with spectral differentials [82] is a novel VC technique without vocoding (analysis-synthesis) processes. As explained in **Section 2.5**, a GMM is trained using speech feature vectors of the input speech parameter sequence, $\boldsymbol{x} = [\boldsymbol{x}_1, \cdots, \boldsymbol{x}_t, \cdots, \boldsymbol{x}_T]$, and output speech parameter sequence, $\boldsymbol{y} = [\boldsymbol{y}_1, \cdots, \boldsymbol{y}_t, \cdots, \boldsymbol{y}_T]$[46] in conventional GMM-based VC. In synthesis, speech parameter sequence $\boldsymbol{y}'$ is generated through the trained GMM. In GMM-based VC with spectral differentials, spectral differential sequence, $\boldsymbol{d}' = \boldsymbol{y}' - \boldsymbol{x}$, is generated from a GMM, which is analytically derived from the original GMM used in the conventional GMM-based VC. Because the input speech waveform is directly filtered with $\boldsymbol{d}'$, we can avoid parameterization errors.

A MS-based post-filter proposed in Chapter 4 is available in this conversion framework as shown in Fig. 108. After generating $\boldsymbol{d}'$ in the standard manner, $\boldsymbol{x}$ is added to $\boldsymbol{d}'$. Then, its MSs are converted by the post-filtering process in order to make it close to natural MS of the output speech parameters. The filtered spectral differential sequence is calculated by subtracting $\boldsymbol{x}$ from the filtered $\boldsymbol{d}' + \boldsymbol{x}$.

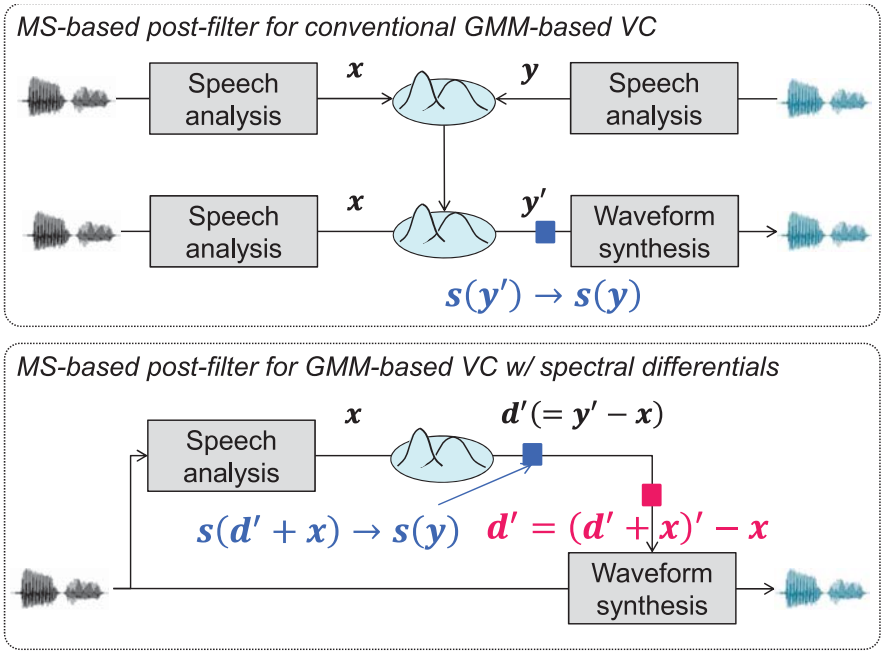


Figure 108. A MS-based post-filter for GMM-based VC with spectral differentials. $\boldsymbol{s}(\cdot) \rightarrow \boldsymbol{s}(\cdot)$ indicates MS-based post-filtering process.

[46] The delta feature is also used, but here we omit it.

A.9 Modulation spectrum-based post-filter using deep neural nets

For MS-based post-filtering process, deep Neural Nets are used as more complicated models than Gaussian distributions that are originally used in Chapter 4. The data used for constructing the DNN is the same as that of the segment-level post-filter (**Section 4.4**) using Gaussian distributions. As shown in Fig. 109, temporally aligned natural and generated speech parameters[47] are prepared first, then, the MSs of the windowed parameter segments are used to train the DNNs. From the result of initial investigations, instead of converting MSs of the generated parameter segment into those of the natural speech parameter segment, we convert into those of a MS differential that means a difference between MSs of the natural and generated speech parameter segments. Let $\boldsymbol{y}$ and $\hat{\boldsymbol{y}}_{\hat{\boldsymbol{q}}}$ be natural and generated speech parameter segments[48], the MS differential is given as $\boldsymbol{s}(\boldsymbol{y}) - \boldsymbol{s}\left(\hat{\boldsymbol{y}}_{\hat{\boldsymbol{q}}}\right)$, where $\boldsymbol{s}(\boldsymbol{y})$ is the MS of $\boldsymbol{y}$. The DNNs are trained to predict $\boldsymbol{s}(\boldsymbol{y}) - \boldsymbol{s}\left(\hat{\boldsymbol{y}}_{\hat{\boldsymbol{q}}}\right)$ from $\boldsymbol{s}\left(\hat{\boldsymbol{y}}_{\hat{\boldsymbol{q}}}\right)$, and The finally used MS is calculated by adding the MS differential and the original non-filtered MS.

For evaluation, we built two post-filters, the original post-filter using Gaussian distributions and that using DNNs. The speaker, training/evaluation data, speech parameters, filter-related parameters (e.g., window length) were the same to to those used in **Section 4.6** and **Section 4.6.4**. The post-filters were applied to spectral parameters. The dimensionality of the MS was 425. For DNN training, 1-hidden-layer feed forward neutral nets were constructed[49]. The hidden layer included 1275 nodes whose activation function was Relu [163]. The activation function of the output layer is linear function. The input and output features were normalized to a range of $[0.01, 0.99]$. The weights of the DNN were randomly initialized, then optimized to minimize the mean squared error between output features of the training data and the predicted values using a GPU implementation of mini-batch training. The number of epochs and mini-batch size were 50 and 500, respectively. The learning rate of the stochastic gradient descent-based back-propagation was scheduled by Adam [164] algorithm. The dropout [165] rate was set to 0.5. The DNN post-filter was implemented on Chainer [166]. We have conducted a preference AB test on speech quality. 6 listeners have participated.

Fig. 110 shows the result of the preference test. We can find that the use of DNNs causes slight improvements, but we have observed some partly buzzy sounds. We expect the use of recurrent structures is required.

[47] The generated speech parameters are generated using natural state duration.

[48] Note that these variables were defined for the speech parameter sequences in Chapter 2-through-5, but here they indicate its segments.

[49] We initially investigated the number of hidden layers, and the cost function using DNNs with 1 hidden layer is smaller than others for the evaluation data.

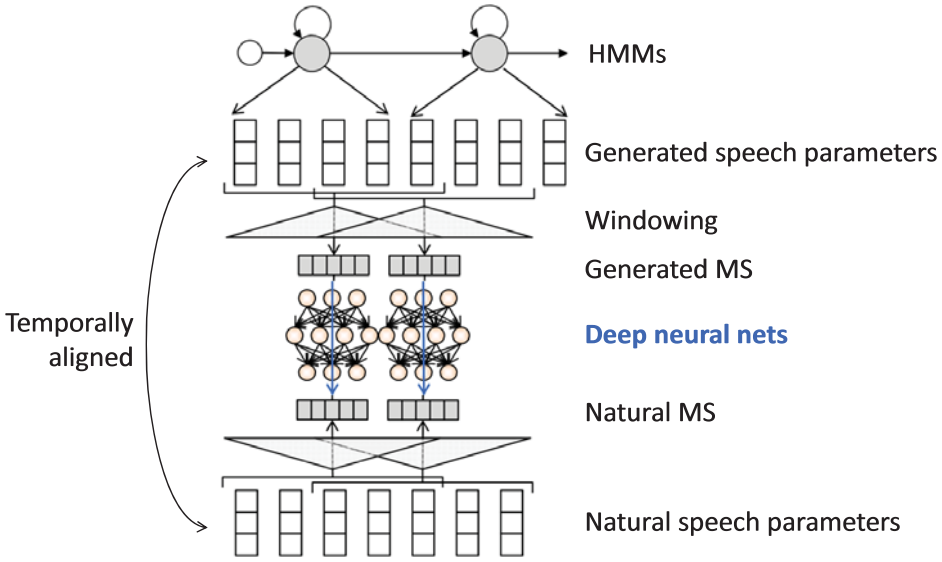


Figure 109. A Modulation Spectrum (MS)-based post-filter using deep neural nets. The training data is the same to the post-filter using Gaussian distributions.

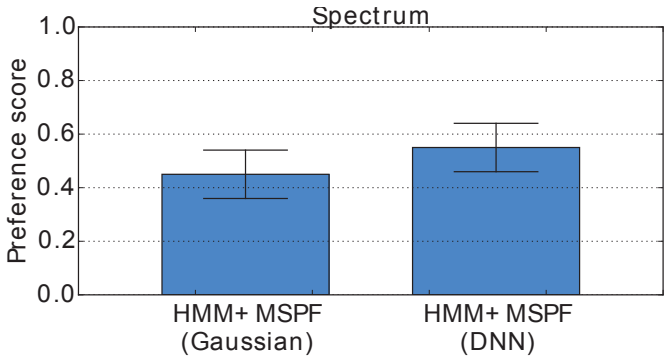


Figure 110. A result of preference test on speech quality for comparing MS-based post-filters using Gaussian distributions or DNNs. We can find that the use of DNNs causes slight improvements.

A.10 Effect of the modulation spectrum on speech quality

Related work [129, 134] investigated impacts of lower modulation frequency components on intelligibility, but the effect in speech quality is not yet investigated. Therefore, we investigated perceptual effects of the modulation spectrum on speech quality. We applied a Low Pass Filter (LPF) to remove the higher modulation frequency components of natural speech parameters, and conducted a listening test using the LPFed analysis-synthesized speech samples.

We used an English male speaker "RMS" and an English female speaker "SLT" from the CMU ARCTIC database [147]. Speech signals were sampled at 16 kHz. The shift length was set to 5 ms. The 0th-through-24th mel-cepstral coefficients were extracted as spectral parameters and log-scaled F_0 and 5 band-aperiodicity [74, 127] were extracted as excitation parameters. The STRAIGHT analysis-synthesis system [10] was employed for parameter extraction and waveform generation. We used 50 sentences from subset A for evaluation. We used Butterworth LPF to remove higher modulation frequency components of the spectral parameters. The cut-off frequency of the LPF is selected from 30, 40, 50, 60, and 70 Hz[50]. Additionally, non-filtered analysis-synthesized speech samples (100 Hz cut-off) were used. We conducted a 5-scaled MOS test on speech quality by 8 listeners using the LPFed analysis-synthesized speech samples.

Fig. 111 shows the result. We can see that there is no significant difference in quality between speech samples of 50 Hz and 100 Hz (non-filtered analysis synthesized speech).

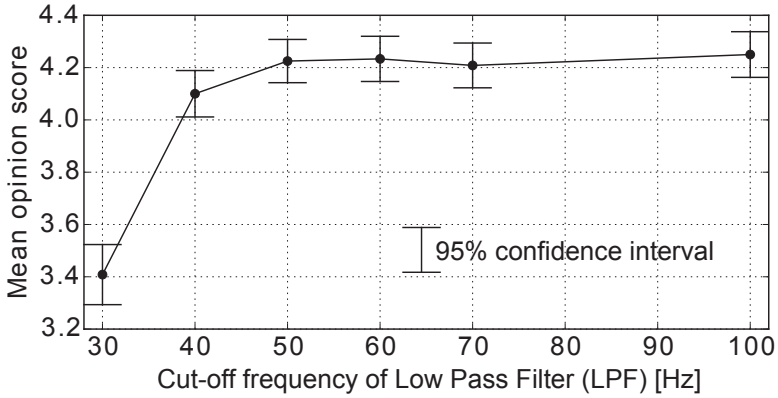


Figure 111. Mean opinion scores on speech quality with LPFed analysis-synthesized speech samples. We can find that MOS scores of cut-off frequency lower than 40Hz are significantly degraded compared to that of non-filtered samples.

[50] We didn't use speech samples with 80 and 90 Hz-cut-off LPFs because we expected that there is no significant difference in quality between these samples and non-filtered samples.