Gesture Design
for a Real-time Gesture-to-Speech Conversion System Based on Space Mapping Between a Gesture Space and an Acoustic Space

音響空間からジェスチャ空間への写像に基づく リアルタイム音声生成系におけるジェスチャ設計

2011年 12月 5日

37-097093　　國越 晶
指導教員　　峯松 信明 准教授

Abstract

Nowadays, most of speech synthesizers are those which require symbol inputs, such as TTS (Text-to-Speech) converters. The quality of synthesized speech sample produced by those speech synthesizers is improving. However, it still has some drawbacks, for example, in emotional speech synthesis or in expressive pitch control. On the other hand, synthesis methods which do not require symbol inputs, such as articulatory synthesis, are effective for continuous speech synthesis and pitch control based on dynamic body motion. Therefore they attract research interest and several applications have been proposed.

A dysarthric engineer, Ken-ichiro Yabu, developed a unique speech generator by using a pen tablet. The F1-F2 plane is embedded in the tablet. The pen position controls F1 and F2 of vowel sounds and the pen pressure controls their energy. Another example of speech generation from body motions is Glove Talk proposed by Sidney Fels. With two data gloves and some additional devices equipped to the user, body motions are transformed into parameters for a formant speech synthesizer. In this study, we consider the process of speech production as media conversion from body motions to sound motions.

Recently, GMM-based speaker conversion techniques have been intensively studied, where the voice spaces of two speakers are mapped to each other and the mapping function is estimated based on a GMM. This technique was directly and successfully applied to estimate a mapping function between a space of tongue gestures and other speech sounds. This result naturally makes us expect that a mapping function between hand gestures and speech can be estimated as well. People usually use tongue gesture transitions to generate a speech stream. But previous works showed that tongue gestures, which are inherently mapped to speech sounds, are not always required to speak. What is needed is a voluntarily movable part of the body whose gestures can be technically mapped to speech sounds. However, Yabu and Fels use classical synthesizers, i.e. formant synthesizers. Partly inspired by the remarkable progress of voice conversion techniques and voice morphing techniques in this decade, we are developing a GMM-based Hand-to-Speech conversion system (H2S system). Unlike the current techniques, our new synthesis method does not limit the input media. Therefore, our technique would be useful in assistive technology, in which devices are tuned for person to person, and in performative field, in which people pursue the human capability of expression.

In this study, we focus attention on the design of the system. As an initial trial, a mapping between hand gestures and Japanese vowel sounds was estimated so that topological

features of the selected gestures in a feature space and those of the five Japanese vowels in a cepstrum space are equalized. Experiments showed that the special glove can generate good Japanese vowel transitions with voluntary control of duration and articulation. We also discussed how to extend this framework to consonants. The challenge here was to figure out appropriate gestures for consonant sounds when the gesture design for vowels is given. We found that inappropriate gesture designs for consonants result in a lack of smoothness in transitional segments of synthesized speech. We have considered the reason to be: (1) the positional relation between vowels and consonants in the gesture space and that in the speech space were not equivalent, (2) parallel data for transition parts from consonants to vowels did not correspond well. In order to solve those problems, we have developed a Speech-to-Hand conversion system (S2H system, the inverse system of H2S system) trained from parallel data for vowels only to infer the gestures corresponding to consonants. Listeners evaluated that an H2S system, which exploits gesture data for consonants derived from an S2H system, can generate more natural sounds than those trained with heuristic gesture design for consonants.

Those natural speech generated by H2S system trained exploiting data generated by S2H system were, however, obtained only when input gestures were the same as the one which generated by S2H system. S2H system sometimes output gestures whose dynamic range is too large or which is not smooth enough. In those cases, it was difficult for users to form those gestures in realistic time. In this thesis, we compensated those problems with two ways: (1) reduce the dynamic range by setting the optimal weight for the gesture model (2) smooth the gesture trajectories by considering delta features. Exploiting parallel data for consonants derived from a S2H system, we also implemented a real-time Hand-to-Speech conversion system and evaluated the effectiveness. Subjective user evaluations showed that almost a half of the phonemes, which are generated by our H2S system are perceived correctly and that this system is effective enough to generate emotional speech.

Abstract

音声合成技術は，TTS に代表される，文字や記号を入力とする合成方式と，調音音声合成に代表される，文字や記号を介さない合成方式に大別される．前者と比較して後者は，運動の連続性に基づく滑らかな合成音の生成や，合成音における持続時間やピッチのリアルタイム制御においてその有効性が注目されており，芸術的歌声生成，教育応用，構音障害者支援など，様々なアプリケーションが提案されている．本研究では文字や記号を介さない音声生成として，構音器官以外の身体運動から直接音声を生成する新しいシステムを提案する．

近年，与えられた二話者間パラレルデータに対して，統計的に空間写像を設計する手法が話者変換の分野で用いられている．本研究ではこの手法を応用し，文字や記号を介さない音声合成方式として，身体運動から音声を生成する過程をメディア変換として捉え，身体運動の特徴量空間から音声の特徴量空間への空間写像に基づく新しい音声生成系を提案している．従来の手法と異なり，提案するシステムは入力メディアを限定しないため，個々のユーザーの能力に合わせてチューニングされることの多い障害者支援機器や，豊かな表現方法を追究するアートの世界などにおける応用が期待される．

本論文では，まず日本語五母音を対象として，身体運動の特徴量空間から音声の特徴量空間への写像に基づく音声生成系を構築する．初めに予備的検討として，二母音間遷移中に別の母音が混入しないように母音とジェスチャーとを対応させ，連結母音音声の生成系を構築し，手の運動から音声生成が可能であることを確認した．次に，母音とジェスチャーとのより良い対応を求めるために,「ジェスチャー空間におけるジェスチャー群の配置」と「母音空間における母音群の配置」の等価性を，より保証できる空間写像を設計した．実験の結果，両メディア間の等価性を考慮した空間写像によって，より明瞭な音声を生成することが可能となった．

一方，子音に適当なジェスチャーを割り当て，母音に対して用いた提案手法を子音の合成に拡張した場合，合成音において遷移部分が適切に知覚されないなどの問題が指摘された．その原因として，ジェスチャー空間における母音と子音に対応するジェスチャーの位置関係が，音響空間における母音と子音の位置関係に対応していないこと，また静的な位置関係が適切に対応づけられた場合でも，ジェスチャーと音声の動的な軌跡において適切な対応付けが取れていない可能性があるといった点が挙げられる．それらを回避するため，母音のみのパラレルデータを用いて音声–ジェスチャー変換システム（Speech-to-Hand system, 以下 S2H システム）を構築し，それに子音音声を入力することにより，子音に相当するジェスチャーを推定する手法を検討した．この手法によって，子音に相当するジェスチャーを推定した場合，適当に設定した子音のジェスチャーを用いた場合と比較して，より自然性

の高い音声を出力する H2S システムが構築されることが確認された．

しかしその音声は，S2H システムが推定したジェスチャーを入力した場合に得られるものであった．S2H システムは，ダイナミックレンジが大きすぎたり，充分に滑らかではないジェスチャー遷移を推定することがあり，その場合，リアルタイム H2S システムにおいては，実用的な速度では形成が困難であることも指摘されていた．本稿ではこの問題に対し，(1) ジェスチャーモデルの重みを設定することにより，出力されるジェスチャーベクトル時系列のダイナミックレンジを下げる，(2) 動的特徴量を考慮することにより，出力ジェスチャーベクトル時系列を平滑化する，の 2 通りの方法によって改善を試みた．そして，提案する S2H システム構築の枠組みにおいて，適切なジェスチャーモデルの重みを動的特徴量を考慮して S2H システムを構築した場合，より滑らかなジェスチャー遷移が与えられることを実験的に検証した．さらに改善された S2H システムによって得られたジェスチャーを学習データとして用い，リアルタイムジェスチャー–音声変換システム（Hand-to-Speech system, 以下 H2S システム）を構築し，その有効性を評価した．聴取実験の結果，本 H2S システムによって合成された音声は，およそ半分の音素が正しく知覚されること，入力するジェスチャーによって合成音声の感情が制御可能であることが示された．

Acknowledgement

This thesis was conducted while I was belonging to Hirose & Minematsu laboratory, the Univerity of Tokyo, in five years. This work was supported in part by the fund for young researchers of Global COE Program "Secure-Life Electronics" from 2009 to 2012.

First of all, I greatly appreciate Prof. Nobuaki Minematsu and Prof. Keikichi Hirose for their great guidance and encouragement in the accomplishment of this work. They have been teaching me how fabulous speech technology is. Especially Prof. Minematsu, he pulled me up to much higher level than I could have reached by myself. He also dealt with lots of troubles I made, such as Rinko failure, internship in China, and even about job-hunting. Prof. Minematsu, I hope that I could be a researcher like you and that will be able to repay the obligation.

I would like to thank Dr. Qiao Yu, currently in Shenzhen Institute of Advanced Technology, for his supervision since this research has started. Many of ideas in this thesis are given by him. Not only because of his great knowledge but also because of his tolerance, he is the researcher I really look up to. The code for MIVC, the nucleate after Chapter 5, is written by Dr.Saito, currently Assistant Professor of Sagayama & Ono laboratory in the University of Tokyo. He greatly supported me not only for research but also for the life at the lab. Mr. Masayuki Suzuki has helped me with the code to calculate structural distortion. Although he is younger than me, I have learnt many things from him as a researcher.

As for English, I very much appreciate Mr. Brett Heckerthorn - a teacher of Berliz, Mr. Josef Novak and Mr. Greg Short - PhD students of Hirose & Minematsu laboratory, and Mr. Adrian Leemann and Mr. Arnaud van Galen. I have been asking them lots of papers to go through in a short period, but they always kindly accepted.

I would like to thank Prof. Yoshihiko Nakamura at the mechanical department of the University of Tokyo, for allowing us to use his DataGlove and Prof. Hideki Banno at the department of information engineering in Meijo university, for variable inputs about his real-time STRAIGHT.

This research has not always been smooth sailing. Especially after I succeeded to develop real-time H2S system for Japanese five vowels, the research broke down and Rinko failure and rejection of international paper followed it. When I was the second grade of the doctoral course, I got totally unhealthy physically as well as mentally. At that time, Dr. Qiao Yu introduced me the internship program at Microsoft Research Asia (MSRA). In

order to refresh myself, I have decided to work in MSRA in Beijing for three months, since May to August, 2010. F0 conversion introduced in Section 3.4 is the topic I had worked on at that time. People in MSRA are top level researchers and students in China. They are however very modest and simply love research. They emphasized that everyone is equal in research and welcomed any discussion or questions. On the other hand, as the slogan " Work hard, Play harder " tells, people in MSRA do not forget to play a lot. I have learnt from them to enjoy my own time at home even I feel pressure for research at the lab. Thanks to the life in MSRA, I got back my health both mentally and physically. After I came back to Japan, unlike before I left, everything went nicely. I would like to say thank you to Dr. Frank Soong, Dr. Qian Yao and the friends in the speech group, who reminded me how great and how fun the research is. Without that period in Beijing, this research would not have published.

I successfully received PhD degree with this thesis. It is not only because my research went well but also because of being mentally healthy thanks to many friends. Especially I would like to thank 5 friends, Mrs. Alisa Sakurai, Mr. Miao Yimin, Mrs. Gao Yang, Dr. Adrian Leemann, and Mr. Arnaud van Galen. I have been talking about various things with Mrs. Alisa Sakurai during I had been in junior high school and via mail after graduation. Those discussion are treasures for a life and I will keep them in my mind. I believe that she understands me most and has given me the biggest influence. When I entered Tokyo university of Science, I did not know anything about communication. At that time, Chinese students, Miao Yimin and Gao Yang helped my university life and I have learnt many things from them. Even after the graduation, they have been caring for me. In my mind, they are my older brother and older sister. Dr. Adrian Leemann, who was a researcher in Hirose & Minematsu lab., has talent to make everything fun and attractive. Thanks to him, things which are not very much interested me changed into very interesting things, such as linguistics, English and Swiss culture. He is not only a good friend but also a great teacher. A programmer of Delta-N, Mr. Arnaud van Galen has taught me the important skills for research, such as programming and computer maintenance. His positive thinking keeps supporting me and changing me a lot. Your encouragement and your smile as well as your English check and programming help are priceless.

At last, parents, grandparents , granduncle and grandaunt, thank you very much for raising me up for 28 years. Although my family is not very rich, I was given sight, plenty of love, and high education until the doctoral course. I do not think that any word can express my thankful enough. I hope to be a person, who returns their love to the world.

Aki Kunikoshi
December 5, 2011

修士在学中に亡くなった両祖父と松井の叔父に.

神戸のおじいちゃん，山梨のおじいちゃん，松井のじいじ
ずっとお見守り下さり，ありがとうございました.
一番に感謝申し上げます.

本論文は，私が東京大学広瀬・峯松研究室に所属していた５年間の研究成果です．本研究の一部は2009年から2012年まで，グローバルCOEプログラム「セキュアライフ・エレクトロニクス」の若手研究ファンド制度の支援を受けました.

この研究が論文になるまで，指導教官の峯松准教授ならびに広瀬啓吉教授には，多大なるご指導，ご鞭撻を賜りました．両先生は音声工学の可能性と素晴らしさを教えてくださいました．特に，私の実力以上の場所まで引っ張りあげて下さった峯松准教授には，感謝の言葉もありません．研究以外でも，輪講の失敗や中国におけるインターン，卒業後の就職先に至るまで，ご迷惑のかけ通しでした．いつか峯松准教授のような研究者になって，少しでもご恩返しができればと思います.

特別研究員の喬宇博士（現・深セン先進科学技術院）には，本研究当初よりご指導いただきました．この論文に記載した多くのアイデアは喬博士よりご教授いただいたものです．豊富な知識，先を見通したご助言，寛大な振る舞い，喬宇博士は私の尊敬する研究者でした．本論文第5章以降の核となったMIVCのプログラムは齋藤大輔博士（現東京大学嵯峨山・小野研究室助教）によるものです．博士には本研究のみならず研究室生活においても，先輩として多大なるご支援をいただきました．博士課程2年の鈴木雅之氏には，第4章で使った構造間距離を求めるプログラムを作っていただきました．後輩でありながら，いつも教えられることばかりでした.

英語に関しては，修士在学中に英語を教えてくださったBrett Heckethorn先生，博士課程2年のJosef Novak氏，同Greg Short氏，そしてAdrian Leemann氏，Arnaud van Galen氏の校閲に深く感謝いたします．いつも無理なスケジュールでお願いしてばかりでしたが，快くお引き受け下さり，本当に多くの文章を丁寧にチェックしていただきました.

そして高価なデータグローブを快くお貸しくださった機械系研究科の中村仁彦先生，リアルタイム版STRAIGHTについてご相談に乗っていただいた名城大学の坂野先生に，心からお礼を申し上げます.

この研究はここに至るまで，いつも順調が進んでいた訳ではありませんでした．特に日本語五母音連続発話のリアルタイムH2Sシステムに成功した後，研究は行き詰まりました．そして輪講の失敗や論文の不採録が続き，博士課程２年次初頭には心身ともに非常に不健康な状態にありました．そのようなときに喬宇博士からMicrosoft Research Asia(MSRA)でのインターンのお話を伺い，2010年5月から同年8月までの三ヶ月を北京の研究所で過ごしました．3.4節でご紹介したF0変換の手法は，その成果です．MSRAは中国トップレベルの研究者や学生さんばかりでしたが，誰もが謙虚で，研究の上ではみな対等であると強調され，どんな質問や議論も歓迎されました．またWork hard, Play harderのスローガ

ンのとおり，MSRA では研究だけに打ち込むのではなく，余暇に精一杯遊ぶことも学びました．インターンの三ヶ月で，私は心身の健康を取り戻し，帰国後は帰国前と打って変わって何もかもがうまくいくようになりました．研究の素晴らしさ・楽しさを思い出させてくださった Frank Soong 博士ならびに Qian Yao 博士，そして Speech group の友人達に心からお礼申し上げます．北京で過ごしたあの一時期がなければ，間違いなくこの研究が論文になることはありませんでした．

また決して人並みはずれた実力や才能を持っている訳ではない私がここまで来られたのは，研究のみならず，精神面でも多くの人々に支えられてきたからです．中でも桜井亜莉沙氏，高揚氏，繆軼旻氏，Adrian Leemann 博士，Arnaud van Galen 氏の 5 人には，ここで深くお礼申し上げたいと思います．桜井亜莉沙氏とは，中学 3 年間を通して，卒業後はお手紙を通して，さまざまなことを語り合いました．私にとって最高の理解者であり，これまでの学生生活にもっとも大きな影響を与えてくれた友人です．東京理科大学に入学して，初めて親元を離れ，生活の仕方も人との付き合い方も分からなかった私は，同大学に留学生として来日していた繆軼旻氏，高揚氏から多くのことを学びました．3 年間で HSK(漢語水平考試)6 級を取得するまでの中国語を身につけることができたのも，両氏のおかげです．大学院に進学してからも現在に至るまで見守り続けてくれた両氏は，私にとって兄，姉のような存在です．東京大学大学院に研究員として在籍していた Adrian Leemann 博士には，どんなことでも楽しく魅力的にする才能を持っていました．言語学や語学，それにスイスドイツ語やスイスの文化など，それまであまりなじみのなかったことが Leemann 博士のおかげで興味の対象に変わりました．素晴らしいお友達でありながら，同時に素晴らしい先生でありました．Delta-N プログラマーの Arnaud van Galen 氏からは，プログラミングやアプリケーションの知識，コンピュータのメンテナンスなど，研究に必要不可欠な技術を学びました．また van Galen 氏の前向きな姿勢は，自信をなくして落ち込みがちな私を支え続けてくれました．特に重要な発表の直前，不安に押しつぶされそうになっているとき，真夜中でも早朝でも（オランダと 8 時間の時差があるため）必ず送ってくれたメッセージには，どんなに励まされたことかわかりません．

最後に，28 年間育ててくださった父と母，そして両祖父母と松井の叔父叔母にお礼を申し上げます．私が育った家庭は，経済的には決して豊かではありませんでした．しかしその中で私は，光と，溢れるほどの愛情を与えられ，博士号を取得するまでの教育を受けさせていただきました．どんな言葉も感謝の気持ちを充分に表現することはできません．将来，注いでいただいた愛情を少しでも社会に還元できるような人間になりたいと思います．

2011 年 12 月 5 日
國越 晶

Contents

Table list

Figure list

Chapter1

Introduction

1.1 Overview

As Astroboy in“ Astroboy ”(1952), HAL 9000 in“ 2001: A Space Odyssey ”(1968), No.5 in“ Short Circuit ”(1986), etc. show, making robots or computers that talk with humans like humans do is one of humanities dreams. In order to realize that dream, scientists and engineers in a wide range of fields have been building new technologies. Among them, speech synthesis technologies would be responsible for the output part of those desired robots/computers - creating natural and intelligible voice.

Speech synthesis methods are roughly divided into two groups depending on its input. When the input is symbols such as a plain text or characters, the system is called Text–to–Speech (TTS) system. Nowadays speech synthesis technologies are not only used in movies or in laboratories but also in our daily life. MacOSX and Windows are equipped by default with functions that read text aloud for visually impaired people and speaking disabled people can use VOCA (Voice Output Communication Aids) machines to talk [1, 2].

Most of those practical speech synthesizers are TTS converters. The reason that TTS converters are widely used would be because of the easiness of their input and the quality of the synthesized speech. With the preparation of large amounts of speech corpora and the development of statistical learning theories and approaches, the quality of synthesized speech by TTS synthesizers has been improved astoundingly. They reached the level that people cannot easily distinguish between synthesized speech by TTS and natural speech. Some high quality TTS demonstrations are found online [3]. A very good synthesizer may be able to even deceive speaker verification systems [4]. TTS technology, however, still has some drawbacks, for example in speech rate/pitch control.

On the other hand, methods that do not require symbol inputs, such as articulatory synthesis, are effective for speech rate/pitch control and smooth speech synthesis based on dynamic body motion, while synthesized sounds are often less articulate than that of TTS. Therefore they attract research interest and several applications have been proposed in performative singing voice synthesis [5], speech education [6, 7], assistive technologies [8, 9] etc.

A dysarthric engineer, Yabu, developed a unique speech generator for dysarthric people, that uses a pen tablet for input [8]. Another example of speech generation from body motions is Glove Talk proposed by Fels [5]. With two data gloves and some additional devices attached to the user, body motions are transformed into parameters required for a speech synthesizer.Unlike TTS converters, those applications give unlimited vocabulary and directly control fundamental frequency and volume. Trained users can even give artistic performance with them. Their methodologies are, however, difficult to apply to other applications, because these applications are designed to make optimal use of the characteristics specific to the input device. For example, Yabu's speech generator exploits the continuity of pen point and that users are capable to control the pen point and pressure

simultaneously. If the user cannot handle the pen - some dysarthric people have handicaps on other body parts besides their articulatory organs, another input media should be chosen for the system and then that system has to be established again with the new media.

Media dependence is a very important factor in both assistive technology and in art. When assistive devices are developed, the appropriate input media is chosen according to the user's handicap and the remaining capabilities of his body. For example, congenital visual impaired people often do not have difficulty to use braille while most of the acquired visually impaired people desire an assistive device with audio [10]. Several types of electric wheelchairs, which are widely used by orthopedically impaired people, are developed according to the input capability of users, such as head gesture, speech, inner force sense and electromyography [11].

In the field of art, various media are explored to pursue the possibility of expression. Dr. Theremin invented Theremin, an early electronic musical instrument controlled without discernible physical contact from the player. It is said that the Theremin instrument brought about a change in the way of playing instruments [12].

As the examples above show, the possibility of media selection is very important especially in the fields in which media is not chosen by the developers but the users. If a media-independent methodology for speech generation were to be established, it would be applied to a broad area of applications.

In order to establish a media-independent methodology for speech generation, we treat the speech generation process from body motion as the mapping problem from non-speech media to speech media. People usually use tongue gesture transitions to generate a speech stream. This is considered as the inherent mapping between tongue gestures and speech sounds. Yabu and Fels's work, however, showed that tongue gestures are not always required to speak. What is needed is a voluntarily movable part of the body whose gestures can be technically mapped to speech sounds. In this thesis, we consider the mapping problem from hand gesture to speech as one example.

Recently, GMM-based speaker conversion techniques have been intensively studied, where the voice spaces of two speakers are mapped to each other and the mapping function is estimated based on a GMM [13, 14, 15]. This technique was directly and successfully applied to estimate a mapping function between a space of tongue gestures and another of speech sounds. This result naturally makes us surmise that a mapping function between hand gestures and speech can be estimated well and we developed a GMM-based Hand-to-Speech conversion system (H2S system).

For voice conversion, it is not very difficult to obtain some correspondence between two data sequences from input and output space, for example, by Dynamic Time Warping (DTW). On the other hand, how to design the optimal correspondence between hand gestures and speech is one of the biggest challenges for our framework. In this thesis, we focus on the design of that system.

As an initial trial, a mapping between hand gestures and Japanese vowel sounds is estimated. This estimation is optimized through equalization of the topological features of the selected gestures in a feature space and those of the five Japanese vowels in a cepstrum space

Next, we discuss how to extend this framework to consonants. The challenge here is to figure out appropriate gestures for consonant sounds when the gesture design for vowels is given. Preliminary experiment shows that inappropriate gesture designs for consonants result in a lack of smoothness in transitional segments of synthesized speech. We have considered the reason to be: (1) the positional relation between vowels and consonants in the gesture space and that in the speech space were not equivalent, (2) parallel data for transition parts from consonants to vowels did not correspond well. In order to get around those problems, we have developed a Speech-to-Hand conversion system (S2H system, the inverse system of H2S system) trained from parallel data for vowels only to infer the gestures corresponding to consonants. Utilizing parallel data for consonants derived from the S2H system, we also implement a real-time H2S conversion system and examined the effectiveness

1.2 Outline of the thesis

This thesis is organized as follows. In this chapter, the overview of this thesis was described. In Chapter 2, current speech technologies are reviewed. Looking through those technologies, the objectives of our research is also made clear. In Chapter 3, the framework of our proposed system is described. Based on this framework, we implement a speech synthesizer from hand gestures in Chapter 4. Our preliminary experiments shows the challenge of our system as well as the effectiveness of our proposed system. In order to solve the challenge, we derive the quasi-optimal correspondence between gestures and speech. In chapter 5, we propose a framework to derive the gestures for consonants when only the correspondence for vowels is given. The gesture design is evaluated using combined S2H–H2S systems. In chapter 6, the real-time H2S system based on the framework described in Chapter 5 is developed and evaluated. Finally, Chapter 7 reviews this thesis and discusses further works.

Chapter2

Review of speech synthesis systems

2.1 Introduction

In this chapter, we will review techniques of speech synthesis. First, the history of speech synthesis is briefly described. Next we will overview current speech synthesis systems. They are divided into two categories: systems which require symbol inputs and systems which do not. For each category, the mechanism and the applications are introduced. Looking through those technologies, the objective of our speech synthesis system will also be made clear.

2.2 History

To understand a technology, learning how that technology has been developed over time would be useful. Therefore as the first topic in this chapter, the history of speech synthesis is described. It has already been in some good texts [16, 17, 18, 19, 20]. Thus, here, only a few interesting landmarks in them will be introduced.

Speech research started in the 18th century. Dodart, in 1700-1707, found the pitch of the voice being dependent on the tension of the vocal folds. The first speaking machines are considered to be those of Kratzenstein [21] and those of von Kempelen [17, 22, 23]. Kratzenstein's resonators were able to produce five vowels (/a/, /i/, /u/, /e/ and /o/) statically and won the prize in 1779 offered by the Imperial Academy of Sciences at St. Petersburgh.

On the other hand, Wolfgang von Kempelen invented a mechanical synthesizer, "speaking machine." This machine used a slamming reed and hissing whistles as sources that corresponded to the vocal cords of human, and a box as a resonance unit corresponding to the vocal tract. It was the first mechanism that allowed the production of not only some speech sounds, but also entire words and short sentences. It was able to produce nineteen consonants and five vowels [17]. Kempelen's speaking machine is shown in Figure 2.1. While working on his speaking machine, Kempelen demonstrated a speaking chess-playing machine. His real speaking machine was therefore not taken so seriously [18].

In the 19th century, theoretical investigations of the vowels were carried out. Helmholz successfully created the electromechanical speech synthesizer, which uses tuning forks, renowned for their pure tone, to generate a fundamental frequency and the first six overtones which may then be combined in varying proportions [24]. Another one of the first analog synthesizer of the human speech organs was presented by Stewart [17] in 1922. Two resonant circuits were excited by a buzzer in this device, permitting approximations to static vowel sounds by adjusting resonance frequencies to the lowest two natural acoustic resonances of the vocal tract (formants) for each vowel.

The discovery of the X-ray by Roentgen in 1895 opened up new areas of research based on imaging techniques. The X-ray observations of the tongue position during vowel production

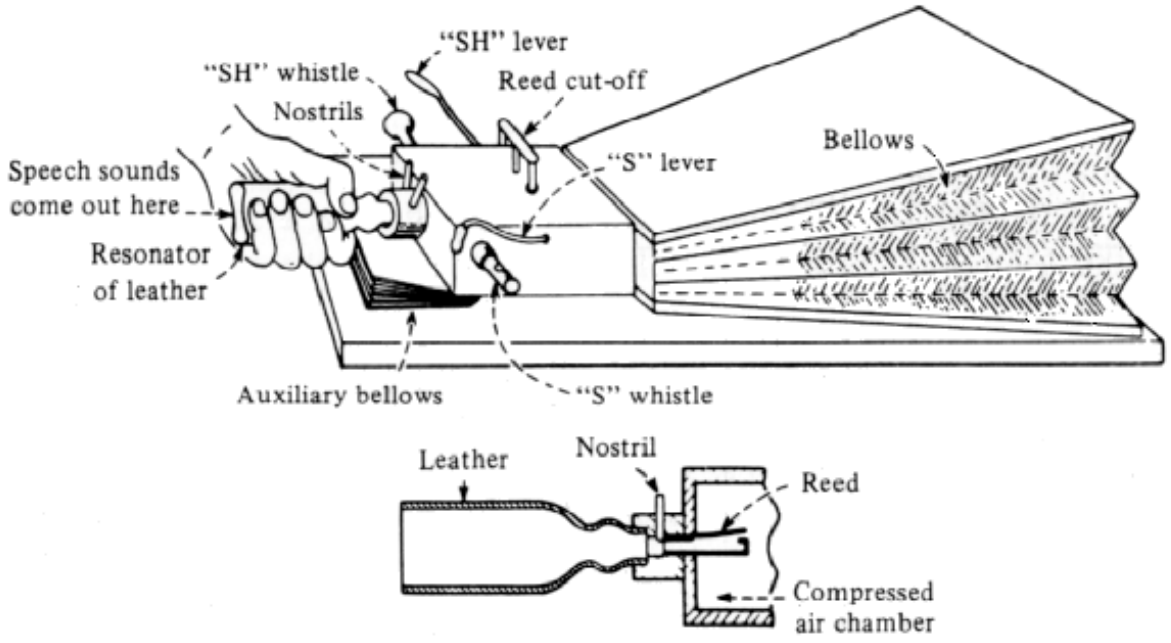


Figure 2.1: Wheatstone's reconstruction of von Kempelen's speaking machine [17].

evoked a debate on the vowel chart. At the time, vowels were described by various types of charts that indicated a triangle or quadrilateral distribution of tongue positions [19]. This dispute was finally ended by Chiba and Kajiyama [25] in 1941. The models they used were reconstructed by Arai with rubber [6] in 2001 and it shows the great educational effect in the class of acoustic phonetics or in the demonstration of speech production mechanism.

In the 20th century, development of various recording and imaging techniques, such as Edison's phonograph and photographic techiniques, contributed to advances in speech research. In 1939, the first device to be considered a speech synthesizer, VODER (Voice Operating Demonstrator), was introduced by Homer Dudley in New York World's Fair.

After VODER finally showed the possibility of artificial speech production, speech synthesis study attracted the interest of researchers and has been widely studied since then. Currently speech synthesis techniques are divided into categories shown in Figure 2.2. In the following sections, those techniques are described in detail.

2.3 Synthesis method which takes symbols as input

Speech synthesis systems, which take symbols as its input, are called Text-to-Speech (TTS) converters. Although they have the word " text " in their name, their input does not need to be a text sequence, it can be a sequence of discrete gestures or pictures. The symbol means those which do not contain phonetic or phonological information. Because

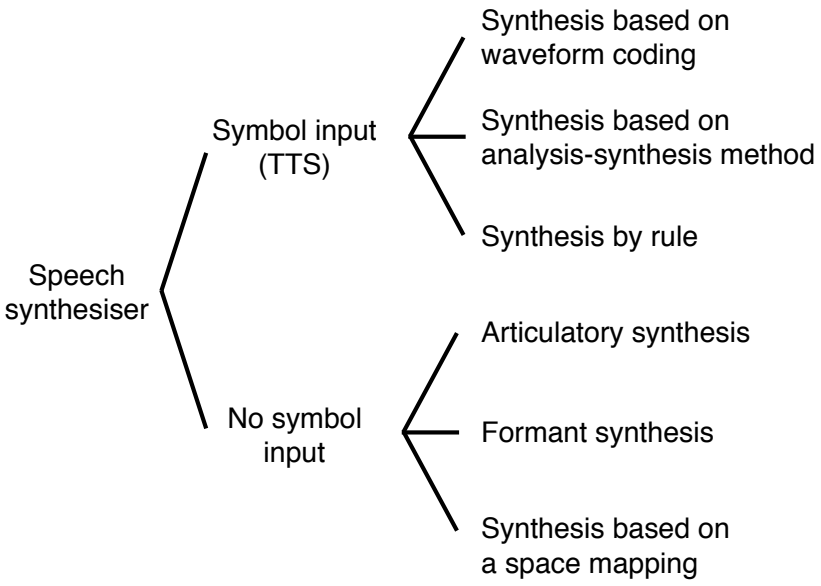


Figure 2.2: Categories of current speech synthesizers.

of the ease of the input method and the quality of the synthesized speech, TTS converters have been widely studied and are used in many practical applications today.

To build a TTS converter, mapping between symbols and sounds needs to be learned from a speech corpus, a set of texts and corresponding speech waveforms. When some text is given, the TTS converter derives the optimal speech waveform corresponding to the text based on the mapping. In other words, TTS can be considered an optimization problem where optimal speech signals X are generated when a word sequence, or linguistic information W is given. From the viewpoint of stochastic theory, this problem is written as:

$$\hat{X} = \underset{X}{\operatorname{argmax}}\, P(X|W). \tag{2.1}$$

Present TTS converting methods can be divided into three types [26]. For the practical use, an appropriate method should be chosen considering the purpose of the application and the performance of the devices.

Synthesis based on waveform coding

This method stores waveform in the form of short segmental units, typically words or phrases. When the input text W is given, appropriate samples are chosen from the database and are concatenated to generate signals X. Although this method provides high quality speech, the continuity of chosen samples affects the quality. As the size of units in the database, such as phrases, sentences, syllables or phonemes, get larger, the quality of synthesized speech, in turn, improves. When the size of units in the database increases, however, synthesizable words decrease. In this method,

therefore, the quality and the size of units in the database is a trade-off. There are several studies based on unit-selection such as [27, 28, 29, 30, 31].

Synthesis based on the analysis-synthesis method

This method stores the time sequences of acoustic parameters extracted from recorded speech waves or acoustic models, which model the mapping of words or phonemes with accoustic parameters. When the input text W is given, appropriate parameter sequences are concatenated and the speech synthesizer generates speech using the concatenated parameter sequences to obtain output signals X. In terms of the naturalness, synthesized speech samples by this method are inferior to those by waveform coding. This method has however, three significant advantages. First, this method only requires the parameters for synthesis, the amount of information is small. Secondly, it is possible to modify synthesized speech waveforms by modifying the parameters. Thirdly, this method shares in the bounty of recent developments of statistical approaches to generation. Since HMM (Hidden Markov Models) is a generative model, it can be used to model the feature parameter production process [32]. Recently HMM-based speech synthesis has become hot topic in speech synthesis studies [33, 34].

Synthesis by rule

In this method, speech is produced based on phonetic and linguistic rules from letter sequences or sequences of phoneme symbols and prosodic features. Synthesizers based on this method are highly complicated, however they have great versatility.

TTS has been widely studied and now reaches the practical level. Nowadays, most of the speech synthesizers are TTS converters. They have however, still some drawbacks, such as continuous speech generation and emotional speech synthesis. Some researchers have been working on those challenges in the framework of TTS. Although they are very interesting topics to study, in this thesis we will put more focus on them with the other synthesis method which does not require symbol input.

2.4 Synthesis methods which do not require symbol inputs

Systems which does not require symbol inputs are effective for smooth speech synthesis and speech rate / pitch control, based on dynamic body motion, while synthesized sounds are often less articulate than those of TTS. They could also be used for basic speech research because of the strong association between input movements and output speech.

Three methods have been proposed to construct synthesizers of this category. One is articulatory synthesis, which simulates the acoustic wave propagation in the vocal tract. The second one is the formant synthesis, which simulates the resonance and antiresonance

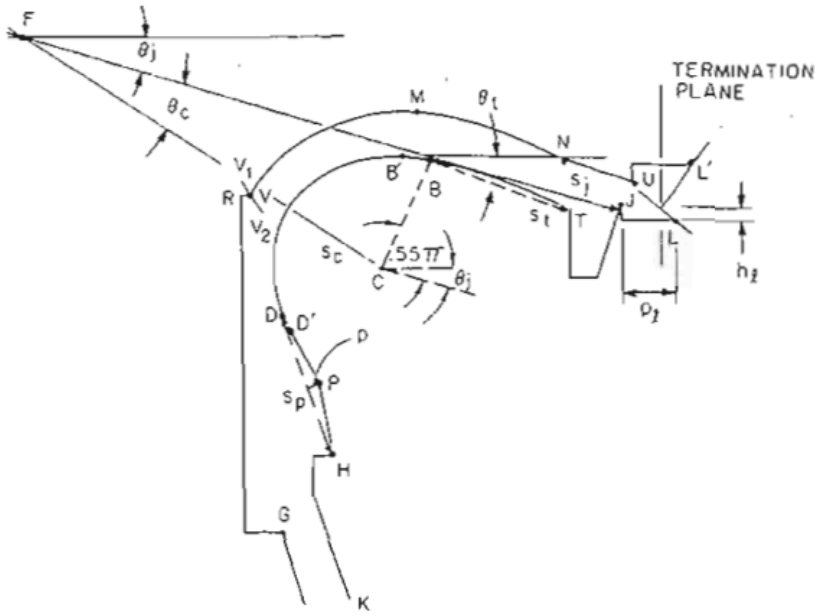


Figure 2.3: Mermelstein's articulatory model [36].

characteristics of the vocal tract, and, as a result, reproduces articulation function. The last one is synthesis based on space mapping. In this method, the mapping function between the non-acoustic space and the acoustic space is learned. The details of each method will be introduced in the following section.

2.4.1 Articulatory synthesis

Articulatory speech synthesis is a synthesis method inspired by the speech production process of human beings. In terms of the quality of the synthesized speech, articulatory synthesis would not be the best method. For centuries however, this method has been attracting researcher's interests. Articulatory speech synthesis may be used as a tool in basic speech research and is in itself a subject of basic speech research [16]. If we understand the human speech production mechanism perfectly, the synthesized speech would be indistinguishable from human voice. In this section, we overview the current technology of articulatory synthesis.

Articulatory synthesis can be divided into two steps: a data acquisition step and an acoustic synthesis step. The first step can be said to have been developed along with devices to measure human articulatory movements. Until the 1980's, Cineradiography has been widely used to capture human articulatory organs. There is no doubt that it contributed to the birth of well used models, such as the source-filter model by Fant [35], Mermelstein's model shown in Figure 2.3 [36], and Maeda model [37]. Cineradiography however, has gradually been replaced with other devices because of its significant X-ray dosage. Cur-

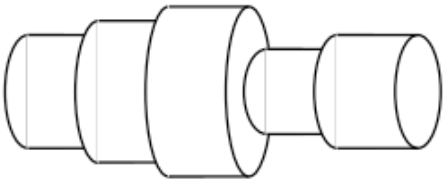


Figure 2.4: The discrete tube model of the vocal tract.

rently, Magnetic Resonance Imaging (MRI) and Electromagnetic Articulography (EMA) are widely used instead. Dang *et al.*, for example, tried to model the movement of articulatory organs using the images from MRI and EMA [38]. Ultrasound, Electropalatography (EPG), motion capture and Electromyography (EMG) etc are other possibilities to capture the posture or the movement of the articulatory organs. Appropriate devices should be chosen, depending on the vocal tract model or the purpose of a project.

Using the data acquired during the first step, a speech signal is obtained in the next step. Since speech is described as a distribution of air pressure, the problem here is to derive the distribution of air pressure in the vocal tract model when the vocal tract shape is given. As assuming that the air in the vocal tract is an ideal gas and there is no mass source in that, the vocal tract geometry can be considered to satisfy the condition of Webster's Horn Equation described as follows:

$$\frac{\partial^2 p(x,t)}{\partial t^2} = c^2 \frac{1}{A(x,t)} \frac{\partial}{\partial x} \left[A(x,t) \frac{\partial p(x,t)}{\partial x} \right], \tag{2.2}$$

where x is the displacement following the axis of the vocal tract and t denotes time. p and A are perturbation pressure (sound pressure) and the area of the vocal tract's cross-section, respectively. c denotes the sound speed. Equation 2.2 has no analytical solution, as long as A is a function of both x and t. It can be solved however, if A is a function of only x.

One way of avoiding the problem in solving Equation 2.2, is to assume the shape of vocal tract as concatenated tubes and let the diameter of tubes change at only specified points. This means the vocal tract is approximated by the model shown in Figure 2.4. The acoustic tube model has been widely used. Ogata *et al.* tried to model the speech production process using a 20 node acoustic tube model [39]. They also modeled the movement of the vocal cords with a two mass model proposed by Ishizaka [40]. Another example of the speech synthesizers based on the acoustic tube model is the Articulatory SYnthesis program (ASY) [41], developed in Haskins laboratory. The Mermelstein model [36] is embodied in ASY (see Figure 2.3). There are 6 key parameters in this model: the tongue body center (C, 2 degrees of freedom (df)), the tongue tip (T, 2 df), the jaw (J, 1 df), the lips (L, 2 df), the velum (V, 1 df), and the hyoid (H, 2 df, controlling larynx height and pharynx width). The tongue tip is a structure that rests on the tongue body, which is implemented as a ball. Once these parameters have been specified, the vocal tract is then converted into a series of uniform tube sections.

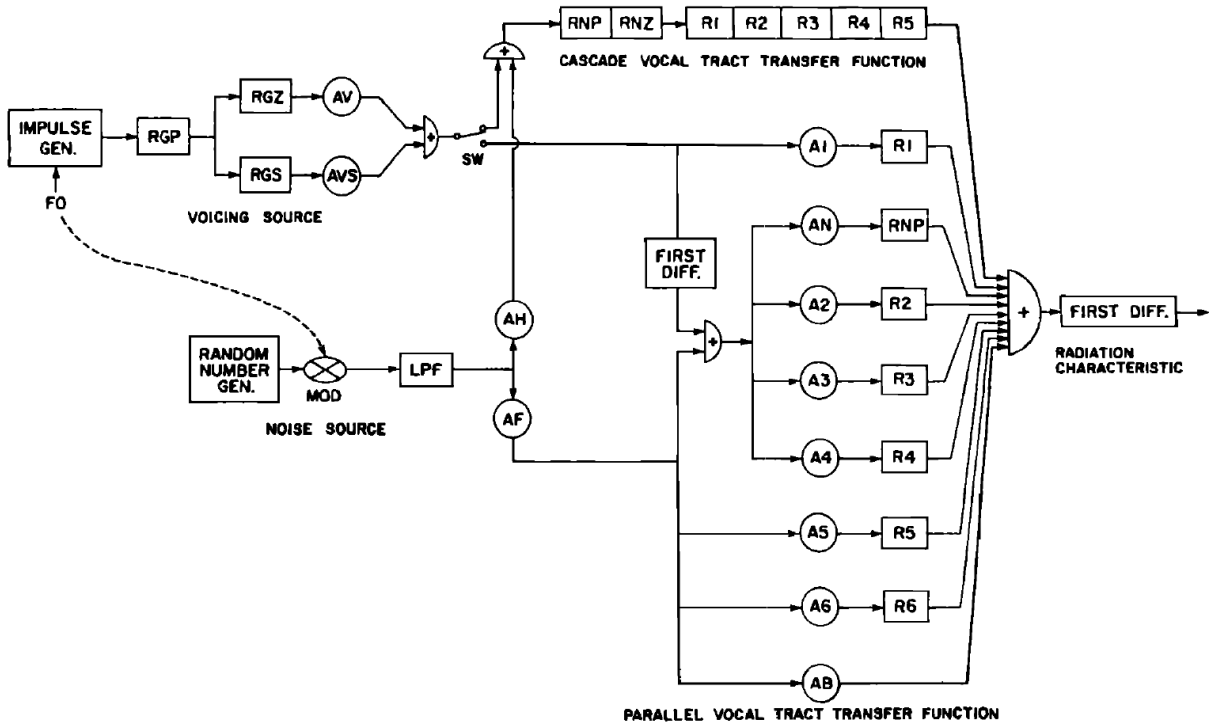


Figure 2.5: One structure of a formant speech synthesizer [46].

Recently, along with the development of measurement devices and computers, the obtained data have become more precise and thus some of the idealizations in the vocal tract models used previously have been removed, for example vocal tracts do not need to be symmetrical (2D model). As a result, Engwall [42] and Birkholz *et al.* [43] modeled the vocal tract shapes in 3D. Fels et al. developed Artisynth, which is a 3D biomechanical simulation platform directed toward modelling the vocal tract and upper airway [44].

2.4.2 Formant synthesis

Fant named the spectral peaks of the sound spectrum " formant " [45]. Formants are broken down into the first, second, third $\cdots$ formant from the lowest frequency component and written as F_1, F_2, F_3 $\cdots$. Formant frequencies express resonant frequencies of the vocal tract. The speech generation method using an electrical structure consisting of the cascade or parallel connection of several resonance (formant) and anti-resonance (anti-formant) circuits is called the formant-type synthesis method or the terminal analog method [26]. In this method, the resonance and anti-resonance frequency and bandwidth of each circuit are variable. Figure 2.5 shows a typical example of the structure of a synthesizer which is constructed based on these considerations [46]. Here F_0 =fundamental frequency, Rx =resonance circuits controlled by resonance frequency and band-width, PGP, RGS =glottal resonance circuits, RGZ =glottal anti-resonance circuit, RNP,

RNZ =resonance and anti-resonance circuits for nasals, Ax =amplitude.

Formant-type speech synthesis has three major advantages. First, formant-type synthesis can generate intelligible speech even at very high speeds. Secondly, unlike corpus-based synthesis, formant-type synthesis does not require a database of speech samples. Thirdly, the relation between body motion and generated speech is explicit. Taking advantage of those points, several unique applications are proposed.

As formants are resonant frequencies of the vocal tract, when the vocal tract changes its shape, formants also change. For example, the formant frequencies of /a/ and those of /i/ are different because the vocal tract shape /a/ differs from that of /i/. Figure 2.6 shows the first and the second formant frequencies of five Japanese vowels. Since the vocal tract shape changes according to the speaker's age and sex, formant frequencies are widely distributed and those distributions are overlapped slightly. When F_1 and F_2 are given however, it is possible to roughly derive which vowel sound those F_1 and F_2 would correspond to.

A dysarthric engineer, Yabu, embedded the F_1-F_2 plane in a pen tablet[8]. The pen position controls F_1 and F_2 of vowel sounds and the pen pressure controls their energy. They reported that rapid formant transition observed in the beginning of consonants could be created only with pen movement on the pen tablet and as a consequence, it is possible for users to make consonant-like sounds. Another example is Glove-Talk II proposed by Fels [5]. With several devices equipped to the user (including a Cyberglove, a ContactGlove, a three-space tracker, and a foot pedal) and three neural networks, body motions are transformed into parameters for a formant speech synthesizer. One subject who was trained on how to speak with Glove-Talk II is able to create far more natural sounding pitch variations than a text-to-speech synthesizer available of the time when Glove-Talk II was proposed.

The explicit relationship between formant and vocal tracts is used not only as speech generation systems but also as speech education systems. With the comparison of the vowel chart shown in Figure 2.7, correspondence between F_1 and the tongue location, and correspondence between F_2 and the rounding the lips, are expected. Stevens *et al.* mentioned the examples of the relationship between vocal tract shapes and vowel spectral envelopes as shown in Figure 2.8 [47]. This characteristic opens up the possibility of a real-time speech-learning tool for hearing impaired people. It is difficult for hearing impaired people to learn how to articulate speech because they cannot get feedback from their own articulation. By checking the location of formants in real time however, they can know how their speech is perceived and how to adjust their vocal tract shape to improve their elocution. Sakata *et al.* proposed audio-visual real-time feedback of vowel speech based on the normalized articulation space implemented using formants [7].

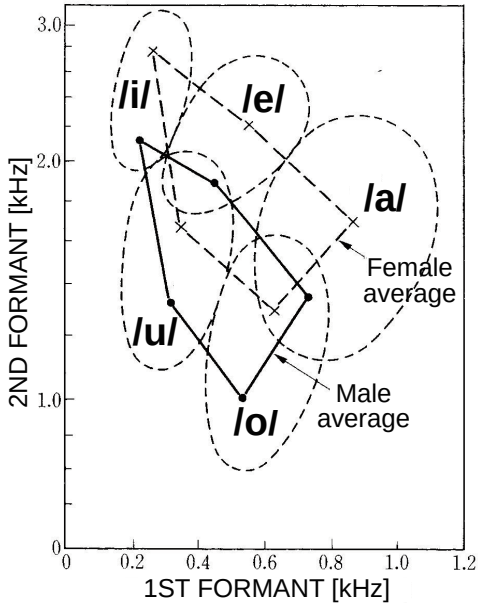


Figure 2.6: The five Japanese vowels in the F_1–F_2 plane.

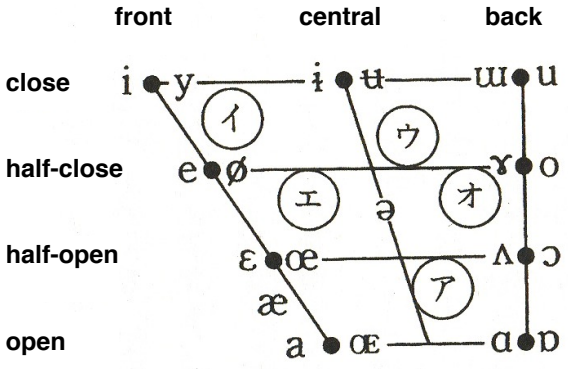


Figure 2.7: The vowel chat for Japanese.

2.4.3 Synthesis based on space mapping

In the previous section, the explicit relationship between formant and vocal tracts was mentioned. Synthesis based on space mapping creates a more direct relationship between feature sequences and sounds. People usually use tongue gesture transitions to generate a speech stream. This is considered as the inherent mapping between tongue gestures and speech sounds. If the tongue gesture space can be replaced with another media in this process, we could speak using that media just like we speak using tongue movements. To realize this, the mapping problem between two media spaces has to be solved.

In the field of speaker conversion, GMM-based conversion techniques have been widely

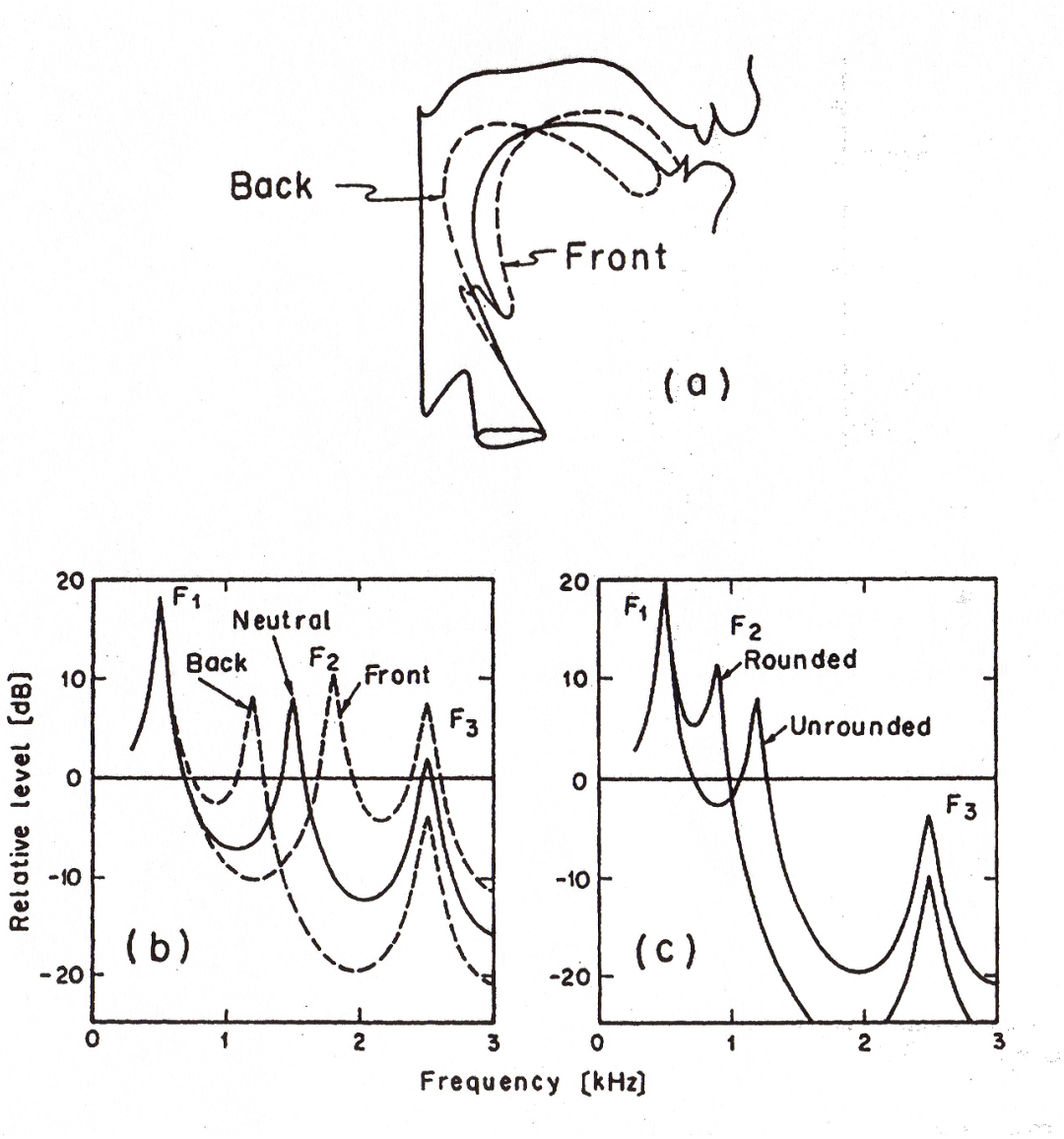


Figure 2.8: The relationship between vocal tract shapes and spectral envelopes [47].

studied. This is where the voice spaces of two speakers are mapped to each other and the mapping function is estimated based on a GMM [13, 14, 15]. Recently, this technique has been directly applied to estimate a mapping function between non-acoustic media and speech media. Toda and Tokuda mapped the articulatory movement with speech [48], Hueber *et al.* used this method to map the acoustic space to the articulatory space and vice versa [49]. Nakamura *et al.* synthesized speech using this method from electromyography signals [50].

This method requires a parallel data set between two spaces to learn the correspondence between two spaces. That is to say, which features should be used for mapping and how to design the correspondence between those spaces are key issues. When the input is

media which have an explicit relationship with the movements of articulatory organs, it is relatively easy to get a good correspondence between the feature spaces of that media and speech. For example, we know which lip shapes should correspond to which sounds. When the input is media that have no explicit relationship with the movements of articulatory organs, such as hand gestures or pen movements, any correspondence between input and target spaces is possible. Of course, the correspondence which synthesizes intelligible sound as well as provides ease-of-use is ideal. To choose such a correspondence is however difficult. Some correspondences will not work as desired. Media which have no explicit relationship with the movements of articulatory organs have not often been chosen in this framework.

2.5 The goal of our research

In this thesis, we focus on speech generation systems which do not require symbol input. As we have seen above, several unique and practical applications have already been proposed. Their sophisticated methodologies are however, difficult to apply to other applications. For example, most articulatory synthesis methods are based on data from one person and it is not easy to adapt to another person. Applications described in Section 2.4.2 are designed to make optimal use of the features specific to the input device. Once the input device is changed for some reason, those systems would need to be established again with the new media. The possibility of media selection is, however, very important especially in the fields in which media should not be chosen by developers but by users, such as assistive technology and art. Our goal is to establish the media-independent methodology for speech generation system.

Considering the freedom of media selection, we extend the speech synthesis framework based on space mapping to desired input media. As we have seen above, the effectiveness of this framework has already been proven in some works such as [48], [49] and [50] when the correspondence between the input media and speech is relatively explicit. If this framework works even when the correspondence is not explicit, it would be a media-independent synthesis framework and would be able to applied to a wide range of applications.

In our research, speech synthesis based on this framework, when the input media does not have explicit relationship to speech, has been implemented and the methodology how to derive the optimal correspondence between the input media and speech has been studied. As one example of such a conversion, a hand gesture to speech conversion system has been adopted.

2.6 Summary

This chapter described the history of the speech synthesis technologies. In addition, we have seen two categories of conventional speech synthesis systems; systems which require

symbol inputs and systems which do not require symbol inputs. Three synthesis methods belongs to the former: synthesis based on waveform concatenation, synthesis based on the analysis-synthesis method and synthesis by rule. The latter also consists of three methods: articulatory synthesis, formant synthesis and media conversion based on space mapping. Looking through those technologies, we have made clear the objective of our speech synthesis system - a media-independent speech synthesis. Among the current speech synthesis techniques, we have found that media conversion based on space mapping is the most suitable framework for the objectives. In the next chapter, the basic idea of our proposed speech synthesis system under the media conversion framework is described.

Chapter3

Framework of the GMM-based media conversion

3.1 Introduction

As we mentioned in the previous chapter, the objective in our study is to establish a media-independent methodology for a speech generation system. In order to reach the objective, we have to consider the speech synthesis framework based on space mapping to desired input media. When the correspondence between the input media and speech is relatively explicit, the effectiveness of this framework is already evident. In this thesis therefore, we are considering the speech synthesis based on this framework when the input media does not have explicit relationship to speech. As an example of such media, hand gesture has been chosen and a Hand gesture to Speech (H2S) converter is developed.

In this chapter, the framework of our H2S system is described. Firstly, the acoustic features, which are used in our framework, are introduced. Next as the basic framework of our system, an overview of voice conversion based on space mapping is given. Then the framework of voice conversion is extended into media conversion, in which an acoustic feature spaces of source speakers is replaced with hand gesture. Finally the challenge of our H2S system is described.

3.2 Acoustic features

When we speak, lungs send air and it causes the periodical vibration of a vocal cord. This movement creates periodical pressure wave. When this wave goes through the tube towards the mouth - this part is called vocal tract - resonant frequency components of the vocal tract are amplified and human speech sounds are produced. This speech production process can be considered as two parts, a source part by the vocal code and a filter part by the vocal tract. This model is called a source-filter model [35]. According to the source-filter model, the frequency distribution of the produced speech $S(\omega)$ is written as follows:

$$S(\omega) = G(\omega)H(\omega). \tag{3.1}$$

Here, $G(\omega)$ and $H(\omega)$ denote the frequency distributions of the vocal code and the vocal tract, respectively. Roughly speaking, information by vocal cords, $G(\omega)$, corresponds to para-linguistic information, and information by vocal tract, $H(\omega)$, includes both linguistic and non-linguistic. Since human speech production mechanism includes some non-linear factors in reality, it cannot be perfectly explained by Equation (3.1). It shows however, abundant performance in speech recognition as well as speech synthesis and is widely used in the speech processing.

In speech signal processing, acoustic features based on the source-filter model, are widely used. Fundamental frequency of vocal code (F0) is used for the information by vocal code. To describe the information by vocal tract, an acoustic feature, called cepstrum, is used.

Figure 3.1 shows the process of speech analysis where cepstral features are extracted

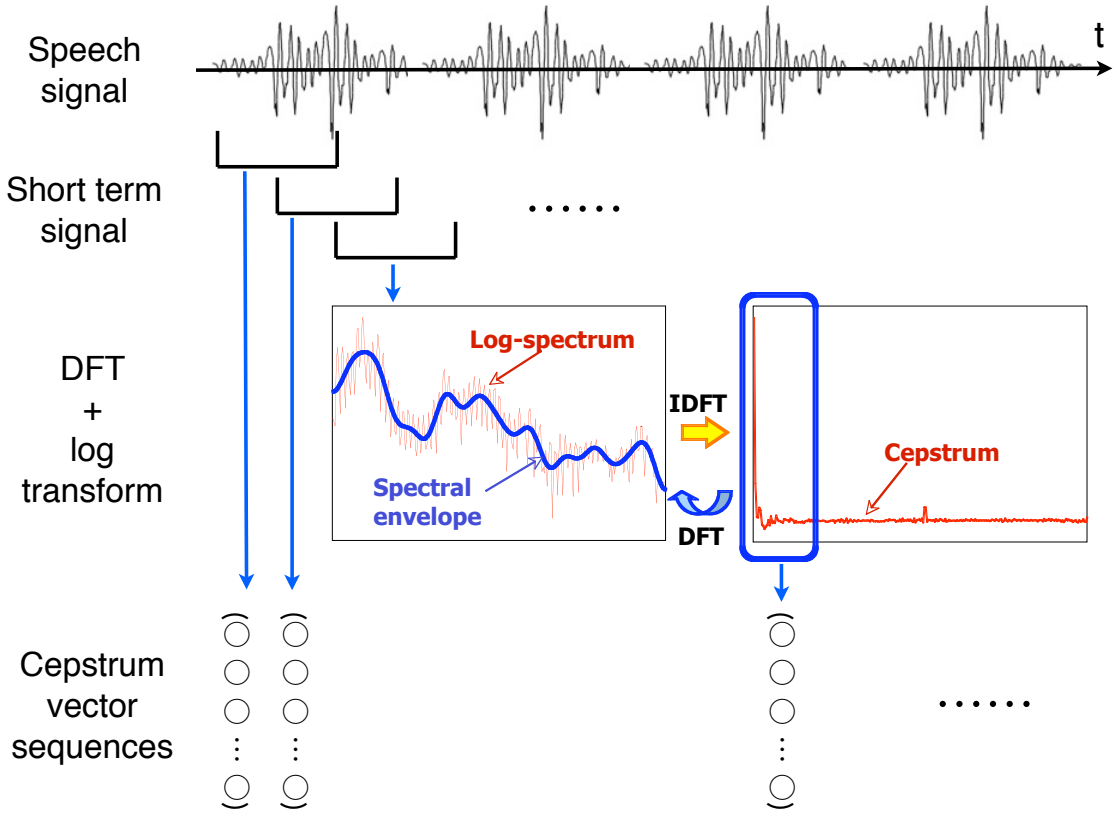


Figure 3.1: Speech analysis.

from the input speech signal. After the input speech signal is segmented into different frames, the acoustic features are extracted from every frame. For each windowed signal, a short time discrete Fourier transform (STDFT) is applied to convert the time domain signal into the frequency or spectral domain. Then cepstrum can be obtained using the inverse DFT of the logarithm of the power spectrum. Through this process, speech signals are expressed with cepstrum vector time sequences.

3.3 Framework of voice conversion

Speech synthesis based on media conversion can be considered as an extension of voice conversion. Thus, in this section, the basic framework of voice conversion is introduced. Voice conversion is the technique that modifies the source speaker's speech as if the target speaker had spoken it. Based on the source-filter model, voice conversion is roughly divided into two parts: spectral conversion for voice quality and F0 conversion for intonation. The details of each part will be introduced in the following section.

3.3.1 Spectral conversion

In this section, statistical spectral conversion technique is introduced.

Let $\boldsymbol{x} = [\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_p]$ be the feature vector sequence, such as cepstrum vector sequence, of a source speaker and $\boldsymbol{y} = [\boldsymbol{y}_1, \boldsymbol{y}_2, ..., \boldsymbol{y}_p]$ be that of the target speaker. The aim of spectral conversion is to get a mapping function $\mathcal{F}(\cdot)$ so that the converted $\mathcal{F}(\boldsymbol{x}_t)$ matches the best to the target vector $\boldsymbol{y}_t$ for all training data.

Many techniques have been proposed to solve this problem. Abe *et al.* made mapping codebooks which represent the correspondence between different speaker's codebooks based on hard clustering [51]. The mapping codebooks for spectrum parameters, power values, and pitch frequencies are separately generated using training utterances. Converted feature vector $\hat{\boldsymbol{y}}$ is described using a corresponding centroid vector $\boldsymbol{c}_m$ of the mapping codebook as follows:

$$\hat{\boldsymbol{y}}_t = \boldsymbol{c}_m^{(y)}. \tag{3.2}$$

Abe *et al.* reported that in the male–to–female conversion, all converted utterances are judged as female, and in the male–to–male conversion, 65% of them are identified as the target speaker. Their method however, can obtain only one centroid vector in the codebook for each frame.

To alleviate this problem caused by hard clustering, Nakamura and Shikano introduced Fuzzy vector quantization [52]. Fuzzy vector quantization represents an input vector as a weighted combination of fuzzy membership function and code-vectors. Replacing the input speaker's code-vectors with the mapped code-vectors, mapping based on the soft clustering is realized as follows:

$$\hat{\boldsymbol{y}}_t = \sum_{m=1}^{M} \omega_{m,t}^{(x)} \boldsymbol{c}_m^{(y)}, \tag{3.3}$$

where M is the number of centroid vectors. Matsumoto *et al.* modeled the difference vector between the source and target vectors [53] as follows:

$$\hat{\boldsymbol{y}}_t = \sum_{m=1}^{M} \omega_{m,t}^{(x)} (\boldsymbol{c}_m^{(y)} - \boldsymbol{c}_m^{(x)}). \tag{3.4}$$

Their method is based on the correspondence of centroid vectors described in codebooks. In order to directly model the correspondence between points in two spaces, Linear Regression Method (LRM) written in the following form is proposed in [54].

$$\hat{\boldsymbol{y}}_t = \boldsymbol{A}_m x_m + \boldsymbol{b}_m, \tag{3.5}$$

where $\boldsymbol{A}_m$ and $\boldsymbol{b}$ are regression parameters.

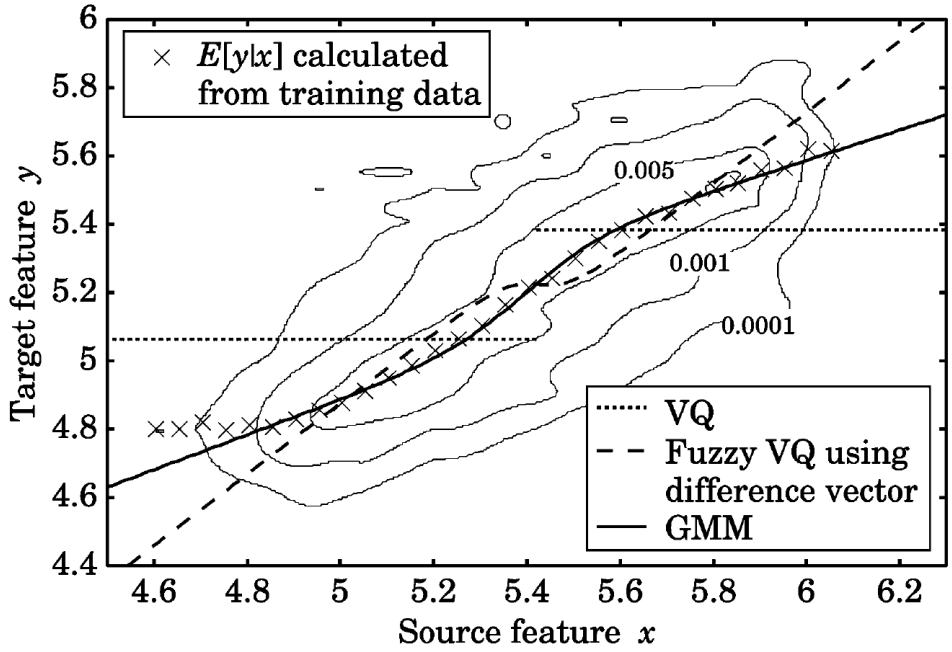


Figure 3.2: Mapping function on the joint feature space [15].

Stylianou [14] assumed the distribution of the source vectors under the form of a continuous probability distribution based on a Gaussian Mixture Model (GMM). Then he extended the idea of LRM and realized the continuous mapping, described as follows.

$$\hat{\boldsymbol{y}}_t = \sum_{m=1}^{M} \omega_{m,t}(\boldsymbol{A}_m x_m + \boldsymbol{b}_m), \tag{3.6}$$

where $\omega_{m,t}, \boldsymbol{A}_m$ and $\boldsymbol{b}_m$ are provided by GMMs. Equation (3.3) is the special form of Equation (3.6), where $\boldsymbol{b}_m$ is zero vector.

Some mapping functions on a joint space of one-dimensional source and target features is shown in Figure 3.2 [15]. The contour lines show normalized frequency distribution of training data samples. As the figure shows, the GMM-based mapping function is the closest to the expectation. Thus the GMM-based mapping becomes the most popular method for the voice conversion and has been widely studied.

3.3.2 GMM-based mapping

Although Stylianou used GMMs to model the distribution of source vectors when the GMM-based mapping was proposed in [14], GMMs are now generally used to model the distribution of the augmented vectors $\boldsymbol{z}_t = [\boldsymbol{x}_i^\top, \boldsymbol{y}_t^\top]^\top$, as Kain *et al.* proposed in [13]. The notation $\top$ denotes transposition. Using GMMs, the joint probability density is modeled

as follows:

$$P(\boldsymbol{z}_t|\boldsymbol{\lambda}^{(z)}) = \sum_{m=1}^{M} \omega_m \mathcal{N}(\boldsymbol{z}_t; \boldsymbol{\mu}_m^{(z)}, \boldsymbol{\Sigma}_m^{(z)}), \tag{3.7}$$

where $\boldsymbol{\lambda}$ is a parameter set of the GMM, M is the number of mixtures, ω_m is the weight of the m-th mixture component, $\mathcal{N}(;\boldsymbol{\mu}_m^{(z)}, \boldsymbol{\Sigma}_m^{(z)})$ is the normal distribution with mean $\boldsymbol{\mu}_m^{(z)}$ and covariance $\boldsymbol{\Sigma}_m^{(z)}$, which are written as:

$$\boldsymbol{\mu}_i^{(z)} = \begin{bmatrix} \boldsymbol{\mu}_i^{(x)} \\ \boldsymbol{\mu}_i^{(y)} \end{bmatrix}, \boldsymbol{\Sigma}_i^{(z)} = \begin{bmatrix} \boldsymbol{\Sigma}_i^{(xx)} & \boldsymbol{\Sigma}_i^{(xy)} \\ \boldsymbol{\Sigma}_i^{(yx)} & \boldsymbol{\Sigma}_i^{(yy)} \end{bmatrix}. \tag{3.8}$$

Conditional probability density of $\boldsymbol{y}_t$ given $\boldsymbol{x}_t$ is described as follows:

$$P(\boldsymbol{y}_t|\boldsymbol{x}_t, \boldsymbol{\lambda}^{(z)}) = \sum_{m=1}^{M} P(m|\boldsymbol{x}_t, \boldsymbol{\lambda}^{(z)}) P(\boldsymbol{y}_t|\boldsymbol{x}_t, m, \boldsymbol{\lambda}^{(z)}). \tag{3.9}$$

The first and second terms describe the component weight and the probability density of each component, respectively. Using these parameters in Equation (3.8), they are described as follows:

$$P(m|\boldsymbol{x}_t, \boldsymbol{\lambda}^{(z)}) = \frac{\omega_m \mathcal{N}(\boldsymbol{x}_t; \boldsymbol{\mu}_m^{(x)}, \boldsymbol{\Sigma}_m^{(xx)})}{\sum_{n=1}^{M} \omega_n \mathcal{N}(\boldsymbol{x}_t; \boldsymbol{\mu}_n^{(x)}, \boldsymbol{\Sigma}_n^{(xx)})}, \tag{3.10}$$

$$P(\boldsymbol{y}_t|\boldsymbol{x}_t, m, \boldsymbol{\lambda}^{(z)}) = \mathcal{N}(\boldsymbol{y}_t; \boldsymbol{E}_{m,t}^{(y)}, \boldsymbol{D}_m^{(y)}), \tag{3.11}$$

where

$$\boldsymbol{E}_{m,t}^{(y)} = \boldsymbol{\mu}_m^{(y)} + \boldsymbol{\Sigma}_m^{(yx)} \boldsymbol{\Sigma}_m^{(xx)-1} (\boldsymbol{x}_t - \boldsymbol{\mu}_m^{(x)}), \tag{3.12}$$

$$\boldsymbol{D}_m^{(y)} = \boldsymbol{\Sigma}_m^{(yy)} - \boldsymbol{\Sigma}_m^{(yx)} \boldsymbol{\Sigma}_m^{(xx)-1} \boldsymbol{\Sigma}_m^{(xy)} \tag{3.13}$$

Stylianou and Kain *et al.* converted source feature $\boldsymbol{x}_t$ vector into a target vector $\boldsymbol{y}_t$ as below, so as to minimize the mean-square error (MMSE) $\sum [|y - \mathcal{F}(x)|^2]$ [14, 13]:

$$\hat{\boldsymbol{y}} = \sum_{m=1}^{M} P(m|\boldsymbol{x}_t, \boldsymbol{\lambda}^{(z)}) \boldsymbol{E}_{m,t}^{(y)}. \tag{3.14}$$

Here, as replacing $P(m|\boldsymbol{x}_t, \boldsymbol{\lambda}^{(z)})$ with $\omega_{m,t}^{(x)}$, $\boldsymbol{\Sigma}_m^{(yx)} \boldsymbol{\Sigma}_m^{(xx)-1}$ with $\boldsymbol{A}_m$ and $\boldsymbol{\mu}_m^{(y)} - \boldsymbol{\Sigma}_m^{(yx)} \boldsymbol{\Sigma}_m^{(xx)-1} \boldsymbol{\mu}_m^{(x)}$ as $\boldsymbol{b}_m$, Equation (3.6) is obtained.

Toda *et al.* performed the conversion so that the following likelihood will be maximized:

$$\hat{\boldsymbol{y}} = \operatorname{argmax} P(\boldsymbol{y}_t|\boldsymbol{x}_t, \boldsymbol{\lambda}^{(z)}). \tag{3.15}$$

Conversion based on Maximum Likelihood (ML) criterion is written as follows [15]:

$$\hat{\boldsymbol{y}} = \left(\sum_{m=1}^{M} \gamma_{m,t} \boldsymbol{D}_m^{(y)-1} \right)^{-1} \left(\sum_{m=1}^{M} \gamma_{m,t} \boldsymbol{D}_m^{(y)-1} \boldsymbol{E}_{m,t}^{(y)} \right), \tag{3.16}$$

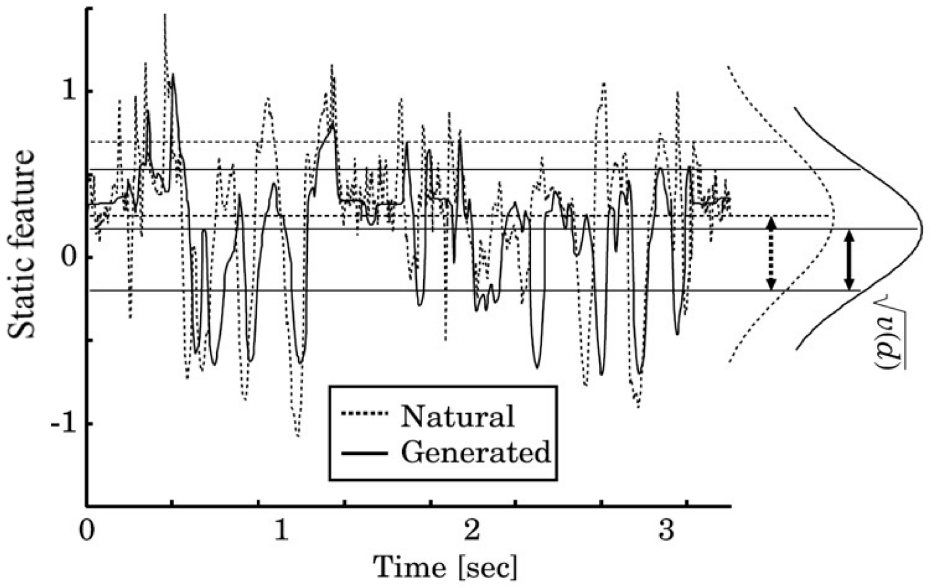


Figure 3.3: Natural and generated mel-cepstrum sequences [55].

here $\gamma_{m,t}$ denotes $P(m|\boldsymbol{x}_t, \boldsymbol{\lambda}^{(z)})$.

GMM-based mapping has two drawbacks [55]. First, the statistical modeling mitigates precipitous peaks in spectra. Consequently, generated spectra are smoothed compared with the natural ones. Figure 3.3 shows the comparison between natural and generated mel-cepstrum[1] sequences [55]. In order to compensate for this problem, Toda and Tokuda introduced the Global Variance (GV). The GVs are calculated utterance by utterance and tend to be smaller after the conversion. A square root of the GV of each sequence is shown as a bidirectional arrow in Figure 3.3. Toda and Tokuda reported that a perceptual evaluation shows that considerably large improvement in the naturalness of synthesized speech generated by the conversion considering the GV [55].

The other problem is that in this frame-by-frame mapping framework, the correlation of the feature vectors is ignored. In order to compensate for this problem, Toda *et al.* proposed a conversion method based on the maximum likelihood estimation of a spectral trajectory [15]. In their framework, not only static but also dynamic features are used.

There are methods to improve the quality of GMM-based mapping other than those two. Thanks to the constant challenge to improve the quality, GMM-based mapping techniques realize a quite high performance spectral conversion. These techniques are however, not directly applied to F0 conversion, because F0 is characterized by a discontinuity in the transition between voiced and unvoiced sounds that presents an obstacle to GMM modeling for use in voice conversion. In the next section, some methods for F0 conversion are

[1]Mel-cepstrum is a kind of cepstrum, considering perceptual aspects of listeners [56, 57]

introduced.

3.4 F0 conversion

F0 is a highly variable acoustic feature. Speaker difference in F0 could be determined by a variety of factors, e.g. age, gender, dialectal background, health condition, education and personal style. In voice conversion however, F0 sequences are usually converted by following a simple linear function:

$$p_t^{(Y)} = \frac{p_t^{(X)} - \mu^{(X)}}{\sigma^{(X)}} \times \sigma^{(Y)} + \mu^{(Y)}, \tag{3.17}$$

where $p_t^{(X)}$ and $p_t^{(Y)}$ are input and converted F0 values, respectively. $\mu^{(\cdot)}$ and $\sigma^{(\cdot)}$ are the same mean and the standard deviation of F0, respectively. It assumes that the F0 has a single Gaussian distribution and converted F0 has the same distribution as the target speaker in the conventional voice conversion approach. This assumption is not appropriate for F0 conversion in converting the characteristics of source speaker to target speaker.

As a result special models have been proposed for F0 modeling in HMM-based speech synthesis. A widely used model, the Multi-Space Distribution (MSD) has been proposed by Masuko *et al.* [58]. MSD-HMM models F0 with a discrete subspace for the unvoiced regions and a continuous subspace for the voiced F0 contours. Another popular model, the Globally Tied Distribution (GTD) has been proposed by Yu *et al.* [59]. GTD-HMM assumes that F0 still exists in unvoiced regions and it is distributed according to an underlying globally tied continuous probability distribution field. Recently, Zhang *et al.* proposed to use voicing strength as an additional feature in F0 modeling and for voiced/unvoiced (v/u) decision in [60].

In voice conversion, Yutani *et al.* proposed a simultaneous modeling of spectrum and F0 for voice conversion, where the MSD models unvoiced region and continuous voiced F0 contour in a linearly weighted mixture [61]. However, two incompatible probabilistic spaces, the continuous probability density for voiced observations or the discrete probability for unvoiced observations, may incur an imprecise v/u conversion in a maximum likelihood (ML) sense.

3.4.1 Our proposed F0 model

As a part of the research for media-independent speech synthesis, we have worked on the improvement of F0 modeling and generation in voice conversion [62]. In our proposed method, we extended the method proposed by Zhang *et al.* to F0 conversion, i.e., using voicing strength as an additional feature for improving F0 modeling and the v/u decision.

Voicing Strength (VS) is characterized by the normalized correlation coefficient (NCC) magnitude, which is calculated during F0 feature extraction on a short-time basis by ap-

plying the Robust Algorithm for Pitch Tracking (RAPT)[63]. The NCC magnitude is described in the following formula:

$$\phi_{i,k} = \frac{\sum_{j=m}^{m+n-1} s_j s_{j+k}}{\sqrt{c_m c_{m+k}}}, \tag{3.18}$$

where

$$c_m = \sum_{l=m}^{m+n-1} s_l^2. \tag{3.19}$$

s_j is a sampled speech signal: $i = 0, 1, ..., M-1$ represents a frame index; $k = 0, 1, ..., K-1$ is the lag; n is the sample number in an analysis window; $m = iz$ and z represents the sample number in a frame.

The procedure of our approach to voice conversion is as follows: In the training phase, F0s in unvoiced regions are first interpolated by the spline function. The entire F0 sequence after interpolation and VS sequence extracted by Equation (3.18) are then smoothed with a low-pass filter. Finally, a GMM is trained with both continuous F0 features (F0 and its first order time derivatives) and VS features (NCC and its first derivatives) as well as spectral features. In other words, the source and target feature vectors, $\boldsymbol{x}$ and $\boldsymbol{y}$ in Section 3.3.2, both contain F0, VS and spectral features and these three features are simultaneously modeled by GMM. The optimal number of mixtures for the spectral part and the F0 and VS parts are calculated independently. In F0 conversion phase, both F0 and VS trajectories are generated in the maximum likelihood sense.

In our approach, Global Variance (GV) [55], which was described in Section 3.3.2, is also applied for the generated F0 as follows:

$$v_d^{(w)} = \sqrt{\frac{\sigma^{(Y)}}{\sigma^{(X)}}}(v_d^{(v)} - \mu^{(X)}) + \mu^{(Y)}, \tag{3.20}$$

$$y_i^{(w)} = \sqrt{\frac{v_d^{(w)}}{\sigma^{(v)}}}(y_i^{(v)} - \mu^{(v)}) + \mu^{(Y)}, \tag{3.21}$$

where $v_d^{(w)}$ is the predicted global variance of the converted sentence, $v_d^{(v)}$ is the global variance of the generated sentence, $\mu^{(X)}$ and $\mu^{(Y)}$ are the means of source and target sentence's global variances over all training data respectively, $\sigma^{(X)}$ and $\sigma^{(Y)}$ are the variance of source and target sentence's global variances respectively. $y_i^{(w)}$ is the predicted F0 value, $\mu^{(v)}$ is the mean F0 value of generated sentence. Here, $\mu^{(Y)}$ is assumed to be the same as $\mu^{(w)}$, the mean of the predicted F0 value.

The generated voicing strength for each frame indicates the probability of whether a frame is voiced or not. Frames with larger values are more likely to be voiced. According to a preset threshold, voiced or unvoiced decisions can be made consequently. The threshold value can be fixed regardless of the source data to be the optimal value obtained by Brute force method.

Table 3.1: Comparison with the conventional method and our method for Mandarin VC.

	RMSE	CorrCoef [%]	v → u error [%]	u → v error [%]
Conventional	41.6	75.5	2.13	2.93
Proposed	37.2	77.9	1.81	2.11
Improvement rate	10.4	3.15	14.9	27.9

Table 3.2: Comparison with the conventional method and our method for English VC.

	RMSE	CorrCoef [%]	v → u error [%]	u → v error [%]
Conventional	14.2	75.8	1.10	3.35
Proposed	22.6	77.9	1.25	1.89
Improvement rate	18.9	2.79	-13.1	43.4

3.4.2 Results and discussion

The results of objective comparison with the conventional F0 conversion based on Equation (3.17) and our proposed method are shown in Table 3.1 and 3.2. Root Mean Square Error (RMSE), average correlation, v → u error and u → v error between predicted F0 sequences and target F0 sequences were used for the evaluation. The objective results show that our framework performs better than the conventional method in terms of all four evaluations for Mandarin Chinese. For English, not all four evaluations are improved. Our method significantly improves performance evaluation matrices: RMSE, correlation and u → v error rate, while v → u error is slightly degraded. Mandarin is known as a syllabically paced tonal language. Compared with English, Mandarin has a more restricted pitch contour pattern due to its lexical meaning. The variation of F0 contour among different speakers is less than that in English. We think this is the possible reason that the objective measure improvement of our method is larger in English than in Mandarin.

Although this work is not explicitly applied to an H2S system in this thesis, this mdel would be useful in future when our system reaches at practical level, because generating natural F0 is an important issue to synthesize natural speech.

3.5 Speech synthesis based on space mapping

Speech synthesis based on space mapping can be considered as a kind of voice conversion, in which an acoustic feature spaces of source speakers is replaced with another media. In our a hand gesture to speech (H2S) conversion system, the input media is replaced with hand gestures. In this section, features for hand gestures are described firstly. Then the challenge of our system is made clear.

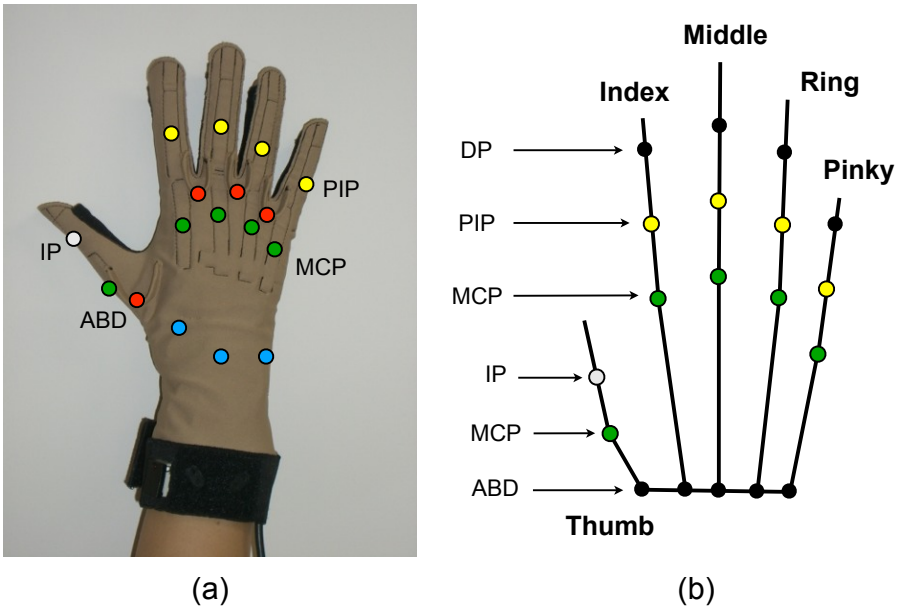


Figure 3.4: The location of 18 sensors on CyberGlove.

3.5.1 Features for hand gestures

In our study, hand postures are measured by 18 sensors embedded in a glove (CyberGlove; Visual Technologies, Palo Alto, CA). The location of those 18 sensors are illustrated in Figure 3.4 (a). We measured the angles at the metacarpalphalangcal (MCP) and proximal interphalangeal (PIP) joints of the four fingers and the angle of abduction (ABD) between adjacent fingers. For the thumb, the MCP, ABD and interphalangeal (IP) angles are measured. MCP, PIP and ABD are depicted with green, yellow and red points in the figure. Location of those joints are also shown in Figure 3.4 (b). In addition, CyberGlove is equipped with a palm arch sensor, a wrist flexion sensor and a abduction sensor (blue points in the figure (a)). The sampling period of the CyberGlove is 10-20 ms.[2]

3.5.2 The challenge of Hand-to-Speech conversion system

In our framework, mapping between gesture space and acoustic space is learned based on Equation (3.14), using parallel data sets which consist of gesture vector sequences and cepstrum vector sequences. When the input is the media which have explicit relationship with the movements of articulatory organs, the parallel data between two data sequences can be obtained, for example, by DTW. Conversely, hand gestures can be corresponded to

[2]Since the sampling period is variable, recorded data was interpolated linearly in such a way that the sampling period would be constant.

any sounds. How to design the optimal correspondence between vowels and hand gestures is, therefore, one of the most important issues in our research. In the next chapter, we will make an H2S system for the five Japanese vowels as a preliminary experiment. Then we will discuss how to evaluate the correspondence between the gestures space and the speech space.

3.6 Summary

In this chapter, the basic framework of current voice conversion technique has been reviewed. In statistical voice conversion, the correspondence of two feature spaces of the source speaker and the target speaker has been modeled with GMMs. This technique is applied to the conversion between different media. Our H2S system can be considered a kind of voice conversion system in which the speech space of a source speaker is replaced by the hand gesture space. It was also mentioned that the correspondence between two spaces are not explicit in our research and that will be one of the most important issues. In the next chapter, we will make an H2S system for the five Japanese vowels as a preliminary experiment. Then we will discuss how to evaluate the correspondence between the gestures space and the speech space.

Chapter4

Hand-to-Speech system for the five Japanese vowels

/a/ /i/ /u/ /e/ /o/

Figure 4.1: Gestures of the five Japanese vowels.

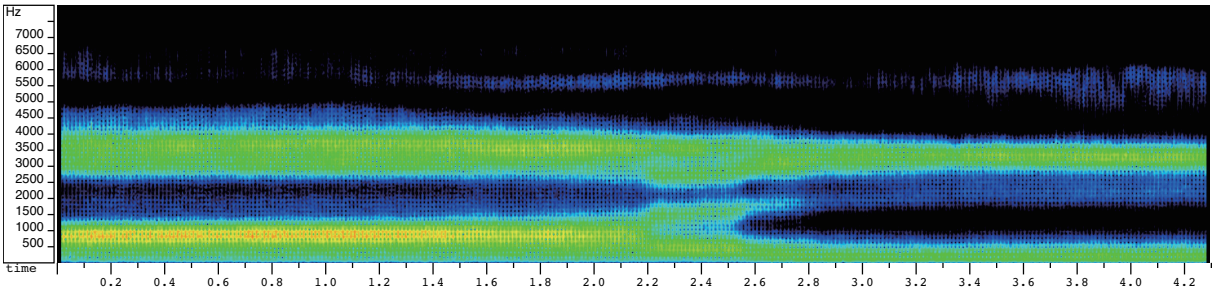
4.1 Introduction

In the previous chapter, the statistical voice conversion technique was described. Our H2S system can be considered a kind of voice conversion system in which the speech space of a source speaker is replaced by a hand gesture space. The most important issue is how to design the correspondence between the gesture space and the acoustic space. In this chapter, we will make an H2S system for the five Japanese vowels as a preliminary experiment. Then how to evaluate the correspondence between the gestures space and the speech space will be discussed.

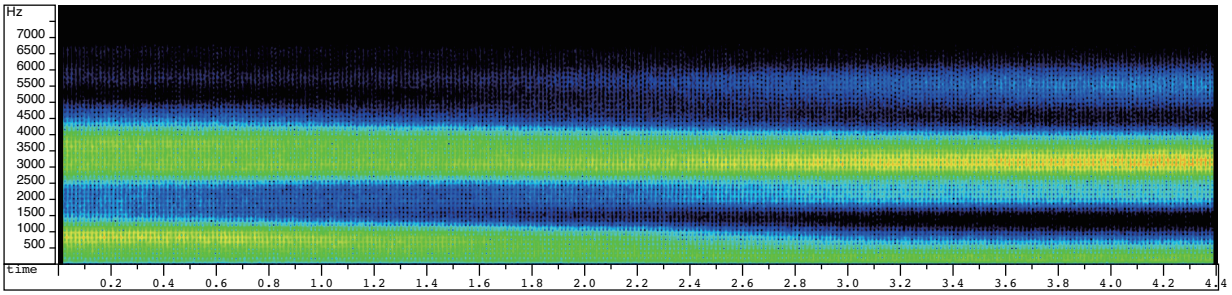
4.2 A preliminary experiment

As a preliminary experiment, an H2S system was implemented for vowel transitions such as /ai/ and /oe/. The correspondence between hand gestures and the five Japanese vowels is shown in Figure 4.1. These gestures are designed so that a transition between any pair of vowels will not generate a third vowel. Hereafter, the correspondence between gestures and speech will be called "gesture design".

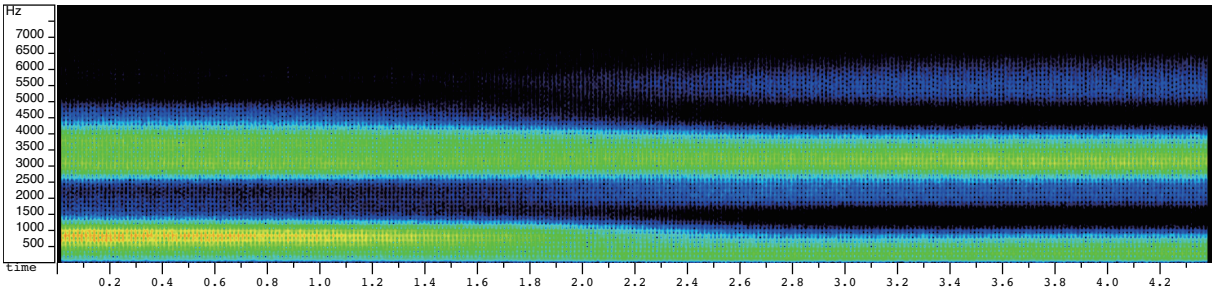
In order to train GMMs, a female adult recorded gesture data for the isolated vowels and ${}_5P_2$=20 transitions for each permutation of two vowels. Every gesture was recorded three times. The total number of gestures was (5+20)×3=75. In addition, a male adult speaker recorded speech for the five vowels and ${}_5P_2$=20 transitions between every two vowels. Speaking rate was adjusted to the transition rate of hand gestures. Each recording was done five times. The total number of speech samples was (5+20)×5=125. 18 dimensional cepstral coefficients were extracted using STRAIGHT [64], where the frame length was 40 ms and the frame shift was 1 ms. After every possible combination between a gesture sequence and its corresponding cepstral sequence were linearly aligned, the distribution of the augmented vector $\boldsymbol{z}$ was estimated based on a GMM for them, where the number of Gaussians was set to be one. Finally, the regression function $\mathcal{F}(\boldsymbol{x})$ was estimated based on Equation (3.14).



(a) resynthesized speech.



(b) hand to speech conversion with closed data as input



(c) hand to speech conversion with open data as input

Figure 4.2: Synthesized speech for vowel transition of /ai/.

Figure 4.2 shows the results for /ai/. (a) indicates a resynthesized speech sample for vowel transition /ai/, (b) is a sample synthesized by using closed hand gesture data as input, and (c) shows a synthesized sample by using open hand gesture data. We used STRAIGHT for waveform generation, where F0 was fixed to be 140 Hz. Through a simple listening test of all the kinds of vowel transitions, we found that the sounds of /i/, /u/, and /o/ were often confused. In the following section, we design the correspondence between the five vowels and hand gestures so vowel sounds will have a distinct difference between them.

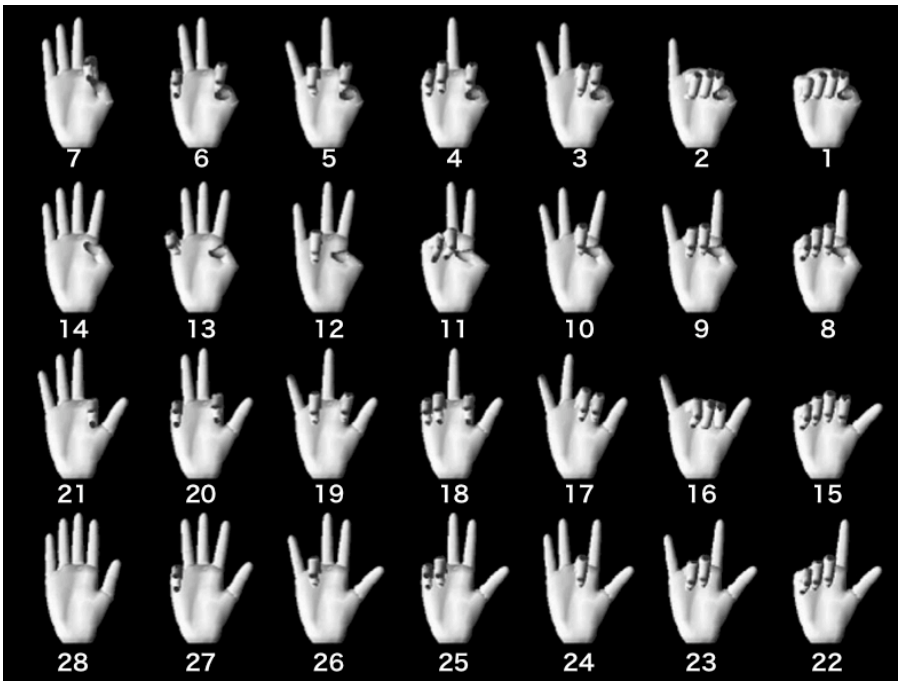


Figure 4.3: The 28 basic hand gestures [65].

4.3 The optimal gesture design

4.3.1 Variation of human hand gestures

What kind of hand gestures are possible and what kind of combination of five gestures is optimal for Japanese vowel production? In the preliminary experiment, sounds /i/, /u/, and /o/ were often confused. This leads us to believe that the gestures for these sounds are close to each other in the hand gesture space.

In [65], 28 basic hand gestures were defined, which are shown in Figure 4.3. These 28 gestures were generated as follows. As a hand has five fingers, each of which has two positions, high and low, we have 2^5=32 combinations for the five fingers. Among which, some are physically impossible to form, for example not everyone is able to bend the pinky without bending the ring finger. By removing those impossible-to-form gestures, we obtained 28 gestures.

A female adult recorded gesture data for these 28 gestures twice, 2×28=56 data in total. Using these data, Principal Component Analysis (hereafter PCA) was conducted to project 18 dimensional gesture data onto a two dimensional plane. The five gestures of the preliminary experiment, each of which had plural samples, were plotted on a plane (Figure 4.4). The five ovals represent regions for the five gestures and a sample trajectory of /aiueo/ is also plotted. As mentioned above, it is clear that the hand gestures of /i/, /u/, and /o/ are very close to each other. To generate distinct sounds for the individual vowels, we have to design an appropriate correspondence between vowels and gestures.

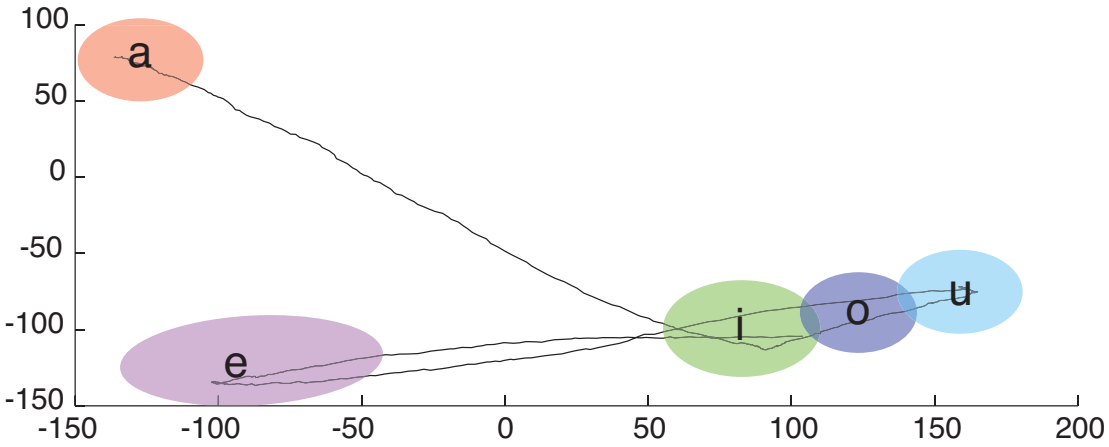


Figure 4.4: The five vowels in the preliminary experiment.

4.3.2 The location of the 28 gestures in a gesture space

We performed the same PCA analysis for the 28 gestures shown in Figure 4.3 with the results shown in Figure 4.5. The numbers in Figure 4.5 correspond to those in Figure 4.3. Postural synagies implied by the first principal component (PC1) and the second principal component (PC2) are depicted in Figure 4.6. This figure shows three-dimensional hand postures along the PC1 and the PC2 axes reconstructed from the data. For simplicity, those images were drawn only using MCP and PIP joints of four fingers and the MCP and IP angles of thumb, i.e., ABDs and angles captured with 3 sensors on the wrist were ignored. The hand posture in the center of the PC axes was rendered using the average of 28 gestures. The other four postures were computed by adding the minimum or maximum values of PC1 and PC2 to the average posture (for which the values of the PC coefficients are all set to be zero).

According to the Figure 4.6, PC1 expresses the closure of all five fingers and PC2 expresses extension of the index and middle fingers and closure of the ring and pinky fingers. Similar tendency of PC1 and PC2 is reported by Santello *et al.* [66] with another gesture data, captured with CyberGlove when subjects grasped 57 imaginary objects.

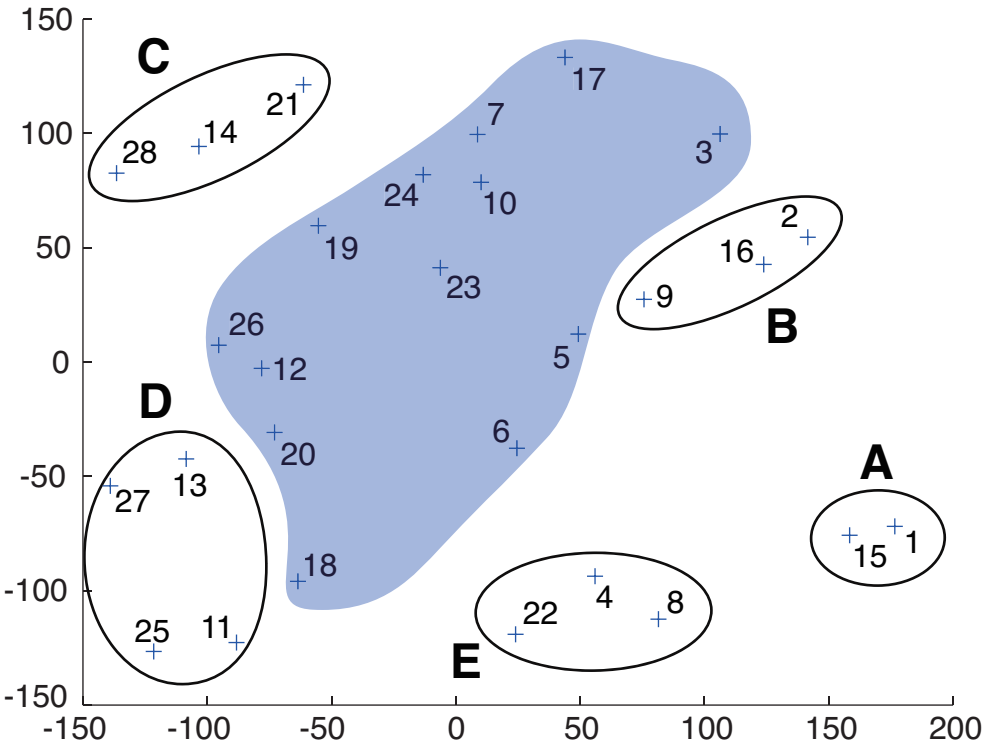


Figure 4.5: The location of the 28 gestures in the PCA space.

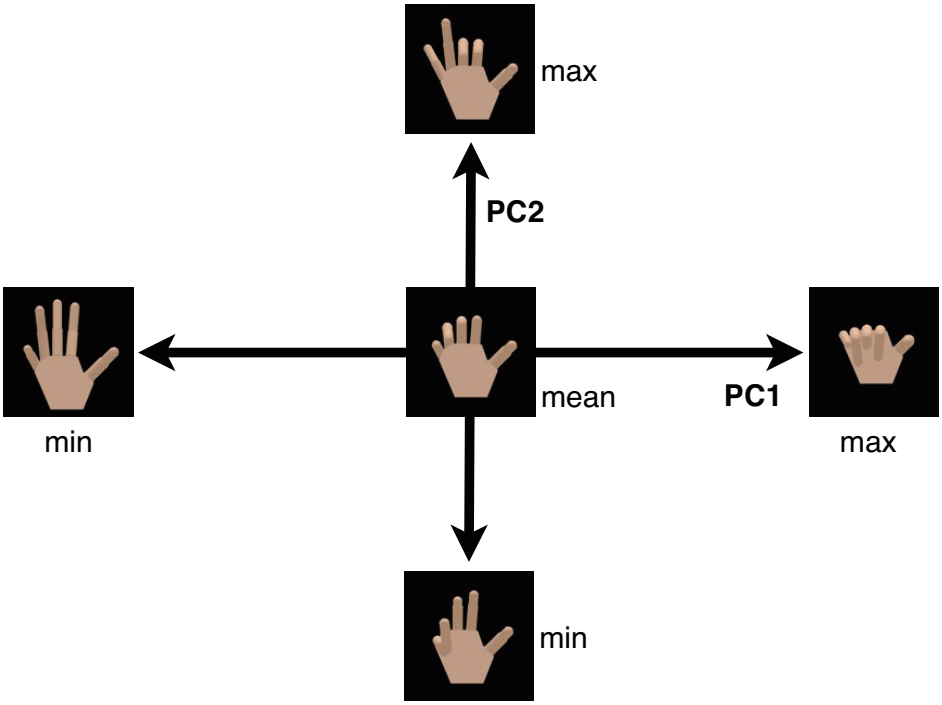


Figure 4.6: Postural synergies defined by the 1st and 2nd PCs.

Table 4.1: Proposed 16 combinations of hand gestures

No.	/a/	/i/	/u/	/e/	/o/	No.	/a/	/i/	/u/	/e/	/o/
1	8	14	2	11	1	9	22	14	2	11	1
2	8	14	2	13	1	10	22	14	2	13	1
3	8	14	16	11	1	11	22	14	16	11	1
4	8	14	16	13	1	12	22	14	16	13	1
5	8	28	2	11	1	13	22	28	2	11	1
6	8	28	2	13	1	14	22	28	2	13	1
7	8	28	16	11	1	15	22	28	16	11	1
8	8	28	16	13	1	16	22	28	16	13	1

4.3.3 Candidate sets of five hand gestures

In 4.5, the gestures in the central blue region require special efforts to form. Since those gestures are considered to be inappropriate for practical systems, they are removed from the candidate gestures for vowels and the remaining gestures are divided into five groups, A to E. By referring to the F_1-F_2 vowel chart of Japanese (See Figure 2.6), we designated those of the five vowels to five regions such that the topological features of the five gestures in gesture space and the five vowels would be equalized. For simplicity, we chose No.1 from group A and, from each of the other groups, we selected two easy-to-form gestures in that group. Thus, the number of gestures we chose was nine in total. Table 4.1 shows all of the $16(=2^4)$ combinations we selected and, out of these, we had to select the optimal one. To compare two topological patterns in different media, we used the structural representation of sequence data [67, 68, 69].

4.3.4 Structural representation and comparison

Since speaker differences can be characterized as a space mapping, mapping invariant features can be used as robust speech features for speech systems such as speech recognizers. [67, 70] showed that f-divergence between two distributions is invariant with any kind of invertible and differentiable transform. In [67, 70], using the Bhattacharyya distance (BD) as one of the f-divergence based distance measures, an utterance is structually represented as shown in Figure 4.7. BD between two Gaussian distributions $p_1(\boldsymbol{x}|\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ and $p_2(\boldsymbol{x}|\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$ is calculated as follows:

$$\begin{aligned} BD(p_1, p_2) &= -ln \int_{\infty}^{\infty} \sqrt{p_1(\boldsymbol{x}) p_2(\boldsymbol{x})} dx \\ &= \frac{1}{8}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^{\top} \left(\frac{\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2}{2} \right)^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) + \frac{1}{2} ln \frac{|(\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2)/2|}{|\boldsymbol{\Sigma}_1|^{1/2} |\boldsymbol{\Sigma}_2|^{1/2}} \end{aligned} \quad (4.1)$$

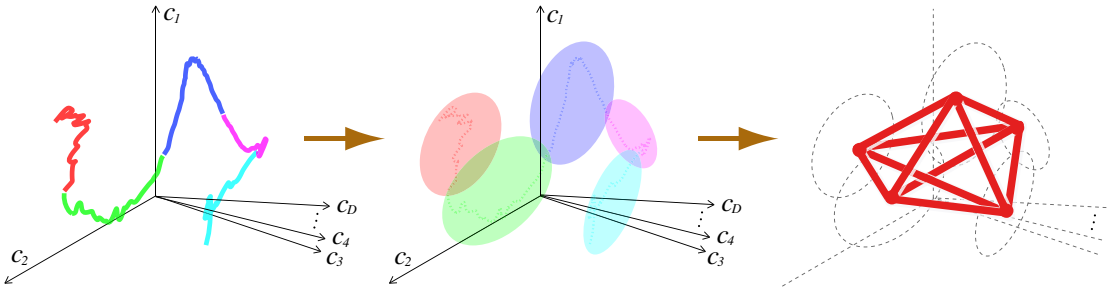


Figure 4.7: Structural representation of an utterance.

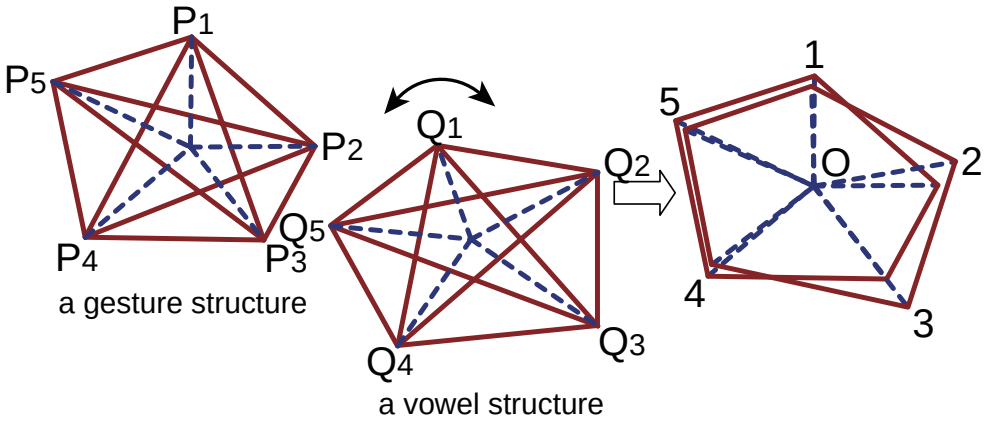


Figure 4.8: Structural matching between two matrices.

A cepstrum sequence is automatically segmented and converted into a distribution sequence. Subsequently, an utterance is characterized as a total set of BDs, namely, a distance matrix. Although this distance matrix is mapping invariant, by imposing some constraints, we introduced constrained invariance [71]. For example, if a distribution is assumed to be a Gaussian, the matrix is invariant only with linear transforms.

In this study, a hand gesture sequence is represented as a structure (distance matrix) and a vowel sequence is represented as another structure. Here, we assumed that the mapping function should be approximately linear. Then, we tentatively investigated whether the structural difference [67] of an utterance matrix and a gesture matrix calculated with each of the 16 candidates in Table 4.1 could work as an evaluation function. The smaller the difference is, the better the candidate will be. Here, the utterance /aiueo/ was used. Its distance matrix was compared to all 16 gesture matrices for the 16 candidates. Following [71], the number of distributions was set to 25.

The structural difference between two matrices is calculated as the Euclidean distance between two vectors, each of which is formed by using all the elements of the upper triangle of a distance matrix. This simple measure can accurately approximate the minimum total distance between the corresponding two points after shifting and rotating a structure (matrix) so that the two structures are overlapped optimally [67] (See Figure 4.8).

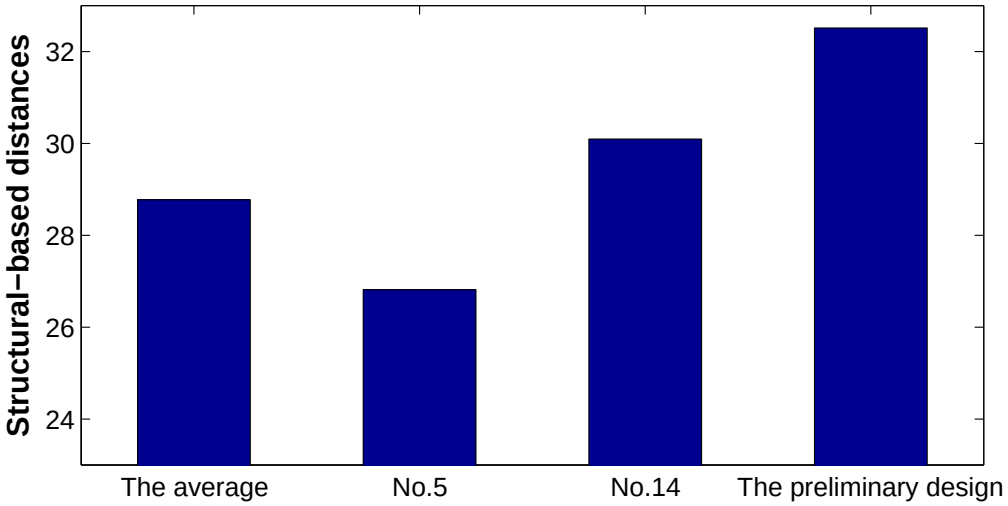


Figure 4.9: The structural distances for several sets of gestures.

4.3.5 Results and discussions

Figure 4.9 shows the structural distances between an /aiueo/ utterance and a few candidates. The average distance over the 16 candidates and the distance of the hand gestures used in the preliminary experiment are also shown. Among the 16 candidates, No.5 shows the smallest distance and No.14 the largest.

10 Japanese adults participated in a listening test with five nonsense words, all of which were composed of the Japanese vowels. The subjects were asked to transcribe the individual vowels. For each word, four versions, a re-synthesized sample, two synthesized samples with No.5 and No.14, and another synthesized one with the preliminary design were presented. The total number of nonsense word utterances was 20 and the total number of vowel sounds was 100. The order of presentation was randomized and the 20 words were presented through headphones. The vowel-based intelligibility was 100%, 99.6%, 99.2%, and 95.2% for the re-synthesized sample, No.5, No.14, and the preliminary design, respectively.

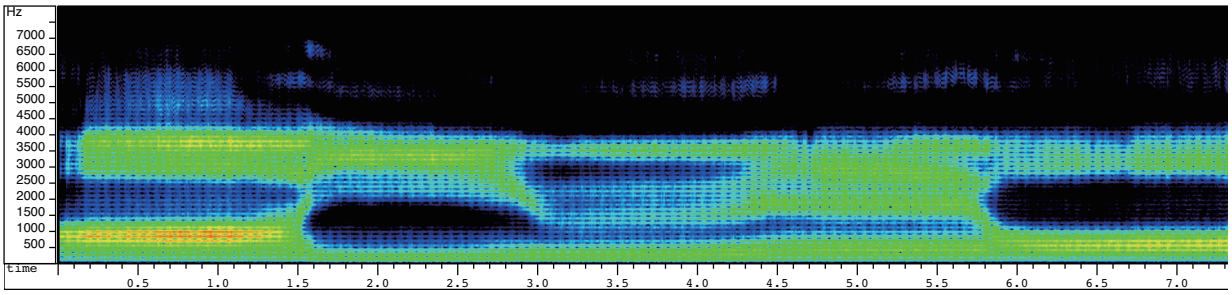
Figure 4.10 shows the spectrograms of (a) re-synthesized, (b) No.5, (c) No.14, and (d) the preliminary design. A small difference is visible between (b) and (c) but a large one between the two and (d).

The above results indicate that an adequate selection of hand gestures improves the intelligibility and the distinctness of synthesized vowel sounds. A visible difference was found between No.5 and No.14 in Figure 4.9 but the difference was not well perceived auditorily and visually. We are unable to claim that the structural difference is sufficient to select a gesture set out of candidates. A certain measure to estimate the goodness of gestures is however needed, because without that, a large number of listening tests are required to decide the optimal set of gestures.

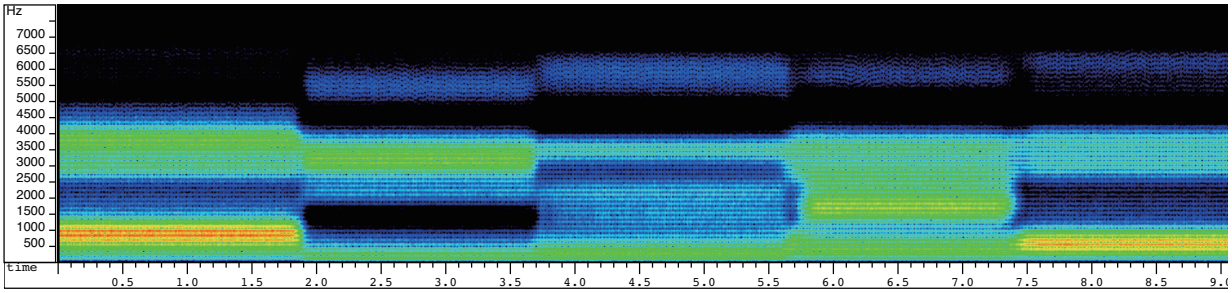
Finally, Figure 4.11 illustrates the spectrograms which were generated by using (a) distinct (articulate) hand gestures and (b) ambiguous (inarticulate) hand gestures. These speech samples were synthesized based on No.5. By comparing (a) with (b) visually and auditorily, we can claim that our hand-to-speech generator can control the degree of articulation well.

4.4 summary

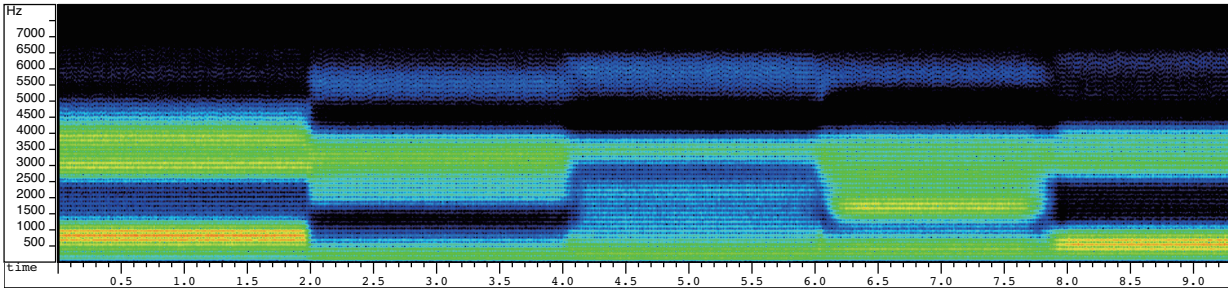
We implemented a speech synthesizer from hand gestures based on space mapping. By considering the topological equivalence between the structure of hand gestures in a gesture space and that of vowel sounds in the vowel space, we demonstrate how a quasi-optimal correspondence can be obtained. In the next chapter, we will discuss how to generate consonant sounds.



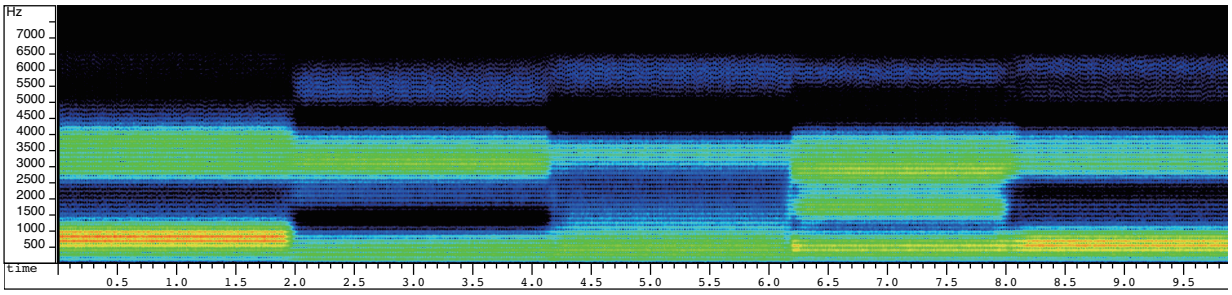
(a) Resynthesized speech.



(b) Synthesized speech by No. 5.

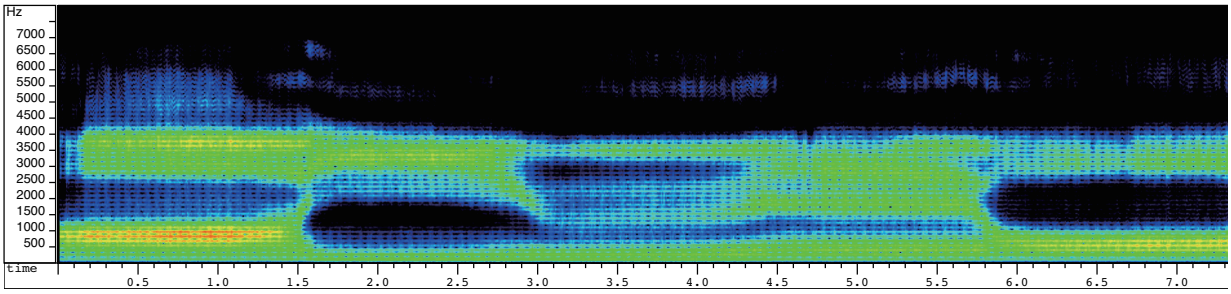


(c) Synthesized speech by No. 14.

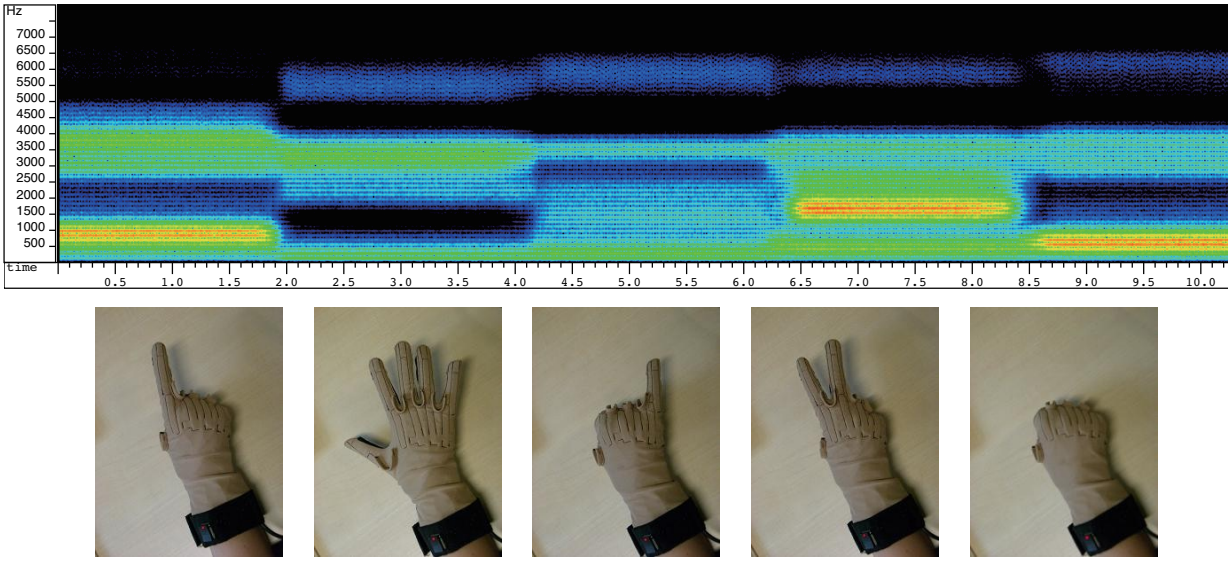


(d) Synthesized speech by preliminary design.

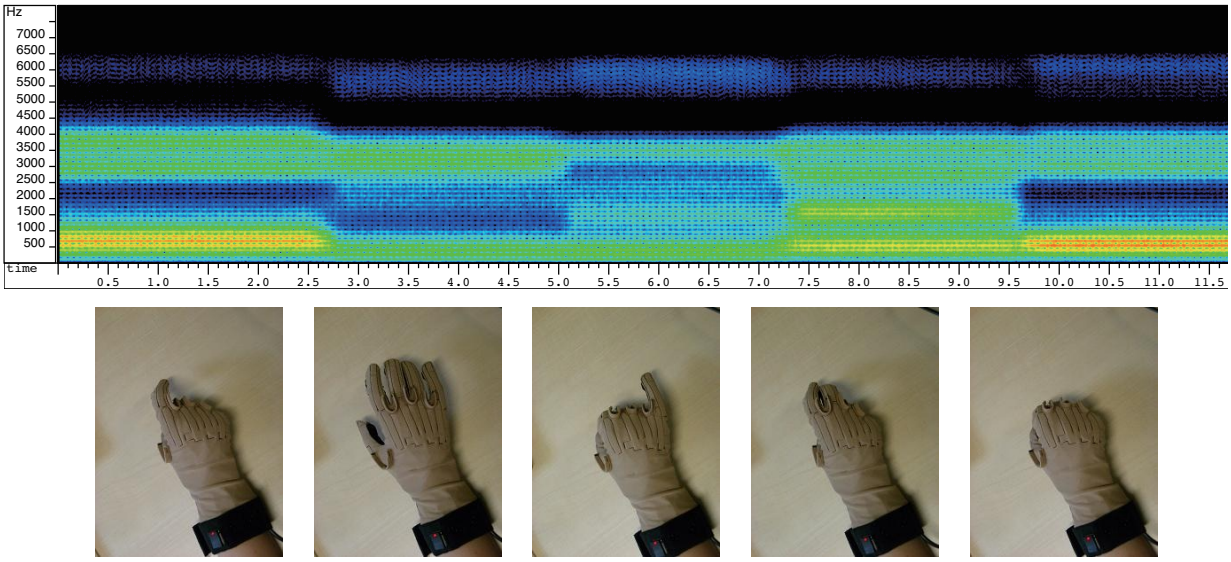
Figure 4.10: Comparison between proposed designs for /aiueo/.



(a) Resynthesized speech



(b) Synthesized speech using articulate hand gestures



(c) Synthesized speech using inarticulate hand gestures

Figure 4.11: Synthesized speech for /aiueo/

Chapter5

Consonant Generation

5.1 Introduction

We have seen that the proposed method is effective for generating the five Japanese vowels. In this chapter, we will discuss how to generate consonants in our system.

5.2 Classification of the Japanese consonants

Table 5.1 shows the classification of the Japanese consonants based on the place and the manner of articulation. In our study, these are divided into three groups: (1) semivowels, (2) fricatives, affricates and plosives, and (3) nasals and 'tap and flips'. For each group, we consider the synthesis methods below.

Semivowels are realized with vowel transitions. For example, /wa/ is expressed with the transition from /u/ to /a/. Yabu *et al.* reported that people percieve several Japanese words which include semivowels, such as /ohajo/ and /konbanwa/ ("good morning" and "good evening" in Japanese, respectively), only with the formant transitions of vowels [8]. Considering their results, we also synthesize semivowels with continuously changing vowel speech generated by continuously changing gestures for vowels.

Fricatives, affricatives and plosives are not affected by the speaking rate or succeeding vowels [73]. That is to say, unlike vowels and semivowels, users do not need to change the waveform or its length using body gestures or the following vowels. Thus, waveforms extracted from the recorded speech in this group can be preset in the system. On the other hand, VOT (Voice Onset Time) largely affects the perception of those consonants. It is well known that /t/ and /p/ are possible to be perceived as /d/ and /b/ respectively, depending on VOT [74]. Yabu *et al.* reported that the signal /sa/ can be perceived as /tsa/ and /ta/. Based to these facts, in our system, preset waveforms for fricatives, affricatives and plosives should be generated from preset waveform depending on VOT, which is controlled with body gestures. Then, the following vowel sounds, which are generated with hand gestures, will be concatenated to the consonant.

Based on the correspondences between the Japanese phones and Japanese phonemes, nasals and 'tap and flips' in Table 5.1 are described with phonemes as follows: [m] is /m/, [ɲ], [ŋ], [n] are /n/ or /N/, and tap and flip [r] is /r/. Hereafter, we simply call those 4 phonemes, /m/, /n/, /N/ and /r/, 'nasals' for convenience. Nasals are, like vowels, described with resonance and anti-resonance characteristics. We therefore extend the vowel generation framework to nasals. In other words, gestures are allocated for nasals as well as vowels, and nasal speech is generated with gesture motion based on space mapping.

Among the above three groups, semivowels are generated using the same framework as for vowels. The time structure of fricatives, affricatives and plosives is different from that for vowels and a new discussion will be needed. In this thesis therefore, we do not address this topic. In the following sections in this chapter and the following chapters, we will discuss nasal synthesis and how to extend the vowel generation framework to handle this.

Table 5.1: Japanese consonants classification [72].

	Bilabital		Alveolar		Alveolo -palatal		Palatal	Velar	Glottal	Synthesis methods in our system
Fricatives	ɸ		s	z	ɕ	ʑ	ç		h	waveform concatenation
Affricates			ts	dz	tɕ	dʑ				
Plosives	p	b	t	d			k	g		
Nasals	m		n		ɲ		ŋ			Expansion of S2H framework used for vowel sound generation
Tap and Flips			ɾ							
Semivowels							j	w		Vowel's transition

5.3 A preliminary experiment

In order to generate nasal sounds, gestures are allocated to nasals as well as to vowels, and nasal speech is generated with gesture motion based on space mapping. As a preliminary experiment, we focused on /n/ as a nasal sound and developed a H2S system for the five Japanese vowels and /n/. The challenge was how to derive a gesture for /n/. For initial trial, we designed the gestures for the five Japanese vowels and /n/ as follows.

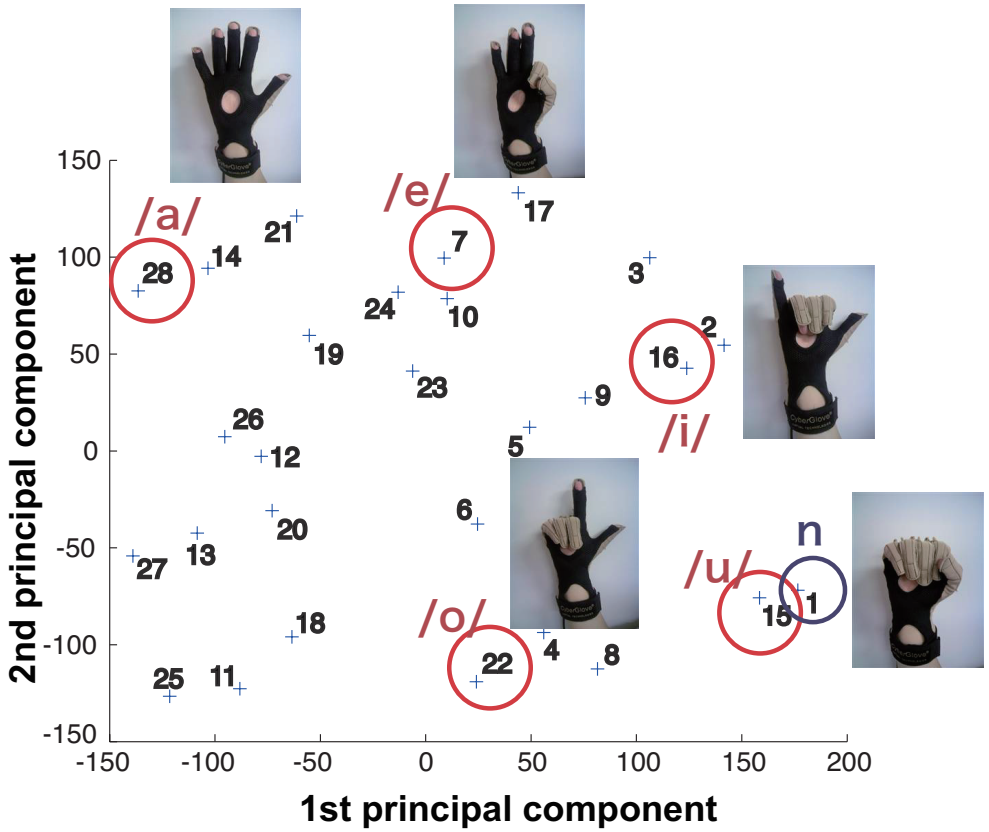
Generally, /n/ keeps 0.07–0.10 seconds [73]. Gestures for /n/ should therefore, be shifted to another gesture in 0.10 seconds order. Since the main function of the hand is to grasp objects, the movement to grasp is expected to be one of the easiest to form and quickest to shift to another gestures. Thus, as a preliminary gesture for the transition from /n/ to vowels, the movement to grasp objects was chosen and gesture No.1 in Figure 4.3 was set to be the gesture for /n/.

Next, gestures for vowels were chosen among 15 candidates gestures in Figure 4.5, considering topological equivalence between gesture space and speech space, as we considered for vowels in the previous chapter. In the previous chapter, a PCA plane and the F_1–F_2 plane were used to match the structures of vowels and of gestures. Consonants are however, not described only with F_1 and F_2. Therefore in the experiment, the F_1–F_2 plane cannot directly be used but Euclidean distances between /n/ and vowels in a gesture space and acoustic space were considered.

A male adult speaker recorded speech for /na/, /ni/, /nu/, /ne/, /no/, and the five vowels and ${}_5P_2$=20 transitions between every two vowels. The number of speech data was $5 + 5 = 10$, in total. Speech data for /na/, /ni/, /nu/, /ne/, /no/ are manually divided into three parts: a consonant part, a transition part and a vowel part. These five samples for consonant parts were considered /n/. The Euclidean distances between /n/ and vowels in the cepstral space were: $d_s(n, a) = 0.075, d_s(n, i) = 0.092, d_s(n, u) = 0.057, d_s(n, e) = 0.078, d_s(n, o) = 0.099$. Here $d_s(x, y)$ denotes the average Euclidean distance between cepstral vectors of phoneme /x/ and /y/. As $d_s(n, u)$ is the minimum among those five, the gesture No.15 that Euclidean distance in gesture spaces $d_h(n, u)$ is the minimum among $d_h(n, a), d_h(n, i), d_h(n, u), d_h(n, e), d_h(n, o)$, was considered the gesture for /u/. Then gestures for the other 4 vowels, /a/, /i/, /e/, /o/, were chosen so that the topological features of the five gestures in gesture space and the five vowels would be equalized, as we described in the previous chapter. Gesture designs for the five Japanese vowels and /n/ chosen in the above manner are shown in Figure 5.1.

5.3.1 The H2S system based on the proposed design

Based on the gesture design proposed in the previous section, an H2S system for the five Japanese vowels and /n/ was developed. A female adult recorded gesture data for /na/, /ni/, /nu/, /ne/, /no/, the isolated vowels and ${}_5P_2$=20 transitions between every two

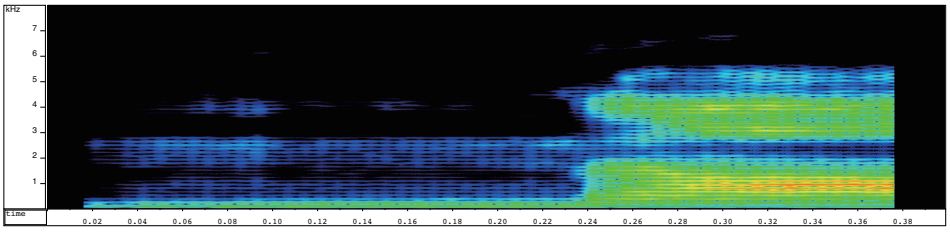


The location of the five Japanese vowels and /n/ in the gesture space.

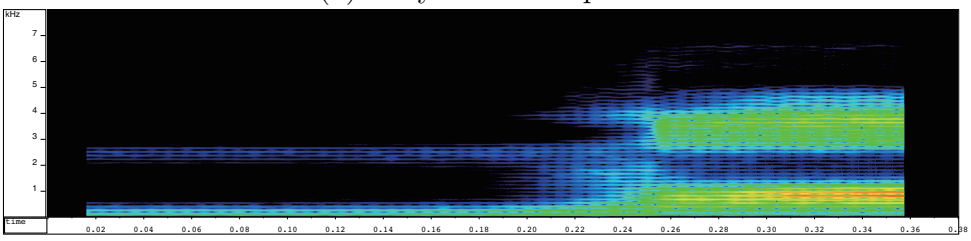
n /a/ /i/ /u/ /e/ /o/

Figure 5.1: Gesture design for the five Japanese vowels and /n/.

vowels using CyberGlove. Every gesture was recorded three times. The total number of gestures was (5+5+20)×3=90. In addition, a male adult speaker recorded speech for /na/, /ni/, /nu/, /ne/, /no/ for 10 times, and the five vowels and ${}_5P_2$=20 transitions between every two vowels five times. The number of speech data was $5 \times 10 + (5 + 25) \times 5 = 175$, in total. In order to make appropriate correspondence, gesture data and speech data for /na/, /ni/, /nu/, /ne/, /no/ were manually divided into three parts: a consonant part, a transition part and a vowel part. Then re-sampling and the 18 dimensional cepstrum coefficient extraction were performed in the same way as performed in Section 4.2. For /na/, /ni/, /nu/, /ne/, /no/, all combinations between 3 sets of gesture data and 10 sets of cepstral data, $3 \times 10 = 30$ combinations in total, for vowels, all combinations between



(a) Resynthesized speech.



(b) Synthesized speech with the H2S system.

Figure 5.2: Synthesized speech for /na/.

3 sets of gesture data and 5 sets of cepstral data, $3 \times 5 = 15$ combinations in total, were used to make augmented vectors. For simplicity, one vector was chosen from each set of 8 augmented vectors. Using these data, the GMM was trained. The number of mixtures were set to be 2. The spectrogram of the generated sound is illustrated in Figure 5.2.

5.3.2 A subsective evaluation

In order to evaluate the H2S system, an intelligibility test was carried out by 14 native Japanese speakers. Samples were composed of re-synthesized speech and synthesized speech with the H2S system for /na/, /ni/, /nu/, /ne/, /no/, 10 samples in total. As dummy samples, re-synthesized speech for /a/, /i/, /u/, /e/, /o/ and /ma/, /mi/, /mu/, /me/, /mo/, i.e., $2 \times 5 = 10$ samples were also included. They were randomized and subjects were asked to write them down with Roman letters. Subjects were noted that every sample is 1 mora speech of Japanese. Results are shown in Table 5.2.

Synthesized /ni/, /nu/ and /ne/ by the proposed H2S system were not perceived at all while more than a half of synthesized /na/ were perceived correctly. The most present error was to replace /n/ with /m/ or /w/ and they occupied almost half of all errors. This error was not found for re-synthesized speech. If this error is allowed, the intelligibility was largely improved as shown in Table 5.3.

It is said that formant transition from a consonant to a vowel has some acoustic characteristics which influences perception of the consonant [74]. Perception error between /n/ and /m/, /n/ and /w/ are considered to be a result of a lack of proper transition parts in synthesized speech. We have considered the reason to be: (1) the positional relation between vowels and consonants in the gesture space and that in the speech space were not

Table 5.2: Intelligibility

Synthesis method	/na/	/ni/	/nu/	/ne/	/no/	Average
Re-synthesis	100	93.8	93.8	81.3	100.0	93.8
Proposed	56.3	0.0	0.0	0.0	18.8	15.0

Table 5.3: Intelligibility when replacement with /m/ and /w/ are allowed

Synthesis method	/na/	/ni/	/nu/	/ne/	/no/	Average
Re-synthesis	100	100	93.8	100	100	98.8
Proposed	87.5	56.3	12.5	18.8	75.0	70.6

equivalent, (2) parallel data for transition parts from consonants to vowels did not correspond well. In order to get around those problems, we have developed a Speech-to-Hand conversion system (S2H system, the inverse system of H2S system) trained from parallel data for vowels only to infer the gestures corresponding to consonants. In the next section, we propose the method to which compensates these problems.

5.4 Probabilistic Integration Model

In the previous section, we have seen that inappropriate gesture designs for consonants result in a lack of smoothness in transitional segments of synthesized speech. In this section, we discuss how to determine appropriate gestures for nasal sounds when the gesture design for vowels is given.

In statistical machine translation studies, the maximization problem of $P(\boldsymbol{y}|\boldsymbol{x})$ for $\boldsymbol{y}$ is often solved using Bayes'rule [76]. Here $\boldsymbol{x}$ stands for Japanese and $\boldsymbol{y}$ for English, for example. In order to directly model and maximize $P(\boldsymbol{y}|\boldsymbol{x})$, a large amount of parallel data are needed. By considering $P(\boldsymbol{y}|\boldsymbol{x})$ as $P(\boldsymbol{x}|\boldsymbol{y})P(\boldsymbol{y})$ however, high performance machine translation is realized with a small amount of parallel data for $P(\boldsymbol{x}|\boldsymbol{y})$ and an accurate model $P(\boldsymbol{y})$ trained with a large English corpus. This framework has also been applied to voice conversion studies. Saito *et al.* proposed a new technique for voice conversion using a joint density model trained by a small amount of parallel data and a target speaker model trained by a large amount of speech from the target speaker, Model-Integrated Voice Conversion (hereafter MIVC) [77].

Our aim is to figure out the appropriate gestures for consonants. In order to achieve this aim, we propose using an S2H system $P(\boldsymbol{h}|\boldsymbol{s})$, the inverse system of the H2S system, and apply Bayes'rule to obtain the gestures. In the other words, we consider the gesture estimation problem given speech $\boldsymbol{s}$ as maximization problem of $P(\boldsymbol{h}|\boldsymbol{s})$ described as follows:

$$\hat{\boldsymbol{h}}_t = \underset{h}{\operatorname{argmax}}\, P(\boldsymbol{h}|\boldsymbol{s}) = \underset{h}{\operatorname{argmax}}\, \frac{P(\boldsymbol{s}|\boldsymbol{h})P(\boldsymbol{h})}{P(\boldsymbol{s})} = \underset{h}{\operatorname{argmax}}\, P(\boldsymbol{s}|\boldsymbol{h})P(\boldsymbol{h}), \tag{5.1}$$

here, $P(\boldsymbol{s}|\boldsymbol{h})$ is the joint density model for parallel data of vowels, $P(\boldsymbol{h})$ is the statistical gesture model trained using large amount of gesture data. A system only using $P(\boldsymbol{h}|\boldsymbol{s})$ may derive hard-to-form gestures. Ours with the well trained $P(\boldsymbol{h})$ is however anticipated to take account of the naturalness of gestures.

In order to solve this problem in MIVC, Saito *et al.* defined a likelihood function based on Equation (5.1). In our study, that function is written as follows [77]:

$$\mathcal{L}(\boldsymbol{h}_t; \boldsymbol{s}_t, \boldsymbol{\lambda}^{(z)}, \lambda^{(g)}) \triangleq P(\boldsymbol{s}_t|\boldsymbol{h}_t, \boldsymbol{\lambda}^{(z)}) P(\boldsymbol{h}_t|\boldsymbol{\lambda}^{(g)})^{\alpha}, \tag{5.2}$$

where $\boldsymbol{\lambda}^{(g)}$ is the model parameter of the gesture model, α is the weight of the gesture model. In MIVC, α means the weight of the speaker's model and corresponds to the weight of the language model in speech recognition.

The equation (5.2) can be written as:

$$\begin{aligned} log\mathcal{L}(\boldsymbol{h}_t) &= log \sum_{m=1}^{M} P(\boldsymbol{s}_t, m|\hat{\boldsymbol{h}}_t, \boldsymbol{\lambda}^{(z)}) + \alpha \times log \sum_{n=1}^{N} P(\hat{\boldsymbol{h}}_t, n|\boldsymbol{\lambda}^{(g)}) \\ &= log \sum_{m=1}^{M} P(m|\hat{\boldsymbol{h}}_t, \boldsymbol{\lambda}^{(z)}) \frac{P(\boldsymbol{s}_t, m|\hat{\boldsymbol{h}}_t, \boldsymbol{\lambda}^{(z)})}{P(m|\hat{\boldsymbol{h}}_t, \boldsymbol{\lambda}^{(z)})} + \sum_{n=1}^{N} logP(n|\hat{\boldsymbol{h}}_t, \boldsymbol{\lambda}^{(g)}) \frac{P(\hat{\boldsymbol{h}}_t, n|\boldsymbol{\lambda}^{(g)})}{P(n|\hat{\boldsymbol{h}}_t, \boldsymbol{\lambda}^{(g)})}, \end{aligned} \tag{5.3}$$

where M and N denote the number of mixtures of a GMM. According to Jensen's inequality,

$$\begin{aligned} & log \sum_{m=1}^{M} P(m|\hat{\boldsymbol{h}}_t, \boldsymbol{\lambda}^{(z)}) \frac{P(\boldsymbol{s}_t, m|\hat{\boldsymbol{h}}_t, \boldsymbol{\lambda}^{(z)})}{P(m|\hat{\boldsymbol{h}}_t, \boldsymbol{\lambda}^{(z)})} + \sum_{n=1}^{N} logP(n|\hat{\boldsymbol{h}}_t, \boldsymbol{\lambda}^{(g)}) \frac{P(\hat{\boldsymbol{h}}_t, n|\boldsymbol{\lambda}^{(g)})}{P(n|\hat{\boldsymbol{h}}_t, \boldsymbol{\lambda}^{(g)})} \\ \geq & \sum_{m=1}^{M} P(m|\hat{\boldsymbol{h}}_t, \boldsymbol{\lambda}^{(z)}) logP(\boldsymbol{s}_t, m|\hat{\boldsymbol{h}}_t, \boldsymbol{\lambda}^{(z)}) + \sum_{n=1}^{N} P(n|\hat{\boldsymbol{h}}_t, \boldsymbol{\lambda}^{(g)}) P(\hat{\boldsymbol{h}}_t, n|\boldsymbol{\lambda}^{(g)}) \\ = & \sum_{m=1}^{M} P(m|\hat{\boldsymbol{h}}_t, \boldsymbol{\lambda}^{(z)}) (logP(m|\hat{\boldsymbol{h}}_t, \boldsymbol{\lambda}^{(z)}) + logP(\boldsymbol{s}_t|m, \hat{\boldsymbol{h}}_t, \boldsymbol{\lambda}^{(z)})) \\ & + \sum_{n=1}^{N} P(n|\hat{\boldsymbol{h}}_t, \boldsymbol{\lambda}^{(g)}) P(\hat{\boldsymbol{h}}_t, n|\boldsymbol{\lambda}^{(g)}) \\ = & Q_{z1}(\boldsymbol{h}_t, \hat{\boldsymbol{h}}_t) + Q_{z2}(\boldsymbol{h}_t, \hat{\boldsymbol{h}}_t) + \alpha\, Q_g(\boldsymbol{h}_t, \hat{\boldsymbol{h}}_t), \end{aligned} \tag{5.4}$$

where,

$$Q_{z1}(\boldsymbol{h}_t, \hat{\boldsymbol{h}}_t) = \sum_{m=1}^{M} \lambda_{m,t}\, logP(m|\hat{\boldsymbol{h}}_t, \lambda^{(z)}), \tag{5.5}$$

$$Q_{z2}(\boldsymbol{h}_t, \hat{\boldsymbol{h}}_t) = \sum_{m=1}^{M} \lambda_{m,t}\, logP(\boldsymbol{s}_t|m, \hat{\boldsymbol{h}}_t, \lambda^{(z)}), \tag{5.6}$$

$$Q_g(\boldsymbol{h}_t, \hat{\boldsymbol{h}}_t) = \sum_{n=1}^{N} \lambda_{n,t}\, logP(\hat{\boldsymbol{h}}_t, n|\lambda^{(g)}), \tag{5.7}$$

$$\gamma_{m,t} = P(m|\boldsymbol{h}_t, \boldsymbol{\lambda}^{(z)}), \gamma_{n,t} \quad = \quad P(n|\boldsymbol{h}_t, \boldsymbol{\lambda}^{(g)}). \tag{5.8}$$

We assume that Equation (5.5) does not rapidly change, i.e., we ignore the differential coefficient of Q_{z1} against $\hat{\boldsymbol{h}}$. Then the optimum solution $\hat{\boldsymbol{h}}$ can be obtained by maximizing the following function:

$$Q'(\boldsymbol{h}_t, \hat{\boldsymbol{h}}_t) = Q_{z2}(\boldsymbol{h}_t, \hat{\boldsymbol{h}}_t) + \alpha\, Q_g(\boldsymbol{h}_t, \hat{\boldsymbol{h}}_t). \tag{5.9}$$

When $\hat{\boldsymbol{h}}_t$ is the optimum value, the differential coefficient of the Equation (5.9) against $\hat{\boldsymbol{h}}_t$ is 0. Thus we obtain the following update function:

$$\begin{aligned} \hat{\boldsymbol{h}} \quad = \quad & \Big(\sum_{m=1}^{M} \gamma_{m,t} \boldsymbol{D}_m'^{(h)-1} + \alpha \sum_{n=1}^{N} \gamma_{n,t} \Sigma_n^{-1}\Big) \\ & \times \Big(\sum_{m=1}^{M} \gamma_{m,t} \boldsymbol{D}_m'^{(h)-1} \boldsymbol{E}_{m,t}'^{(h)} + \alpha \sum_{n=1}^{N} \gamma_{n,t} \Sigma_n^{-1} \boldsymbol{\mu}_n\Big), \end{aligned} \tag{5.10}$$

where $\boldsymbol{\mu}_n$ and $\boldsymbol{\Sigma}_n$ are the n^{th} mean vector and the covariance matrix of the gesture model GMM. $\boldsymbol{E}_{m,t}'^{(h)}$ and $\boldsymbol{D}_m'^{(h)-1}$ are described as follows:

$$\boldsymbol{E}_{m,t}'^{(h)} \quad = \quad \boldsymbol{\mu}_m^{(h)} + \boldsymbol{\Sigma}_m^{(hh)} \boldsymbol{\Sigma}_m^{(sh)+} (\boldsymbol{s}_t - \boldsymbol{\mu}_m^{(s)}), \tag{5.11}$$

$$\boldsymbol{D}_m'^{(h)-1} \quad = \quad \left[\boldsymbol{\Sigma}_m^{(hh)} - \boldsymbol{\Sigma}_m^{(hs)} \boldsymbol{\Sigma}_m^{(ss)-1} \boldsymbol{\Sigma}_m^{(sh)}\right]^{-1} - \boldsymbol{\Sigma}_m^{(hh)-1}. \tag{5.12}$$

Here, $(\cdot)^+$ denotes the pseudo-inverse matrix. As for the initial value of Equation (5.8), first we apply Equation (3.14) to our H2S system and convert input speech $\boldsymbol{s}_t$ into hand gestures $\boldsymbol{h}_t = \mathcal{F}(\boldsymbol{s}_t)$, then it is used as the initial value of Equation (5.8).

In order to obtain the optimal gestures for consonants, we develop an S2H system based on the framework described above. Then the gestures should be obtainable by inputting the consonants into the S2H system. In the next section, the gestures derived with S2H system are evaluated to determine their effectiveness for an H2S system.

5.5 Experiments

In order to verify that the S2H system is able to derive the gestures for speech, that are not included in the parallel data, the experiments below were carried out. The procedure for the experiments in Figure 5.3. First, an S2H system was made with the gesture model and the conversion model trained with parallel data for vowels. Then consonants speech were input to the S2H system and gestures corresponds to those speech are obtained. In order to check if those gestures were effective for generating consonant speech, they were input to the H2S system which was made in the same manner as the S2H system. Finally the input speech of the S2H system and the derived speech of the H2S system were compared. They were expected to be the same, ideally.

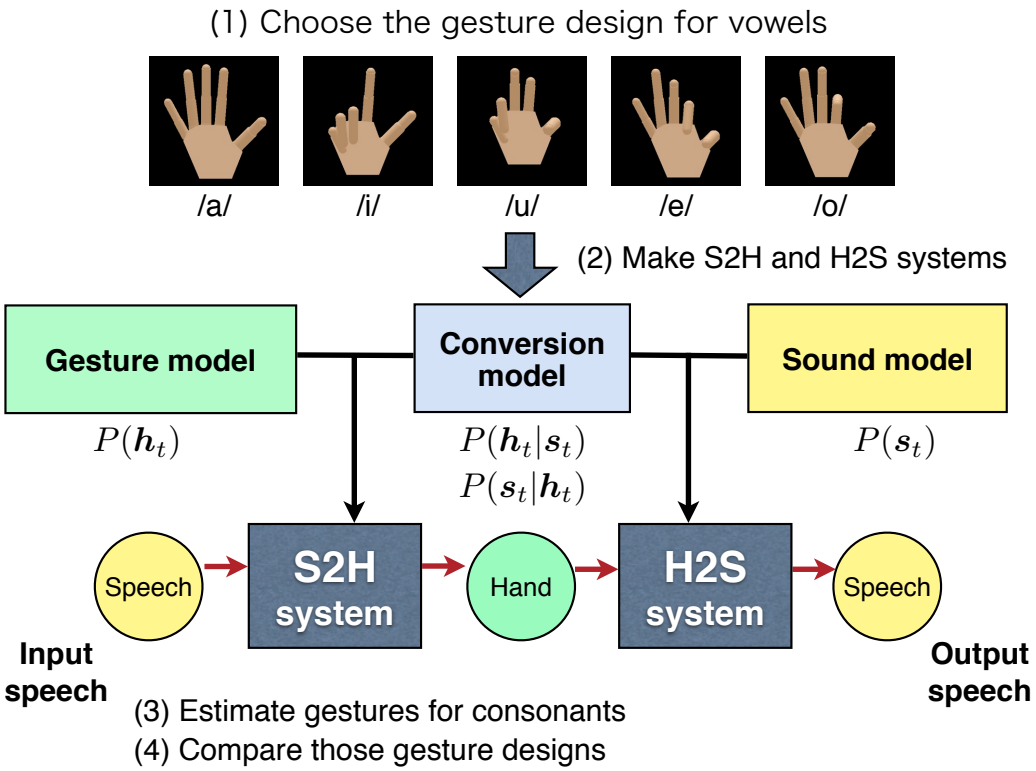


Figure 5.3: The procedure of the experiments.

5.5.1 Speech-to-Hand conversion sytem

First off, an S2H system was developed using parallel data based on the gesture design for vowels and the statistical gesture model. By inputting the features for consonant speech, the gestures for consonants were obtained.

For the gesture model, No.1, No.2, No.4, No.7, No.8, No.9, No.11, No.13, No.14, No.15, No.16, No.21, No.22, No.25, No.27, No.28 were chosen from 28 gestures depicted in Figure 4.3. These are gestures that are formed easily by the female adult who records gesture data for the experiments. Recorded gesture data from herfor the isolated gestures and the transitions of all pairs, $16 + {}_{16}P_2 = 256$ in total. The gesture model was developed using this gesture data. The number of mixtures was set to 64.

Next, the gesture designs for the five Japanese vowels were chosen from among those 16 gestures. For simplicity, we chose No.28 for /a/. Considering the discussion in Chapter 4, we did not use gesture designs satisfying the condition that the Euclidean distances in the gesture space $d_h(a,i) < d_h(a,e)$ and $d_h(a,u) < d_h(a,o)$. We obtained 8190 candidates for the gesture designs for the five Japanese vowels. Then the joint density models for the S2H system were obtained for every gesture design as follows.

The gestures for the isolated five Japanese vowels and ${}_5P_2 = 20$ transitions between every

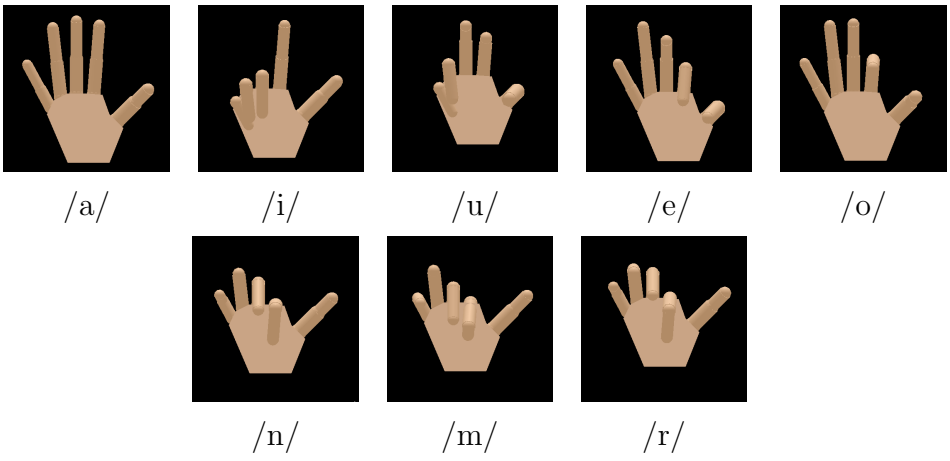


Figure 5.4: Derived gesture design for consonants (sample 1).

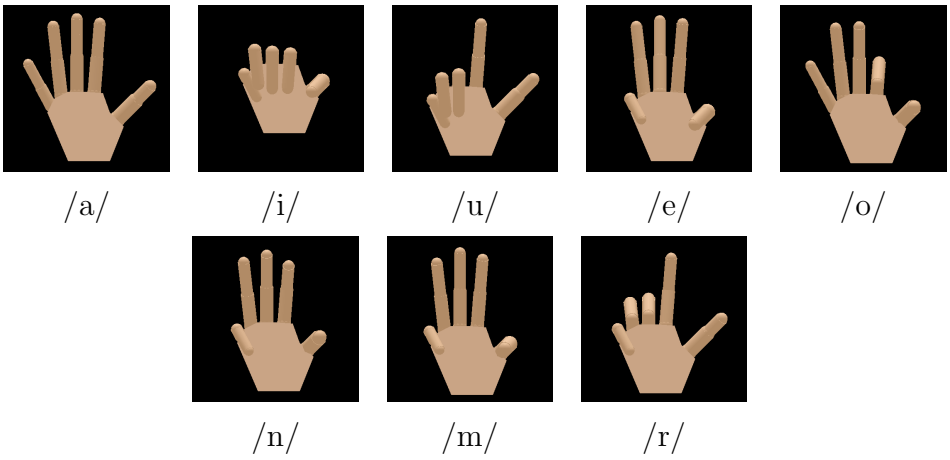


Figure 5.5: Derived gesture design for consonants (sample 2).

two vowels, $5 + 20 = 25$ gesture data in total, were extracted from the gesture data set above. In addition, a male adult speaker recorded speech for the five Japanese vowels and ${}_5P_2 = 20$ transitions between every two vowels. Each recording was done once. The total number of speech samples was $5 + 20 = 25$. Then, cepstrum extraction and interpolation were carried out in the same way as in section 4.2. Using these gesture sequences and cepstral vector sequences, augmented vectors were made and the joint density model was trained with them. The number of mixtures of the joint density model was set to 8.

By inputting the /n/, /m/, /r/ sounds, the gestures for those consonants were obtained. Thus, we got 8190 gesture designs for /a/, /i/, /u/, /e/, /o/, /n/, /m/, /r/. Some of the obtained gesture designs are shown in Figure 5.4, 5.5. In the next section, we chose the best one from these candidates.

5.5.2 Hand-to-Speech conversion sytem

In order to compare those 8190 gesture designs obtained in the previous section, the following experiments were carried out.

First, for every gesture design, the H2S system was developed using the method described in Section 5.4. The joint density models $P(\boldsymbol{h}|\boldsymbol{s})$ were trained with the same training data and the same number of mixtures as the S2H systems in the previous section. The speech model $P(\boldsymbol{s})$ was trained using A set of 50 sentences from the ATR phoneme-balanced sentences recorded by the speaker who recorded the speech data for the joint density model. The number of mixtures was 64, the same as that of the gesture model.

Then, gestures estimated by the S2H systems were input into those H2S systems. If the S2H and H2S systems are both ideal, input speech for the S2H systems and output speech of the H2S systems will be identical. The joint density models $P(\boldsymbol{h}|\boldsymbol{s})$ and $P(\boldsymbol{s}|\boldsymbol{h})$ however, do not completely describe the correspondence between the gesture space and the speech space. There will therefore be a distortion between the input speech for the S2H systems and the output speech of the H2S systems. We considered using this distortion as a criteria to design the gestures. That is to say, we thought that the smaller the cepstrum distortion between the input speech for the S2H systems and the output speech of H2S systems is, the better the gesture designs will be.

Figure 5.6 shows the average and standard deviation of 8190 cepstral root mean square errors (RMSE). For simplicity, we only focused on the /n/ consonant. Input speech for the S2H systems were the five Japanese vowels in the training data and /na/, /ni/, /nu/, /ne/ and /no/ recorded by one speaker, $5+5 = 10$ samples in total. Thus, 10 mora-unit cepstral RMSEs were calculated for every gesture design [1]. Then gesture designs were sorted for mora-unit cepstral RMSE. Every gesture design has therefore 5 ranks for each mora. The quasi-optimal design, the one where the summation of the ranks was minimal, was No. 28 for /a/, No. 22 for /i/, No. 11 for /u/, No. 7 for /e/ and No. 21 for /o/. Figure 5.6 also shows the cepstral RMSE of the quasi-optimal design.

Experimental evaluation showed that the cepstral RMSE depends on the mora. Of the five Japanese vowels, the smallest one was /u/ and the largest one was /i/. Overall, compared with the vowels included in the training data, consonants which are not included in the training data tended to show larger cepstral RMSE. On the other hand, the cepstral RMSEs of vowels were almost the same as those of consonants in the quasi-optimal design. Figure 5.7 shows re-synthesized speech and the output of the quasi-optimal S2H–H2S combined system [2].

[1] Mora is a unit of speech production and speech rhythm of Japanese, which is usually composed of CV or V.

[2] Smoothing was performed before visualization

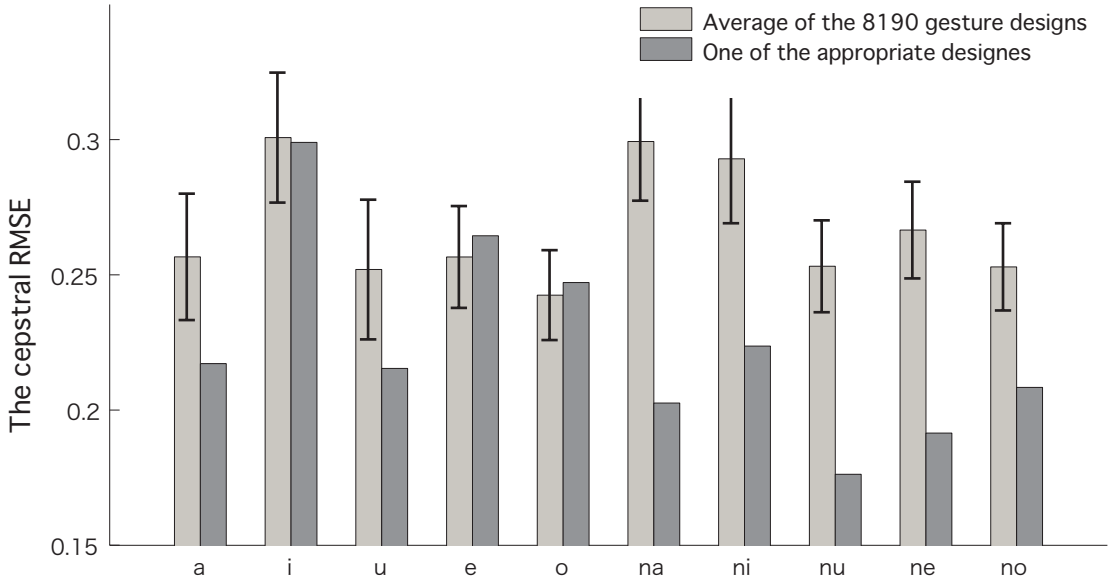


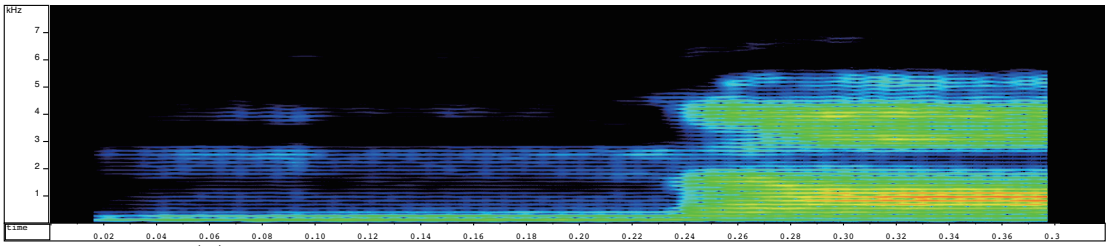
Figure 5.6: The cepstral RMSE between input and output speech.

5.5.3 Subjective evaluations

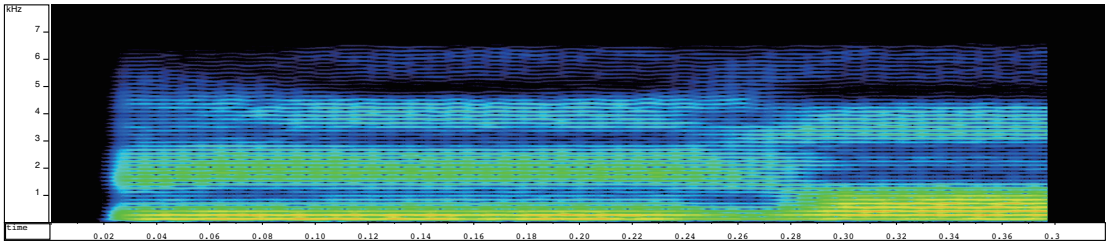
The effectiveness of our approach was evaluated subjectively through an AB preference test, in which 15 Japanese native speakers selected the preferred one from a pair of synthesized speech in terms of naturalness. A and B were output speech from the H2S systems which were trained by the design, which was designed based on the way described in Section 5.3.1 (hereafter conventional design) and the proposed design, respectively. For both designs, the gesture design for vowels was the quasi-optimal one described in the previous section and the same training data for vowels as for those for the S2H system was used. In the conventional design, the gesture for /n/ was chosen as No.8. Then gesture data for /na/, /ni/, /nu/, /ne/, /no/ were recorded once and added to the training data. In the proposed design, 10 sets of the gesture data for /na/, /ni/, /nu/, /ne/, /no/, which were obtained by a combined S2H–H2S system, were added to the training data. The total number of frames for /n/ of both designs were approximately the same. The mixture number of GMM was set to 64 for both designs.

15 Japanese native speakers participated in this test. A preference test was conducted separately for each case of /a/, /i/, /u/, /e/, /o/, /na/, /ni/, /nu/, /ne/ and /no/. Preference of the proposed design was 48%, no preference was 27% and the preference of the conventional design was 24%.

In Section 5.3, we found that inappropriate gesture designs for a consonant result in the lack of transitions in synthesized speech. For example, the one-mora synthetic speech of /na/ was perceived as two sounds /n/+/a/. On the other hand, synthesized speech



(a) re-synthesized speech which input to S2H system.



(b) the output of the S2H–H2S combined system developed by the quasi-optimal design

Figure 5.7: Synthesized speech for /na/

based on our proposed method was perceived as one unit in the simple listening test. This result shows that the proposed method is effective for deriving appropriate gestures for consonants.

5.6 Summary

In this chapter, we proposed a framework to derive the gestures for consonants when only the correspondence for vowels is given. According to the listeners evaluations, an H2S system, which exploits gesture data for consonants derived from an S2H system, can generate more natural sounds than those trained with heuristic gesture designs for consonants. In the next chapter, a real-time H2S system is developed based on the framework described in this chapter and is evaluated by subjective users.

Chapter6

Real-time H2S system

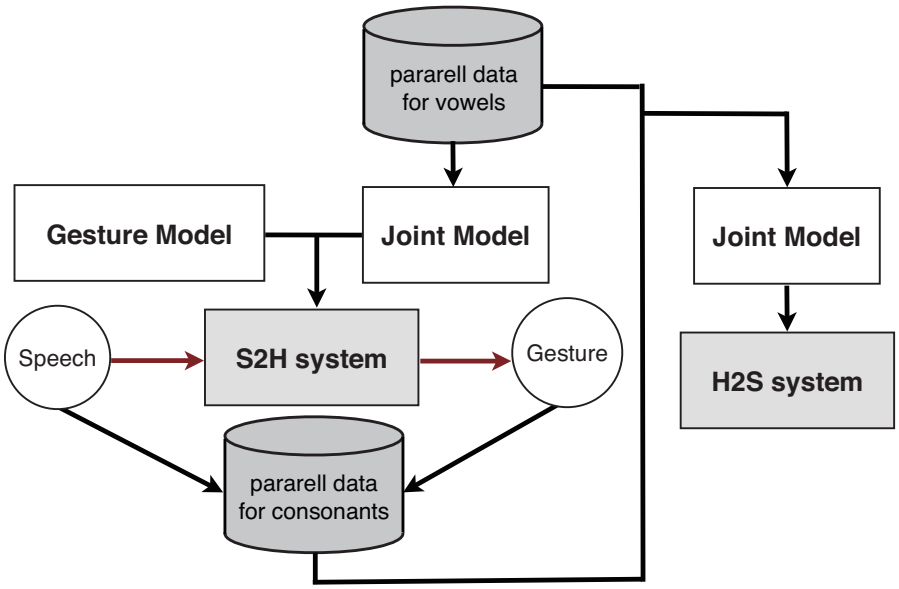


Figure 6.1: The procedure to establish real-time H2S system.

6.1 Introduction

In the previous section, we verified that an H2S system which exploit gesture data for consonants derived from an S2H system can generate more natural sounds than those trained with gesture designs which are chosen from given candidates. Natural speech generated by an H2S system trained by exploiting data generated by an S2H system is, however, obtained only when input gestures are the same as the ones which were generated by the S2H system. The S2H system sometimes outputs gestures with dynamic ranges that are too large or that are not smooth enough. In those cases, it is difficult for users to form such gestures in real time. In this section, we compensate for those problems in two ways: (1) reducing the dynamic range by setting the optimal weight for the gesture model (2) smoothing the gesture trajectories by considering delta features. We also develop a real-time H2S system exploiting the gestures generated by the improved S2H system.

6.2 How to develop a real-time H2S system

The challenge in this chapter is to develop a real-time H2S system based on the gesture design. The proposed procedure is shown in Figure 6.1. First, a conversion model $P(\boldsymbol{s}|\boldsymbol{h})$ is trained with parallel data for vowels. Then an S2H system is developed based on the method described in the previous chapter, with the joint model and the gesture model $P(\boldsymbol{h})$. Inputting speech for consonants to the S2H system, the gesture vector output corresponds

to the speech that is obtained. Since this conversion is performed frame-by-frame, every input speech frame corresponds to the derived gesture frame. We therefore easily obtain parallel data for consonants by simply joining input speech vectors and the converted output gesture vectors frame by frame. This parallel data for consonants is added to those for vowels and the conversion model $P(\boldsymbol{h}|\boldsymbol{s})$ is trained. With this conversion model, real-time H2S system is developed based on the method described in Chapter 3.

As a preliminary experiment, we developed real-time H2S system following the flow above.

6.2.1 Dataset for the gesture model

For the dataset to train the gesture model, we used the same ones which are used in Section 5.5.1. In this dataset, sensor No.17 and 18 were not used explicitly. Removing these sensor data, data from 16 sensors were used to train gesture models.

The movement of the different joints of the hand are not independent. For example, when the 2nd joint of the pinky finger is bent, the 2nd joint of the ring finger is also bent. The dimensions of the 16 dimensional gesture data are therefore interrelated. In order to ensure gesture space corresponds to the cepstral space, for which each dimension is independent, PCA was performed on all the gesture data using all the data in the gesture dataset. 16 dimensional data points after PCA were used to train the 64 mixture gesture model $P(\boldsymbol{h})$.

6.2.2 Dataset for the conversion model

The gesture designs for the five Japanese vowels were chosen from among the 16 gestures described in Section 5.5.1 and an S2H system was developed with them. Experiments in Chapter 4 showed that a quasi-optimal correspondence can be obtained by considering the topological equivalence between the topological features (structure) of hand gestures in the gesture space and those of vowel sounds in the vowel space. Considering the F_1–F_2 plain therefore, we chose the gesture designs which satisfy the condition that the Euclidean distances in the gesture space $d_h(a,i) < d_h(a,e)$ and $d_h(a,u) < d_h(a,o)$. In this preliminary experiment, No. 28, No.2, No. 1, No.9 and No.25 were used for gestures for /a/, /i/, /u/, /e/ and /o/.

Then, an S2H system was developed in the same manner as the previous chapter. Input speech for the S2H systems were the five Japanese vowels in the training data and /na/, /ni/, /nu/, /ne/ and /no/ recorded by one speaker. $5+5=10$ samples in total. Inputting these consonants to the S2H system, gestures for consonants were derived.

10 sets of the gesture data for /na/, /ni/, /nu/, /ne/, /no/, which were obtained by the S2H system, were added to the training data. The number of mixtures for the GMM was set to 64.

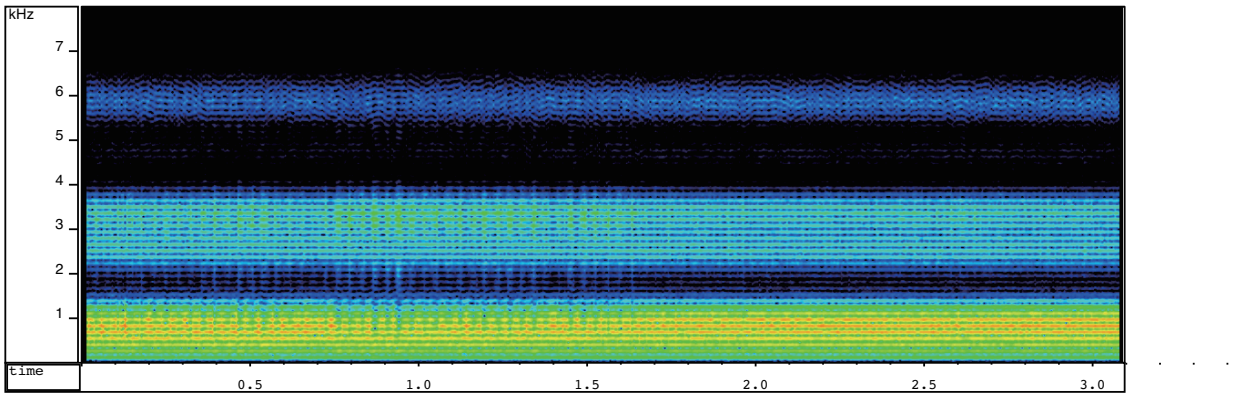


Figure 6.2: Consonant part of /nu/ sound generated by the realtime H2S system.

6.3 The problem of gestures derived from an S2H system

Inputting gesture movement to the real-time H2S system developed above, speech sounds were generated in real-time. The consonant part of the /nu/ sound generated by the realtime H2S system are shown in Figure 6.2. Jagged parts can be seen in the spectrum. This problem was not found in the spectrums of speech sounds generated by the H2S system in the previous chapter. Those two H2S systems were developed with the same framework. The only difference was their inputs. For the H2S system in Section 5.5.2 generated speech from the gestures derived from the S2H system, while the H2S system here needs real gesture inputs. In other words, we input into the H2S system in section 5.5.2 the optimal gestures to get the closest speech to the original speech, while we input into the H2S system here the gestures which we can form realistically. Figure 6.3 shows No.6 sensor output of the DataGlove for the gesture for /na/ generated by S2H and of real gesture data. Thanks to the gesture model $P(\boldsymbol{h})$, S2H systems seldom recognize impossible-to-form gestures at all the frames. Gesture transitions derived from the S2H systems are, however sometimes difficult to form in a realistic time frame, because of their exceedingly large dynamic range and exceedingly rapid changes. In the real-time H2S system, the conversion model was trained with the parallel data of impossible-to-form gestures and natural speech, therefore the system outputs natural speech only when unrealistic gesture sequences were input while jagged spectrums were derived when real gesture sequence are input.

In order to mitigate the above problem, the following ideas were considered: (1) reducing the dynamic range by setting the optimal weight for the gesture model (2) smoothing the gesture trajectories by considering delta features. These methods were verified in the following sections.

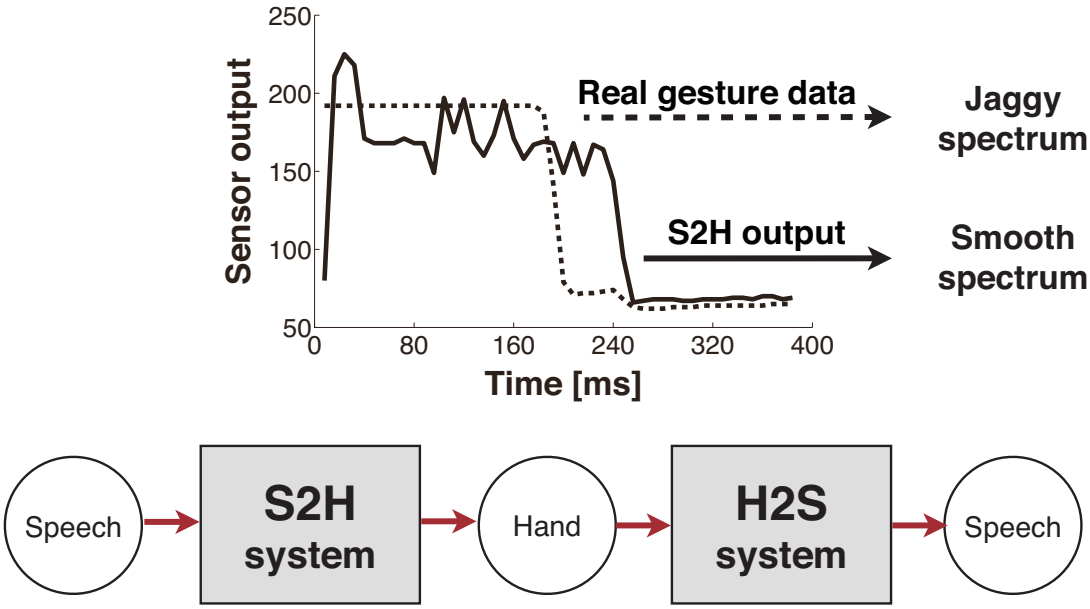


Figure 6.3: Comparison between S2H output and real gesture data.

6.3.1 The weight factor of the gesture model

The framework described in Chapter 3 estimates gestures only using $P(\boldsymbol{s}|\boldsymbol{h})$. An S2H system based on this framework therefore, would derive impossible-to-form gestures when consonants sounds, which are not included in the training data, are input.

On the other hand, in the framework based on MIVC, which is described in chapter 5, gestures are derived with the weighted gesture model $P(\boldsymbol{h})$ as well as the conversion model $P(\boldsymbol{s}|\boldsymbol{h})$. On account of the gesture model, derived gestures are expected to be more natural (see Equation (5.2)).

When α is small, the naturalness of the gestures is not considered. It may result in the derivation of difficult-to-form gestures. Meanwhile when α gets larger, derived gesture trajectories become flat since the gesture model $P(\boldsymbol{h})$ is modeled independent of each input speech sequence.

In the previous section, we assumed $\alpha = 1$ as well as in the voice conversion task performed in [77]. The appropriate α is, however, expected to be different in a Gesture→Speech conversion task than in a Speech→Speech conversion task. We expected that the appropriate α would reduce the dynamic range of derived gesture trajectories and consequently those gesture trajectories would be easy-to-form.

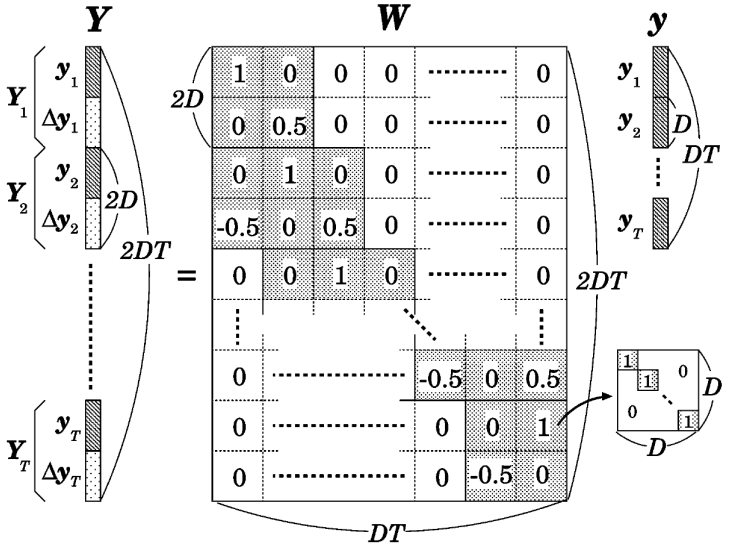


Figure 6.4: Relationship between a sequence of the static and dynamic feature vectors [15].

6.3.2 Conversion considering dynamic features

In the frame-by-frame mapping where the correlation of the feature vectors is ignored, the discontinuity of the parameter trajectory becomes a problem [15]. In MIVC, this is more serious than conventional voice conversion, which is described in Chapter 3, because even slight skips affect the perceptual quality of the whole sentence badly since each frame in the sequence is converted more precisely. Saito *et al.* reported that this problem is mitigated by considering dynamic features in the framework of MIVC [78]. Their method was applied to our S2H system.

Here, denote input speech sequences and the target gesture vector sequences as $\boldsymbol{S} = [\boldsymbol{S}_1^\top, \boldsymbol{S}_2^\top, ..., \boldsymbol{S}_T^\top]^\top$ and $\boldsymbol{H} = [\boldsymbol{H}_1^\top, \boldsymbol{H}_2^\top, ..., \boldsymbol{H}_T^\top]^\top$, respectively. $\boldsymbol{S}$ and $\boldsymbol{H}$ consist of static and dynamic features and described as $\boldsymbol{S}_t = [\boldsymbol{s}_t^\top, \Delta\boldsymbol{s}_t^\top]^\top, \boldsymbol{H}_t = [\boldsymbol{h}_t^\top, \Delta\boldsymbol{h}_t^\top]^\top$. $\boldsymbol{\lambda}^{(Z)}$ and $\boldsymbol{\lambda}^{(G)}$ are trained by these features as well as the conventional MIVC, described in chapter 5. A time sequence of the converted gesture vectors is derived as follows:

$$\hat{\boldsymbol{h}} = \underset{\boldsymbol{h}}{\operatorname{argmax}}\, P(\boldsymbol{S}|\boldsymbol{H}, \boldsymbol{\lambda}^{(Z)}) P(\boldsymbol{H}|\boldsymbol{\lambda}^{(G)}), \tag{6.1}$$

where $\boldsymbol{H} = \boldsymbol{W}\boldsymbol{h}$. $\boldsymbol{W}$ denotes the matrix that extends the static feature sequence to the static and dynamic feature sequence (see Figure 6.4). In a similar manner to that in [15]

and [77], we derive the following updating equations:

$$\hat{\boldsymbol{h}} = \left(\boldsymbol{W}^{\top}\overline{\boldsymbol{D}^{(H)-1}}\boldsymbol{W}\right)\boldsymbol{W}^{\top}\overline{\boldsymbol{D}^{(H)-1}\boldsymbol{E}^{(H)}} \tag{6.2}$$

$$\overline{\boldsymbol{D}^{(H)-1}} = diag\left[\overline{\boldsymbol{D}_1^{(H)-1}}, ..., \overline{\boldsymbol{D}_T^{(H)-1}}\right] \tag{6.3}$$

$$\overline{\boldsymbol{D}^{(H)-1}\boldsymbol{E}^{(H)}} = \left[\overline{\boldsymbol{D}_1^{(H)-1}\boldsymbol{E}_1^{(H)}}^{\top}, ..., \overline{\boldsymbol{D}_T^{(H)-1}\boldsymbol{E}_T^{(H)}}^{\top}\right]^{\top} \tag{6.4}$$

$$\overline{\boldsymbol{D}_t^{(H)-1}} = \left(\sum_{m=1}^{M}\gamma_{m,t}\boldsymbol{D}_m^{\prime(H)-1} + \sum_{n=1}^{N}\gamma_{n,t}\boldsymbol{\Sigma}_n^{(G)-1}\right) \tag{6.5}$$

$$\overline{\boldsymbol{D}_t^{(H)-1}\boldsymbol{E}_t^{(H)}} = \left(\sum_{m=1}^{M}\gamma_{m,t}\boldsymbol{D}_m^{\prime(H)-1}\boldsymbol{E}_{m,t}^{\prime(H)} + \sum_{n=1}^{N}\gamma_{n,t}\Sigma_n^{(G)-1}\boldsymbol{\mu}_n^{(G)}\right) \tag{6.6}$$

$$\lambda_{m,t} = P(m|\boldsymbol{H}_t, \lambda^{(Z)}), \lambda_{n,t} = P(n|\boldsymbol{H}_t, \lambda^{(G)}) \tag{6.7}$$

The S2H system developed based on this framework (hereafter S2H-delta) is expected to derive smoother trajectories than the previous S2H system.

6.4 Experiments

6.4.1 Experimental setup

In order to verify the effects of α and the dynamic features, an S2H system and the S2H-delta system were developed according to the flow of Figure 6.1.

For the gesture model and the conversion model, the same data sets for section 6.2.1 and section 6.2.2 were used. Gesture models $P(\boldsymbol{h})$ and $P(\boldsymbol{H})$ were trained with a 64-mixture GMM and a 512-mixture GMM, respectively. Conversion models $P(\boldsymbol{s}|\boldsymbol{h})$ and $P(\boldsymbol{S}|\boldsymbol{H})$ were trained with an 8-mixture GMM and a 16-mixture GMM, respectively. The number of mixtures were chosen to be the optimum based on ML criterion. Due to the existence of silence, dynamic features of the first and the last frames would be outliers. Thus they were removed when $P(\boldsymbol{H})$ was trained.

6.4.2 Results

i) The effect of α

Using the conversion model $P(\boldsymbol{s}|\boldsymbol{h})$ and the gesture model $P(\boldsymbol{h})$, an S2H system was developed based on the framework described in Chapter 5. α takes a value between 0.0 and 1.0. Figure 6.5 shows the No.6 sensor output of the DataGlove for the /na/ gesture when α changes. Gestures for /n/, /m/, /r/ derived from S2H systems ($\alpha = 1.0$) were also shown in Figure 6.7.

Frames after 800ms correspond to the vowel parts of /na/. Because vowels were included in the training data for the conversion model $P(\boldsymbol{s}|\boldsymbol{h})$, they were estimated properly despite

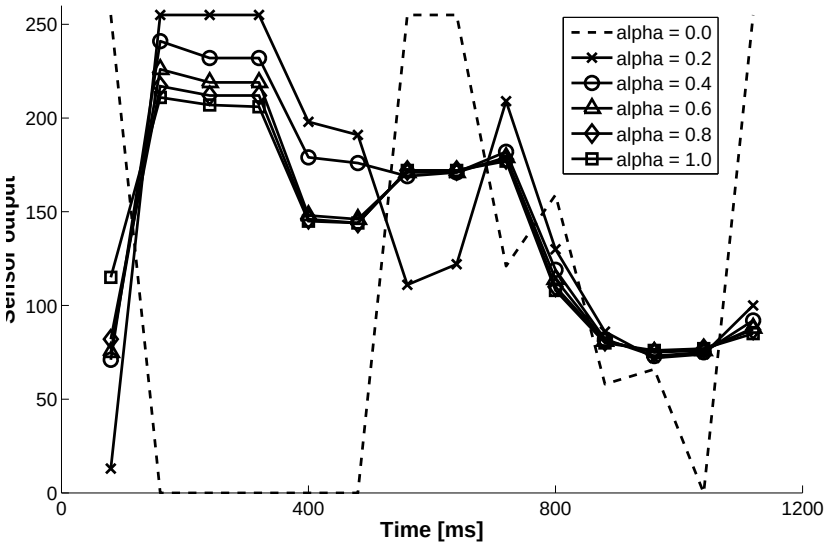


Figure 6.5: The effect of α.

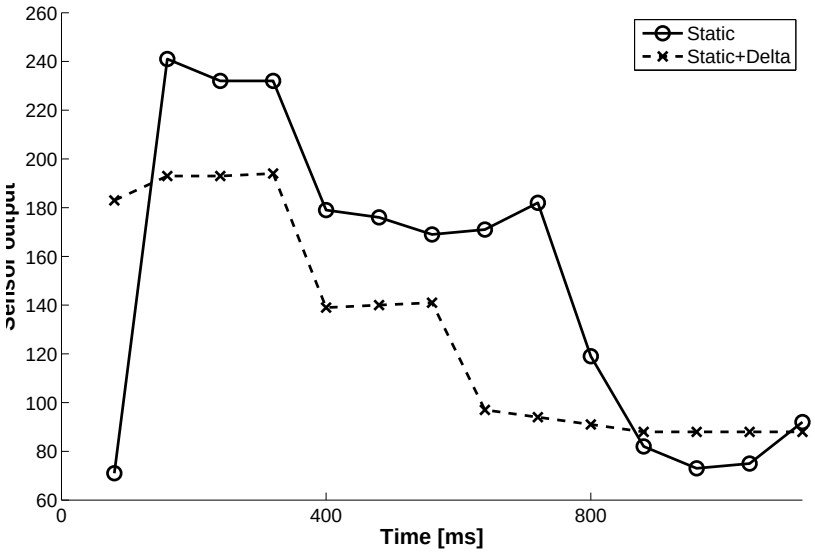


Figure 6.6: The effect of dynamic features.

the value of α. On the other hand, frames before 800ms, which correspond to consonants and transition parts, largely change depending on α.

When $\alpha = 0$, the S2H system does not consider the gesture model $P(\boldsymbol{h})$ at all, i.e., it uses inverse conversion based on the framework described in Chapter 3 (see Equation (5.2)). This system derives gestures that exceed the dynamic range of sensors (0–255). As α increases, the naturalness of gestures is considered thanks to $P(\boldsymbol{h})$ and the decrease of rapid changes. When α gets too large however, the influence of the conversion model

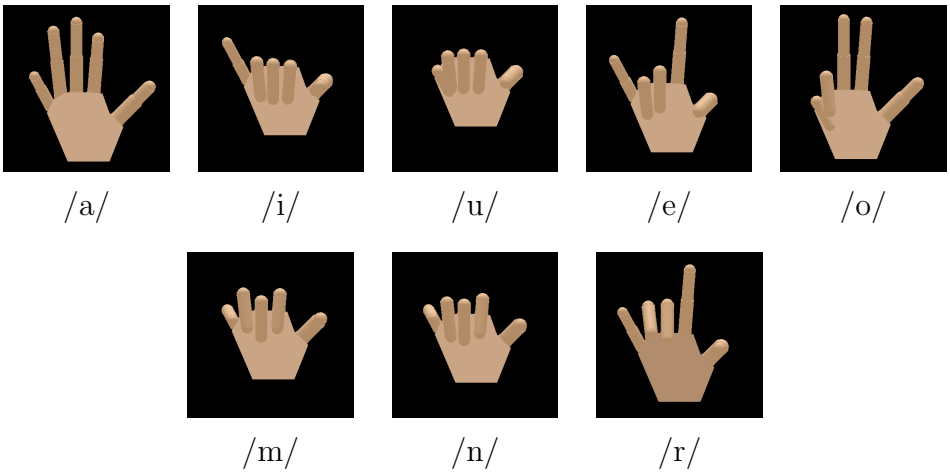


Figure 6.7: Gesture design for the real-time S2H system.

$P(\boldsymbol{s}|\boldsymbol{h})$ is ignored and the derived gesture trajectories become flat. Other sensors and other gestures showed similar tendencies. In our research, the smallest α 0.6, with which all frames of derived gestures were included in the dynamic range of sensors, was considered appropriate.

ii) The effect of dynamic features

Using the conversion model $P(\boldsymbol{s}|\boldsymbol{h})$ and the gesture model $P(\boldsymbol{s})$, an S2H system was generated based on the framework described in Chapter 5. In addition, an S2H-delta system was developed based on Section 6.3.2 with the conversion model $P(\boldsymbol{S}|\boldsymbol{H})$ and the gesture model $P(\boldsymbol{S})$. In both systems, α was set to 0.6. We input /n/ to these systems and compared the outputs. Figure 6.6 shows No.6 sensor output of DataGlove for the /na/ gesture.

Compared with the S2H system, the rapid changes of the beginning part and of the transition parts at about 800ms are mitigated in the S2H-delta system. According to the above results, S2H-delta with $\alpha = 0.6$ was used for S2H part in Figure 6.1.

Theoretically, a conversion method considering dynamic features would be effective in H2S systems as well as S2H systems. In a real-time H2S system however, using dynamic features is not a good idea. This is because a 20–40ms will be consumed to calculate $\Delta\boldsymbol{h}$, creating a time lag, due to the sampling rate of the DataGlove being about 10–20ms, and this is critical for the real-time system. Taking this into account, an H2S system was developed based on the framework described in Chaper 3, while an S2H-delta system was used to derive gestures for consonants.

6.5 Real-time H2S system

6.5.1 System condition

Following the flow of Figure 6.1, the prototype of the real-time H2S system is established.

At first, an S2H-delta system was developed based on Section 6.3.2 with the conversion model $P(\boldsymbol{S}|\boldsymbol{H})$ and the gesture model $P(\boldsymbol{S})$, trained in Section 6.2.2 and in Section 6.2.1. Input speech for the S2H systems were Japanese nasal speech, /na/, /ni/, /nu/, /ne/ and /no/, recorded by one speaker ten times. $5 \times 10 = 50$ samples in total. Then, cepstrum extraction and interpolation are carried out in the same way as in section 4.2. Inputting this consonant speech to the S2H-delta system, gestures for consonants were derived.

Using input speech and derived gesture data with an S2H-delta system, parallel data for a consonants were obtained. They were added to the parallel data for vowels and the conversion model $P(\boldsymbol{s}|\boldsymbol{h})$ is trained. The mixture number for GMMs was set to 8.

6.5.2 Pitch and volume control

Our system controls parameters corresponding to vocal tract shape with hand gestures. We introduced hand direction as parameters to control pitch and volume. A sensor module kit TDS01V was used to capture hand direction. 3 angles captured with TDS01V were shown in Figure 6.8. Since angles can be stably controlled compared with acceleration, Azimuth and pitch of the arm was used to control F0 and volume, respectively.

In order to mitigate the error, every time the user puts on the TDS01V, calibration was performed. When the user straightens his or her arm with the palm facing the bottom, Azimuth and Pitch were set to 0. As the arm moves in the horizontal direction, Azimuth changes from $-60°$ up to $+120°$. In our system, F0 was set to $1.2^{(\theta-30)/30} \times 115$, where θ denotes azimuth, so that F0 generated F0 changes roughly between 90–200 [Hz]. As the arm is moved in the vertical direction, pitch changes from $0°$ up to $+70°$. In our system, Volume was set to $1.22^{(\sigma/45} \times const.$. The working real-time H2S system is shown in Figure 6.9.

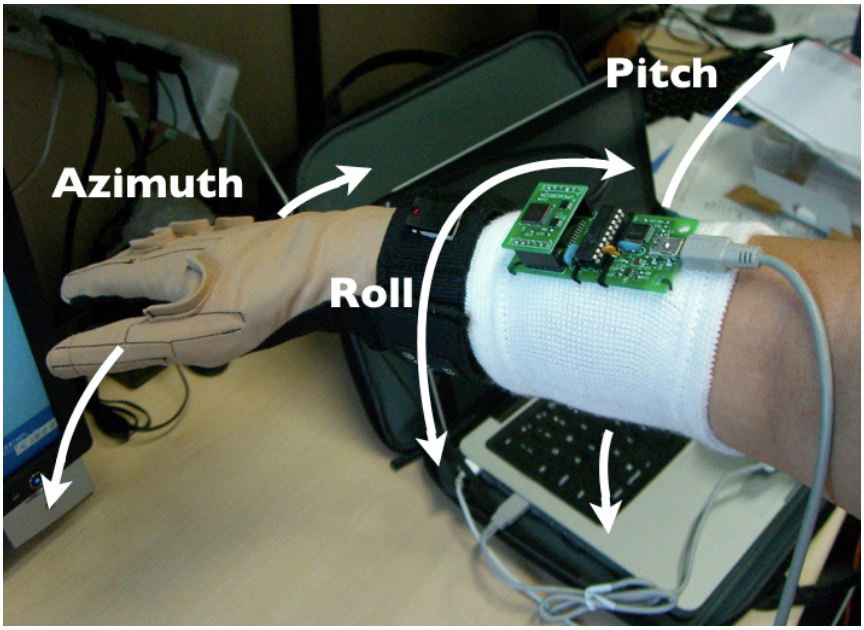


Figure 6.8: Azimuth, pitch and roll of TDS01V.

Figure 6.9: The working real-time H2S system.

6.6 Subjective user evaluations

Our goal is to establish a methodology of speech synthesis which does not require symbols inputs. In order to claim that the new framework is effective, it should have at least the same quality as the conventional ones. Current synthesis methods which do not require symbol inputs are, as we saw in Chapter 2, effective on emotional speech synthesis based on dynamic pitch and duration control. Although they are less articulate than TTS, they still show more than 50 % intelligibility. Considering these points, subjective user evaluations for our real-time H2S system in terms of intelligibility and expressiveness were carried out.

The same female adult who recorded gesture set for the system was trained for about an hour on the H2S system consisting of a CyberGlove and a TDS01V. After the training, she generated the following Japanese words:

1. /nani/, means "What" in Japanese, when you do not understand and want to ask the speaker (nani1)
2. /nani/, means "What" in Japanese, when you are upset with the speaker (nani2)
3. /iina/, means "OK" in Japanese, as the back-channel feedback (iina1)
4. /iina/, means "Jealous of you" in Japanese, when you are jealous of the speaker (iina2)
5. /iie/, means "No" in Japanese, without any emotion (iie)

Speech was recorded 10 times for each word. She chose one sample from a total of 10 samples for each word by herself, which best meets the above conditions.

The speech rate of generated samples were about 1.2 morae/sec. This is much slower than the appropriate speaking rate (about 6.5 morae/sec, 1.3 times faster than average NHK news 8.5 morae/sec) [79] and the appropriate speed of the finger alphabet (2 morae/sec) [80]. In order to mitigate the influence of the speaking rate on impression for the listeners, double speed 5 samples were made from those 5 samples for the listening tests.

These $5+5=10$ samples were randomized and shown 6 native Japanese speakers. They were asked to answer (1) what did the speaker say (2) what do you think about the emotion of the speaker. They were informed that the samples are Japanese words used in daily conversation. Results are shown in Figure 6.6 and Table 6.1. When phoneme-based Intelligibility was calculated based on the answers to question (1), prolonged sound and duplication of /n/ were ignored. For example, 「んあ (na)」「なー (na-)」「んな (nna)」 were considered as /na/. For question (2), multiple answers were allowed. Numbers in the parentheses show the number of people who answered. For simplicity, the answers "no emotion" and similar were omitted.

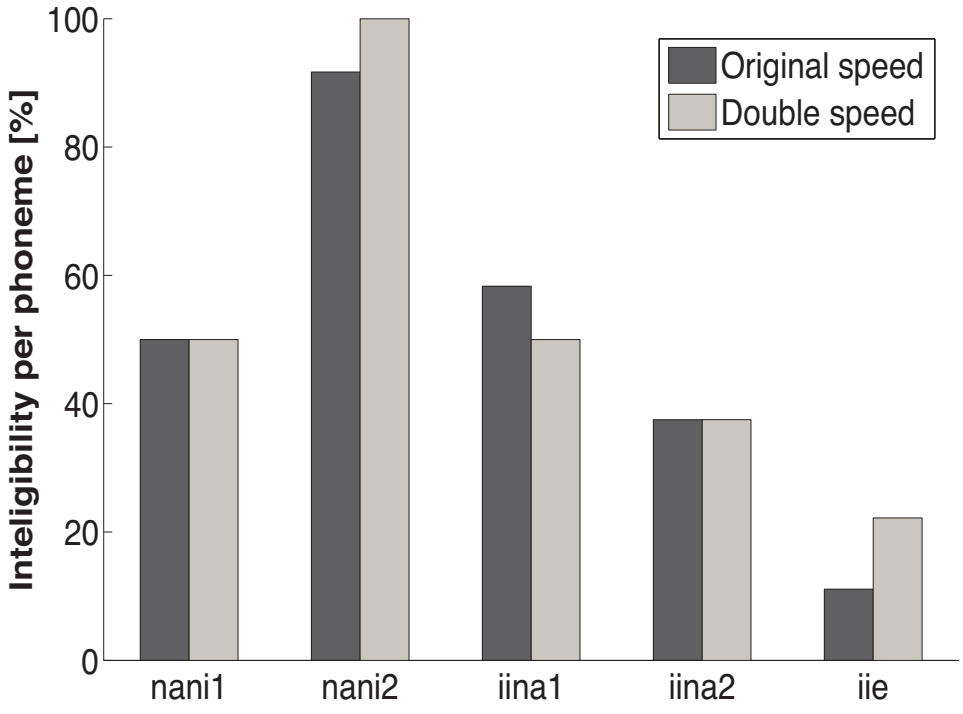


Figure 6.10: Phoneme-based intelligibility.

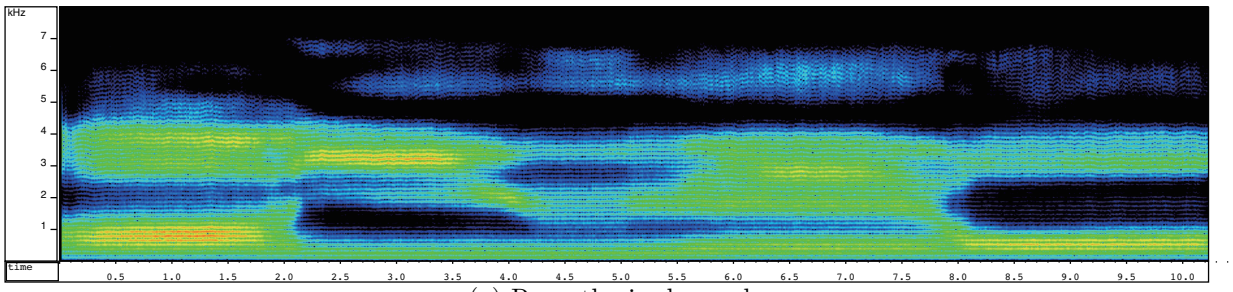
6.7 Discussion

According to Figure6.6, nearly half of the phonemes generated by our H2S system were perceived properly. It is almost the same quality as Yabu's speech synthesizer (65% for nasals), described in Chapter 2. The most intelligible sample was nani2 and almost all people perceived /nani/ correctly. The worst one was iie, whose intelligibility was 10〜20%. The most common error was to add /n/ at the beginning of the word, e.g. /iina/ was written as /nina/. The second most was to drop /i/, such as writing /iina/ as /na/. We have considered the reason for this to be that the gestures for /i/ and /n/ are very similar (see Figure 6.7) in our H2S system.

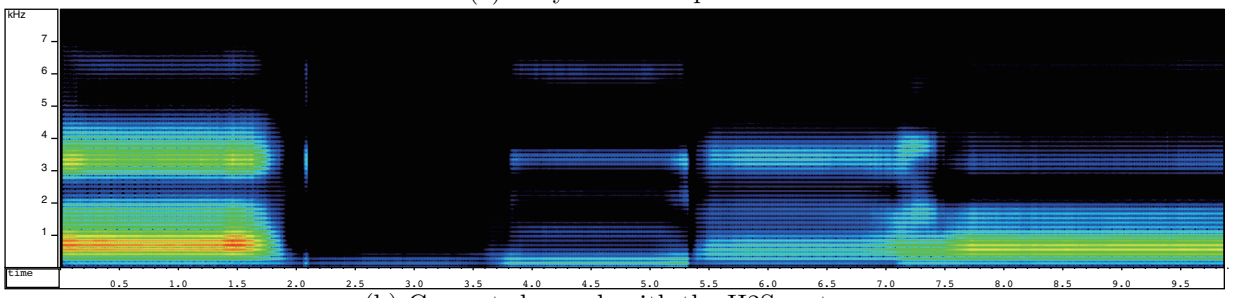
Energy for vowels are larger than that for consonants. By considering this point, the conversion models for the S2H-delta and H2S systems were trained with 0 degree of cepstrum, which express the energy of speech. In this way, vowels and consonants are expected to be discriminated in the conversion model for S2H-delta systems. On the other hand, in the conversion model for the H2S system, /i/ and /n/ are handled similary due to the similarity of their gestures. In the result, the energy of /i/ was smaller than that of other vowels. Figure 6.11 shows the spectrograms for /aiueo/, one resynthesized and one generated with the H2S system.

Table 6.1: Emotions estimated from the generated speech.

Samples	Impression
nani1	asking(2), confirming, disappointed, bad temper
nani1x2	asking(2), blunt, bored, small anger
nani2	surprised(3), sleepy(3), happy
nani2x2	surprised(2), anger(2), happy
iina1	sleepy(2), at ease, anger, jealous
iina1x2	sleepy, sad
iina2	sleepy(2), sad, exhausted
iina2x2	hurry, anger
iie	sleepy, at ease, cheerful, jealous
iiex2	sad



(a) Resynthesized speech.



(b) Generated speech with the H2S system.

Figure 6.11: Comparison of synthesized speech for /aiueo/.

For the same reason as in the case of /i/, the energy of /u/ is also small compared with /a/, /e/ and /o/. In order to generate more articulate speech, gestures should be designed so that all gestures can be discriminated.

Speaking rate does not influence the intelligibility a lot, while it does influence the emotion estimated from the speech. At the original rate, generated speech tend to be perceived as "sleepy", "at ease" and "exhausted". At double rate however, this tendency was mitigated and the generated speech was perceived as it was intended. As the speaking rate got closer to that which people hear in daily conversation, the emotion of the speaker

was more easily perceived. In this experiment, the user is not capable of forming gestures quickly. People using sign language use a finger alphabet in conversation to express new words. Considering this point, conversation with our H2S system may be possible if the user is well trained in the gesture design.

6.8 Summary

In this chapter, we developed a real-time H2S system and evaluated its effectiveness. An H2S system based on the framework described in the previous chapter sometimes derives difficult-to-form gesture transitions in a realistic period, because of their large dynamic range or rapid changes. We compensated for those problems by setting the optimal weight for the gesture model and by considering delta features. Subjective user evaluations showed that a real-time H2S system trained by exploiting those data is effective enough to generate emotional speech.

Chapter7

Conclusions

7.1 Review of work

The objectives in our study is to establish the media-independent methodology for speech generation system. In order to reach the objective, we considered the speech synthesis framework based on a space mapping to desired input media. When the correspondence between the input media and speech is relatively explicit, the effectiveness of this framework is already proven. In this thesis therefore, we considered the speech synthesis based on this framework when the input media does not have explicit relationship to speech. As an example of such media, hand gesture is chosen and Hand gesture to Speech converter is developed. In this section, the contents of this thesis are quickly summarized.

Chapter 2 described the history of the speech synthesis technologies. Then we have seen two categories of conventional speech synthesis systems; systems which require symbol inputs and systems which do not require symbol inputs. The former one has three synthesis methods: synthesis based on waveform coding, synthesis based on analysis-synthesis method and synthesis by rule. The latter one also has three methods: articulatory synthesis, formant synthesis and media conversion based on a a space mapping. Looking through those technologies, we made clear the the objective of our speech synthesis system - media independent speech synthesis.

In Chapter 3, the basic frameworks for current voice conversion technique were reviewed. In statistical voice conversion, the correspondence of the two feature spaces of the source speaker and the target speaker are modeled with GMMs. This technique is applied to the conversion between different media. Our H2S system can be considered a kind of voice conversion system in which the speech space of a source speaker is replaced by the hand gesture space. It was also mentioned that the correspondence between two spaces are not explicit and that will be one of the most important issues in our research.

In Chapter 4, we implement a speech synthesizer from hand gestures based on space mapping. By considering the topological equivalence between the structure of hand gestures in a gesture space and that of vowel sounds in the vowel space, we demonstrate how a quasi-optimal correspondence can be obtained.

In chapter 5, we proposed a framework to derive the gestures for consonants when only the correspondence for vowels is given. According to the listeners evaluations, an H2S system, which exploits gesture data for consonants derived from an S2H system, can generate more natural sounds than those trained with heuristic gesture designs.

In chapter 6, we developed a real-time H2S system and evaluated its effectiveness. An H2S system based on the framework described in the previous chapter sometimes derives difficult-to-form gesture transitions in a realistic period, because of their large dynamic range or rapid changes. We compensated for those problems by setting the optimal weight for the gesture model and by considering delta features. Subjective user evaluations showed that a real-time H2S system trained by exploiting those data is effective enough to generate emotional speech.

7.2 Future work

We developed a Hand-to-Speech converter based on media conversion framework and proposed a methodology to derive the optimal correspondence between two feature spaces. This thesis claims that our proposed real-time H2S conversion system is effective enough to generate emotional speech in the five Japanese vowels and nasals. Some problems were however also found.

One of the problems of our system is gesture similarity. When similar gestures are derived with the S2H system, speeches for those gestures are often confused. In order to generate more articulate speech, gestures should be designed so that every gesture can be easily discriminated.

The second problem is F0. In current framework, F0 is generated frame-by-frame. F0 is however, a piecewise feature. More natural speech would be obtained if an F0 generation system in which longterm changes of F0 is taken account is introduced.

Most importantly, this system was not evaluated by non-expert users. The needs of users are often difficult to determine for researchers at the lab. As soon as more practical systems are realized, subjective user evaluations should be carried out with normal/non-expert users.

Bibliography

[1] Say-it! SAM, Words+, Inc.
http://www.words-plus.com/website/products/syst/say_it_sam2.htm

[2] Voice Aids, Arcadia, Inc.
http://www.arcadia.co.jp/VOCA/ (in Japanese)

[3] The world's top Speech Synthesis websites
http://allwebhunt.com/dir-wiki.cfm/Top/Computers/Speech_Technology/Speech_Synthesis

[4] T. Masuko, K. Tokuda, and T. Kobayashi, "Imposture using synthetic speech against speaker verification based on spectrum and pitch," *ICSLP2000*, pp. 302–305 (2000).

[5] K. I. Nordstrom, S. Fels, C. D. Hassall, and B. Pritchard, "Developing vowel mappings for an interactive voice synthesis system controlled by hand motions ", *Journal of Acoustic Society of America*, Vol.127, Issue 3, pp. 2021 (2010).

[6] 荒井 隆行, "声道形状を単純化した模型による音声の音響教育 ", 電子情報通信学会技術研究報告, Vol. 109, No. 10, pp. 7–12 (2009).

[7] 坂田 聡, 佐伯 勇哉, 柴田 航, 上田 裕市, "母音発声のリアルタイム視聴覚フィードバックのための正規化構音空間の検討とその応用 ", 電子情報通信学会技術研究報告, Vol. 111, No. 225, pp. 55–60 (2011).

[8] 藪 謙一郎, 伊福部 達, 青村 茂, "ポインティングデバイスを利用した音声生成方式 : 発話障害者のための支援機器として", 日本保健科学学会誌, Vol. 12, No. 1, pp. 49–57 (2009).

[9] A. Kunikoshi, Y. Qiao, N. Minematsu and K. Hirose, "Speech generation from hand gestures based on space mapping, "*in Proceedings of INTERSPEECH*, pp. 308–311 (2009).

[10] 市川 熹、手嶋 教之：福祉と情報技術, オーム社 (2006).

[11] 井上 剛伸, "重度障害者の自立移動を支援する技術の開発－自ら移動し、社会に出て行く力のために－", 第 23 回国立身体障害者リハビリテーションセンター業績発表会 (2006).

[12] 竹内 正実：テルミン エーテル音楽と 20 世紀ロシアを生きた男, pp.28, 岳陽舎 (2000).

[13] A. Kain and M. W. Macon, "Spectral voice conversion for text-to-speech synthesis," *in Proceedings of ICASSP*, vol. 1, pp. 285–288 (1998).

[14] Y. Stylianou, O. Cappé and E. Moulines, "Continuous Probabilistic Transform for Voice Conversion," *IEEE Transaction of Speech Audio Processing*, vol. 6, pp. 131–142 (1998).

[15] T. Toda, A. W. Black and K. Tokuda, "Voice Conversion Based on Maximum-Likelihood Estimation of Spectral Parameter Trajectory," *IEEE Transaction of Audio Speech Language Processing*, vol. 15, pp. 2222–2235 (2007).

[16] P. Palo, "A Review of Articulatory Speech Synthesis," *Master's Thesis of University of Helsinki* (2006).

[17] D. H. Klatt, "Review of text-to-speech conversion for English," *Journal of Acoustic Society of America*, Vol. 82, No. 3, pp. 737–783 (1987).

[18] S. Lemmetty, "Review of Speech Synthesis Technology," *Master's Thesis of Helsinki University of Technology* (1999).

[19] K. Honda, "Evolution of vowel production studies and observation techniques," *Acoustical Science and Technology*, Vol. 23, No. 4, pp. 189–194 (2002).

[20] B. H. Story, "TubeTalker: An airway modulation model of human sound production," *in Proceedings of First International Workshop on Performative Speech and Singing Synthesis* (2011).

[21] J. P. Cater, "Electronically Speaking: Computer Speech Generation," *Howard M. Sams & Co.* (1983).

[22] L. R. Rabiner and R. W. Schafer, "Digital processing of speech signals, " *Prentice Hall* (1978).

[23] Museum of Speech Analysis and Synthesis,
`http://mambo.ucsc.edu/psl/smus/smus.html`

[24] Helmholtz's apparatus for the synthesis of sound: an electrical 'talking machine', Whipple Museum of the History of Science

`http://www.hps.cam.ac.uk/whipple/explore/acoustics/hermanvonhelmholtz/helmholtzssynthesizer/`

[25] T. Chiba and M. Kajiyama, "The vowel: Its Nature and Structure," *Tokyo-Kaiseikan* (1941).

[26] S. Furui, "Digital Speech Processing, Symthesis, and Recognition," *Marcel Dekker Inc,* (2001).

[27] 西澤 信行, 河井 恒, "素片接続型音声合成における最良優先探索に基づく素片選択", 電子情報通信学会技術研究報告, Vol. 105, No. 272, pp. 67–72 (2006).

[28] N. Campbell and A. W. Black, "Chatr : a multi-lingual speech re-sequencing synthesis system," *Technical Report of IEICE*, Vol. 96, No. 39, pp. 45–52 (1996).

[29] W. Hamza, R. Bakis, Z. W. Shuang and H. Zen, "On building a concatenative speech synthesis system from the Blizzard Challenge Speech Databases," *in Proceedings of EUROSPEECH*, pp. 97–101 (2005).

[30] D. Tihelka, J. Kala, and J. Matousek, "Enhancements of Viterbi Search for Fast Unit Selection Synthesis," *in Proceedings of INTERSPEECH*, pp. 174–177 (2010).

[31] V. Strom and S. King, "A classifier-based target cost for unit selection speech synthesis trained on perceptual data," *in Proceedings of INTERSPEECH*, pp. 150–153 (2010).

[32] T. Nose, J. Yamagishi, T. Masuko and T. Kobayashi, "A style control technique for HMM-based expressive speech synthesis," *IEICE Transaction on Information and Systems,* Vol. E90-D, No. 9, pp.1406–1413 (2007).

[33] A. Black, P. Taylor and R. Caley: "The festival speech synthesis system, " `http://www.festvox.org/festival/`

[34] M. Schröder and J. Trouvain: "The German text-to-speech synthesis system MARY: A tool for research, development and teaching," *International Journal of Speech Technology*, Vol. 6, pp.365–377 (2003).

[35] G. Fant, "Acoustical Theory of Speech production: With Calculations based on X-Ray studies of Russian Articulations," *The Hague Mouton* (1970).

[36] P. Mermelstein, "Articulatory model for teh study of speech production," *Journal of the Acoustical Society of America*, Vol. 53, No. 4, pp. 1070–1082 (1973).

[37] S. Maeda, "Compensatory articulation during speech: evidence from the analysis and synthesis of vocal-tract shapes using an articulatory model," *W. J. Hardcastle and A. Marchal (Eds.), Speech Production and Speech Modelling, Kluwer Academic, Dordrecht*, pp. 131–149 (1990).

[38] J. Dang, K. Honda, "Estimation of vocal tract shape from sounds via a physiological articulatory model," *J.Phonetics,* Vol.30, pp. 511–532 (2002).

[39] 緒方 公一, 園田 頼信, "調音に基づく音声合成システム–GUI を用いたシステムの開発–", 電子情報通信学会技術研究報告, Vol. 102, No. 292, pp. 29–34 (2002).

[40] K. Ishizaka and J. L. Flanagan, "Synthesis of voiced sounds from a two-mass model of the vocal cords,", *Bell Syst. Tech. J.,* 51, pp. 1233–1268 (1972).

[41] P. Rubin, T. Baer, and P. Mermelstein. "An articulatory synthesizer for perceptual research," *Journal of Acoustic Society of America*, Vol. 70, pp. 321–329 (1981).

[42] O. Engwall, "Combining MRI, EMA and EPG measurements in a three-dimentional tongue model," *Speech Communication*, Vol. 41, pp, 303–329 (2003).

[43] P. Birkholz, D. Jackel, B.J.Kroeger, "Construction and control of a three-dimensional vocal tract model," *in Proceedings of ICASSP*, pp. 873–876 (2006).

[44] I. Stavness, J. Lloyd, Y. Payan, and S. Fels, "Coupled Hard-Soft Tissue Simulation with Contact and Constraints Applied to Jaw-Tongue-Hyoid Dynamics," *International Journal of Numerical Methods in Biomedical Engineering* (2010).

[45] Fant, G., "Acoustic Theory of Speech Production. ", *Mouton & Co, The Hague, Netherlands* (1960).

[46] Klatt, D. H., "Software for a cascade/parallel formant synthesizer", *Journal of Acoustic Society of America*, Vol. 82, No. 3, pp.737–793 (1980).

[47] Stevens, K. N., Keyser, S. J., and Kawasaki, H., "Toward a phonetic and phonological theory of redundant features in Invariance and Variability in Speech Processes", *Lawrence Erlbaum Associates, New Jersey* (1986).

[48] T. Toda and K. Tokuda, "Statistical approach to vocal tract transfer function estimation based on factor analyzed trajectory HMM," *in Proceedings of ICASSP*, pp. 3925–3928 (2008).

[49] T. Hueber, P. Badin, and G. Bailly, "Statistical Mapping between Articulatory and Acoustic Data, Application to Silent Speech Interface and Visual Articulatory Feedback, "*Proceedings of First International Workshop on Performative Speech and Singing Synthesis* (2011).

[50] K. Nakamura, M. Janke, M. Wand, and T. Schultz, " Estimation of fundamental frequency from surface electromyographic data: EMG–to– F0, "*Proceedings of ICASSP*, pp. 573–576 (2011).

[51] M. Abe, S. Nakamura, K. Shikano and H. Kuwabara, "Voice conversion throught vector quantization," *Proc. of International Conference of Acoustics, Speech and Signal Processing*, pp. 655–658 (1998).

[52] S. Nakamura and K. Shikano, "Speaker adaptation applied to HMM and neural networks", *Proc. ICASSP*, pp. 89–92, Glasgow, UK (1989).

[53] H. Matsumoto and Y. Yamashita, "Unsupervised speaker adaptation from short utterances based on a minimized fuzzy objective function", *Journal of Accoustic Society of Japan*, Vol. 14, No. 5, pp. 353–361 (1993).

[54] H. Valbret, E. Moulines and J.P.Tubach, "Voice transformation using PSOLA technique", *Speech Communication*, Vol. 11, No. 2–3, pp. 175–187 (1992).

[55] T. Toda and K. Tokuda, "A speech parameter generation algorithm considering Global Variance for HMM-based speech synthesis," *IEICE Transaction on Information and Systems*, Vol. E90–D, No. 5, pp. 816–823 (2007).

[56] S. S. Stevens, J. Volkmann, and E. B. Newman, "A scale for the measurement of the psychological magnitude pitch," *The Journal of the Acoustical Society of America*, Vol. 8, pp. 185–190 (1937).

[57] S. Davis, and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *in IEEE Transaction on Acoustic, Speech and Signal Processing*, Vol. 28, No. 4, pp. 357–366 (1980).

[58] T. Masuko, K. Tokuda, N. Miyasaki and T. Kobayashi, "Pitch pattern generation using multi-space probability distrinutioin HMM," *IEICE Transaction*, Vol. J83-D-II, no. 7, pp. 1600–1609 (2000).

[59] K. Yu, T. Toda, M. Gasic, S. Keizer, F. Mairesse, B. Thomson and S. Young, "Probablistic modelling of f0 in unvoiced regions in HMM based speech synthesis," *Proceedings of ICASSP* (2009).

[60] Q. Zhang, F. Soong, Y. Qian, J. Pan and Y. Yan, "Improved Modeling for F0 Generation and V/U Decision in HMM-based TTS, *Proceedings of ICASSP*, pp. 4606–4609 (2010).

[61] K. Yutani, Y. Uno, Y. Nankaku, A. Lee and K. Tokuda, "Voice Conversion based on Simultaneous Modeling of Spectrum and F0, *Proceedings of ICASSP*, pp. 3897–3900 (2009).

[62] A. Kunikoshi, Y. Qian, F. Soong and M. Minematsu, " Improved F0 modeling and Generation in Voice Conversion, "Proceedings of ICASSP, pp. 4568–4571 (2011).

[63] D. Talkin, W. Kleijn and K. Paliwal, "A robust algorithm for pitch tracking (RAPT) in Speech Coding and Synthesis," *Eds. Elsevier*, pp. 495–518 (1995).

[64] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigné, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction," *Speech Communication*, Vol. 27, pp.187–207 (1999).

[65] Y. Wu, J. Lin, and T. S. Huang, "Analyzing and Capturing Articulated Hand Motion in Image Sequences," *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol. 27, No.12, pp. 1910–1922 (2005).

[66] M. Santello, M. Flanders, and J. F. Soechting, "Postural Hand Synergies for Tool Use," *The Journal of Neuroscience*, Vol. 18, No. 23, pp. 10105–10115 (1998).

[67] 峯松 信明, 朝川 智, 広瀬 啓吉, "線形・非線形変換不変の構造的情報表象とそれに基づく音声の音響モデリングに関する理論的考察, "日本音響学会春季講演論文集, 1–P–12, pp. 147–148 (2007).

[68] N. Minematsu, "Mathematical evidence of the acoustic universal structure in speech," *in Proceedings of ICASSP*, pp. 889–892 (2005).

[69] N. Minematsu, T. Nishimura, K. Nishinari, and K. Sakuraba, "Theorem of the invariant structure and its derivation of speech Gestalt," *in Proceedings of SRIV*, pp. 47–52 (2006).

[70] Y. Qiao and N. Minematsu, "Structual representation with a general form of invariant divergence", *in Proceedings of Autumn Meeting of Acoustic Society of Japan*, 2–P–1, pp.105–108 (2008).

[71] S. Asakawa, N. Minematsu, and K. Hirose, "Automatic recognition of connected vowels only using speaker-invariant representation of speech dynamics," *in Proceedings of INTERSPEECH*, pp. 890–893 (2007).

[72] 町田 健 編，猪塚 元・猪塚 恵美子著：日本語音声学のしくみ，研究社 (2003).

[73] 匂坂 芳典, 東倉 洋一, "規則による音声合成のための音韻時間長制御 ", 電子情報通信学会論文誌 A, Vol. J67–A, No. 7, pp. 629–636 (1984).

[74] ジャック・ライアルズ：音声知覚の基礎，海文堂 (2003).

[75] 薮 謙一郎, 伊福部 達, 青村 茂, "発話障害者支援のための音声合成器の基礎的設計", 電子情報通信学会技術研究報告, Vol. 105, No. 686, pp. 59–64 (2006).

[76] P. F. Brown, J. Cocke, S. A. Della Pietra, V. J. Della Pietra, F. Jelinek, J. D. Lafferty, R. L. Mercer, and P. S. Roossin, "A statistical approach to machine translation," *Computational Linguistics*, Vol. 16, No. 2, pp. 79–85 (1990).

[77] D. Saito, S. Watanabe, A. Nakamura, and N. Minematsu, "Probabilistic Integration of Joint Density Model and Speaker Model for Voice Conversion," *in Proceedings of INTERSPEECH*, pp.1728–1731 (2010).

[78] Saito, D., Watanabe, S., Nakamura, A., and Minematsu, N., "High accurate Model-Integration-based Voice Conversion using dynamic features and model structure optimization," Proc. ICASSP, pp. 4576–4579 (2011).

[79] 今井 篤, 池沢 龍, 清山 信正, 中村 章, 都木 徹, 宮坂 栄一, 中林 克己, "ニュース音声を対象にした時間遅れを蓄積しない適応形話速変換方式", 電子情報通信学会論文誌 A, Vol.J83–A, No.8, pp.935–945 (2000).

[80] 平井 花奈, 松本 充司, "指文字アニメーションの携帯電話への適用性に関する一検討", 電子情報通信学会技術研究報告. Vol.104, No.280, pp.7–12 (2004).

List of Publications

Journal papers

[1] **A. Kunikoshi**, Y. Qiao, D. Saito, N. Minematsu and K. Hirose, “空間写像に基づく母音と鼻子音を対象としたジェスチャー–音声変換システム (The Speech–to–Hand Conversion System for Vowels and Nasals based on Space Mapping),” (in Japanese), Journal of Information Processing Society of Japan, (submitted).

International conferences

[1] **A. Kunikoshi**, Y. Qiao, N. Minematsu and K. Hirose, “Speech generation from hand gestures based on space mapping, ”In Proc. INTERSPEECH, pp.308–311 (2009).

[2] **A. Kunikoshi**, Y. Qian, F. Soong and M. Minematsu, “Improved F0 modeling and Generation in Voice Conversion, ”In Proc. ICASSP, pp.4568-4571 (2011).

[3] **A. Kunikoshi**, Y. Qiao, D. Saito, N. Minematsu and K. Hirose, “Gesture Design of Hand-to-Speech Converter derived from Speech-to-Hand Converter based on Probabilistic Integration Model, ”In Proc. INTERSPEECH (2011).

Technical Reports and domestic conference papers

[1] **A. Kunikoshi**, Y. Qiao, N. Minematsu, K. Hirose, “空間写像に基づく手の動きを入力とした音声生成系 (Speech generation from hand motions based on space mapping), ” (in Japanese), In Proc. Autumn meeting of Acoustic Society of Japan, pp.375–376 (2008).

[2] **A. Kunikoshi**, Y. Qiao, M. Suzuki, N. Minematsu, K. Hirose, “空間写像に基づく手の動きを入力とした音声生成系の構築 (Development of a speech generator from hand motions based on space mapping), ” (in Japanese), In IEICE Technical Report, SP2007-30, pp.37–42 (2008).

[3] **A. Kunikoshi**, Y. Qiao, M. Suzuki, N. Minematsu, K. Hirose, H. Banno, “ジェスチャー空間と音響空間の写像に基づくリアルタイム音声生成系 (Speech generation from

hand motions based on space mapping), ” (in Japanese), In Proc. Spring meeting of Acoustic Society of Japan, pp.445–448 (2009).

[4] **A. Kunikoshi**, Y. Qiao, N. Minematsu, K. Hirose, “手の動きを入力としたリアルタイム音声生成系における鼻音の合成とピッチ制御に関する検討 (Nasal sound generation and pitch control for the real-time hand to speech system), ” (in Japanese), In IEICE Technical Report, SP2009-56, pp.43–48 (2009).

[5] **A. Kunikoshi**, Y. Qiao, N. Minematsu, K. Hirose, “手の動きを入力としたリアルタイム音声生成系における鼻音の合成に関する検討 (Nasal sound generation for the real-time hand to speech system), ”(in Japanese), In Proc. Spring meeting of Acoustic Society of Japan, pp.419–422 (2010).

[6] **A. Kunikoshi**, D. Saito, Y. Qiao, N. Minematsu, K. Hirose, “手から声のメディア変換も出ると手のジェスチャーモデルの確率的統合に基づく異メディア空間の対応付けの検討 (Gesture Design of Hand-to-Speech Conversion Based on Probabilistic Integration of Speech-to-Hand Joint Density Model and Hand Gesture Model), ” (in Japanese), In IEICE Technical Report, SP2010-127, pp.73–78 (2011).

[7] **A. Kunikoshi**, Y. Qiao, D. Saito, N. Minematsu, K. Hirose, “手から声のメディア変換モデルと手のジェスチャーモデルの確率的統合に基づく異メディア空間の対応付け (Gesture design of Hand-to-Speech Conversion based on Probabilistic Integration of Speech-to-Hand Joint Density Model and Hand Gesture model), ” (in Japanese), In Proc. Spring meeting of Acoustic Society of Japan, 1–Q–18(c), pp.371–374 (2011).

[8] **A. Kunikoshi**, D. Saito, Y. Qiao, N. Minematsu, K. Hirose, “Gesture Design of Hand-to-Speech Conversion Based on Probabilistic Integration of Speech-to-Hand Joint Density Model and Hand Gesture Model, ”(in Japanese), In Proc. International Workshop on Performative Speech and Singing Synthesis (2011).

[9] **A. Kunikoshi**, Y. Qiao, N. Minematsu, K. Hirose, “手の動きを入力とした音声生成系における有声／無声の明確化に関する検討 (Differentiation of voiced/unvoiced gestures for Hand-to-Speech system), ” (in Japanese), In Proc. Autumn meeting of Acoustic Society of Japan, 3–Q–33 (2011).

Awards

Poster Presentation Award at Fall Meeing of ASJ 2008