NAIST-IS-DD0761022

Doctoral Thesis

Speaking-Aid Systems Using Statistical Voice Conversion for Electrolaryngeal Speech

Keigo Nakamura

March 24, 2010

Department of Information Processing
Graduate School of Information Science
Nara Institute of Science and Technology

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

Keigo Nakamura

Thesis Committee:

Professor Kiyohiro Shikano	(Supervisor)
Professor Yuji Matsumoto	(Co-supervisor)
Associate Professor Hiroshi Saruwatari	(Co-supervisor)
Assistant Professor Tomoki Toda	(Co-supervisor)

Speaking-Aid Systems Using Statistical Voice Conversion for Electrolaryngeal Speech*

Keigo Nakamura

Abstract

Speaking impairment is a serious problem for people trying to communicate using speech utterances. Laryngectomees are one type of speaking-impaired people due to losing their vocal folds because of accidents, disease, and so on. An electrolarynx (EL) is an external medical device that is easily used by attaching it to the lower jaw.

This thesis addresses two issues of the existing EL: (1) unnaturalness of the produced electrolaryngeal speech (EL speech), and (2) noisy sounds radiated from the attaching location of the EL. The production mechanism of the laryngeal sound of human voices is too complex to be completely modeled mechanically. Therefore, an EL has not been developed that enables laryngectomees to speak comparably to normal speech. Moreover, a current EL generates sound source signals with large powers. The radiated sound source signals from the attaching location might be noisy for listeners.

In order to improve the EL speech quality, a voice conversion (VC) technique is introduced, which consists of training and conversion procedures. Two sets of speech data, which are the source and the target speech, are set in this VC method so that the speech of the source speaker is converted so that it sounds like that of the target speaker. The training and conversion procedures are conducted based on the maximum likelihood criterion. Gaussian mixture models are used to describe the acoustic feature spaces between the source and the target speech data.

*Doctoral Thesis, Department of Information Processing, Graduate School of Information Science, Nara Institute of Science and Technology, NAIST-IS-DD0761022, March 24, 2010.

This thesis first proposes two kinds of speaking-aid systems. One of the aid systems converts EL speech to normal speech by estimating both target spectra and F_0 contours from only the source spectral information. The other system also converts the EL speech to whispered voice to avoid the problem of estimating natural F_0 contours.

To address the other problem of noisy radiated sound source signals, this thesis employs another sound source unit that generates signals with extremely small power so that the source signals are not captured by listeners. The produced small-powered EL speech is recorded using a special microphone called a Non-Audible Murmur (NAM) microphone by attaching it to the user's skin directly. Although the speech is recorded, the voice quality is extremely poor. In order to make the small-powered EL speech audible, this thesis also proposes two other aid systems that convert the small-powered EL speech to whispered or normal speech.

From experimental evaluations, it is demonstrated that the proposed systems dramatically improve the naturalness of the EL speech using the VC. Moreover, although the intelligibility of the converted speech is slightly degraded than that of the source EL speech, the converted speech is preferred to the source EL speech.

Keywords:

Laryngectomee, Electrolarynx, Voice conversion, Silent sound source, Non-Audible Murmur microphone, Air-pressure sensor

電気音声に対する統計的声質変換を用いた発声支援システム*

中村 圭吾

内容梗概

音声に障害がある発声障害者は，他者との音声コミュニケーションにおいて深刻な困難を伴う．喉頭摘出者は，事故や病気などを理由に声帯を失った発声障害者である．喉頭摘出者の音声再建は彼らの日常生活において極めて深刻な課題であり，古くから研究されている．電気式人工喉頭（以下，電気喉頭）は，喉頭摘出者による発声を容易に実現する医療用の外部音源機器である．ユーザは電気喉頭を手で保持し，下あごに圧着してスイッチで音源の発生または停止を切り替えることで音源を出力し，音声を発声する．

本論文では，(1) 電気喉頭をのような外部機器を用いて発声された音声（以下，電気音声）の不自然性，(2) 電気喉頭の圧着位置から漏れる音源の騒音性という 2 つの問題を解決する．人の声帯音源は非常に複雑であり，機械的に模擬することは容易ではない．特に，電気喉頭を用いた自然な抑揚の実現は，電気喉頭に関する主要な課題として以前から研究されている．現在では，自然な抑揚を生成するような電気喉頭は少なく，多くの電気喉頭の振動数はあらかじめ内蔵されたものに固定されている．そのため，発声される電気音声の抑揚は極めて乏しく，人間が発声する音声として極めて不自然である．さらに，既存の電気喉頭は通常音声と同等の大きさで発声することを前提としているため，十分大きなパワーを持った音源信号を出力する必要がある．一方で，電気喉頭の圧着位置から漏れる音源信号の音は，静環境下では周囲の者にとって耳障りとなり，雑音環境下では電気音声の明瞭性を下げる一因になっている．

本論文では，電気音声の音質を改善するために，統計的声質変換技術を用いた発声支援システムを提案し，評価する．本論文で用いる統計的声質変換技術を

*奈良先端科学技術大学院大学 情報科学研究科 情報処理学専攻 博士論文，NAIST-IS-DD0761022, 2010 年 3 月 24 日.

電気音声に適応して音質を改善する試みは過去になされておらず，電気音声に対する統計的声質変換の可能性を調査するという点が，本論文の最大の貢献である．本論文ではまず電気音声を通常音声あるいはささやき声に変換するシステムを提案する．電気音声を通常音声に変換するシステムでは，電気音声のスペクトルのみから通常音声のスペクトルと基本周波数（以下，F_0）を推定する．ささやき声への変換は，自然な F_0 推定の困難さを回避する狙いがある．本論文では，より自然な F_0 を推定するために，ユーザの呼気で電気喉頭の振動数を制御する呼気センサーを導入し，呼気センサーを用いた電気音声から通常音声に変換するシステムも提案する．さらに，音源信号の騒音性問題を解決するために，本論文では周囲の者に聴取されないほど微弱な音源信号を出力する音源信号を用いる．発声される微弱な電気音声は，非可聴つぶやきマイクロフォンを用いて話者の体表から直接収録される．収録された微弱な電気音声の音質は極めて悪いため，本論文では微弱な電気音声を通常音声またはささやき声に変換する発声支援システムも提案し，微弱な電気音声の品質改善を目指す．

実験結果から，統計的声質変換を用いた提案システムによって，電気音声の自然性を劇的に改善可能であるおとを示す．さらに，変換音声の明瞭性は元音声と比べて若干劣化するが，変換音声は元音声と比べて好ましいことを確認する．また，呼気センサーを用いることで，従来の電気喉頭では達成できなかった F_0 推定精度が実現されることを示す．

キーワード

喉頭摘出者, 電気式人工喉頭, 統計的声質変換, 微弱音源, 非可聴つぶやきマクロフォン，呼気センサー

Acknowledgements

I would like to express my appreciation to Professor Kiyohiro Shikano of Nara Institute of Science and Technology, my thesis advisor, for his constant guidance through my master's and doctoral course.

I would also like to express my gratitude to Professor Yuji Matsumoto, and Associate Professor Hiroshi Saruwatari of Nara Institute of Science and Technology, my thesis advisor, for their invaluable comments to this thesis.

I would like to thank Assistant Professor Hiromichi Kawanami of Nara Institute of Science and Technology, for their beneficial comments. I would like to particularly express my deepest appreciation to Assistant Professor Tomoki Toda of Nara Institute of Science and Technology, my thesis advisor, for his constant guidance and considerate encouragement through my master's and doctoral course. His great efforts for voice conversion play the most important key technique in this thesis. This work could not have been accomplished without his skills and carefully designed exact direction of this research.

I want to thank all members of the Speech and Acoustics Laboratory in Nara Institute of Science and Technology for providing fruitful discussions.

I would like to thank Japan Society for the Promotion of Science for making me a research fellowship for young scientists.

I would like to thank Dr. Yoshitaka Nakajima who is a cooperative researcher of Nara Institute of Science and Technology for developing key device of Non-Audible Murmur (NAM) microphone. I would like to thank Professor Hiroshi Hosoi and Assistant Professor Takefumi Sakaguchi of Nara Medical University for providing the idea of small sound source signals that is one of key ideas in the proposed system. I would like to thank Professor Tatsuya Hirahara of Toyama Prefectural University for providing the idea of converting small-powered electro-

laryngeal speech captured using a NAM microphone that is the fundamental idea of the proposed system in this thesis. The concept of this thesis originally comes from him.

I would like to thank Mr. Rikiya Hanawa who is a cooperative researcher of Nara Institute of Science and Technology for providing important advices about medical terms and ideas about F_0 detection. I would like to thank Dr. Andreas Maier, Chair of Computer Science 5 (Pattern Recognition) of the Friedrich-Alexander University Erlangen-Nuremberg, for providing novel ideas about practical speech recognition for impaired speech data and for many interesting discussions.

I am indebted to many subjects. I specially appreciate Mr. Tamioki Yamauchi, a laryngectomee, for his well-grounded comments and great efforts about many times of electrolaryngeal speech recording. Core results in this thesis are derived from his electrolaryngeal speech data. The work of this thesis could not have been accomplished without his great contribution. I would also like to express my sincere appreciation to Dr. Hiroshi Sunaga, medical doctor about ear, nose, and throat, for recording many alaryngeal speech utterances of many kinds of laryngectomees. Many alaryngeal speech samples he recorded would be useful for future work of many aid systems. I also appreciate Shimada welfare industry Ltd. for providing speech samples of various speech-impaired people. Experimental results of the appendix in this thesis could not been accomplished without their great efforts. I would like to appreciate Mr. Yoshiaki Ito, a speech-impaired people of cerebral palsy, for his great efforts of speech recording. His constant recording would be important data to establish a new system in the future.

Finally, I would like to acknowledge my family and friends for their support.

Contents

Acknowledgements **v**

1 Introduction **1**
1.1. Background and Problem Definition 1
1.2. Thesis Scope 4
1.3. Thesis Overview 8

2 Laryngectomees and Conventional Researches **11**
2.1. Introduction 11
2.2. Laryngectomees 12
2.3. Speaking Methods 14
2.4. Conventional Speaking-Aid Systems for Laryngectomees 17
2.4.1 Developing a new artificial larynx 18
2.4.2 Enhancing alaryngeal speech from software approaches . . 23
2.4.3 Speech enhancement for esophageal speech signals 25
2.4.4 Novel devices for speech communication 26
2.4.5 Applied research for VC and speech recognition 30
2.5. Summary 32

3 Statistical Voice Conversion **33**
3.1. Introduction 33
3.2. Voice Conversion Using Joint Probability Density 34
3.3. Employing Dynamic Features and Global Variances 36
3.4. Summary 42

4 Proposed Speaking-Aid System for EL Speech **44**
4.1. Introduction 44
4.2. Speaking-Aid System for EL Speech 46
4.3. Voice Conversion for EL Speech 48
4.4. Preliminary Experimental Evaluation of the Speaking-Aid System for EL speech 50
4.4.1 Experimental conditions 50
4.4.2 Experimental results 52
4.5. Speaking-Aid System for EL(air) Speech 56
4.6. Voice Conversion for EL(air) Speech 57
4.7. Conclusion 59

5 Proposed Speaking-Aid System for EL(small) Speech **60**
5.1. Introduction 60
5.2. Speaking-Aid System for EL(small) Speech 62
5.3. Voice Conversion for EL(small) Speech 70
5.4. Preliminary Experimental Evaluation of the Speaking-Aid System for EL(small) Speech 72
5.4.1 Effectiveness of enhancing auditory feedback 73
5.4.2 Robustness of VC for several small-powered sound source signals 75
5.5. Conclusion 81

6 Experimental Evaluations **83**
6.1. Introduction 83
6.2. Objective Evaluations of the Speaking-Aid Systems 84
6.2.1 Experimental conditions 84
6.2.2 Experimental results 86
6.3. Subjective Evaluations of the Speaking-Aid Systems 88
6.3.1 Experimental conditions 88
6.3.2 Experimental results 93
6.4. Discussion 96
6.4.1 Experimental conditions 96
6.4.2 Experimental results 98

6.5. Conclusion 101

7 Conclusion **107**

7.1. Summary of This Thesis 107

7.2. Future Work 110

Appendix **114**

A. Case Studies of Speech Recognition for Impaired Speech 114

A.1 Introduction 114

A.2 Speech Recognition Using Hidden Markov Models 115

A.3 Experimental Evaluation 119

A.4 Conclusion 128

References **132**

List of Publications **143**

List of Figures

1.1 Speech chain of inter-personal speech communication. 2
1.2 Problems addressed in this thesis. 5
1.3 Comparison between conventional and one of speaking-aid systems proposed in this thesis. 6

2.1 Anatomical image of non-laryngectomees and total laryngectomees. 14
2.2 Major alternative speaking methods for laryngectomees. 15
2.3 Basic structure and examples of existing electrolarynxes. 18
2.4 Example of EL speech produced by laryngectomee who is proficient to produce EL speech. 19
2.5 Recording scene of EL speech using air-pressure sensor. 21
2.6 Example of waveforms, spectrograms, and F_0 contours for EL speech using air-pressure sensor produced by laryngectomee. . . . 22
2.7 Attaching location and basic structure of NAM microphone. . . . 29
2.8 Examples of waveforms and spectrograms of normal speech and NAM recorded with NAM microphone. Solid lines in the spectrograms show first and second formants extracted automatically. . . 29

3.1 Overview of statistical VC procedures using Gaussian mixture model in maximum likelihood manner. 35
3.2 Example of over-smoothing of converted features compared to target ones. 40
3.3 Example of converted features considering GV. 43

4.1 Overview of proposed speaking-aid system for conventional EL speech. 46

4.2 Flow chart of constructing segmental feature vectors from static feature vectors. 49
4.3 Example of waveforms, spectrograms, and F_0 contours for source EL speech. 53
4.4 Example of waveforms, spectrograms, and F_0 contours for converted normal speech. 54
4.5 Example of waveforms, spectrograms, and F_0 contours for target normal speech. 55
4.6 Overview of proposed speaking-aid system for EL(air) speech. . . 57
4.7 Flow chart of constructing segmental feature vectors from spectral and F_0 feature vectors. 58

5.1 Overview of proposed speaking-aid system for EL(small) speech. . 62
5.2 Designed sound source signals with different spectra. 64
5.3 Flow chart of designing compensation waves into target speech. . 66
5.4 Example of waveforms, spectrograms, and F_0 contours using pulse train produced by a non-laryngectomee. 67
5.5 Example of waveforms, spectrograms, and F_0 contours using sawtooth waves produced by a non-laryngectomee. 68
5.6 Example of waveforms, spectrograms, and F_0 contours using compensation waves into whispering produced by a non-laryngectomee. 69
5.7 Histograms of normalized powers of EL(small) speech using pulse train, sawtooth waves, and compensation waves into whispering. . 70
5.8 Histograms of normalized powers of EL(small) speech using average-powered sawtooth waves, large-powered sawtooth waves, and EL. 71
5.9 Histograms of normalized powers of EL(small) speech using average-powered sawtooth waves, extremely small-powered sawtooth waves, and articulation in which any sound source signals are not used when speaking. 72
5.10 Speech recording condition under existing background noises. . . . 75
5.11 Mean opinion score of the stability of articulation under existing background noises. 76
5.12 Example of EL(small) speech with or without amplified auditory feedback under existing background noises. 77

5.13 Mel-cepstral distortions without power information under existing background noises. 78
5.14 Mel-cepstral distortion of EL(small) speech without power information using sound sources with different spectra. 79
5.15 Mel-cepstral distortion of EL(small) speech without power information using sound sources with different powers. 80
5.16 Mean opinion score of EL(small) speech using sound sources with different spectra and powers. 82

6.1 Errors of voiced or unvoiced decision for all speaking-aid systems proposed in this thesis. Basic power denotes same power as pulse train. 87
6.2 Correlation coefficients between converted and target F_0 contours about only voiced frames for both data. 'spc $\rightarrow F_0$' denotes target F_0 contours are estimated using source spectral features. 88
6.3 Examples of waveforms, spectrograms, and F_0 contours of EL speech produced by a laryngectomee and those of converted normal speech. 89
6.4 Examples of waveforms, spectrograms, and F_0 contours of EL(air) speech produced by a laryngectomee and those of converted normal speech. 90
6.5 Examples of waveforms, spectrograms, and F_0 contours of EL(small) speech using pulse train produced by a laryngectomee and those of converted normal speech. 91
6.6 Example of waveforms, spectrograms, and F_0 contours of target normal speech. 92
6.7 Mean opinion score with related to intelligibility for all proposed aid systems. 96
6.8 Mean opinion score with related to naturalness for all proposed aid systems. 97
6.9 Mean opinion score with related to preference for all proposed aid systems. 98
6.10 Errors of voiced or unvoiced decision using target normal speech of which F_0 contours similarly represents that of EL(air) speech. . 100

6.11 Correlation coefficients between converted and additionally recorded target F_0 contours about only voiced frames for both data. 101
6.12 Examples of waveforms, spectrograms, and F_0 contours of EL(air) speech of which F_0 contours similarly represents those of normal speech shown in **Figure 6.13**. 103
6.13 Examples of waveforms, spectrograms, and F_0 contours of target normal speech of which acoustic parameters including F_0 contours are same as those of target speech shown in **Section 6.2**. 104
6.14 Examples of waveforms, spectrograms, and F_0 contours of target normal speech of which F_0 contours are produced to similarly represent those of EL(air) speech shown in **Figure 6.12**. 105
6.15 Examples of waveforms, spectrograms, and F_0 contours of converted normal speech of which F_0 contours are estimated to represent those of normal speech shown in **Figure 6.14**. 106

A.1 Overview of speech recognition system. 115
A.2 Example of left-to-right HMM. 117
A.3 Word accuracy of EL speech and converted normal speech signals. 129
A.4 Word accuracy of EL(air) speech and converted normal speech signals. 130
A.5 Word accuracy of EL(small) speech using the sawtooth waves and converted normal speech signals. 131

List of Tables

2.1 Benefits and defects of alternative speaking methods for total laryngectomees 16

4.1 Input and output acoustic features for EL-to-Whisper and EL-to-Speech 47

4.2 Averaged mel-cepstral distortion. Values in front of and behind the slash shows distortions considering and not considering power information (i.e. 0^{th} coefficient), respectively 52

4.3 Voiced or unvoiced error rates and correlation coefficients between voiced frames of converted F_0 values and those of target ones. '$x \rightarrow y$' denotes the rate of x frames regarded as y frames. The label 'U' and 'V' denote unvoiced and voiced frames, respectively. For example, V $\rightarrow$ U means rate of voiced frames regarded as unvoiced frames 52

4.4 Input and output acoustic features for EL(air)-to-Speech 57

5.1 Averaged mel-cepstral distortion for EL-to-Speech in which imitated EL speech produced by a non-laryngectomee is used as the source speech 80

5.2 Voiced or unvoiced error rates and correlation coefficients for voiced frames for EL-to-Speech between converted F_0 values estimated from imitated EL speech produced by a non-laryngectomee and target ones. Notations are same as those in **Table 4.3** 81

6.1 Advantages and effective use of individual speaking-aid systems proposed in this thesis 84

6.2 Averaged mel-cepstral distortions for all kinds of speaking-aid systems proposed in this thesis. Values in front of and behind the slash respectively shows distortions considering and not considering power information (i.e., 0^{th} coefficient). 'Sawtooth waves 1' means averaged power of sawtooth waves is same as that of pulse train, and 'Sawtooth waves 2' means another sawtooth waves including larger power compared to 'Sawtooth waves 1' 86
6.3 Averaged mel-cepstral distortions for all proposed aid systems using target normal speech of which F_0 contours similarly represents those of additionally recorded EL(air) speech. Format of this table is same as **Table 6.2** 100

A.1 Information of speaking-impairment, adaptation data and test data for individual speakers. Note that the number in the parenthesis after the words shows that the number of the represented words are recorded in one file. For example, "words(3)" and "words(all)" show 3 words and all words are recorded in one file, respectively. NS and AG notes "The North Winds and the Sun" and "The Ants and the grasshopper", respectively 123
A.2 Information of speaking-impairment, adaptation data and test data for the remaining individual speakers. Notations are same as **Table A.1** 124
A.3 Word accuracy of various kinds of speaking-impaired people. Acoustic model after 10 iterations of MLLR is used for the SD-AM . . . 128

Chapter 1

Introduction

1.1. Background and Problem Definition

Much important information for humans is expressed by natural language, and such information is easily transferred by human voices. An advantage of speech is that speech conveys not only linguistic information but also para-linguistic information that does not appear in texts. The human voice has been one of the most traditional and powerful methods for people to transfer information and communicate with each other.

When we communicate with each other, many segmentalized processes are continuously conducted as **Figure 1.1** shows, and the whole of the connections is called the speech chain. A speaker first establishes an abstract concept that he or she wants to utter. This step is the level of consciousness. Next, the speaker encodes the abstract image established in the brain to the corresponding sequence of symbols called language. This coding step is the linguistic level. Next, individual articulatory organs including vocal folds, oral cavity, tongue, and so on are moved to generate speech sounds corresponding to the coded language sequence. This step is the physiological level. The generated speech sounds are transferred through the air to the listener. This step is the acoustic level. The ears and the brain of the listener sense the speech sounds at the physiological level. The listener decodes the speech sounds to the language sequence in the linguistic level, and understand the contents in the level of consciousness. These processes in the speech chain are iterated between the speaker and the listener

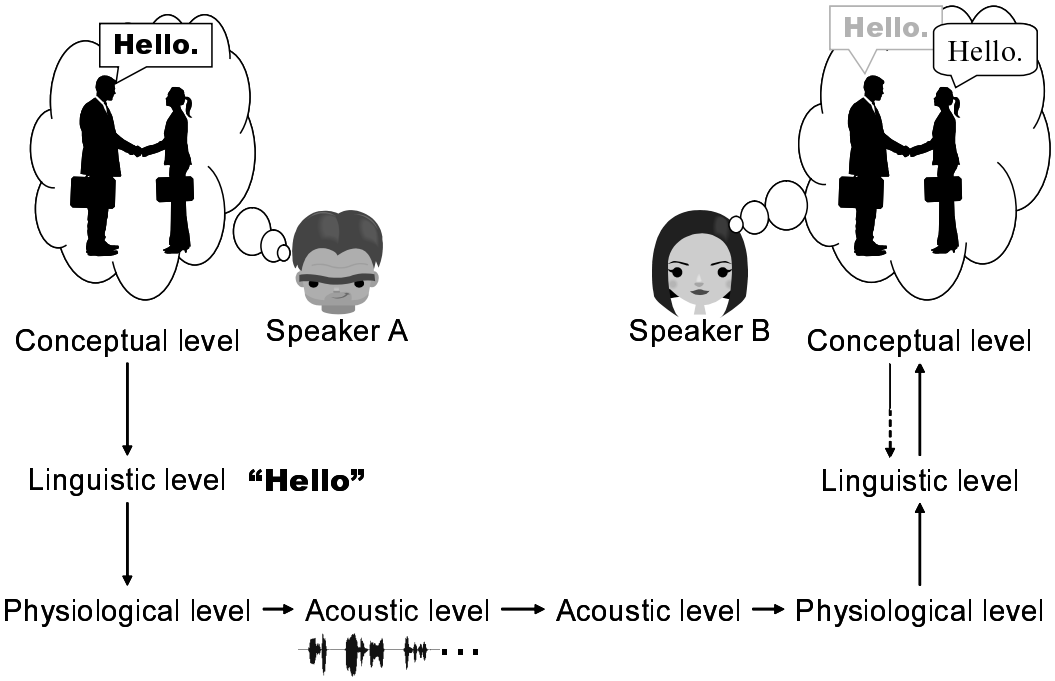

Figure 1.1. Speech chain of inter-personal speech communication.

to work out their speech communication.

There are many kinds of disorders related to speech communication, and therefore, it is difficult to know the entire scope of communication disorders [1]. Communication disorders can be classified from several points of view. Communication disorders occur wherever the speech chain breaks, and one classification from the viewpoint of the speech chain is considered.

In disorders of the conscious level, linguistic disorders due to developmental language delay or mental retardation are concerned. In disorders of the linguistic level, aphasia is the most major disorder, which is total or partial disorder of language ability including listening and speaking related to speech function and also including reading and writing related to text function due to cerebral vascular disorders such as cerebral infarction caused by disorders of the linguistic function in the brain. Disorders of the physiological level are also known as speech disorders. Speech disorders are due to hearing difficulty or due to disorders of peripheral speech organs such as vocal folds, the tongue, and so on. Speech

disorders include sub-groups from the viewpoint of which organ is affected and another viewpoint of how the organ is affected. The latter case includes organic, motor, and functional disorder as sub-groups. Organic disorders include physical change of the organ such as loss or hypertrophy due to diseases, accidents, and so on. Motor dysfunction, which is often also known as dysarthria, includes disorders of smoothly using the organ due to cerebral palsy, for example. When the function of the organ is impaired, and the organ has no physically problem, the disorders are regarded as mental problems and categorized as functional disorders. Some patients might have disorders in plural categories, and therefore, not all the patients are always classified in those categories. The author recognizes that this classification might be clinically or scientifically imprecise; however, the classification of disorders is not within the scope of this thesis. It is important simply to understand the disorders focused on in this thesis.

The disorder focused on in this thesis is due to loss of the whole larynx including vocal folds, which is classified as an organic disorder at the physiological level. Laryngectomy is a major surgical operation to cure the laryngeal cancer, in which the larynx and the vocal folds are removed. A patient who has undergone a laryngectomy is called a laryngectomee. This thesis assumes that laryngectomees do not have disorders related to phonation. The term laryngectomee includes both partial and total laryngectomees, in whom vocal folds are partially or totally removed. The target of this thesis, however, is total laryngectomees. Therefore, the word laryngectomee refers total laryngectomees in this thesis.

Laryngectomees can speak by obtaining alternative sound source signals although they have completely lost their vocal folds. Alaryngeal speech includes all kinds of alternative speech after total laryngectomy. Major alaryngeal speech is 1) esophageal speech, 2) Tracheo-Esophageal (T-E) shunt speech, and 3) speech using an external device [2, 3, 4]. Esophageal speech generates sound source signals at the beginning of the esophagus by air flowing up from the stomach. T-E shunt speech also generates sound source signals at almost the same position of the esophageal speaking method by conveying air through a prosthesis inserted between the trachea and esophagus. Sixty surgical operations to embed voice prostheses were reported in 1980 [5]. A pneumatic artificial larynx is one of the major external devices, which generates sound source signals by exciting the vi-

brator by air flowing up from the lungs. An electrolarynx (EL) is another major external device that generates sound source signals by exciting the vibrator by pushing a button of the EL.

This thesis is interested in the speaking method using an EL. Many laryngectomees speak using an EL, although the current trend in Japan is esophageal speech and that in foreign countries is T-E shunt speech. The alternative speaking method using an EL is easy to learn. Moreover, users need less physical power to produce electrolaryngeal speech (EL speech) compared to other alternative speaking methods. On the other hand, the generated EL speech is mechanical and artificial because the frequency of the vibration is pre-defined.

1.2. Thesis Scope

This research finally aims to provide laryngectomees more smooth speech communication by modifying produced EL speech at the acoustic level of the speech chain. Given this motivation, this thesis addresses two problems as shown in **Figure 1.2**. One problem is the unnaturalness of the EL speech, and the other is radiated noises of the EL itself leaked from the attaching location of the lower jaw. These problems are carefully considered through the author's experience of speaking using only an EL for 21 days, and the author believes that addressing these two problems dramatically improves the quality of life of laryngectomees [6].

Conventional studies have tried to address the unnaturalness of EL speech at the physiological level in the speech chain. Some conventional studies have so far striven to make the alternative sound source signals of the EL close to those of natural vocal fold vibration so that laryngectomees can speak naturally with the EL. For example, the hardware approach is employed to develop a new EL [7], or a pitch control mechanism is introduced [8]. From the viewpoint of software, some problems still remain, such as acceptable content is limited and the generated speech quality is not satisfied. Other conventional studies have tried to address the radiated noise at the acoustic level in the speech chain. For example, spectral subtraction is employed [9, 10] to reduce the radiated noise.

Voice conversion (VC) is a technique in which speech signals of a speaker (so-called source speaker) are modified so that it sounds as if it has been spoken

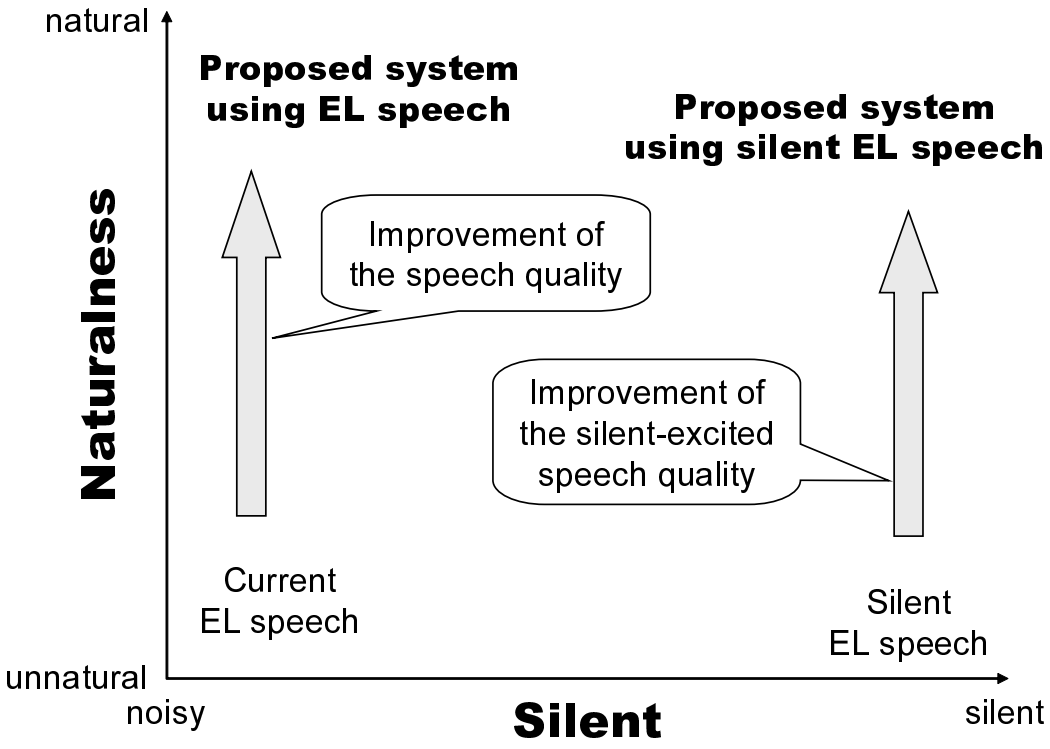

Figure 1.2. Problems addressed in this thesis.

by a different speaker (so-called target speaker) while preserving the linguistic information. Recent efforts regarding VC have made a great contribution to present natural synthesized speech. In particular, the statistical approach of VC attracts interest all over the world. As a result, VC is applied to many applications such as voice response, text-to-speech that synthesizes speech waveforms from input text information, and so on.

The problem of unnatural EL speech due to the pre-defined pitch is one of the most classical and important ones. Japanese is a kind of tone language in which pitch plays an important role to inform listeners of exact meaning. However, intonations in EL speech are lost, and therefore, the information conveyed by EL speech is limited. The key issue with the unnaturalness is fundamental frequency (F_0) contours of EL speech. Many studies have been conducted so far to address this problem; however, natural alaryngeal speech using an external

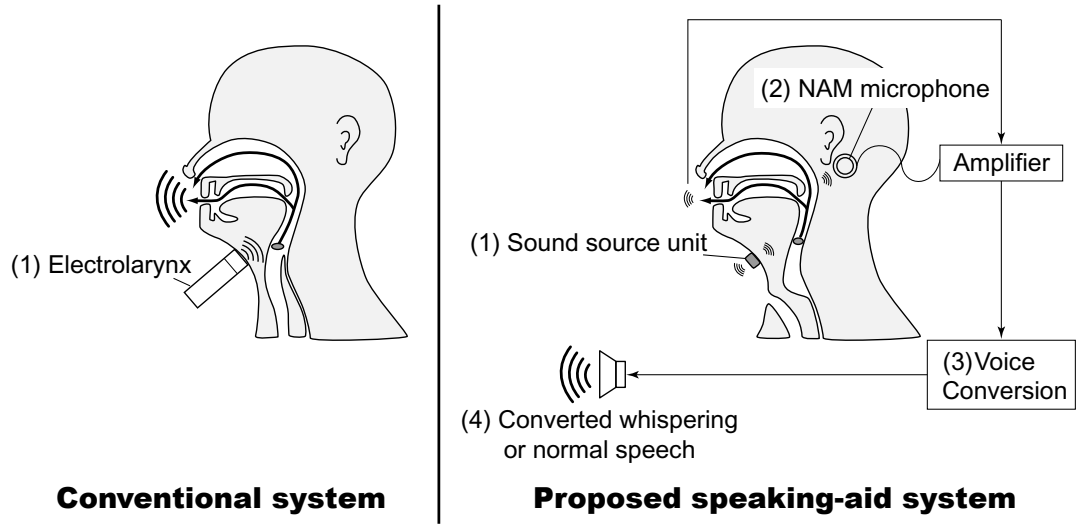

Figure 1.3. Comparison between conventional and one of speaking-aid systems proposed in this thesis.

device has not been achieved.

The radiated noises of the EL itself also constitute an important problem to be addressed. It was reported that the intensity of the radiated noises was about 20-25 dB in the development of a similar device [11] even if the lips and the mouths of the speakers were closed. Moreover, this value varies over 4-15 dB across the subjects for the same device [12]. These steady noises are generated due to the leakage of the vibrations from the attaching location of the lower jaw. This kind of leakage to the air contributes to the background noise for people around the user. As a result, the background noise including the radiated noise annoy people around the speaker in quiet places such as libraries and make the produced EL speech unintelligible in noisy, crowded settings. Moreover, the speaker has a concern that the radiated noise annoys people around the speaker even if the listeners understand the user's concerns because the radiated noises are generated whenever the user speaks. The power of the radiated noises would be larger when the user has recently had the laryngectomy because cells around the tracheostoma are hard, and therefore, the cells play the role of an amplifier. Almost the same case results when the speaker is a non-laryngectomee because some bones such as the thyroid cartilage also play the role of an amplifier.

This thesis proposes a novel speaking-aid system for laryngectomees using VC technique, as **Figure 1.3** shows, to address two problems of unnatural EL speech and radiated noises. In the proposed system, a laryngectomee produces EL speech in the same manner as the conventional method. The produced EL speech is recorded once, and then, it is converted to natural voice by the VC procedure. Finally, the converted speech is presented as the user's voice. The proposed system addresses the unnaturalness of EL speech using the VC technique, and this approach is in the acoustic level in the speech chain. The proposed system also addresses the radiated noises by employing another sound source unit that generates extremely small sound source signals, and this approach is at the physiological level in the speech chain. The proposed speaking-aid system of enhancing EL speech signals using the statistical VC approach has some advantages as follows:

- Users speak by themselves
- Users speak with a natural voice
- Arbitrary content is accepted
- Users easily learn the method
- Users speak without annoying people around them

The application study of statistical VC from EL speech has not been conducted yet, and the EL speech enhancement using statistical VC technique is the most important contribution of this thesis. In particular, this thesis enhances three kinds of EL speech, which are conventional EL speech, another EL speech using an air-pressure sensor (EL(air) speech), and other EL speech using small sound source signals captured by a body-attached microphone [13]. This thesis investigates the acceptable ranges of the small sound source signals for the statistical VC by using different kinds of sound source signals. Moreover, this thesis confirms the effectiveness of using a supporting device capturing information that is applicable to F_0 information of input EL speech.

The conversion from EL speech to normal speech (EL-to-Speech) is one reasonable conversion framework; however, natural F_0 estimation from EL speech

is a difficult problem because EL speech does not include efficient F_0 information. This problem has been revealed in another VC framework from extremely small unvoiced whispering called Non-Audible Murmur (NAM) [14] to normal speech [15]. To avoid this problem, another conversion framework has been proposed, in which EL speech is converted into whispering (EL-to-Whisper). EL-to-Whisper is a conversion between speech data not including efficient F_0 information, and this conversion framework comes from the same motivation of the conversion framework from NAM to whispering [16]. In order to estimate F_0 contours closer to the target F_0 contours, an air-pressure sensor is additionally introduced to extract source F_0 contours. This air-pressure sensor was originally designed to allow laryngectomees to speak with a natural voice using an EL, and the EL speech produced using the air-pressure sensor (EL(air) speech) is converted to normal speech (EL(air)-to-Speech). In order to address the problem of the radiated noises, this thesis introduces a new sound source unit that generates extremely small power compared to existing ELs. The EL speech excited by the small source signals also has only small power, and the small EL speech is extremely difficult to capture using a typical air-conductive microphone such as a head-set or a pin-type microphone. A NAM microphone, which is directly attached to the user's neck behind the ear, is introduced to capture the small EL speech. The small EL speech captured by a NAM microphone (EL(small) speech) is converted into normal speech (EL(small)-to-Speech) or whispering (EL(small)-to-Whisper) for the same reason as EL-to-Speech and EL-to-Whisper. The effectiveness of individual systems is experimentally evaluated in later chapters.

1.3. Thesis Overview

This thesis is organized as follows.

In **Chapter 2**, anatomical description of laryngectomees is given. The current situations of laryngectomees such as numbers of laryngectomees and the way to cure laryngeal cancer are also described. In particular, major alternative speaking methods for laryngectomees after laryngectomy are compared each other. Moreover, conventional studies related to this thesis are overviewed. Applied studies are overviewed, which are about VC and aid systems for speaking-

impaired people including not only laryngectomees but also other people who are speaking-impaired due to dysarthria.

In **Chapter 3**, the statistical VC method that is the core technique of the proposed system is described. The VC method used in this thesis consists of training and conversion parts. Before the training part, the VC method first defines the source and the target speech to prepare parallel data constructed by time-aligned identical utterances. Then, a Gaussian mixture model (GMM) is trained in the training part to model the acoustic features of joint probability density function of the source and the target acoustic features. In the conversion part, the trained GMM is used as the conversion model that outputs target static features based on the conditional probability density given the time sequence of the input feature.

In **Chapter 4**, one speaking-aid system is proposed, which enhances conventional EL speech through the conversion frameworks of EL-to-Whisper or EL-to-Speech. In EL-to-Whisper, only spectral features are estimated. In EL-to-Speech, on the other hand, not only spectral but also F_0 features are estimated from only the spectral information of the source EL speech. These conversion frameworks are experimentally evaluated as the preliminary evaluation using imitated EL speech produced by a non-laryngectomee. In order to extract manipulated source F_0 features, an air-pressure sensor is introduced, and another speaking-aid system is also proposed, which enhances EL(air) speech through the conversion framework of EL(air)-to-Speech. In EL(air)-to-Speech, target spectral features are estimated from source spectral features in the same manner of EL-to-Speech. Target F_0 features are, on the other hand, estimated from only source spectral features or from source spectral and F_0 features.

In **Chapter 5**, the other speaking-aid system is proposed, which enhances EL(small) speech through the conversion frameworks of EL(small)-to-Whisper or EL(small)-to-Speech. In this aid system, a sound source unit generating extremely small signals is employed to address the problem of radiated noises. Estimated acoustic features and the manner of estimation in EL(small)-to-Whisper and EL(small)-to-Speech are the same as EL-to-Whisper and EL-to-Speech, respectively. Because this system has the VC part inside, small sound source signals are designed from different viewpoints compared to conventional ELs. This thesis

designs small sound source signals by changing the spectrum and the power independently in this chapter. This proposed system including two conversion frameworks of EL(small)-to-Whisper and EL(small)-to-Speech are also experimentally evaluated as the preliminary evaluation using imitated EL(small) speech uttered by the same non-laryngectomee as in **Chapter 4**.

In **Chapter 6**, three kinds of proposed speaking-aid systems are experimentally evaluated using one laryngectomee's data. Advantages of the proposed systems are first summarized in this chapter. Next, objective and subjective evaluations are conducted. From the experimental results, the proposed systems dramatically enhance the naturalness of the EL speech using VC procedures. Although the intelligibility of the converted speech is slightly degraded than that of the source EL speech, the converted speech is finally preferred to the source EL speech by the listeners. From the experimental results, the results of EL(air)-to-Speech are almost the same as those of EL-to-Speech and EL(small)-to-Speech. In order to investigate the effectiveness of using the air-pressure sensor, target speech utterances are additionally recorded so that the pitch of the target normal speech is close to that of source EL(air) speech utterances. From the objective result of an additional experimental evaluation, F_0 estimation accuracy is improved by using the air-pressure sensor.

In **Chapter 7**, this thesis is concluded. This thesis is first summarized in this chapter. Future work related to this research is explained second.

In **Appendix A**, speech recognition of impaired speech due to cerebral palsy, acquired hearing impairment or laryngectomy is described. For the first group of impaired speech data due to cerebral palsy or acquired hearing impairment, an acoustic model is adapted using a small amount of adaptation data. After the adaptation, word recognition is conducted to evaluate the adapted acoustic model for individual patients. For the other group of impaired speech data due to the laryngectomy, the same data and the same subject as in **Chapter 6** are used for the dictation task. Finally, the availability of speech recognition systems for those speaking-impaired people is evaluated.

Chapter 2

Laryngectomees and Conventional Researches

This chapter explains laryngectomees, laryngeal cancer that is the major cause of losing the vocal folds, alternative speaking methods, and conventional research related to this thesis. Development of medical techniques allowed doctors to find out about laryngeal problems early on; however, it is said that there are almost 600 thousand speaking-impaired patients due to the loss of the vocal folds throughout the world [17]. Many communication-aid systems including speaking-aid systems for laryngectomees have so far been developed. This chapter describes conventional approaches from the viewpoint of (1) developing a new artificial larynx, (2) enhancing alaryngeal speech with software approaches, (3) enhancing esophageal speech, (4) novel devices for speech communication, and (5) other applied research for VC and speech recognition.

2.1. Introduction

The larynx is an organ located at the position at where the trachea and the esophagus are separated [18]. The important role of the larynx is to prevent aspiration and to ensure the safety of the airway by guiding food and drink to the stomach through the esophagus and air to the lungs through the trachea. Although laryngeal cancer is comparatively easy to discover, the surgical operation to remove the larynx is often selected as the definitive procedure for curing laryngeal

cancer. The subjects, laryngectomees, lose their voices and they need to learn another speaking method to speak without vocal fold vibration. Laryngectomees can again obtain their voices by learning an alternative speaking method; however, the alaryngeal speech quality is often not satisfactory. Therefore, help is needed for laryngectomees to communicate.

This chapter is organized as follows. In **Section 2.2**, laryngeal cancer and anatomical features of laryngectomees are described. In **Section 2.3**, alternative speaking methods for laryngectomees are explained. In **Section 2.4**, conventional aid systems for speaking-impaired people and novel devices used in this thesis are described.

2.2. Laryngectomees

Larynx is an organ located at the position at where it separates the trachea from the esophagus [18]. The important role of the larynx is to prevent aspiration by guiding foods and drinks to the stomach through the esophagus and also guiding the air to the lungs through the trachea. Larynx includes vocal folds, which generate primary tone that is essential sound sources when speaking. Upper and lower organs than glottis are called supraglottis and subglottis, respectively.

Laryngeal cancer is the highest incidence among head and neck cancers although it is a kind of minor disease among all cancers [19, 20]. Laryngeal cancer is categorized according to the location of the tumor; the glottic cancer, the supraglottic cancer, and the subglottic cancer. Although the number of contracting the laryngeal cancer in the 70s was less than two thousand people in Japan, that in 1996 was coming up to almost three thousand people in Japan [21]. It is said that the number of laryngectomees were estimated less than 20 thousand people in more than 20 years ago [22], and the number of them would be more increased these days. Male patients are much more than females, and it is often developed at an advanced age [19]. Major problems causing the disease are smoking, continual drinking of much alcohol, and so on [23]. It is said that there are almost 600 thousands speaking-impaired patients due to the loss of vocal folds all over the world [17].

Although it is a terrible problem for us, early detection of the cancer is com-

paratively easier than other cancers because in the most cases, some troubles of the neck are observed by the output speech utterances [24]. In these days, the ways to cure the disease are becoming diverse according to the progress of the cancer [25][26][27][28]. Radiation therapy is effective, which has fascinating option of keeping the larynx and vocal folds especially in the early stages of laryngeal cancer. It is possible to cure the disease by the radiation therapy in the early stage; however some surgical procedures to directly remove the disease are introduced. There are mainly three types of surgical procedures, which are partial laryngectomy, total laryngectomy, and supracricoid laryngectomy with cricohyoidoepiglottopexy (SCL-CHEP). Partial laryngectomy partially removes the larynx including vocal folds to preserve the patients' voices. Although the partial laryngectomy sounds effective for speech communication and it has become popular from 60s to the end of 80s, it is no longer generally performed because of high possibility of reappearance of the disease and high frequency of aspiration. Total laryngectomy removes all surrounding areas including epiglottis, hyoid bone, arytenoid cartilage, cricoid cartilage, thyroid cartilage, and vocal folds, that is a default surgical procedure for laryngeal cancer in the last stage. SCL-CHEP is a novel surgical procedure, which preserves the patients' voices even though the vocal folds are removed. Many successful procedures have been reported, and it is highly expected for the cure of the laryngeal cancer. On the other hand, there are many patients who have been undergone the treatment of total laryngectomy, and the aid of them socially plays extremely important rolls.

Figure 2.1 shows anatomical images of non-laryngectomees and laryngectomees. Larynx works as a valve so that the trachea carries air and the esophagus does foods. The operation of total laryngectomy completely removes the larynx including vocal folds. To prevent foods flowing into lungs through trachea, total laryngectomees must choose which organ is connected to the mouth; the trachea or the esophagus. Most of laryngectomees connect their mouth to the esophagus. In that case, they have a hole called tracheostoma at the middle of their neck (see **Figure 2.1**) to breath. To keep the tracheostoma clean, it is covered by gauze, and certain constraints such as bath is concerned.

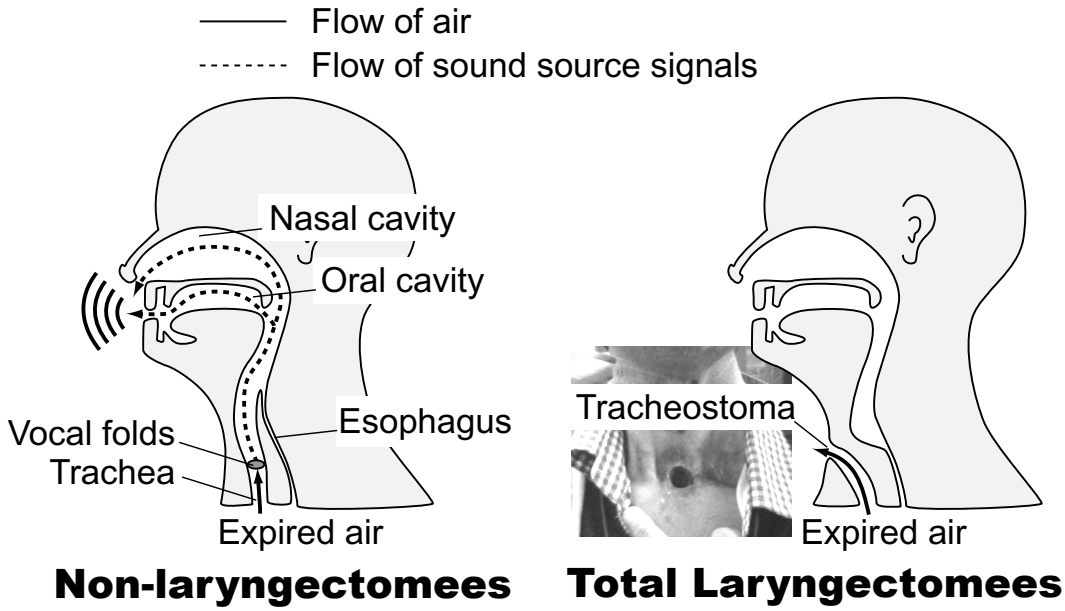

Figure 2.1. Anatomical image of non-laryngectomees and total laryngectomees.

2.3. Speaking Methods

Laryngectomees mainly have three kinds of alternative speaking methods that are different ways of obtaining the sound sources: 1) esophageal speaking method, 2) tracheo-esophageal (TE) shunt speaking method, and 3) a method using an external unit such as an EL or pneumatic artificial larynx [29] [30, 31]. **Figure 2.2** shows the route of floating air from the lungs to expiring. The benefits and defects of these methods are shown in **Table 2.1**.

1 Esophageal speech

The esophageal speaking method is conducted in the following procedure; taking air from the mouth to the beginning of the stomach, exploring the air to the mouth, and vibrating gelled gathers of the beginning of the esophagus to be the sound source vibration. It is said that the esophageal speech is natural compared to other alternative speaking methods because this methods generates the sound source signals in their body. There are many supporting society for esophageal speech in Japan. As the result, the esophageal speaking method is

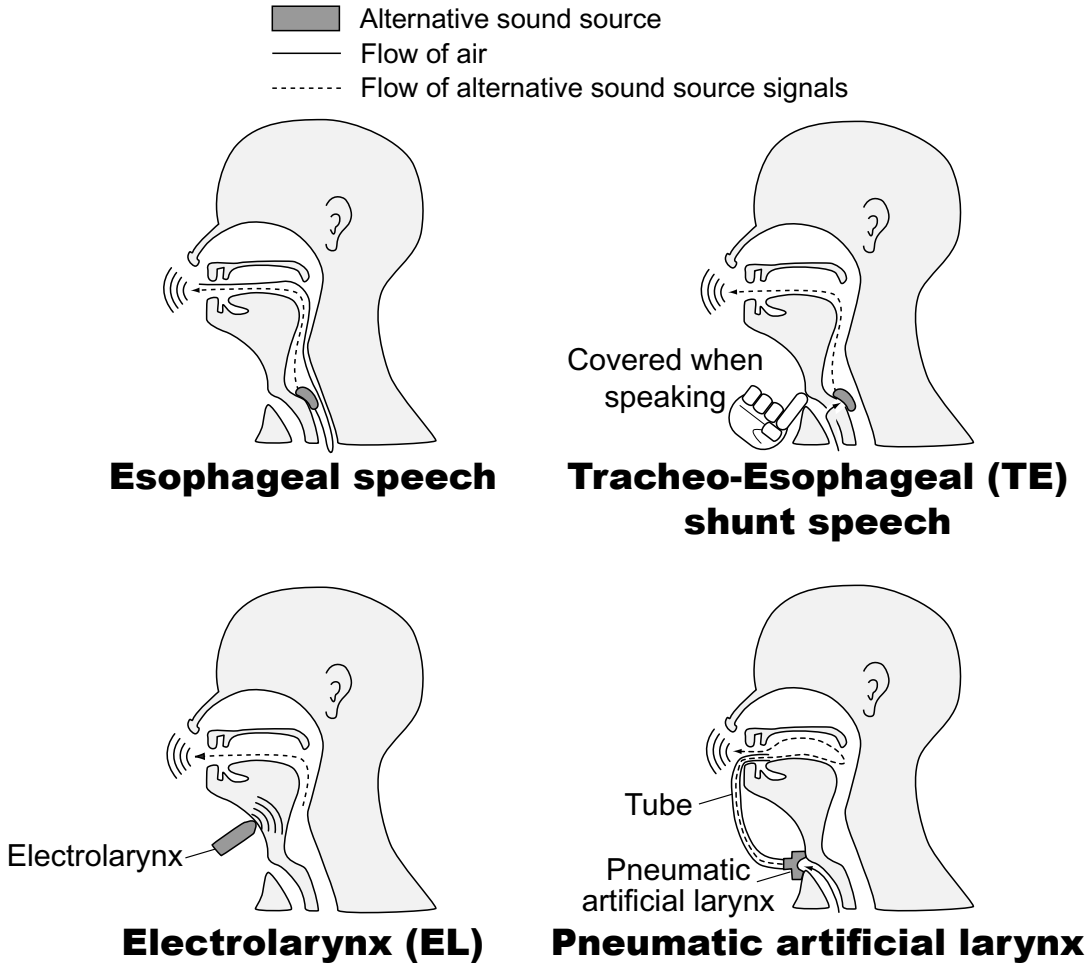

Figure 2.2. Major alternative speaking methods for laryngectomees.

the most likely to use in Japan. On the other hand, the esophageal speaking methods requiring strength for the speakers, and therefore, some aged people are difficult to speak with the esophageal speaking methods to use another method such as an EL.

2 TE shunt speech

To speak with the TE shunt speaking method, a voice prosthesis [32] that is a valve with only one direction is inserted between the trachea and esophagus. In speaking, laryngectomees block the tracheostoma to make the air go to the esophagus through the prosthesis. The air vibrates gelled gathers of the beginning

Table 2.1. Benefits and defects of alternative speaking methods for total laryngectomees

Alaryngeal speech	Naturalness	Difficulty	Popularity (in Japan)
Esophageal speech	Good	Bad	Popular
TE shunt speech	Good	Good	Not yet
EL speech	Awful	Good	Popular
Pneumatic artificial laryngeal speech	Better	Good	Popular in past days

of the esophagus to be the sound source vibration just like the esophageal speech. It is easier to produce the TE shunt speech than the esophageal speech with almost the same quality. It is said, on the other hand, the operation to insert the voice prosthesis is difficult. As the result, it is not popular method in Japan.

3 Pneumatic artificial larynx

One major external speaking device is pneumatic artificial larynx. Pneumatic artificial larynx is used by pushing the vibrator to the tracheostoma and by holding the whistle in the mouth. The fundamental frequency (F_0) is manipulated by the expired air flowed from the tracheostoma, and moreover, the vibration is once taken into the mouth to be articulated so that the voice humanity is added. As the result, pneumatic artificial larynx enables laryngectomees to speak with natural speech compared to an EL. An interesting pneumatic artificial larynx was developed [33]; however, this device is less used in these days even though it seems useful because both of the speaker's hands are used to produce the alaryngeal speech, the visual is not acceptable, and speakers have a sanitary concern about whistle.

4 Electrolarynx

The other major external medical device is an EL. The basic structure of an EL is shown in **Figure 2.3**. An EL is pushed on the lower jaw in speaking, and the on/off is switched by the button. One of the advantages of the EL is

its short learning period. The other advantage is that laryngectomees with less physical power also can use the device. One biggest defect of the EL is its fixed fundamental frequency as the **Figure 2.4** shows deriving artificial and mechanical unnatural speech even though human speaks. Some problems about EL including unnatural speech have been represented so far. The author had tried to speak using only an EL for 21 days to find out which problems should be solved. As the result, this research focuses on two problems; 1) the naturalness of the EL speech and 2) the large sound of the EL itself. The current ELs have to output with large power of the sound source signals because it is assumed that ELs are used for normal speech conversation. It is reasonable; on the other hand, the sound source signals are noisy for people around the speaker especially in the quiet situation such as library. Moreover, the speaker might have a concern that he or she would annoy for other people because of the noisy sound source signals. These defects prevent smooth inter-personal speech communication for laryngectomees.

2.4. Conventional Speaking-Aid Systems for Laryngectomees

Most of conventional studies have been conducted by addressing the unnaturalness of EL speech in the physiological level in the speech chain shown in **Figure 1.1**. Namely, those studies have aimed to enable the laryngectomees to directly speak with natural voice. There are many aid devices and procedures. This section describes those approaches from the view point as followings. **Subsection 2.4.1** describes about the development of a new artificial larynx to generate alternative sound source signals. **Subsection 2.4.2** describes approaches to enhancing EL speech using signal processing procedures. **Subsection 2.4.3** describes other studies to enhance esophageal speech. **Subsection 2.4.4** describes other devices that would be useful for not only laryngectomees but also other people including speaking-impaired and elderly people. Other applied researches for VC and for automatic speech recognition (ASR) are described in **Subsection 2.4.5**.

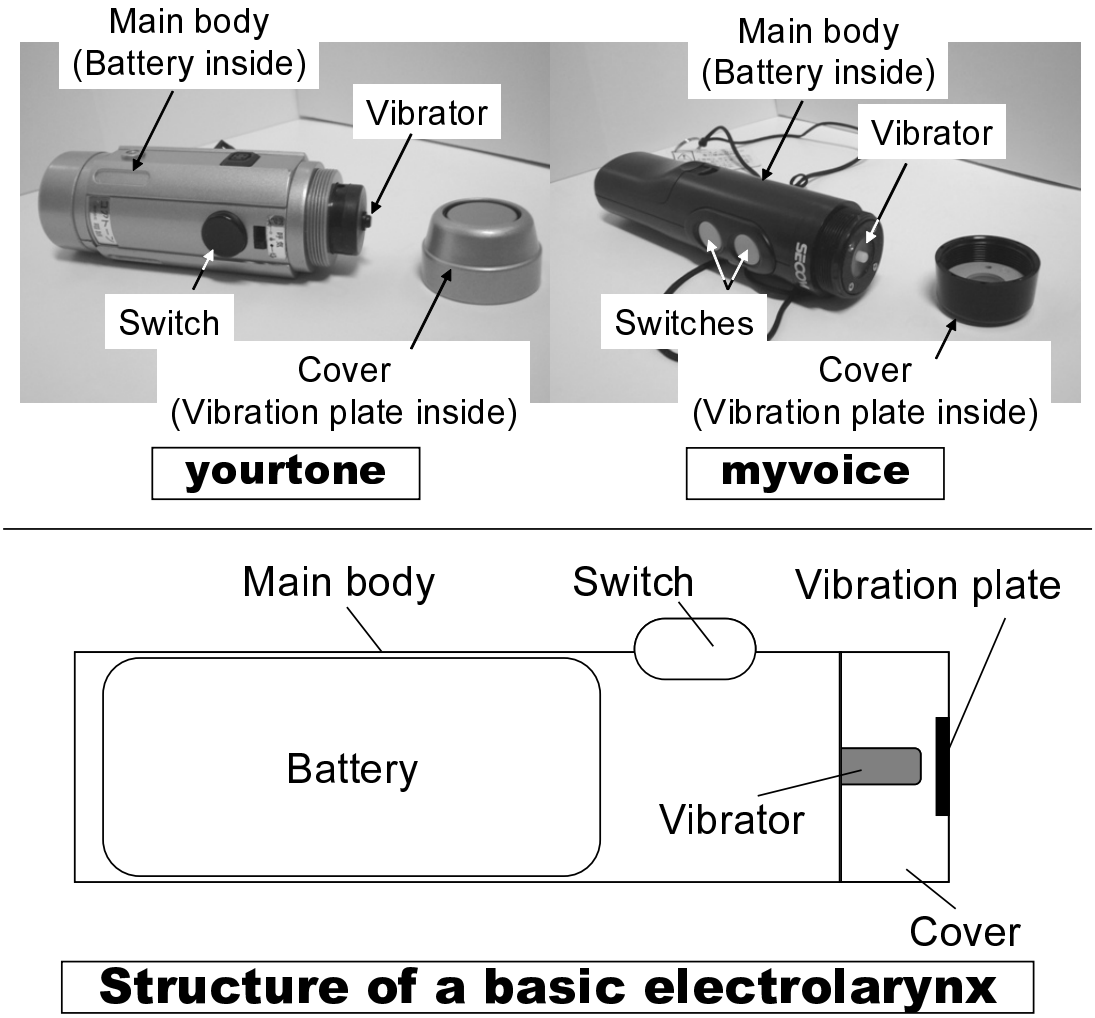

Figure 2.3. Basic structure and examples of existing electrolarynxes.

2.4.1 Developing a new artificial larynx

One EL named 'yourtone' is developed in Japan, which considers acoustic fluctuations of vowels of our normal speech to enable laryngectomees to speak with more natural voice even though it outputs only fixed F_0 [7] [34] [35]. All ELs before developing 'yourtone' was made in abroad, and therefore, technical supports and other advisements were significantly poor for users. The basic idea of 'yourtone' is to produce an EL in Japan to carefully support users of laryngectomees

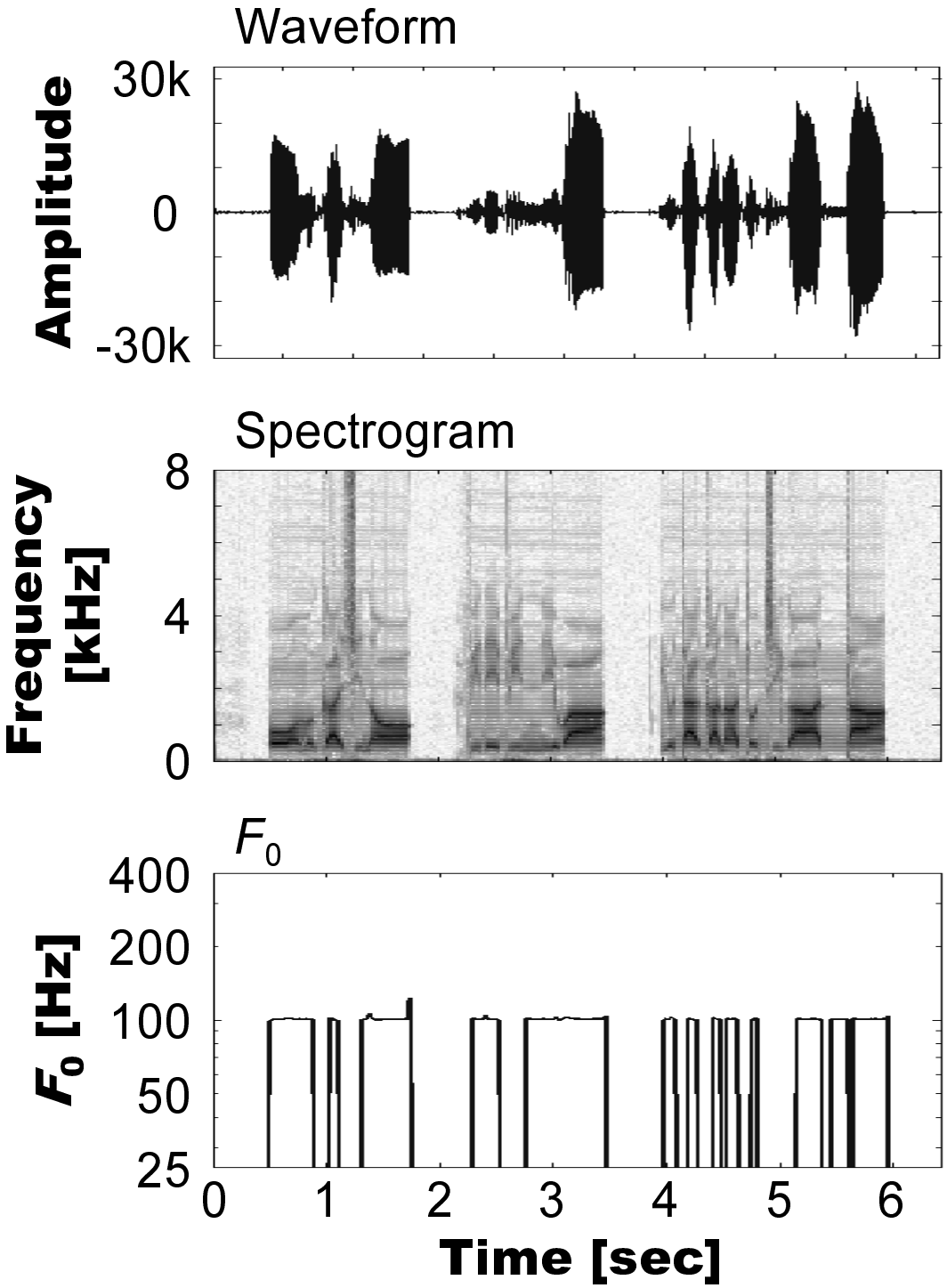

Figure 2.4. Example of EL speech produced by laryngectomee who is proficient to produce EL speech.

in Japan. Moreover, 'yourtone' had tried to enable laryngectomees to speak with more natural speech even though the generated EL speech only has monotone pitch. In the first stage to develop 'yourtone', developers first had analyzed human voices to confirm the impact for the naturalness of human voices affected by sensitive variations appeared in outline shapes of waveforms compared to another sensitive variations caused by durations [34]. As the result, they had found that at least pitch waveforms for 32 cycles in which each cycle is normalized are necessary to enhance the naturalness of EL speech. They confirmed the effectiveness of the acoustic variations using a prototype of pipe-inserted artificial larynx. In the second step of the development the 'yourtone', a novel air-pressure sensor is developed to enable laryngectomees to control the intonations using their breath flowed from the tracheostoma [35]. A recording scene of the EL speech using the air-pressure sensor (EL(air) speech) is shown in **Figure 2.5**, and an example of waveforms, spectrogram, and extracted F_0 contours is shown in **Figure 2.6**. As the figure shows, more rich F_0 contours are obtained compared to the conventional EL generating monotone pitch. The naturalness using EL(air) speech is much higher than conventional type of EL speech. On the other hand, the convenience of speaking using the external device is reduced since users needs both hands to hold the main body of the EL and the air-pressure sensor. More than one thousand of 'yourtone' is used and it is preferred by buyers. This is the first EL made in Japan, and its effectiveness is experimentally and practically confirmed. Therefore, this thesis mainly uses this EL.

Another EL named 'myvoice' is also developed by SECOM Corporation in Japan [36]. 'myvoice' provides users natural voice by gently changing F_0 that once rising and going down as the time goes. One advantage of this EL is that the F_0 patterns are included in the EL in advance; therefore, the action required to the user is to just put the button of the EL to turn on or off the vibrator. Using this EL, laryngectomees can speaks with more natural speech than using EL generating monotone pitch even the action in speaking is equivalent to previous devices. This EL might be regarded as an abridged edition of EL(air). It seems effective; however, the essential problem of EL is not addressed by 'myvoice', since the F_0 pattern is decided and users cannot manipulate it. Moreover, the basic structure of the EL is the same as others, namely the included vibrator vibrates

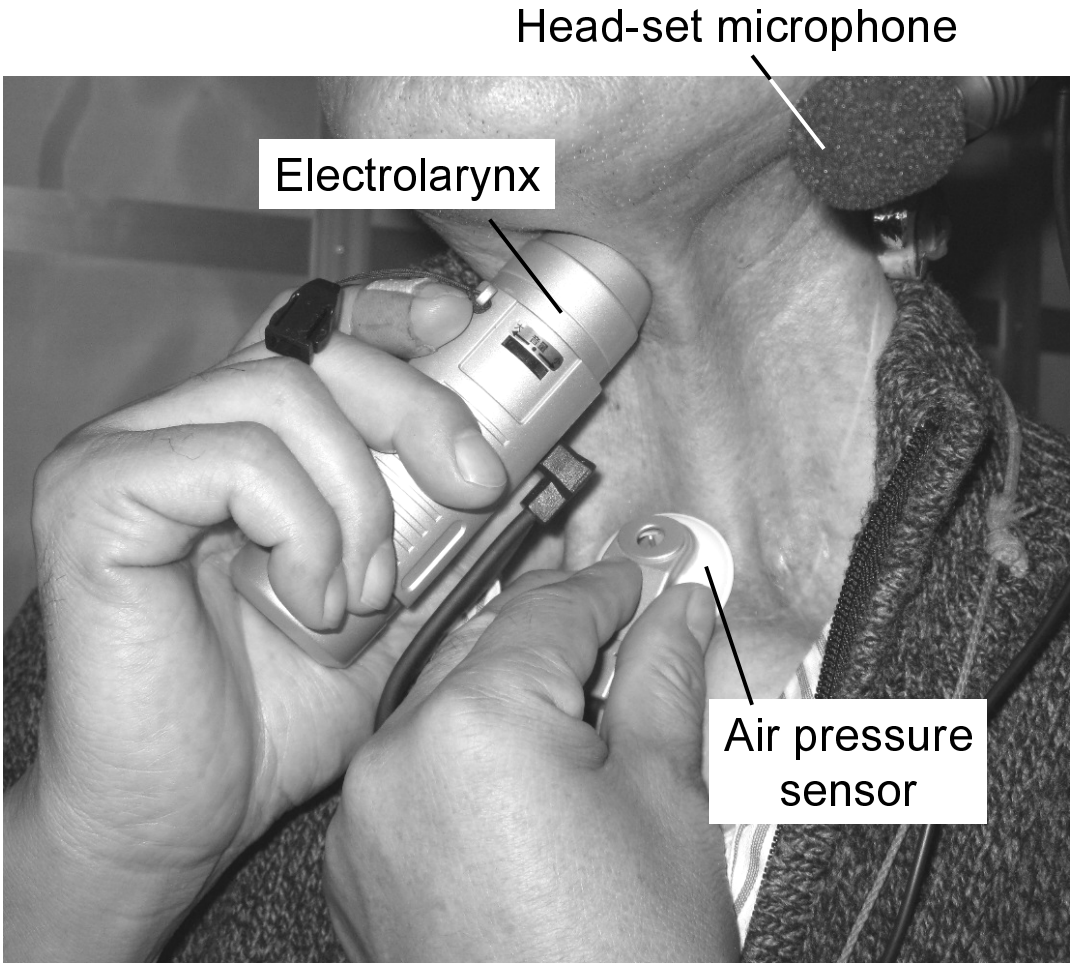

Figure 2.5. Recording scene of EL speech using air-pressure sensor.

to hit the vibration plate in generating the sound source signals, which mechanism is also the same as 'yourtone' even if it uses air-pressure sensor. Therefore, this thesis thinks that it would be able to obtain natural speech with the equivalent quality of normal speech only by enhancing the EL speech out of the basic structure.

Takahashi *et al.* developed an interesting voice generation system using an intramouth vibrator [37]. The vibrator is embedded in artificial teeth to be fixed in the oral cavity. This device had begun developing from wired-type, and a prototype of the wireless vibrator is also developed. The F_0 generation

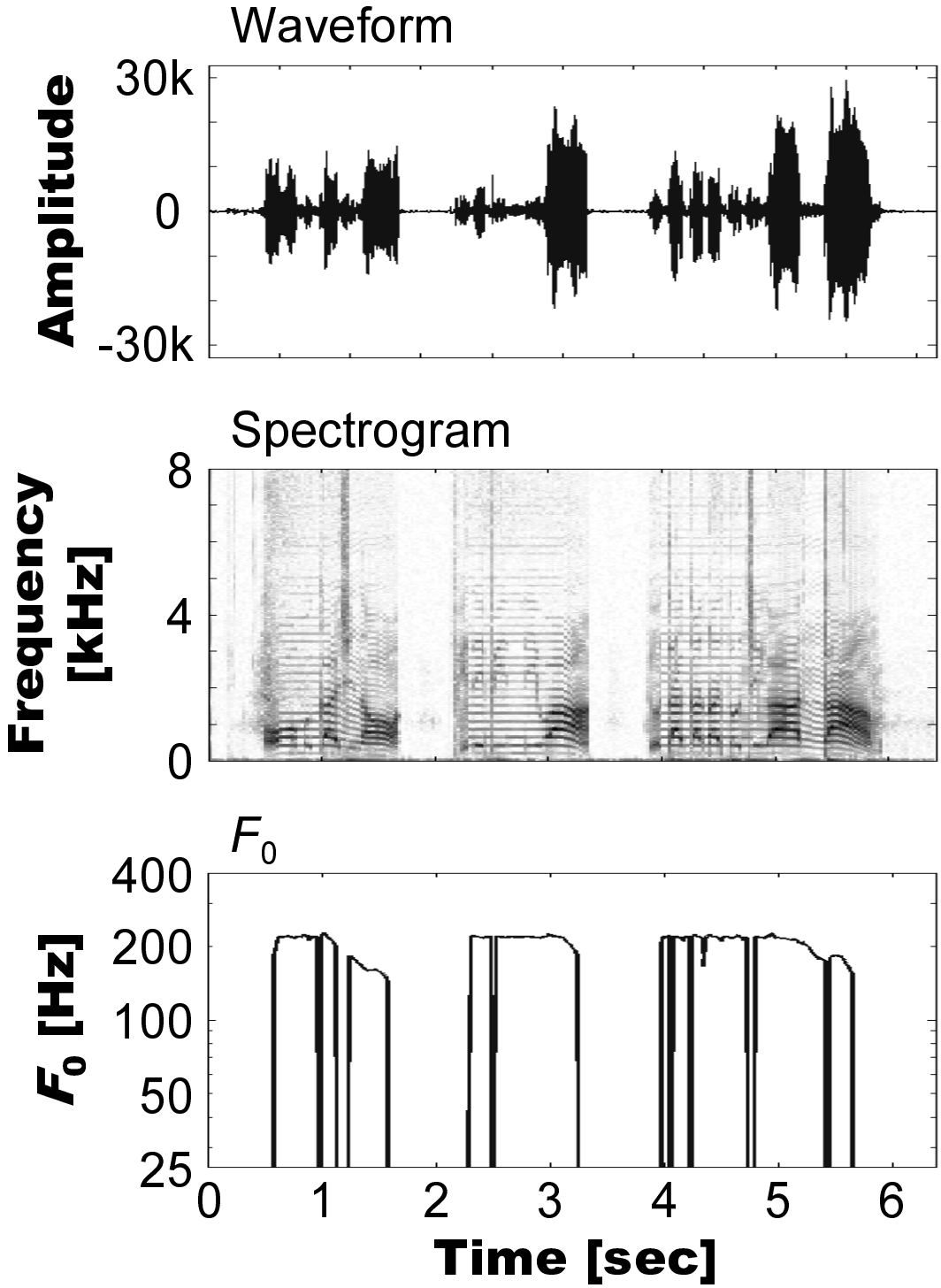

Figure 2.6. Example of waveforms, spectrograms, and F_0 contours for EL speech using air-pressure sensor produced by laryngectomee.

model proposed by Fujisaki *et al.* [38] is employed in this device. Although it is an interesting system, it is concerned that the user can use this system in any situations and that the physical conditions of the user due to sound source generation are no problem. Moreover, not all users available this system since this device is used by embedding the sound source unit in their artificial teeth.

2.4.2 Enhancing alaryngeal speech from software approaches

Murakami *et al.* proposed a speech transformation method from EL speech, which is source speech, into normal speech, which is target speech, by applying transformation rules to the input EL speech [39]. This method consists of two parts of acquiring conversion rules from the training data and applying those rules to the test data. In the step of acquiring conversion rules, the following procedures are conducted utterance by utterance. First, the same utterance pairs of the source EL speech and the target normal speech are prepared, and then, a time-sequence of spectral parameters are extracted from those two speech data. Next, dynamic programming (DP) procedure is independently conducted to both EL and normal speech to split the whole utterance to acoustic difference or common segments based on local directions in the DP. After that, the segments decided as the common parts, namely the segments in which the local direction on the path of DP matching is constantly less than a threshold set in advance, is registered to a dictionary as the transformation rule. Here, if multiple segments are available, only one segment which has the highest confidence measure is registered, which is calculated from the start point, end point, and the local length of the common segment part. In the other step of applying the rules, first, a time sequence of spectral parameters are extracted from the input speech as the same method as the acquiring step. Next, all conversion ruled registered in the conversion dictionary is compared with the input parameters. Here, in order to compare the data, partial patterns are extracted from input parameters by conducting word spotting by continuous DP procedures [40] since the length of the speech segment registered in the conversion dictionary is shorter than that of the input EL speech segment. After the comparison, conversion rules are selected if the input data and registered segments are match more than certain frames. One final conversion rule is selected from the selected conversion rules, and then, the input speech

segment is converted to normal speech using the selected rule. Finally, converted normal speech is output by waveform concatenative speech synthesis. This study is interesting, since the concept of this method is equivalent to that of this thesis from the view point of enhancing the EL speech quality for the produced speech while keeping its linguistic information. Although the effectiveness of this method is experimentally evaluated, the problem of unacceptable samples exists if nothing rules are matched to the input data in conversion applying procedures remains.

Norton *et al.* [41] analyzed the effect of sound-shielding to suppress the radiated noise, and found some improvement. However, later experiments found that the sound-shielding makes the size of the EL increasing, and it makes it inconvenient to hold the EL. The failure of physical approach to suppressing its radiated noises has led researchers to consider the use of signal processing techniques. Spectral subtraction (SS) technique [42] is employed to reduce the additive noise of the direct path noise of the EL [9][10] which is based on the assumption that speech signals and the additive noise is uncorrelated. SS for EL speech provided an improvement the EL speech; however, the subtraction parameters used for the EL speech enhancement are fixed and cannot be adapted frame-by-frame.

Liu *et al.* updated the idea of conventional SS for the EL speech [43]. They introduced auditory masking properties in the enhancement process of EL speech, in which perceptual weighting technique is applied to adapt the subtraction parameters. Auditory masking properties take into account the frequency-domain masking properties of the human auditory system. Frequency masking experiments [44] showed that noise near formant peaks is inaudible to the human ear, where the speech signal has high energy. Using this property, the auditory masking property adapts the subtraction parameters to more subtract near the formant peaks. Subtraction parameters, on the other hand, is set to make the distortions between before and after subtraction lower in order to suppress the generation of musical noises [45] due to the over-subtraction. As the result of this property, the weighting filter for the subtraction parameters is drawn so that it has almost inverse peaks and valleys compared to the spectral envelope in a frame. Subtraction parameters are adapted based on the perceptual weighting filters, and weighted estimated noise parameters are subtracted from the input observation signals. As the result of subjective experimental evaluation, the en-

hanced EL speech by their method is more pleased than that by the conventional PSS method especially in the case of the additive noises such as white Gaussian noise and speech babble noise. Finally, based on the frame-by-frame adaptation of the subtraction parameters, the enhanced EL speech in [43] realized the tradeoff between reducing noise and increasing intelligibility, and achieved the tradeoff between keeping the residual noises and the distortion acceptable to a human listener. The enhancing method in [43] was intended to reduce additive noises including radiated noise that is directly observed from the EL itself during phonation. Their intention to reduce the radiated noise is just the same as one problem focusing in this thesis, and it is interesting.

2.4.3 Speech enhancement for esophageal speech signals

In Japan, there are many laryngectomees who selects esophageal speech as their alaryngeal speech. Proficients of the esophageal speech speak with comparatively natural voice than beginners, and they have reintegrated into society, e.g. getting a job. On the other hand, not all the esophageal speakers can be proficients. It is said that the esophageal speech is further different from normal speech than TE shunt speech [46] and it is meaningful for both laryngectomees and non-laryngectomees to enhance the esophageal speech by signal processing procedures. The loudness of the esophageal speech is almost the same as that of the normal speech. In other words, when esophageal speech signals are enhanced to be output, listeners would listen to not only processed speech but also to produced esophageal speech. The enhancement of the esophageal speech might be difficult to be used in conversations with face-to-face. On the other hand, it would be useful in situations which laryngectomees have to communicate with others using only their voices. For example, in the situation of telecommunication, listeners must understand what the laryngectomees said from only the transmitted speech. Therefore, the transmitted speech plays extremely important rolls to make smooth conversation successful in such situations, and the enhancement for the esophageal speech would be effective to conduct more smooth speech communication.

Hisada *et al.* had introduced a comb filter to clarify esophageal speech while keeping the speaker's individuality preserved. This study is intended to the real-

time clarification that is essential problem for the clarification to be used in out daily lives. Experimentally, this filtering method is confirmed to be effective even though some problems still remain such as echo effects and electronic impressions caused by the misdetection of the pitch.

Another study also aims to enhance esophageal speech by analysis-synthesis approach [47, 48]. In this study, the input esophageal speech is firstly divided into two channels of lower and higher frequency components (at 2.5 [kHz] in the study [48]). Only the lower frequency channel is used for the analysis-synthesis procedures. The higher frequency channel is mixed at the final stage (after synthesizing the lower channel). This filtering has two major effects. One is that blending the higher frequency component would give more natural sounding and intelligible consonants. The other is that errors of voiced or unvoiced decisions using only lower channel would decrease. Voiced or unvoiced speech frames are decided using the power level of the lower channel. The unvoiced frames are not processed. Only voiced frames are analyzed in which linear prediction coding (LPC) analysis is applied to extract formant information, and are synthesized [49]. These algorithms are implemented on a DSP system [48]. This analysis-synthesis method for the esophageal speech is rated by speech therapists.

Another study of enhancing the esophageal speech using the same conversion technique as used in this thesis has been performed [50]. This enhancement approach tries to essentially modify the acoustic features of the esophageal speech by describing joint acoustic features of esophageal and normal speech with GMMs. This study is focused on the converted voice quality than the processing speed. As the result of objective evaluation, the correlation between resulting F_0 contours and the target ones are almost 0.7 with voiced or unvoiced errors was 8.36 [%] even though the correlation between F_0 contours of before processing and those of normal speech was 0.12. From the subjective evaluation, the naturalness of the modified esophageal speech is highly scored than that of the source esophageal speech while keeping the intelligibility almost the same scores.

2.4.4 Novel devices for speech communication

Hanada *et al.* developed an alternative speaking system using PDA to support Japanese daily conversations for speaking-impaired people [51]. Their aid sys-

tem (Voice Output Communication Aid: VOCA in the paper) accepts arbitrary input texts and outputs synthesized speech as their voice. In order to use VOCA, users make an input text by (1) selecting a conversational sentence registered in advance or (2) selecting Japanese characters using a pen. Next, input text is split into pairs of consonant-vowel (CV). Next, CV pairs corresponding to the split CV pairs from the input text are searched from a dictionary. Next, vocal tract parameters (coefficients of linear predictive coding in the paper [51]) on the basis of F_0 generation model that is derived from physiology model of controlling the larynx and vocal source parameters calculated from a glottal waveform model are obtained. Finally, synthesized speech is output by convolving these parameters. The size of the dictionary used in the paper [51] is 1.2 [MB]. VOCA system has important advantages that are results of carefully consideration. For example, the PDA is small (83.5 [mm] wide, 130 [mm] long, 15.9 [mm] thick) and light (185 [g] including batteries) enough to carry it out, and the speech synthesis is conducted within three seconds. Moreover, VOCA is established through a lot of trial and error processes such as employing histories of past input texts to make the input time shorter, allocating highly frequent texts to certain buttons of the PDA, and so on. VOCA is expected to be a greatly effective tool for speaking-impaired people as an alternative speaking system. On the other hand, this thesis thinks that it is natural to conduct speech communication by humans themselves. This thesis intends to establish a speaking-aid system using speech of the user oneself.

Hosoi *et al.* had founded a novel device that generates extremely small power so that it is almost too difficult for people around the speaker to capture the signals [52]. This sound source unit is expected to give people silent communication without annoying any other people while speaking. Actually, people would hear the small sound source signals in silent spaces such as a fully anechoic room; however, it is rare to find such quiet situations in our daily lives. There are almost background noises such as air-conditioner, machine noises and so on even in quiet scenes such as library. As the result, the small-powered sound source signals are almost too difficult for others to be captured.

Nakajima *et al.* had defined Non Audible Murmur (NAM) as articulated respiratory sound without vocal-fold vibration transmitted through the soft tissues

of the head [14]. NAM is acoustically seen as a whispering with extremely small power so that the people around the speaker cannot hear the sounds. NAM is too small-powered to be effectively detected by a usual air-conductive microphones such as head-set microphones or pin-type microphones. NAM microphone is simultaneously developed with NAM to detect NAM signals [13]. **Figure 2.7** shows the attaching location and basic structure of NAM microphone, and **Figure 2.8** shows an example of waveforms and spectrograms of normal speech recorded with a head-set microphone and NAM recorded with the NAM microphone, in which first and second formants are automatically extracted. It is characteristic about its materials. Soft silicone is used as the attaching area, where is safe for human body. The acoustic impedance of the soft silicone is almost the same acoustic impedance as that of muscles of the attaching location so that it can detect the resonance in the oral cavity. The sensor is wrapped by the soft silicone and the soft silicone is wrapped by a noise proof. Resin is introduced as a material for noise proof, which is hard and safe for human body. NAM microphone is also characteristic about its attaching location. Nakajima found the better location to detect the resonance, where is on the sternocleidomastoid behind the ear as shown in **Figure 2.7**. Although the effectiveness of the location has not scientifically revealed, it is concerned to be effective because the microphone detects the resonance occurred in the oral cavity without been prevented by any bones. NAM microphone detects the vibration of the muscle. The vibration data detected by NAM microphone is regarded as speech as a matter of convenience, although some people might disagreement to call the detected data 'speech' because it is not the speech waveform propagated through the air.

Figure 2.8 shows an example of speech signals recorded with a NAM microphone. Higher frequency components over 4000 Hz of those speech signals are almost not observed because of the loss of radiation features and low-pass filtering that is thought to be occurred when the resonances are passed through the muscle. This characteristic is observed for not only NAM signals but also other speech signals recorded with a NAM microphone. Acoustic features in higher frequency components are lost; however, NAM can automatically be recognized [14] and it is expected to be a novel sensor for speech communication.

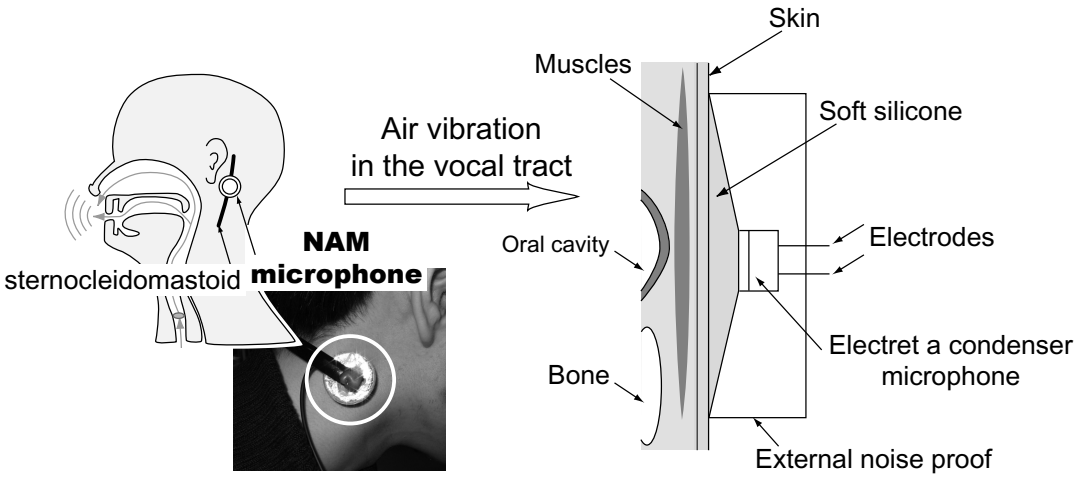

Figure 2.7. Attaching location and basic structure of NAM microphone.

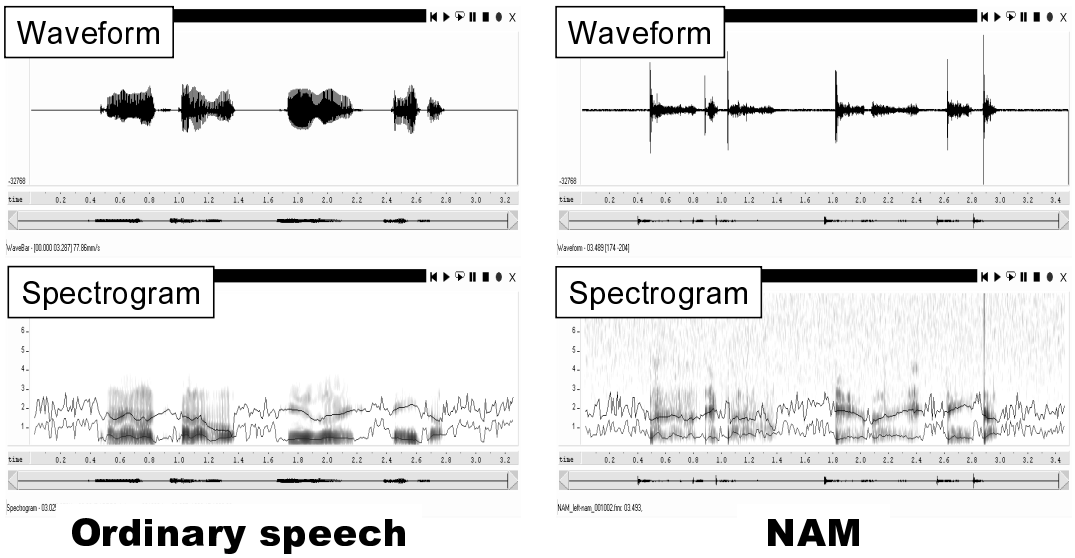

Figure 2.8. Examples of waveforms and spectrograms of normal speech and NAM recorded with NAM microphone. Solid lines in the spectrograms show first and second formants extracted automatically.

2.4.5 Applied research for VC and speech recognition

Detected NAM using a NAM microphone is less intelligible, and therefore, two approaches to enhancing NAM are proposed. One is to convert NAM to normal speech using statistical VC [15], and the other is to convert NAM to whispering using the same statistical VC [16]. NAM does not have effective F_0 information. Therefore, in the conversion from NAM to normal speech, all acoustic features including spectral, F_0, and aperiodic components are estimated from only the spectral information of NAM. Moreover, in order to capture not only statistical and dynamic features, segmental features including the data of several frames are used for the source data. As the result of experimental evaluations, the correlation of F_0 contours between converted and target speech has been over 0.5 [16], and the converted normal speech is more natural than NAM [15].

Tran *et al.* had tried to address the difficulty of estimating natural F_0 contours [53] in the conversion framework from whispering to normal speech. In this study, linear discriminative analysis is introduced to reduce the dimension of feature vectors, although principal component analysis (PCA) is used in the conventional study [15]. Moreover, long frame intervals are considered to capture suprasegmental features. In order to improve the estimating performance from whispering, visual data are additionally used in this study. Target normal speech is estimated using both audio and visual information using Gaussian mixture models (GMMs). They have also proposed hidden Markov model (HMM)-based audio-visual conversion system [54]. In the training procedure of their system, the joint probability densities of source and target parameters and duration probability distribution are modeled by context-dependent phone-sized HMM using aligned training utterances. In the synthesizing procedures, HMM-based recognition is firstly performed using the source stream consisted of audio and visual data to determine phone sequence. And then, HMM-based speech synthesis is conducted using HMM-based speech synthesis system [55, 56]. This system is also expected to help people including impaired-people to communicate with each others.

Although the VC from NAM to normal speech is effective to enhance NAM, it is difficult to estimate natural F_0 contours from only the source spectral information. Japanese is a kind of tone language, and therefore, the estimation of natural

F_0 contours plays important roll in the speech enhancement. In order to avoid this problem, the other approach to enhancing NAM is proposed, in which NAM is converted into whispering that is a hoarse unvoiced speech. Since whispering does not have F_0 information, VC only needs to convert spectral information. As the result of experimental evaluations, converted whispering is dramatically preferred to converted normal speech derived from NAM.

Another interesting research to generate speech waveforms from gestures of the user's hand have been proposed by Kunikoshi *et al.* [57]. In this study, Japanese five vowels have so far been focused to be generated, and it has so far been investigated that features of hand gestures detected using a special glove are able to be converted to corresponding acoustic features of vowels. This study is expected to be a novel speaking-aid interface that is effective for speaking-impaired people such as dysarthria.

Kain *et al.* had intended to enhance the intelligibility of conversational speech produced by any speakers so that hearing-impaired people are easy to understand the contents [58]. It would be an assistive system for hearing-aid people in their speech communication. They have also tried to enhance the intelligibility of speech produced by speaker-impaired people with dysarthria, who have problems in articulation, for example, ataxic, flaccid, and hyperkinetic in their paper, also using conversion techniques. In this case, they had not achieved to the significant improvement of the intelligibility. There are so many speaking- and language-impaired people (in Japan, there are 40 thousand people including speech, language, and mastication impaired [59]). Moreover, the acoustic feature space of the speaking-impaired people would be strongly deviant from that of the normal speaker. These problems make the resolution difficult.

ASR is a technique to extract linguistic information from input speech data. Speech can be recognized with more than 90 % accuracy in the framework of large vocabulary continuous speech recognition (LVCSR), which is the practical system [60]. On the other hand, ASR for speaking-impaired people, especially for dysarthria who have problems in articulation, is dramatically difficult problem compared to normal speech. In order to establish ASR system for speaking-impaired people, generally, several constraints are employed such as the speaker-dependent system using speaker adaptation technique, comparatively small dic-

tionary, and the combination of these constraints.

Matsumasa *et al.* had tested the ASR performance for dysarthria (especially cerebral palsy) in the task of controlling home electronics [61]. In their later work, they have proposed a robust feature extraction method for ASR systems of speaker-impaired people [62]. ASR is also used in a speech training system [63]. Moreover, another aid system using ASR system for physically impaired person is proposed [64]. These studies would be great help for impaired people.

2.5. Summary

This chapter described laryngeal cancer and the way to cure the disease. Moreover, differences between laryngectomees and non-laryngectomees were anatomically described.

Conventional studies for not only laryngectomees but also other speech-impaired people were also described. Moreover, novel devices such as the NAM microphone and small sound source signals were introduced.

Chapter 3

Statistical Voice Conversion

This chapter describes the statistical VC technique used in this thesis. The conversion approach includes the training and conversion parts. In the training part, a GMM is trained using the training data that consists of joint feature vectors of source and target feature vectors. In the conversion part, the conditional probability density function given the source feature vector is output as the estimated target feature. One important feature for improving the naturalness of converted speech is dynamic features. The other important feature described in this thesis is global variance (GV), which is a global variance of acoustic features over a time sequence.

3.1. Introduction

The primary role of speech is to convey the linguistic information; however, secondary information that speech conveys such as speaker individuality also plays an important role in inter-personal speech communication. The VC technique modifies speech signals of a given source speaker so that another speaker speaks while maintaining its linguistic information. This technique is useful for many applications such as voice responses and text reader systems. It is often convenient to specify the desired modification of the acoustic characteristics with reference to an existing speaker (so-called target speaker).

This chapter is organized as follows. In **Section 3.2**, the basic framework of VC used in this thesis is explained. In **Section 3.3**, novel statistic features of

dynamic features and GV are introduced. Finally, this chapter is summarized in **Section 3.4**.

3.2. Voice Conversion Using Joint Probability Density

The global framework of the VC method using GMMs is shown in **Figure 3.1** [65]. This method includes training and conversion part. In the training part, the time sequence of the source and the target training data are automatically aligned by dynamic time warping (DTW) procedure in advance. Then, the joint probability density functions of the source and the target data are modeled by a GMM. Finally, the conditional probability density function of the target data given the source data is to be the converted target data.

In the training procedure, let $\boldsymbol{x}_t = [x_t(1), \cdots, x_t(d_x)]^\top$ and $\boldsymbol{y}_t = [y_t(1), \cdots, y_t(d_y)]^\top$ be a static input and output feature vector at frame t, respectively, where d_x and d_y denote the dimensions of $\boldsymbol{x}_t$ and $\boldsymbol{y}_t$, respectively, and $\top$ denotes transposition. The joint probability density of the source and the target feature vector $\boldsymbol{z}_t = [\boldsymbol{x}_t^\top, \boldsymbol{y}_t^\top]^\top$ is described by a GMM as follows:

$$P(\boldsymbol{z}_t|\lambda^{(z)}) = \sum_{m=1}^{M} \omega_m \mathcal{N}\left(\boldsymbol{z}_t; \boldsymbol{\mu}_m^{(z)}, \boldsymbol{\Sigma}_m^{(z)}\right), \tag{3.1}$$

where m is an index of an mixture component, M is a number of mixture component, ω_m is a weight parameter of mth mixture component, and $\mathcal{N}(\cdot; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ represents a Gaussian distribution including a mean vector $\boldsymbol{\mu}$ and a covariance matrix $\boldsymbol{\Sigma}$. $\lambda^{(z)}$ is a model parameter set including weights, mean vectors, and covariance matrices. The mth mean vector and the covariance matrix is written as

$$\boldsymbol{\mu}_m^{(z)} = \begin{bmatrix} \boldsymbol{\mu}_m^{(x)} \\ \boldsymbol{\mu}_m^{(y)} \end{bmatrix}, \qquad \boldsymbol{\Sigma}_m^{(z)} = \begin{bmatrix} \boldsymbol{\Sigma}_m^{(xx)} & \boldsymbol{\Sigma}_m^{(xy)} \\ \boldsymbol{\Sigma}_m^{(yx)} & \boldsymbol{\Sigma}_m^{(yy)} \end{bmatrix}, \tag{3.2}$$

where $\boldsymbol{\mu}_m^{(x)}$ and $\boldsymbol{\mu}_m^{(y)}$ represent the mean vectors of the mth mixture component for the source and the target features, respectively. $\boldsymbol{\Sigma}_m^{(xx)}$ and $\boldsymbol{\Sigma}_m^{(yy)}$ represent the covariance matrices of the mth mixture component for the source and the target

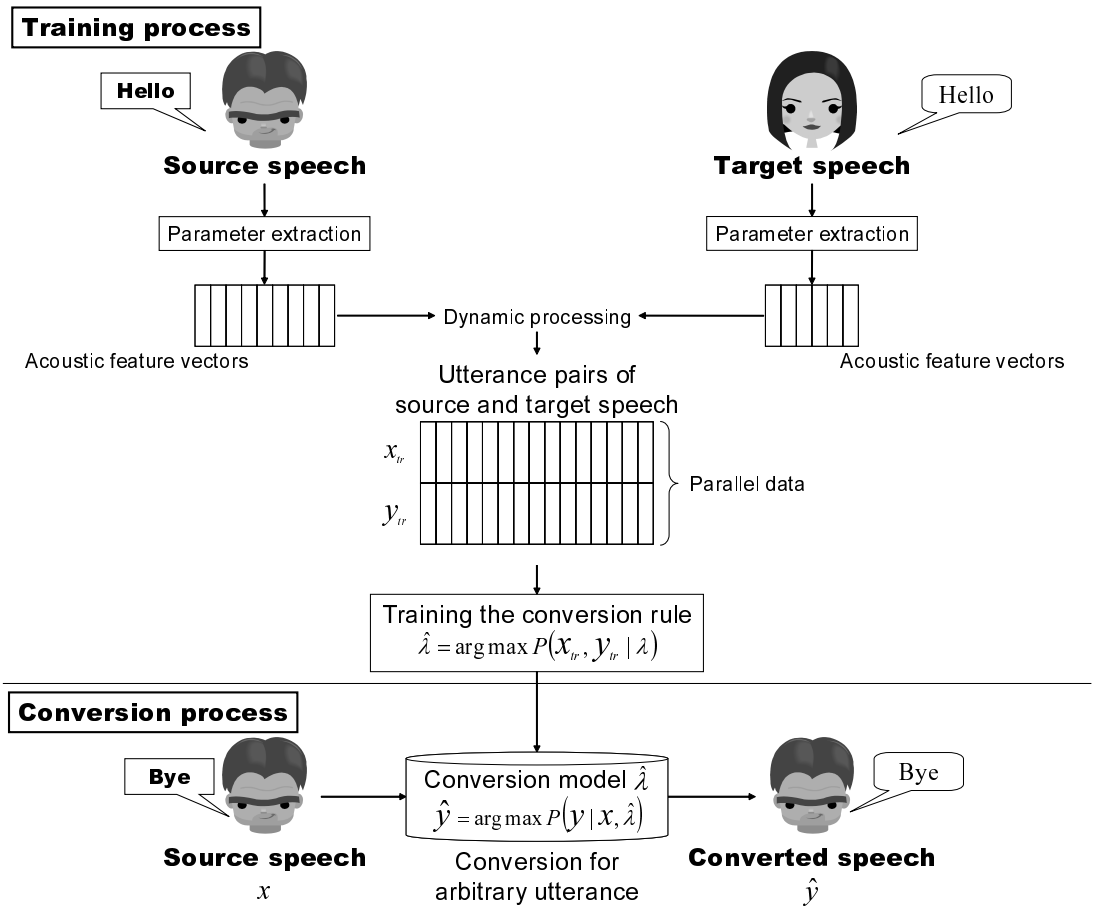

Figure 3.1. Overview of statistical VC procedures using Gaussian mixture model in maximum likelihood manner.

features, respectively. $\mathbf{\Sigma}_m^{(xy)}$ and $\mathbf{\Sigma}_m^{(yx)}$ represent the cross-covariance matrices of the mth mixture component for the source and the target features, respectively. The model parameters are estimated by expectation-maximization (EM) algorithm [66].

In the conversion procedure, the conditional probability density function of $\boldsymbol{y}_t$ given $\boldsymbol{x}_t$ is also described by a GMM as follows:

$$P(\boldsymbol{y}_t|\boldsymbol{x}_t, \lambda^{(z)}) = \sum_{m=1}^{M} P(m|\boldsymbol{x}_t, \lambda^{(z)}) P(\boldsymbol{y}|\boldsymbol{x}_t, m, \lambda^{(z)}), \tag{3.3}$$

where

$$P(m|\boldsymbol{x}_t, \lambda^{(z)}) = \frac{\omega_m \mathcal{N}(\boldsymbol{x}_t; \boldsymbol{\mu}_m^{(x)}, \boldsymbol{\Sigma}_m^{(xx)})}{\sum_{n=1}^{M} \omega_n \mathcal{N}(\boldsymbol{x}_t; \boldsymbol{\mu}_n^{(x)}, \boldsymbol{\Sigma}_n^{(xx)})}, \tag{3.4}$$

$$P(\boldsymbol{y}|\boldsymbol{x}_t, m, \lambda^{(z)}) = \mathcal{N}\left(\boldsymbol{y}_t; \boldsymbol{E}_{m,t}^{(y)}, \boldsymbol{D}_m, ^{(y)}\right). \tag{3.5}$$

The mean vector $\boldsymbol{E}_{m,t}^{(y)}$ and the covariance matrix $\boldsymbol{D}_m^{(y)}$ of the mth conditional probability density function are represented as

$$\boldsymbol{E}_{m,t}^{(y)} = \boldsymbol{\mu}_m, ^{(y)} + \boldsymbol{\Sigma}_m^{(yx)} \boldsymbol{\Sigma}_m^{(xx)^{-1}} \left(\boldsymbol{x}_t - \boldsymbol{\mu}_m^{(x)}\right), \tag{3.6}$$

$$\boldsymbol{D}_m^{(y)} = \boldsymbol{\Sigma}_m^{(yy)} - \boldsymbol{\Sigma}_m^{(yx)} \boldsymbol{\Sigma}_m^{(xx)^{-1}} \boldsymbol{\Sigma}_m^{(xy)}. \tag{3.7}$$

In the conventional study [67], the converted target features are calculated on the basis of the minimum mean-square error as follows:

$$\hat{\boldsymbol{y}}_t = E[\boldsymbol{y}_t|\boldsymbol{x}_t] \tag{3.8}$$

$$= \int p\left(\boldsymbol{y}_t|\boldsymbol{x}_t, \lambda^{(z)}\right) \boldsymbol{y}_t d\boldsymbol{y}_t \tag{3.9}$$

$$= \int \sum_{m=1}^{M} P\left(m|\boldsymbol{x}_t, \lambda^{(z)}\right) P\left(\boldsymbol{y}_t|\boldsymbol{x}_t, m, \lambda^{(z)}\right) \boldsymbol{y}_t d\boldsymbol{y}_t \tag{3.10}$$

$$= \sum_{m=1}^{M} P\left(m|\boldsymbol{x}_t, \lambda^{(z)}\right) \boldsymbol{E}_{m,t}^{(y)}, \tag{3.11}$$

where $\hat{\boldsymbol{y}}_t$ means the converted target feature vector, and $E[\cdot]$ means the expectation. As the equation (3.11) shows, the converted target features are calculated by weighted sum of the conditional mean vectors, where the posterior probabilities of the source vector belonging to individual mixture components are used as weights. Moreover, the equation (3.11) also shows that the converted features are independently obtained frame by frame.

3.3. Employing Dynamic Features and Global Variances

Toda *et al.* have improved the performance of the conversion accuracy by considering dynamic features [68]. When people utter something, articulatory organs

are smoothly moved. This fact, which is not considered in MMSE-based conversion method, indicates that a time sequence of extracted acoustic parameters should have certain correlations between frames over an utterance. The effectiveness of dynamic features is confirmed in hidden Markov model (HMM)-based speech synthesis method [55][56].

Let $\boldsymbol{X}_t = [\boldsymbol{x}_t^\top, \Delta\boldsymbol{x}_t^\top]^\top$ and $\boldsymbol{Y}_t = [\boldsymbol{y}_t^\top, \Delta\boldsymbol{y}_t^\top]^\top$ be the source and the target joint feature vector of static and dynamic feature vector for frame t. Then, time sequences of source and target feature vectors are written as $\boldsymbol{X} = [\boldsymbol{X}_1^\top, \cdots, \boldsymbol{X}_t^\top, \cdots, \boldsymbol{X}_T^\top]^\top$ and $\boldsymbol{Y} = [\boldsymbol{Y}_1^\top, \cdots, \boldsymbol{Y}_t^\top, \cdots, \boldsymbol{Y}_T^\top]^\top$, respectively. Although $\boldsymbol{Y}$ seems a new feature sequence, it is rewritten using an extending matrix from the static feature vector to the joint feature vector of static and dynamic feature vector as

$$\boldsymbol{Y} = \boldsymbol{W}\boldsymbol{y}. \tag{3.12}$$

Using this correlation, a time sequence of the converted feature vectors is determined as follows:

$$\begin{aligned} \hat{\boldsymbol{y}} &= \arg\max P\left(\boldsymbol{Y}|\boldsymbol{X}, \lambda^{(Z)}\right) && (3.13)\\ &= \arg\max P\left(\boldsymbol{W}\boldsymbol{y}|\boldsymbol{X}, \lambda^{(Z)}\right), && (3.14) \end{aligned}$$

where $\hat{\boldsymbol{y}}$ is a time sequence of the converted static feature vectors. Two solutions are considered for the equation (3.14); EM algorithm and an approximation of mixture sequences.

In EM algorithm, the following auxiliary function is considered:

$$Q(\boldsymbol{Y}, \hat{\boldsymbol{Y}}) = \sum_{all\boldsymbol{m}} P\left(\boldsymbol{m}|\boldsymbol{X}, \boldsymbol{Y}, \lambda^{(z)}\right) \log P\left(\hat{\boldsymbol{Y}}, \boldsymbol{m}|\boldsymbol{X}, \lambda^{(z)}\right), \tag{3.15}$$

where $\hat{\boldsymbol{Y}}$ is the time sequence of converted joint feature vectors consisting of static and dynamic feature vectors and $\boldsymbol{m} = [m_1, \cdots, m_t, \cdots, m_T]$ is a mixture

component sequence. The auxiliary function is written as

$$
\begin{aligned}
Q(\boldsymbol{Y}, \hat{\boldsymbol{Y}}) &= \sum_t \sum_m P\left(m|\boldsymbol{X}_t, \boldsymbol{Y}_t, \lambda^{(Z)}\right) \log P\left(\hat{\boldsymbol{Y}}_t, m|\boldsymbol{X}_t, \lambda^{(Z)}\right) && (3.16)\\
&= \sum_t \sum_m \gamma_{m,t} \left\{ \log \gamma_{m,t} - \frac{D}{2}\log(2\pi) - \frac{1}{2}\log\left|\hat{\boldsymbol{D}}_m\right| \right.\\
&\quad \left. - \frac{1}{2}\left(\hat{\boldsymbol{Y}}_t - \boldsymbol{E}_{m,t}^{(Y)}\right)^\top \boldsymbol{D}_m^{(Y)^{-1}} \left(\hat{\boldsymbol{Y}}_t - \boldsymbol{E}_{m,t}^{(Y)}\right) \right\} && (3.17)\\
&= \sum_t \sum_m \gamma_{m,t} \left\{ \log \gamma_{m,t} - \frac{D}{2}\log(2\pi) - \frac{1}{2}\log\left|\hat{\boldsymbol{D}}_m\right| \right.\\
&\quad \left. - \frac{1}{2}\left(\hat{\boldsymbol{Y}}_t^\top \boldsymbol{D}_m^{(Y)^{-1}} \hat{\boldsymbol{Y}}_t - 2\hat{\boldsymbol{Y}}_t^\top \boldsymbol{D}_m^{(Y)^{-1}} \boldsymbol{E}_{m,t}^{(Y)} + \boldsymbol{E}_{m,t}^{(Y)^\top} \boldsymbol{D}_m^{(Y)^{-1}} \boldsymbol{E}_{m,t}^{(Y)}\right) \right\} && (3.18)\\
&= \sum_t \sum_m \gamma_{m,t} \left(-\frac{1}{2}\hat{\boldsymbol{Y}}_t^\top \boldsymbol{D}_m^{(Y)^{-1}} \hat{\boldsymbol{Y}}_t + \hat{\boldsymbol{Y}}_t^\top \boldsymbol{D}_m^{(Y)^{-1}} \boldsymbol{E}_{m,t}^{(Y)} \right)\\
&\quad + \overline{K}, && (3.19)
\end{aligned}
$$

where $\overline{K}$ is independent terms for $\hat{\boldsymbol{Y}}_t$. $\gamma_{m,t}$ is a weight for mth mixture component at frame t written as

$$
\begin{aligned}
\gamma_{m,t} &= P\left(m|\boldsymbol{X}_t, \boldsymbol{Y}_t, \lambda^{(Z)}\right) && (3.20)\\
&= \frac{\omega_m \mathcal{N}\left(\boldsymbol{X}_t, \boldsymbol{Y}_t; \boldsymbol{\mu}_m^{(Z)}, \boldsymbol{\Sigma}_m^{(Z)}\right)}{\sum_n \omega_n \mathcal{N}\left(\boldsymbol{X}_t, \boldsymbol{Y}_t; \boldsymbol{\mu}_n^{(Z)}, \boldsymbol{\Sigma}_n^{(Z)}\right)}. && (3.21)
\end{aligned}
$$

The equation (3.19) is also written as

$$
\begin{aligned}
(3.19) &= \sum_t \left\{ -\frac{1}{2}\hat{\boldsymbol{Y}}_t^\top \overline{\boldsymbol{D}_m^{(Y)^{-1}}} \hat{\boldsymbol{Y}}_t \hat{\boldsymbol{Y}}_t^\top \sum_m \left(\gamma_{m,t} \boldsymbol{D}_m^{(Y)^{-1}}\right) \hat{\boldsymbol{Y}}_t \right.\\
&\quad \left. + \hat{\boldsymbol{Y}}_t^\top \sum_m \left(\gamma_{m,t} \boldsymbol{D}_m^{(Y)^{-1}} \boldsymbol{E}_{m,t}^{(Y)}\right) \right\} + \overline{K}. && (3.22)
\end{aligned}
$$

The time sequence of the converted static features maximizing the auxiliary function is given by

$$
\hat{\boldsymbol{y}} = \left(\boldsymbol{W}^\top \overline{\boldsymbol{D}^{(Y)^{-1}}} \boldsymbol{W}\right)^{-1} \boldsymbol{W}^\top \overline{\boldsymbol{D}^{(Y)^{-1}} \boldsymbol{E}^{(Y)}}, \qquad (3.23)
$$

where

$$\overline{\boldsymbol{D}^{(Y)^{-1}}} = diag\left[\overline{\boldsymbol{D}_1^{(Y)^{-1}}}, \cdots, \overline{\boldsymbol{D}_t^{(Y)^{-1}}}, \cdots, \overline{\boldsymbol{D}_T^{(Y)^{-1}}}\right], \tag{3.24}$$

$$\overline{\boldsymbol{D}^{(Y)^{-1}}\boldsymbol{E}^{(Y)}} = \left[\overline{\boldsymbol{D}_1^{(Y)^{-1}}\boldsymbol{E}_1^{(Y)}}, \cdots, \overline{\boldsymbol{D}_t^{(Y)^{-1}}\boldsymbol{E}_t^{(Y)}}, \cdots, \overline{\boldsymbol{D}_T^{(Y)^{-1}}\boldsymbol{E}_T^{(Y)}}\right], \tag{3.25}$$

$$\overline{\boldsymbol{D}_t^{(Y)^{-1}}} = \sum_{m=1}^{M} \gamma_{m,t}. \tag{3.26}$$

One essential problem in maximum likelihood criterion is that estimated parameters tend to over-smoothed as **Figure 3.2** shows. In order to address the over-smoothing problem, Toda *et al* have proposed a novel feature of GV which is a global variance of acoustic features over a time sequence [68]. GV is expected to effectively capture acoustic variations that are lost in other statistical conversion methods. GV is seen as meta acoustic features since it is the acoustic feature to capture behaviors of acoustic features that are the target of GV. The concept of GV is similar to that of the 'yourtone' described in **Subsection 2.4.1**; sensitive fluctuations (or outline shapes) of speech waveforms give important influences for the naturalness of human voices.

GVs of the target static feature vectors over a time sequence are written as follows:

$$\boldsymbol{v}(y) = [v(1), \cdots, v(d), \cdots, v(D)]^\top, \tag{3.27}$$

$$v(d) = \frac{1}{T}\sum_{t=1}^{T}\left(y_t(d) - \bar{y}(d)\right)^2, \tag{3.28}$$

$$\bar{y}(d) = \frac{1}{T}\sum_{t=1}^{T} y_t(d), \tag{3.29}$$

where $y_t(d)$ is the dth component of the target static feature vector at frame t. In this thesis, GVs are calculated utterance by utterance. A new likelihood function consisting of two probability density function for a sequence of target static and dynamic feature vectors and for the GVs of the target static feature vectors as follows:

$$P\left(\boldsymbol{Y}|\boldsymbol{X}, \lambda^{(Z)}, \lambda^{(v)}\right) = P\left(\boldsymbol{Y}|\boldsymbol{X}, \lambda^{(Z)}\right)^{\omega} P\left(\boldsymbol{v}(y)|\lambda^{(v)}\right), \tag{3.30}$$

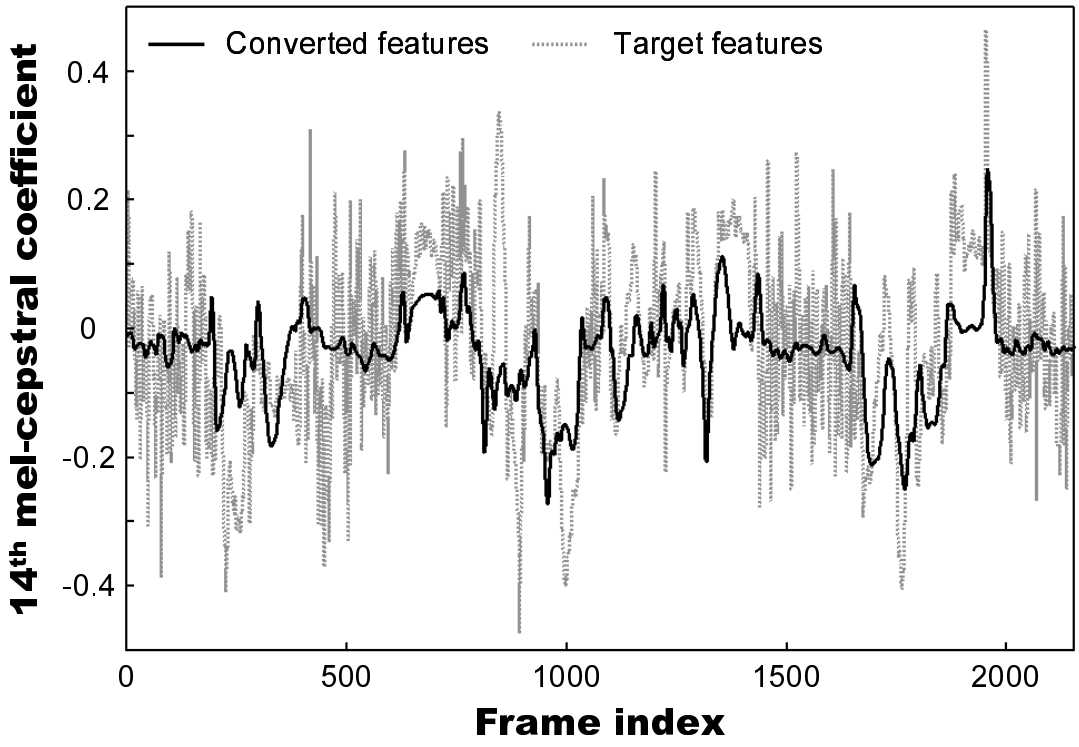

Figure 3.2. Example of over-smoothing of converted features compared to target ones.

where $P\left(\boldsymbol{v}(y)|\lambda^{(v)}\right)$ is described by a single Gaussian with the mean vector $\boldsymbol{\mu}^{(v)}$ and the covariance matrix $\boldsymbol{\Sigma}^{(vv)}$ as follows:

$$P\left(\boldsymbol{v}(y)|\lambda^{(v)}\right) = \mathcal{N}\left(\boldsymbol{v}(y); \boldsymbol{\mu}^{(v)}, \boldsymbol{\Sigma}^{(vv)}\right). \tag{3.31}$$

Two models of the GMM $\lambda^{(z)}$ and the Gaussian distribution $\lambda^{(v)}$ are independently trained using training data. The constant ω denotes the weight factor to control the balance between those two likelihoods. This thesis set ω as the ratio of the number of dimensions between $\boldsymbol{v}(y)$ and $\boldsymbol{Y}$, namely, $\frac{1}{2T}$.

The new time sequence of the converted static feature vectors is determined as follows:

$$\hat{\boldsymbol{y}} = \arg\max P\left(\boldsymbol{Y}|\boldsymbol{X}, \lambda^{(Z)}, \lambda^{(v)}\right) \tag{3.32}$$

$$= \arg\max P\left(\boldsymbol{W}\boldsymbol{y}|\boldsymbol{X}, \lambda^{(Z)}, \lambda^{(v)}\right). \tag{3.33}$$

The new auxiliary function is updated from Eqn. (3.15) as follows:

$$Q(\boldsymbol{Y}, \hat{\boldsymbol{Y}}) \;=\; \omega \mathcal{L}_1 + \mathcal{L}_2, \tag{3.34}$$

where $\mathcal{L}_1$ is equal to an auxiliary function of Eqn. (3.15) and $\mathcal{L}_2$ is written as follows:

$$\begin{aligned}
\mathcal{L}_2 &= \log P\left(\boldsymbol{v}(\hat{\boldsymbol{y}}) | \lambda^{(v)}\right) && (3.35)\\
&= -\frac{1}{2}\boldsymbol{v}(\hat{\boldsymbol{y}})^\top {\boldsymbol{\Sigma}^{(vv)}}^{-1} \boldsymbol{v}(\hat{\boldsymbol{y}}) + \boldsymbol{v}(\hat{\boldsymbol{y}})^\top {\boldsymbol{\Sigma}^{(vv)}}^{-1} \boldsymbol{\mu}^{(v)} + \overline{K'}, && (3.36)
\end{aligned}$$

where $\overline{K'}$ is independent factors for $\hat{\boldsymbol{y}}$. The derivative of $\mathcal{L}_2$ with respect to $\hat{\boldsymbol{y}}$ is given as follows:

$$\begin{aligned}
\frac{\partial \mathcal{L}_2}{\partial \hat{\boldsymbol{y}}} &= \left[{\boldsymbol{v}_1'}^\top, \cdots, {\boldsymbol{v}_t'}^\top, \cdots, {\boldsymbol{v}_T'}^\top \right], && (3.37)\\
\boldsymbol{v}_t' &= \frac{\partial \mathcal{L}_2^\top}{\partial \hat{\boldsymbol{y}}_t} && (3.38)\\
&= \left[v_t'(1), \cdots, v_t'(d), \cdots, v_t'(D) \right]^\top, && (3.39)\\
v_t'(d) &= \frac{\partial \mathcal{L}_2}{\partial \hat{y}_t(d)} && (3.40)\\
&= \frac{\partial \mathcal{L}_2}{\partial v(d)} \frac{\partial v(d)}{\partial \hat{y}_t(d)}. && (3.41)
\end{aligned}$$

In order to obtain $\frac{\partial \mathcal{L}_2}{\partial v(d)}$, let us concern dth factor after obtaining the partial derivative of $\mathcal{L}_2$ with respect to $\boldsymbol{v}(\hat{\boldsymbol{y}})$.

$$\begin{aligned}
\frac{\partial \mathcal{L}_2}{\partial v(d)} &= \frac{\partial}{\partial v(d)} \left(-\frac{1}{2}\boldsymbol{v}(\hat{\boldsymbol{y}}) {\boldsymbol{\Sigma}^{(vv)}}^{-1} \boldsymbol{v}(\hat{\boldsymbol{y}}) + \boldsymbol{v}(\hat{\boldsymbol{y}}) {\boldsymbol{\Sigma}^{(vv)}}^{-1} \boldsymbol{\mu}^{(v)} + \overline{K'} \right) && (3.42)\\
&= -{\boldsymbol{\Sigma}^{(vv)}}^{-1} \boldsymbol{v}(\hat{\boldsymbol{y}}) + {\boldsymbol{\Sigma}^{(vv)}}^{-1} \boldsymbol{\mu}^{(v)} && (3.43)\\
&= -{\boldsymbol{\Sigma}^{(vv)}}^{-1} \left(\boldsymbol{v}(\hat{\boldsymbol{y}}) - \boldsymbol{\mu}^{(v)} \right). && (3.44)
\end{aligned}$$

Let $\boldsymbol{p}_v^{(d)}$ be the dth row vector of ${\boldsymbol{\Sigma}^{(vv)}}^{-1}$. Then, the dth coefficient of Eqn. (3.44), namely $\frac{\partial \mathcal{L}_2}{\partial v(d)}$, is written as follows:

$$\frac{\partial \mathcal{L}_2}{\partial v(d)} \;=\; -\boldsymbol{p}_v^{(d)} \left(\boldsymbol{v}(\hat{\boldsymbol{y}}) - \boldsymbol{\mu}^{(v)} \right). \tag{3.45}$$

In order to obtain $\frac{\partial v(d)}{\partial \hat{y}_t(d)}$, first $v(d)$ is rewritten as follows:

$$
\begin{aligned}
v(d) &= \frac{1}{T}\sum_{t=1}^{T}\left\{\left(\hat{y}_t(d) - \overline{\hat{y}(d)}\right)^2\right\} && (3.46)\\
&= \frac{1}{T}\sum_{t=1}^{T}\left\{\left(\hat{y}_t(d) - \frac{1}{T}\sum_{u=1}^{U}\hat{y}_u(d)\right)^2\right\} && (3.47)\\
&= \frac{1}{T}\sum_{t=1}^{T}\left\{\hat{y}_t(d)^2 - \frac{2}{T}\hat{y}_t(d)\sum_{u=1}^{U}\hat{y}_u(d) + \frac{1}{T^2}\left(\sum_{u=1}^{U}\hat{y}_u(d)\right)^2\right\} && (3.48)\\
&= \frac{2}{T}\left(\hat{y}_t(d) - \frac{1}{T}\sum_{u=1}^{U}\hat{y}(d)\right) && (3.49)\\
&= \frac{2}{T}\left(\hat{y}_t(d) - \overline{\hat{y}(d)}\right). && (3.50)
\end{aligned}
$$

Finally, $v'_t(d)$ is written as follows:

$$
\begin{aligned}
v'_t(d) &= \frac{\partial \mathcal{L}_2}{\partial v(d)}\frac{\partial v(d)}{\partial \hat{y}_t(d)} && (3.51)\\
&= -\frac{2}{T}\boldsymbol{p}_v^{(d)}\left(\boldsymbol{v}(\hat{\boldsymbol{y}}) - \boldsymbol{\mu}^{(v)}\right)\left(\hat{y}_t(d) - \overline{\hat{y}(d)}\right). && (3.52)
\end{aligned}
$$

Figure 3.3 shows an example of converted spectral features considering GV. Comparing the figure with **Figure 3.2**, the parameters are estimated so that is has larger variances than the parameters derived from maximum likelihood estimation. The converted features considering GV is actually out from the parameters of maximum likelihood estimation, on the other hand, the synthesized voice quality is much better than using parameters not considering GV. This comes from the result of effective modeling of acoustic variations that the natural target speaker includes.

3.4. Summary

This chapter described the statistical VC method using GMM based on maximum likelihood criterion. The dynamic characteristic was introduced in order to capture certain correlations between frames. Moreover, other statistics of GV were introduced to suppress over-smoothing of converted features due to maximum likelihood estimation.

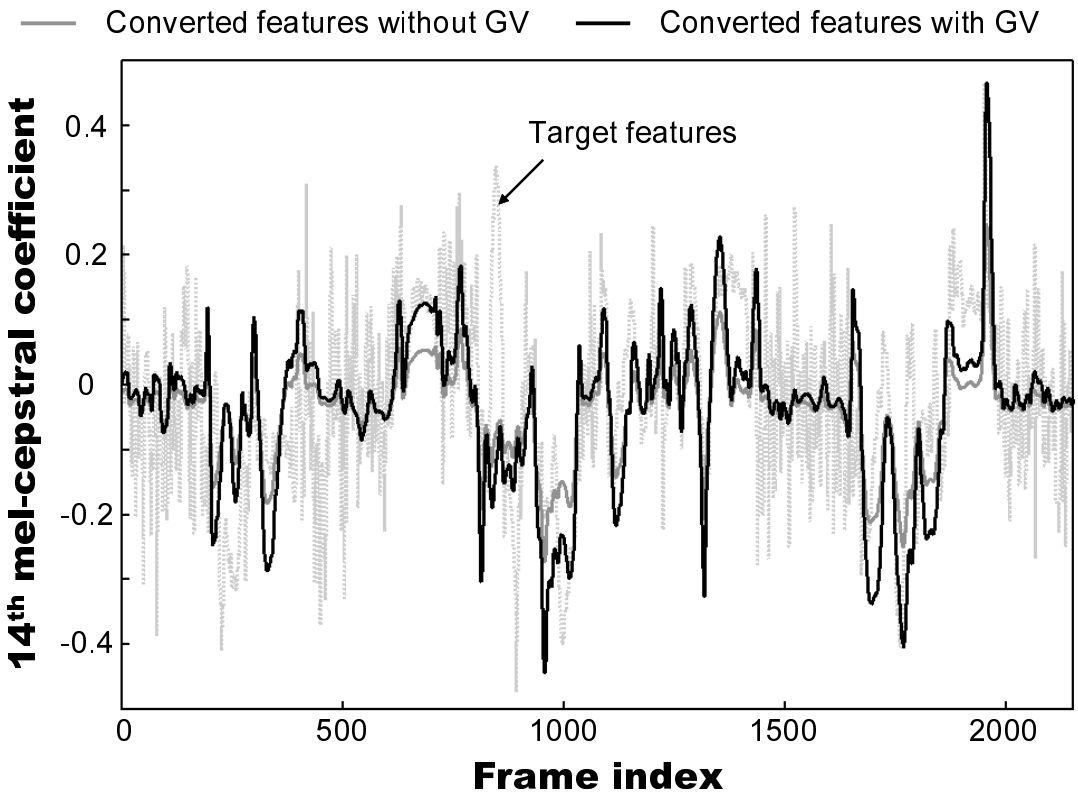

Figure 3.3. Example of converted features considering GV.

Chapter 4

Proposed Speaking-Aid System for EL Speech

This chapter proposs a speaking-aid system that enhances EL speech by statistical VC described in **Chapter 3**. First, the proposed system and the enhancement approach that accepts EL speech and outputs converted natural speech are described. This proposed system is experimentally evaluated as a preliminary test using EL speech imitated by a non-laryngectomee. In order to estimate more natural F_0 contours, an air-pressure sensor is introduced. Then, another speaking-aid system that enhances EL(air) speech is also proposed. Using the air-pressure sensor, the source F_0 values are available from the input EL(air) speech.

4.1. Introduction

This chapter proposes a novel speaking-aid system for laryngectomees to enhance EL speech using statistical VC technique. Although it is no wonder that it is natural for people to transmit something to others by speech in their daily lives, the convenience of speech is not always available to everyone, especially for speaking-impaired people. Laryngectomees described in this thesis can articulate with sound excitations; however, the produced alaryngeal speech is still unnatural and it makes it difficult for them to remain part of the society of non-laryngectomees as they had been. For EL and EL speech, many studies have so far been conducted to enhance the EL speech. Although those studies work well,

the generated speech quality has not been satisfied.

This thesis proposes a speaking-aid system to enhance the EL speech by the statistical VC technique. Basically, F_0 contours of ELs are pre-defined. This indicates that the EL speech does not have sufficient F_0 information except for the number of vibrations. In the VC to normal speech, the loss of F_0 information might derive degradations of F_0 estimation accuracy. This concern comes from VC from NAM to normal speech [15]. In order to avoid this problem, the proposed system enhancing EL speech converts EL speech to not only normal speech as the conversion framework of EL-to-Speech but also to whispering that does not include F_0 information as another conversion framework of EL-to-Whisper. The motivation for EL-to-Whisper comes from similar conversion from NAM to whispering [16]. The effectiveness of EL-to-Speech and EL-to-Whisper are experimentally confirmed. In the conversion from NAM to whispering, the naturalness of the converted whispering is more highly scored than that of the converted normal speech [16]. We often speak with whispering, and therefore, the whispering is an important output of the aid system. On the other hand, the proposed aid system aims to be used in the daily life of laryngectomees. Although we often speak with a whispering, it would be not natural for laryngectomees to speak with whispering all the time in their conversations. To convert the EL speech to more natural speech, an air-pressure sensor to detect F_0 information of EL speech is introduced. EL(air) speech would be converted to normal speech in the conversion framework of EL(air)-to-Speech.

This chapter is organized as follows. In **Section 4.2**, the speaking-aid system for EL speech is described. In **Section 4.3**, VC from EL speech is explained. In **Section 4.4**, the speaking-aid system for EL speech is experimentally evaluated as a preliminary test using imitated EL speech produced by a non-laryngectomee. In **Section 4.5**, another speaking-aid system for EL(air) speech using an air-pressure sensor is described. In **Section 4.6**, VC from EL(air) speech is described. Finally, this chapter is summarized in **Section 4.7**.

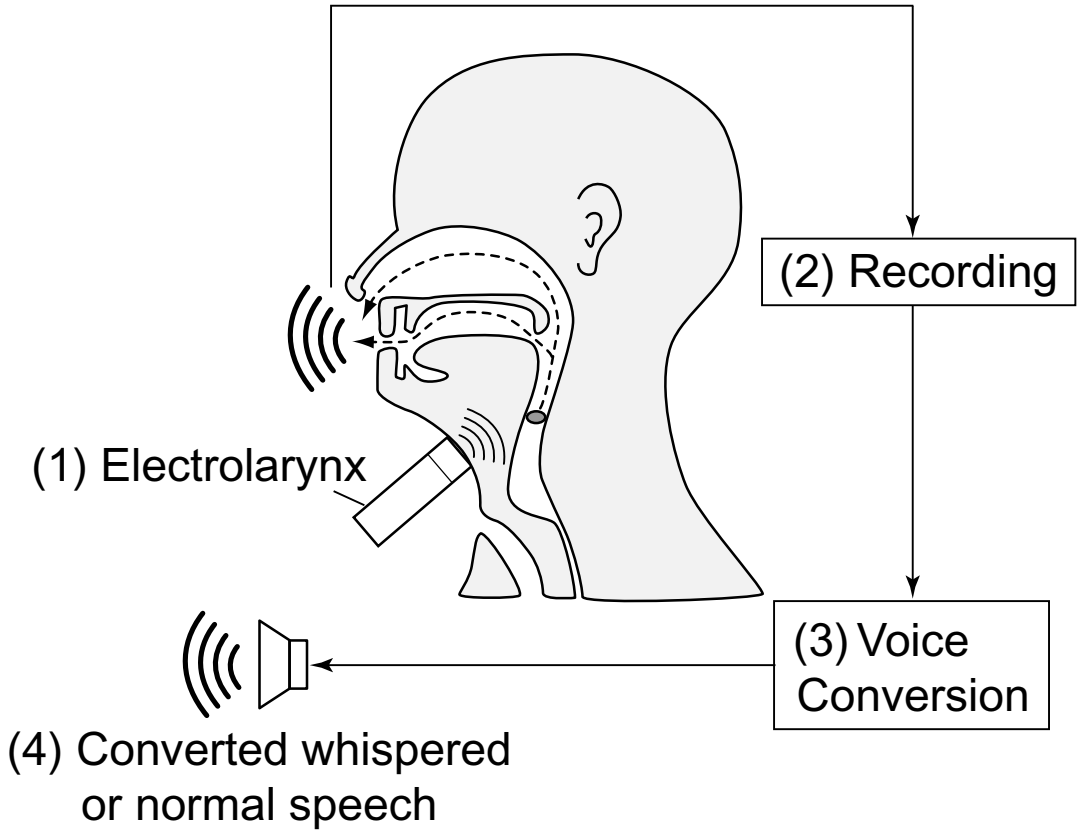

Figure 4.1. Overview of proposed speaking-aid system for conventional EL speech.

4.2. Speaking-Aid System for EL Speech

An overview of the speaking-aid system for EL speech is shown in **Figure 4.1**. This system consists of four parts; (1) generating sound source signals, (2) recording the EL speech, (3) converting the EL speech, and (4) output the converted speech. A current existing EL is supposed to be attached to the same location as a laryngectomee usually does. Conventional studies for ELs have focused on only the sound source signals to enable laryngectomees speak with more natural voice. The novel point of view of this system is that once it records the produced speech and converts the data to achieve more natural voice.

Because of the same reason of the enhancement of the esophageal speech, the proposed system would be significantly effective certain situations in which

Table 4.1. Input and output acoustic features for EL-to-Whisper and EL-to-Speech

System	Input	Output
EL-to-Whisper	spectrum	spectrum
EL-to-Speech	spectrum	spectrum
	spectrum	F_0
	spectrum	aperiodic components

laryngectomees have to communicate with others only by their voices such as telecommunication. The system would be difficult to be used in conversations with face-to-face because listeners would listen not only converted speech but also the produced EL speech. On the other hand, in situations which laryngectomees have to communicate with others only by their voices, listeners must understand what the laryngectomees said from only the transmitted speech, and therefore, the transmitted speech plays extremely important rolls to make smooth conversation successful. In telecommunication, most of the current telephones are specialized for human speech to decrease data quantity to be transmitted. Because of that, the current mechanical EL speech is not suitable for telecommunication and it confines the use of telephone of laryngectomees. In other situations in which EL speech has to be amplified to be transmitted for many people such as lectures, this system would be a powerful tool to help laryngectomees.

One usual VC framework for EL speech is to convert the EL speech into normal speech (EL-to-Speech). Although it is reasonable to convert the EL speech to normal speech, it is difficult to estimate natural F_0 contours. To avoid the difficulty of F_0 estimation, another conversion framework from EL speech to whispering (EL-to-Whisper) is concerned. In this conversion, only the spectral information of the EL speech is converted. Input and output acoustic features in EL-to-Whisper and EL-to-Speech is shown in **Table 4.1** in which aperiodic components [69] shows the strength of noises in each frequency band that is used in constructing mixed excitation signals [70].

4.3. Voice Conversion for EL Speech

The basic idea of the VC for EL speech is the same as the statistical VC described in **Section 3.2**. The training data of a GMM for spectral estimation is joint vectors of the source and the target spectral features. This thesis calculates delta information from one previous and succeeding frame. For the target data, joint feature vectors of static and delta features are constructed. For the source data, this thesis uses segmental feature vector that includes information of multiple frames. Acoustic features of the EL speech are significantly different from those of the normal speech, and therefore, the same feature extraction method might have a risk of degradations of the conversion accuracy. On the other hand, information that the conversion wants is supposed to be included in the current, several previous and succeeding frames. Multivariate analysis powerfully works to extract common factors from observations consisting of multivariable feature vector. There are some methods of multivariate analysis such as on the basis of correlation matrix that is on of most classical methods. This thesis uses PCA as one of major multivariate analyses. PCA procedure extract common factors which are decorrelation each other to explain observations by linear combinations of the extracted common factors, which are called principal components. Principal components are extracted in decreasing order of variances for each dimension in order to catch the data as most as the components can.

Figure 4.2 shows a flow chart of constructing segmental feature vectors from static feature vectors which are spectral feature vectors in the case of EL-to-Whisper and EL-to-Speech. Let $\Theta_{\boldsymbol{x}} = \{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_t, \cdots, \boldsymbol{x}_T\}$ be a set of source static feature vectors where $\boldsymbol{x}_t = [x_t(1), \cdots, x_t(d), \cdots, x_t(D_x)]$ is D_x-dimensional feature vector. Let $\boldsymbol{c}_t^* = [\boldsymbol{x}_{t-L}^\top, \cdots, \boldsymbol{x}_t^\top, \cdots, \boldsymbol{x}_{t+L}^\top]^\top$ be a $D_c = D(2L+1)$-dimensional concatenated feature vector over the current $\pm L (L \geq 1)$ frames. Then, the $\hat{D}(\hat{D} \leq D_c)$-dimensional segmental feature vector $\boldsymbol{X}_t$ at frame t is extracted by PCA.

PCA procedure is regarded as eigenvalue decomposition for the covariance matrix of the data. Usually, the dimension of the segmental feature is less than that of the concatenated feature vector. From the view point of dimension, PCA procedure is regarded as the procedure to reduce dimensions. This procedure is expected to compensate loss information due to laryngectomy. The effectiveness

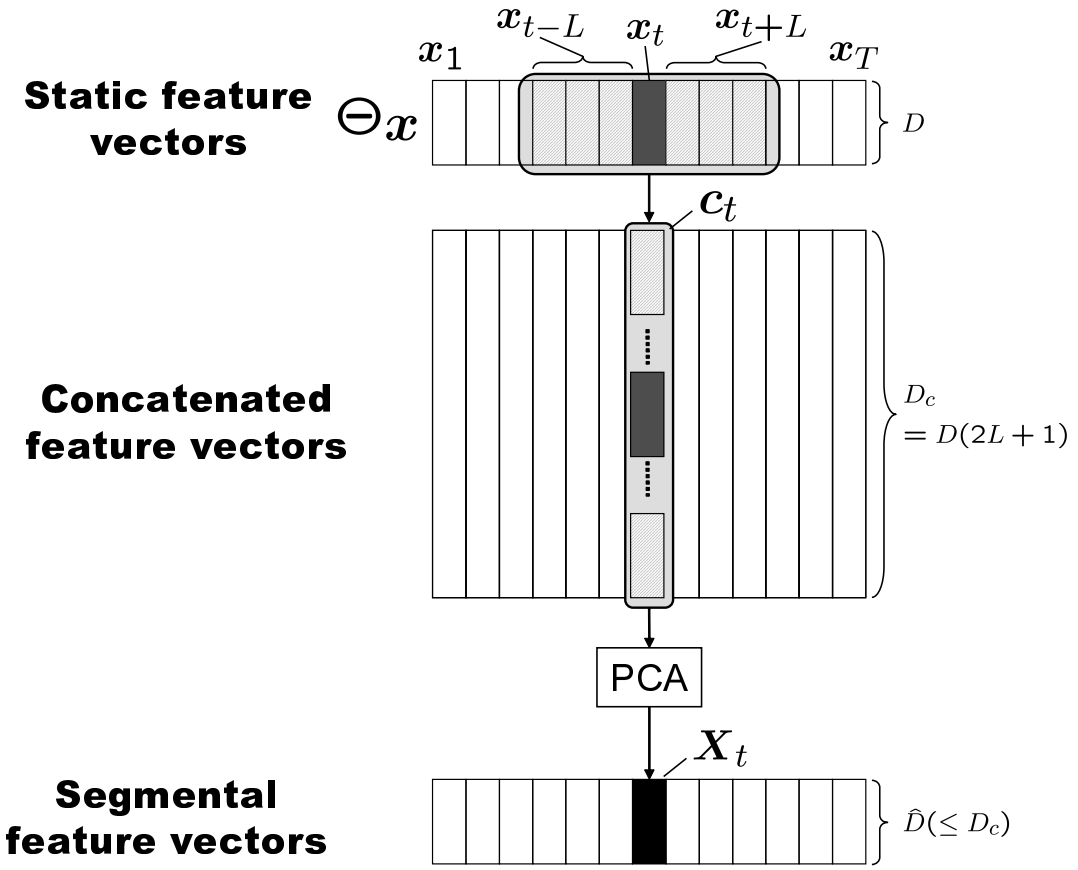

Figure 4.2. Flow chart of constructing segmental feature vectors from static feature vectors.

of using not joint feature vectors of static and delta feature but segmental feature vectors is experimentally confirmed in other studies in which some data such as higher frequency components are lost. Therefore, the segmental feature vector is expected to be effective for the EL speech conversion.

In the conversion procedure, source segmental feature vectors are constructed by the same approach in the training part. Acoustic parameters are estimated by the same method described in **Section 3.2**.

In EL-to-Speech, not only spectral information but also F_0 information should be estimated; however, source EL speech does not have effective F_0 information to be converted. Therefore, the GMM to estimate F_0 is trained using joint feature

vectors of source spectral data and target F_0 data. Source spectral segment is set to the spectral data, which is constructed by the same method as EL-to-Whisper. Joint vector of the target log-F_0 and its delta at frame t are set to the target F_0 data. In the conversion procedure, source spectral segment is given to the trained GMMs for both spectral and F_0 estimation and then, acoustic parameters are generated in maximum likelihood manner.

4.4. Preliminary Experimental Evaluation of the Speaking-Aid System for EL speech

4.4.1 Experimental conditions

The speaking-aid system for EL speech was experimentally evaluated as a preliminary test. The source speaker and the target speaker were the same non-laryngectomee. He had trained to produce EL speech for 21 days; therefore, he had learned the best position to produce EL speech. The speaker recorded 50 newspaper articles for the training data and the other 20 ones for the test data. EL speech was set to the source speech, and whispering or normal speech was set to the target speech. All speech data were recorded using a head-set microphone in a sound proof room. Data format of the recorded data was 16000 Hz sampling with 16 bit for each sample.

For the spectral data of source EL speech, 0th through 24th mel-cepstral coefficients, which were extracted by mel-cepstral analysis [71], were used. Note that 0th coefficient captured power information. For the spectral data of the target whispering, 0th through 24th mel-cepstral coefficients extracted by the same mel-cepstral analysis [71] were used. For the spectral data of the target normal speech, the 0th through 24th mel-cepstral coefficients, which were extracted by Speech Transformation and Representation using Adaptive Interpolation of weiGHTed spectrum (STRAIGHT) analysis [72], were used. F_0 values and aperiodic components of the target speech were extracted by STRAIGHT analysis [72]. The segmental feature vector of source spectral data to estimate target data was constructed by the following procedures; first, the current, previous and succeeding eight frames were concatenated into one vector, and then, the dimension of the

concatenated vector was compressed by PCA procedures. Finally, 50-dimensional segmental feature vector was constructed frame by frame. For F_0 estimation, the frame length to construct segmental feature vectors was set to eight, and the same method as constructing the segmental feature vectors of spectral data was employed. The number of GMM components to estimate spectral, F_0 and aperiodic parameters was set to 32, respectively.

Mel-cepstral distortion was introduced to objectively calculate conversion accuracy about spectrum between converted and target spectra, which was given by

$$Mel - cd\ [dB] = \frac{1}{T}\sum_{t=1}^{T}\frac{10\sqrt{2\sum_{d=1}^{D}\{(tar_t[d] - conv_t[d])^2\}}}{\ln 10}, \tag{4.1}$$

where $tar_d[d]$ and $conv_t[d]$ were dth coefficients of the target and converted mel-cepstrum at the frame t, respectively. Calculating the distortion between source and target spectra, $org_t[d]$ as the dth coefficients of the source mel-cepstrum at the frame t was set instead of $conv_t[d]$. Two measures were introduced to objectively calculate F_0 conversion accuracy between converted and target F_0; (1) unvoiced or voiced decision errors (U/V errors) and (2) correlation coefficient only for voiced frames. U/V errors were calculated as the rate of number of target U/V frames and converted U/V frames. Since source EL speech did not have effective F_0 information for VC, only the combination of converted and target F_0 values were calculated. Because of the same reason as F_0 evaluations, only the distortion of converted and target aperiodic components were calculated by root mean square errors given by

$$RMSE = \sqrt{\frac{\sum_{t=1}^{T} rmse_t}{T}}, \tag{4.2}$$

$$rmse_t = \sum_{d=1}^{D}\left(AP_t^{tar}(d) - AP_t^{conv}(d)\right)^2, \tag{4.3}$$

where $AP_t^{tar}(d)$ and $AP_t^{conv}(d)$ were the dth aperiodic component at frame t of target and converted data, respectively.

Table 4.2. Averaged mel-cepstral distortion. Values in front of and behind the slash shows distortions considering and not considering power information (i.e. 0^{th} coefficient), respectively

System	Source-Target	Converted-Target
EL-to-Whisper	9.09 / 7.57	5.00 / 4.38
EL-to-Speech	9.42 / 8.43	4.73 / 3.99

Table 4.3. Voiced or unvoiced error rates and correlation coefficients between voiced frames of converted F_0 values and those of target ones. '$x \rightarrow y$' denotes the rate of x frames regarded as y frames. The label 'U' and 'V' denote unvoiced and voiced frames, respectively. For example, V $\rightarrow$ U means rate of voiced frames regarded as unvoiced frames

Correlation coefficients	0.317 ± 0.105
V $\rightarrow$ V	41.92 [%]
U $\rightarrow$ U	48.96 [%]
V $\rightarrow$ U	7.09 [%]
U $\rightarrow$ V	2.04 [%]

4.4.2 Experimental results

Table 4.2 shows averaged mel-cepstral distortions utterance by utterance for EL-to-Whisper and EL-to-Speech. **Table 4.3** shows objective results for F_0 estimation, which are correlation coefficients and U/V errors between converted and target F_0 values. **Figure 4.3**, **Figure 4.4**, and **Figure 4.5** show examples of waveforms, spectrograms, F_0 contours of source, converted, and the target normal speech, respectively. Although the correlation coefficient of F_0 contours calculated from only voiced frames of between converted and target normal speech is not high, F_0s with certain tones are estimated.

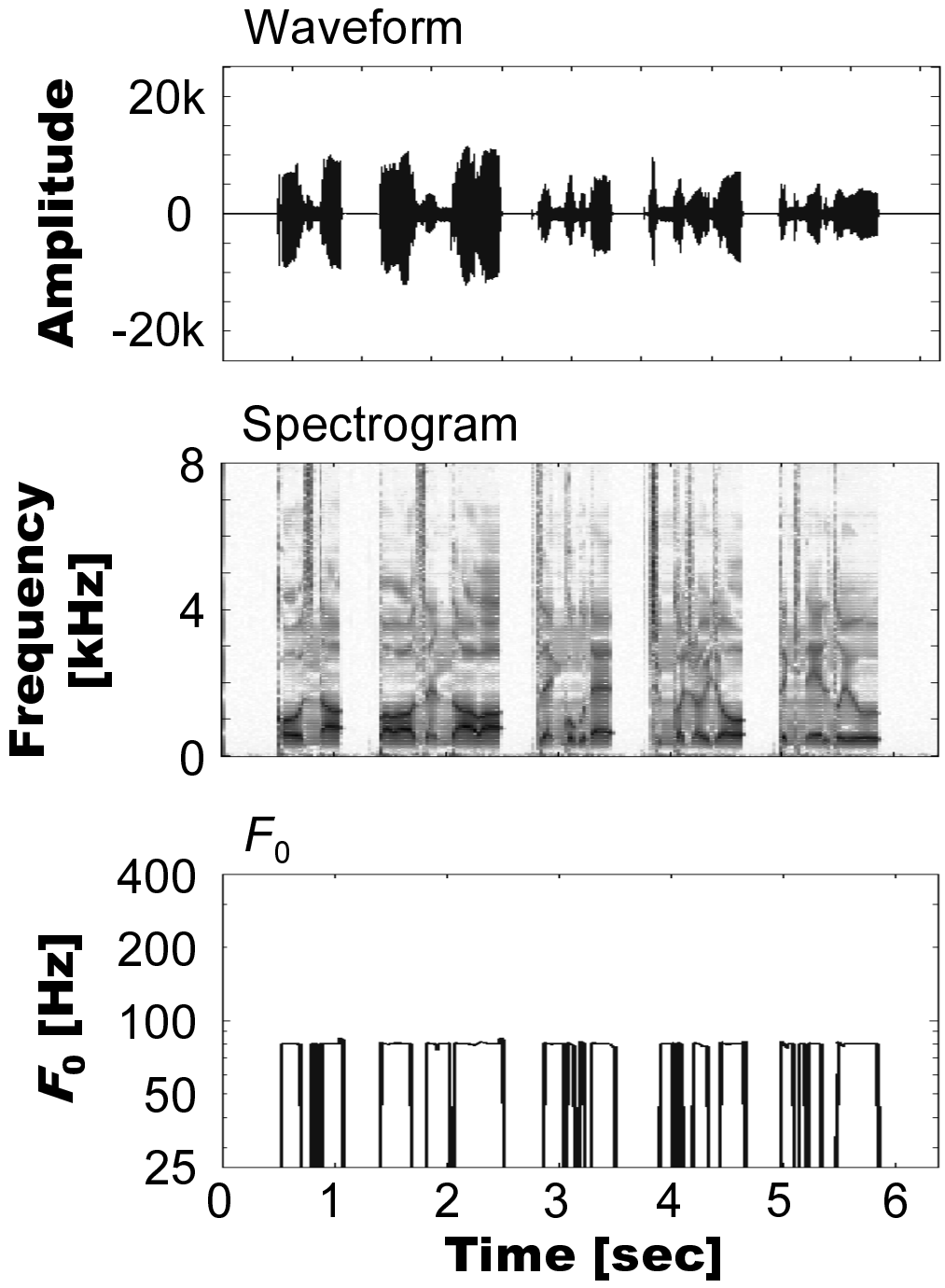

Figure 4.3. Example of waveforms, spectrograms, and F_0 contours for source EL speech.

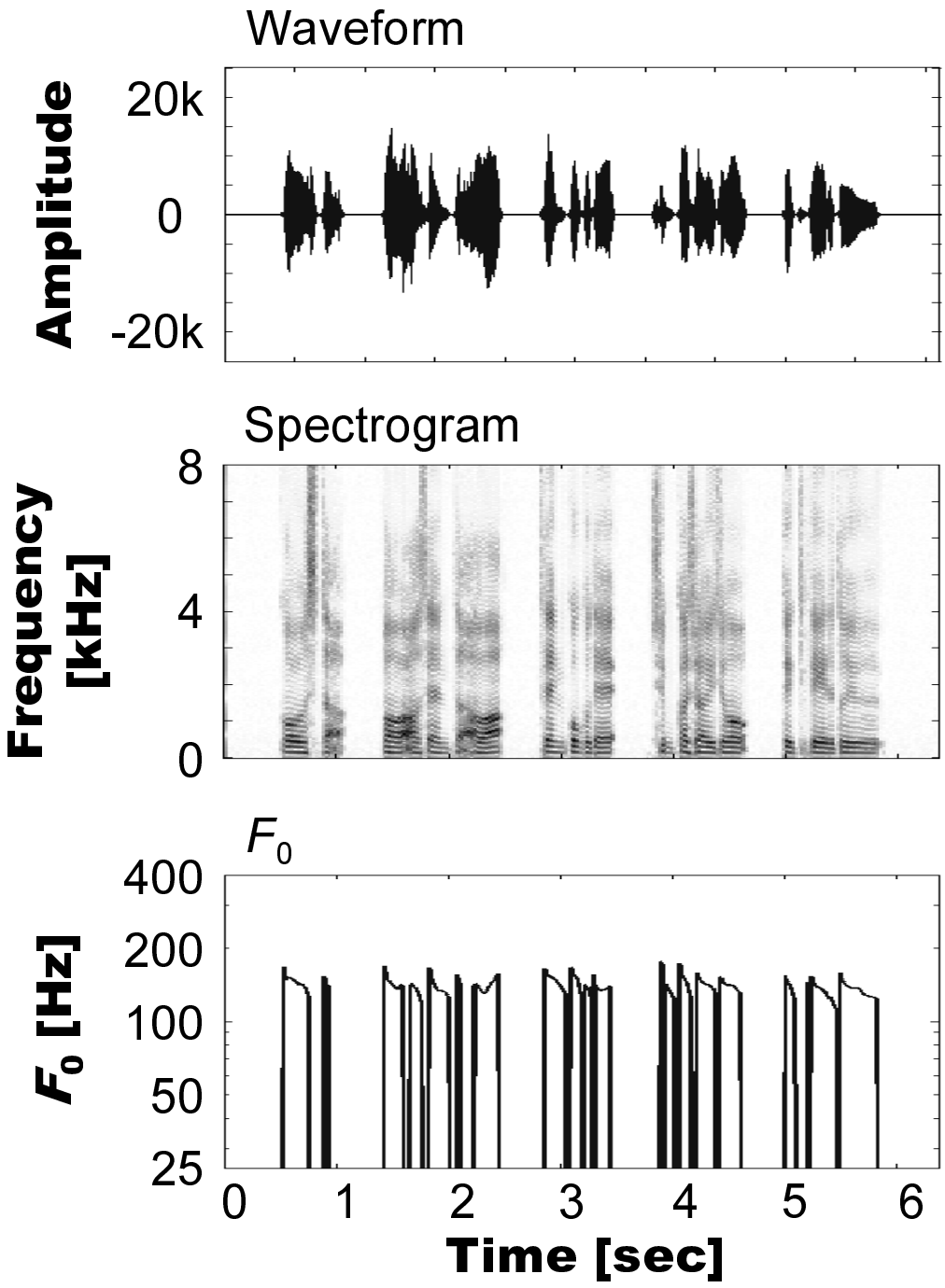

Figure 4.4. Example of waveforms, spectrograms, and F_0 contours for converted normal speech.

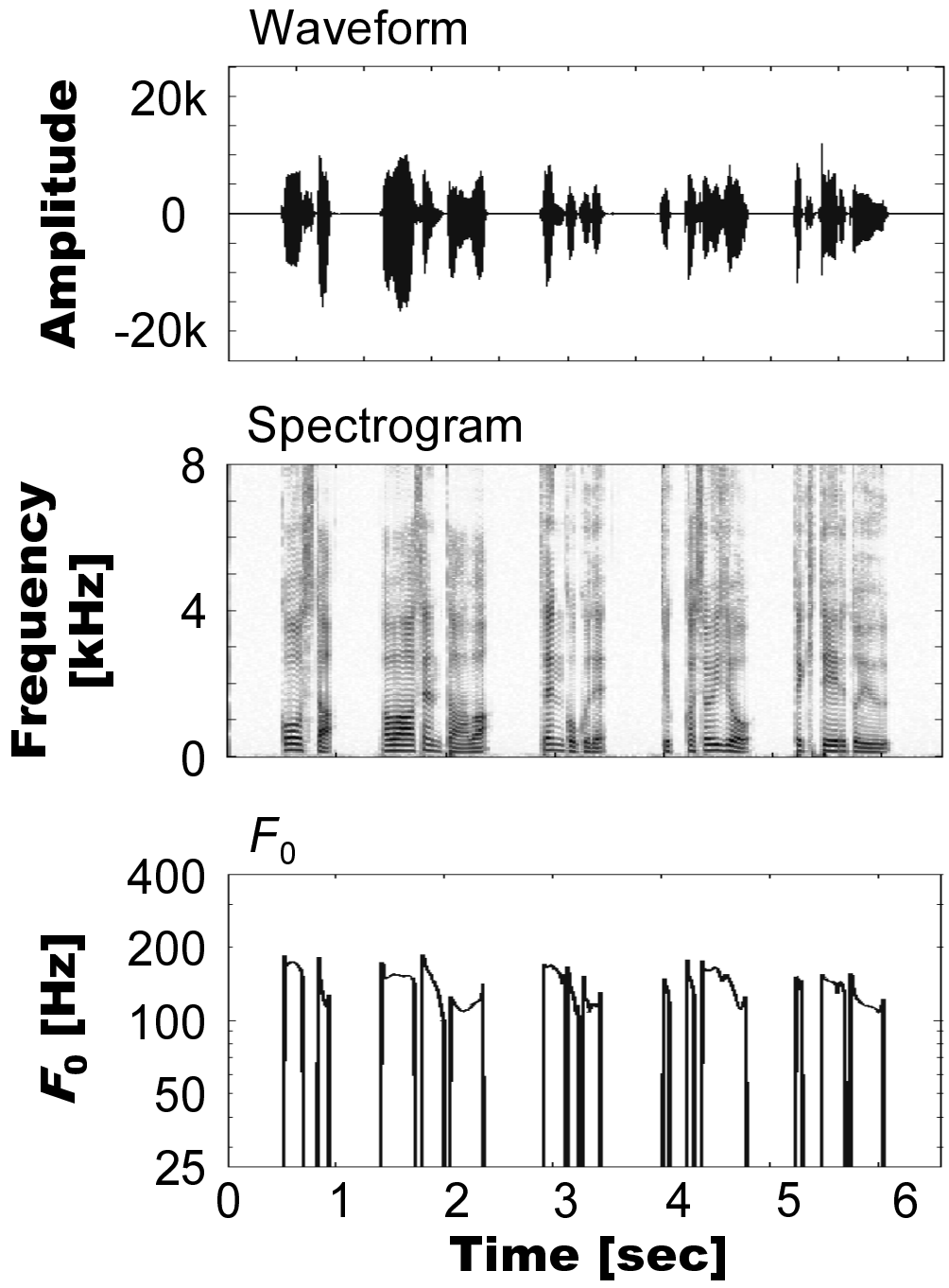

Figure 4.5. Example of waveforms, spectrograms, and F_0 contours for target normal speech.

4.5. Speaking-Aid System for EL(air) Speech

An air-pressure sensor to manipulate F_0 contours has been developed by Uemi *et al* [35]. In order to use this sensor, the air-pressure sensor is connected to the existing EL called 'yourtone'. A user holds both the body of the EL and the air-pressure sensor with his or her both hands. The body of the EL is attached to the speaker's lower jaw. The air-pressure sensor is set to cover the tracheostoma. The current air-pressure sensor is closed-type that does not pass through the air when the sensor covers thetracheostoma. Consequently, the speaker do the following processes to speak using the air-pressure sensor for every pause insertions: 1) taking air into the lungs while attaching only the body of the EL to the lower jaw, 2) putting the air-pressure sensor so that it covers the tracheostoma, 3) expiring the air so that it drives the vibrator of the EL, and 4) articulating the vibrations to speak with EL speech. The circuit that converts the air-pressure to the number of vibration of the vibrator is built in the main body of the EL.

As the **Figure 4.6** shows, this system is similar to **Figure 4.1** except its sound source unit and target speech. The F_0 contours of EL(air) speech are not monotone; however, those are also not natural than those of normal speech. This F_0 information might be effective for VC of converting EL(air) to normal speech (EL(air)-to-Speech). The effectiveness of using the air-pressure sensor and F_0 information is experimentally investigated in **Chapter 6** using a laryngectomee's data. In the enhancement of EL(air) speech, EL(air) speech is converted to only normal speech because source EL(air) speech has not monotone pitch information. In the VC of EL(air)-to-Speech, the same kinds of acoustic features as EL-to-Speech are estimated; spectrum, F_0, and aperiodic components, which are independently estimated. Input and output acoustic features in EL(air)-to-Speech are shown in **Table 4.4**. For the spectral estimation, one GMM is trained. For the F_0 estimation, two types of conversion are concerned in which F_0 estimation from only spectral data, only F_0 data, or both spectral and F_0 data.

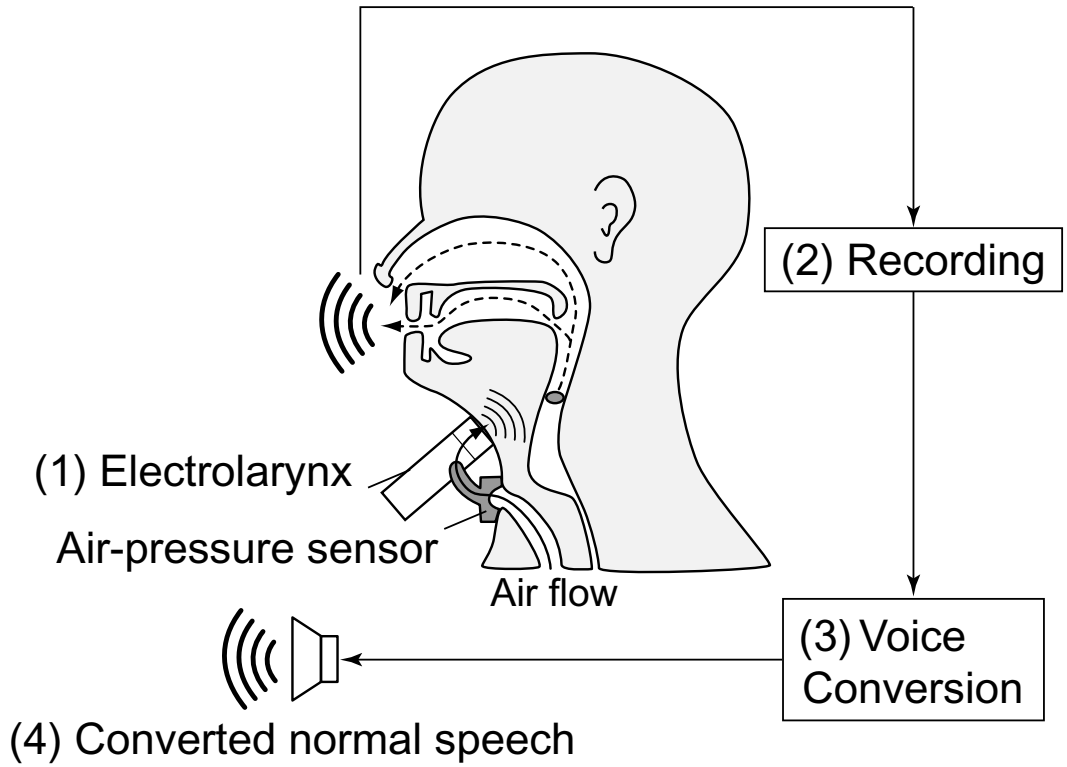

Figure 4.6. Overview of proposed speaking-aid system for EL(air) speech.

Table 4.4. Input and output acoustic features for EL(air)-to-Speech

Input	Output
spectrum	spectrum
spectrum or spectrum and F_0	F_0
spectrum	aperiodic components

4.6. Voice Conversion for EL(air) Speech

In EL(air)-to-Speech, joint feature vectors of static and dynamic target $\log F_0$ values are set to the target F_0 feature vectors. When target F_0 values are estimated from only source spectral features, the same segmental feature vectors as

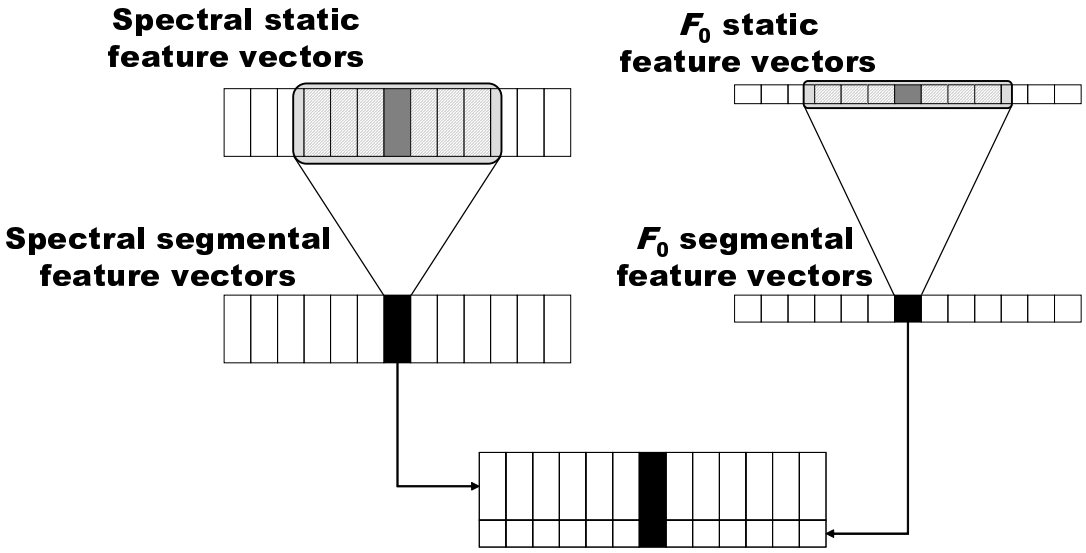

Figure 4.7. Flow chart of constructing segmental feature vectors from spectral and F_0 feature vectors.

the spectral estimation are set to the source data. Next, joint vectors consisting on segmental vectors of source spectral features and joint vectors of target $\log F_0$ features are modeled by a GMM. When estimating target F_0 values using both spectral and F_0 information, two methods of constructing source segmental features are concerned; 1) firstly concatenate two static feature vectors of spectral and F_0 feature vectors, and then, construct the segmental features, or 2) firstly construct spectral and F_0 segmental feature vectors independently for each frame in the same method as **Figure 4.2** shows, and then, concatenate these vectors as **Figure 4.7** shows. What it wants to confirm by using both spectral and F_0 information is the effectiveness of the source F_0 information. In the first method, it is not ensured that information of source F_0 data remains in the segmental feature vectors. Because of that, the second method in which two individual segmental feature vectors are concatenated is employed in this thesis.

4.7. Conclusion

This chapter described a speaking-aid system to enhance EL speech using the statistical VC method. In the aid system, a VC framework of EL-to-Speech was concerned. It is attractive; however, the natural F_0 estimation might be difficult because source EL speech does not have sufficiently rich F_0 information. To avoid the problem of natural F_0 estimation, another VC framework of EL-to-Whisper was considered. To obtain a non-monotone pitch of source EL speech, a novel air-pressure sensor was additionally introduced, with which speakers can manipulate F_0 of the EL by their own air expired from the tracheostoma. This EL speech using the air-pressure sensor was converted to normal speech. In this VC framework, F_0 contours of the converted speech were expected to be closer to those of the normal speech. When estimating target F_0 using EL speech with air-pressure sensor, three kinds of source segmental vectors were used, which were spectral segmental vectors, $\log F_0$ segmental vectors, or joint vectors of spectral and $\log F_0$ segmental features.

Chapter 5

Proposed Speaking-Aid System for EL(small) Speech

This chapter proposes the other speaking-aid system that accepts EL speech using the extremely small-powered sound source signals and outputs natural speech. If the users can select sound source signals, it might be the versatility of the system. This thesis designed sound source signals by changing the spectrum and the power independently. In the case of changing its spectrum, this thesis designed three kinds of sound source signals of pulse train, a sawtooth waves, and the compensation waves into whispering. In the other case of changing its power, the spectrum is fixed to the sawtooth waves that has the largest dynamic range. This system is also experimentally evaluated as a preliminary test using imitated EL speech by a non-laryngectomee.

5.1. Introduction

This chapter describes another speaking-aid system for laryngectomees using the statistical VC approach. The input of the speaking-aid system described in **Chapter 4** is EL speech that has enough power for speech communication in laryngectomees' daily life. On the other hand, the large power of the electrolarynx often annoys people around the speaker especially in quiet settings such as a library. Moreover, the problem is that the speaker would be anxious that he or she makes people around him or her uncomfortable since the sound source sig-

nals are output every utterance. This problem was made obvious by the author's experience producing utterances using only an electrolarynx for 21 days.

To address the noisy sound source signals, this thesis introduces another sound source unit, proposed by Hosoi *et al.*, that outputs any signals with extremely small power so that people around the user have difficulty to catch the sound. The new sound source unit (EL(small)) addresses the noisy sounds of the electrolarynx; on the other hand, another problem rises that not only the sound source signals but also the produced EL speech have only small power so that it is almost too difficult for a usual air-conductive microphone such as head-set microphone to catch the EL speech. To record the EL speech excited by small-powered signals, this paper introduces the NAM microphone.

As the other challenging speaking-aid system for laryngectomees, the small-powered EL speech recorded with a NAM microphone (EL(small) speech) is set to the source signals. In this system, two kinds of speech signals of whispering or normal speech are output as the same idea as the EL-to-Whisper; EL(small)-to-Whisper and EL(small)-to-Speech in this thesis, respectively. EL(small) can generate arbitrary signals. Using these characteristics, three sound source signals are designed; (1) pulse train, (2) sawtooth waves, and (3) compensation waves that compensates for acoustic features of EL(small) speech to whispering. The robustness of VC for these sound source signals is experimentally evaluated in advance. The effectiveness of EL(small)-to-Whisper and EL(small)-to-Speech are experimentally evaluated in the following **Chapter 6** in which the pulse train and the sawtooth waves are used.

This chapter is organized as follows. The speaking-aid system using EL(small) is described in **Section 5.2**. VC for EL(small) speech is described in **Section 5.3**. In **Section 5.4**, the speaking-aid system for EL(small) speech is experimentally evaluated as a preliminary test using imitated EL(small) speech produced by the same non-laryngectomee as **Chapter 4**. This section is summarized in **Section 5.5**.

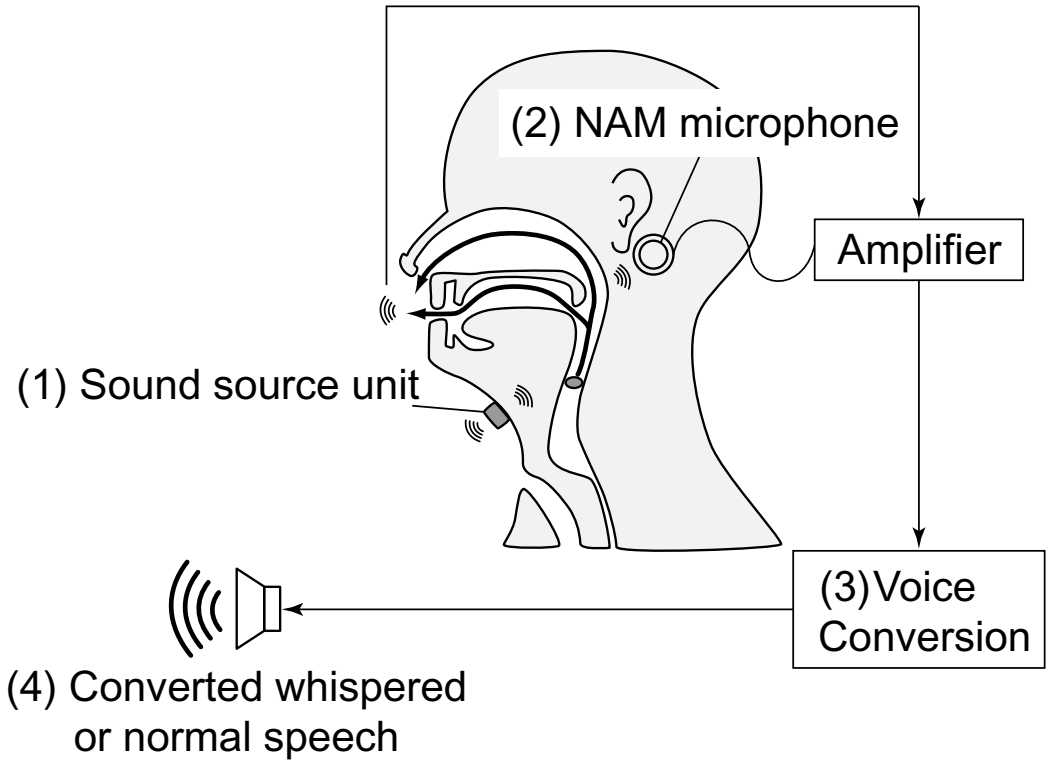

Figure 5.1. Overview of proposed speaking-aid system for EL(small) speech.

5.2. Speaking-Aid System for EL(small) Speech

The speaking-aid system using EL(small) speech is shown in **Figure 5.1**. The basic framework of this system consisting of uttering, recording, converting, and outputting are the same as other aid systems described in **Chapter 4**. The input and output acoustic features in EL(small)-to-Whisper and EL(small)-to-Speech are also the same as EL-to-Whisper and EL-to-Speech (see **Table 4.1**). Although the sound source unit is different as other aid systems, the attaching location of the EL(small) is the same as EL. This system is expected to be used not only for telecommunication but also conversations of face-to-face. Using this system, three concerns would be expected; (1) volumes of source signals and EL(small) speech, (2) auditory feedback to the user oneself, and (3) delay of the conversion.

One powerful feature of this system is the silence of the produced EL(small)

speech. Although the EL(small) is captured in extremely quiet rooms such as sound proof room, it is rare for users to use this system in such too quiet scenes. There are usually some background noises in our lives such as electric furniture of air-conditioner. As the result, it is expected that such background noises would mask the EL(small) speech. Next concern is that such mask effect also works not only for people around the speaker but also the speaker oneself. In other words, the speaker might be difficult to hear one's own voice. Even though the speaker could capture the EL(small) speech, he or she would hear both EL(small) speech and converted speech that would make the speaker confusing since the current VC framework for alaryngeal speech is not specified for real-time framework. As the result, the speaker would too difficult to hear his or her own EL(small) speech or would hear delayed those speech. Feedback mechanism often plays important rolls for a subject to continuously monitor the subject's action by giving back the result to the subjects. Auditory feedback is one of the important feedback information about speech, that carries the implicit idea that speakers listen to the sound of their voice and send the result of this perception back to the brain in a level where this result can be compared with the production the speaker intended to produce [73]. Yates have pointed out at least three feedback information for the subject [74]; kinesthetic and proprioceptive feedback from changes in the muscular and sensory apparatuses involved in speaking and listening, auditory feedback transmitted via the physical structures of bones or muscles in uttering, and another auditory feedback transmitted through the air to the speaker's ears. Since some muscles around the neck of most of laryngectomees would be removed in the laryngectomy, they are difficult to obtain second in-body auditory feedback described above as same as non-laryngectomees do. The mask effect for the EL(small) speech by background noises prevents the third auditory feedback described above. Moreover, even if the speaker could obtain the third auditory feedback, the articulation would be modified by hearing both produced EL(small) speech and converted speech in extremely short but not the same terms since the current VC framework used in aid systems in this thesis is not specified for real-time. As the result, the user needs well perceptible auditory feedback for the produced EL(small) speech to make the user's articulation stable in the use of this EL(small) aid system in the user's daily life. This thesis simply gives

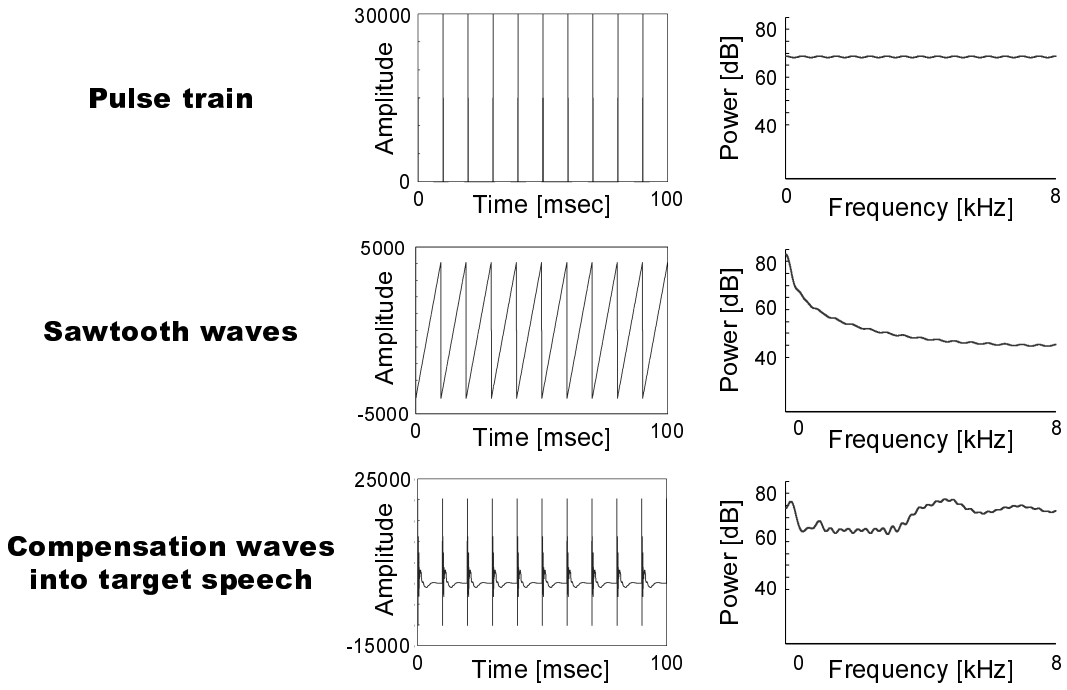

Figure 5.2. Designed sound source signals with different spectra.

the speaker recorded EL(small) speech directly to the user's one ear to make the user's articulation stable. The effectiveness of enhancing auditory feedback is experimentally evaluated in **Section 5.4**.

This thesis designed sound source signals from the view points of spectra and powers, independently. In case of changing the spectra, this thesis designed three types of sound source signals of (1) pulse train, (2) sawtooth waves, (3) compensation waves to whispering. The waveforms and spectra of these signals are shown in **Figure 5.2**. These sound source signals are designed in the following concepts:

1 Pulse train

Pulse train only has the amplitude at the first sample of the cycle. This signal is designed as one of the simplest signals as sound source excitations. As the **Figure 5.2** shows, this signal has equally powers for all components in the frequency domain.

2 Sawtooth waves

Sawtooth wave is one of asymmetric triangular waves that the power is going down as the time goes. Sawtooth waves showed in the **Figure 5.2** are designed since it is said that vocal folds vibrations of humans are approximated by asymmetric triangular waves. As the figure shows, this signal includes larger powers of acoustic features in lower frequency components. Since NAM microphone is suitable to capture lower frequency components, it is expected to be an advantage for users to clearly listen to their own auditory feedback. Moreover, it has certain amplitude over the time sequence, and therefore, the dynamic range of the power is larger than pulse train. As the result, it makes possible to design other sawtooth waves with different powers.

3 Compensation waves

The VC method used in thesis optimizes model parameters by EM algorithm that is related to initial parameters. Compensation waves are designed so that more suitable initial parameters in VC are obtained. As the result, the VC accuracy might be improved by making the initial parameters close to other ones after maximum likelihood estimation. This thesis tries to design compensation waves into whispering. The flow chart of designing the compensation waves is shown in **Figure 5.3**.

Figure 5.4, **Figure 5.5**, and **Figure 5.6** show examples of waveforms, spectrograms, and F_0 contours produced using those sound source signals with different spectra and with the same power, respectively. As these figures show, EL speech using pulse train includes almost the same powers in almost all frequency components. Another EL speech using the sawtooth waves includes much spectral intensity in lower frequency components. On the other hand, the other EL speech using compensation waves into whispering has much spectral intensity over 4000 Hz compared to other two EL speech signals. **Figure 5.7** shows histograms of normalized powers of 50 utterances of EL(small) speech using those three different small sound source signals, in which left groups show power histograms of silence parts and the other groups show those of speaking parts. As the figure shows, although the EL(small) speech using the compensation waves includes much spectral intensity over 4000 Hz, it totally has smaller powers than other EL(small) speech signals. Moreover, EL(small) speech using the sawtooth waves totally has larger power than others. This tendency is related to the spec-

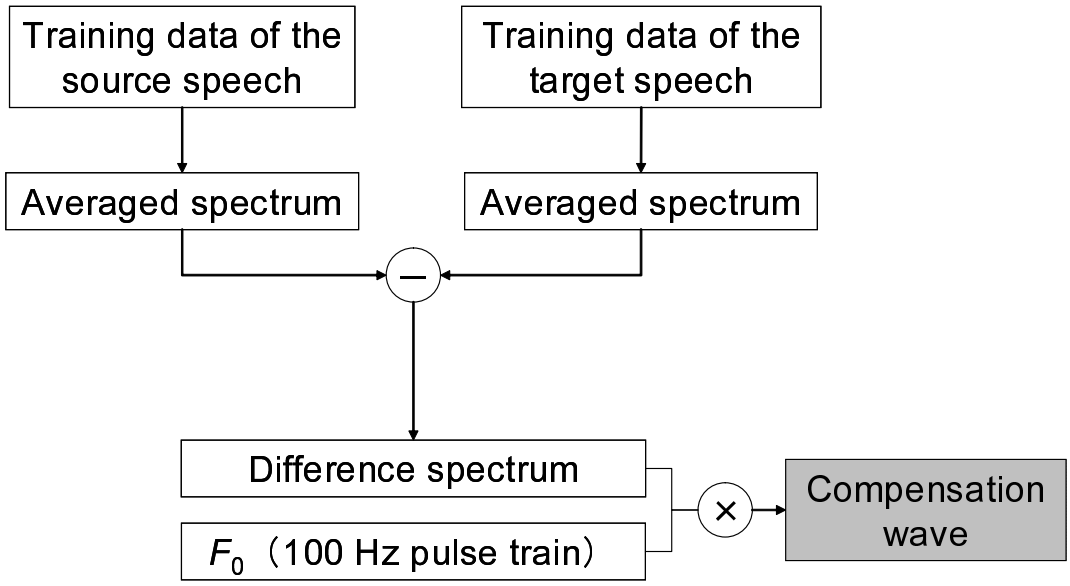

Figure 5.3. Flow chart of designing compensation waves into target speech.

tral intensity included in the lower frequency components of the sound source signals.

In case of changing the powers, the spectrum is fixed to be the sawtooth waves because it has the largest dynamic ranges among those three different spectra. **Figure 5.8** shows histograms of normalized powers of 50 utterances of three kinds of speech signals. One of the histograms comes from EL(small) speech using the sawtooth waves which power is the same as the power of the pulse train. This is the basic power among power variations. Another one comes from another EL(small) speech using sawtooth waves which power is larger than the previous one. The other histogram comes from EL speech using a conventional EL to be the higher limit of the power. As the figure shows, EL(small) speech using the large-powered sawtooth waves has larger powers than the other EL(small) speech using the basic-powered sawtooth waves; however, the power of the EL(small) speech is still small than that of the EL speech. **Figure 5.9** shows other histograms of normalized powers of 50 utterances of three kinds of speech signals. One of the histograms comes from EL(small) speech using the sawtooth waves which power is the same as the power of the pulse train. Another one comes from another

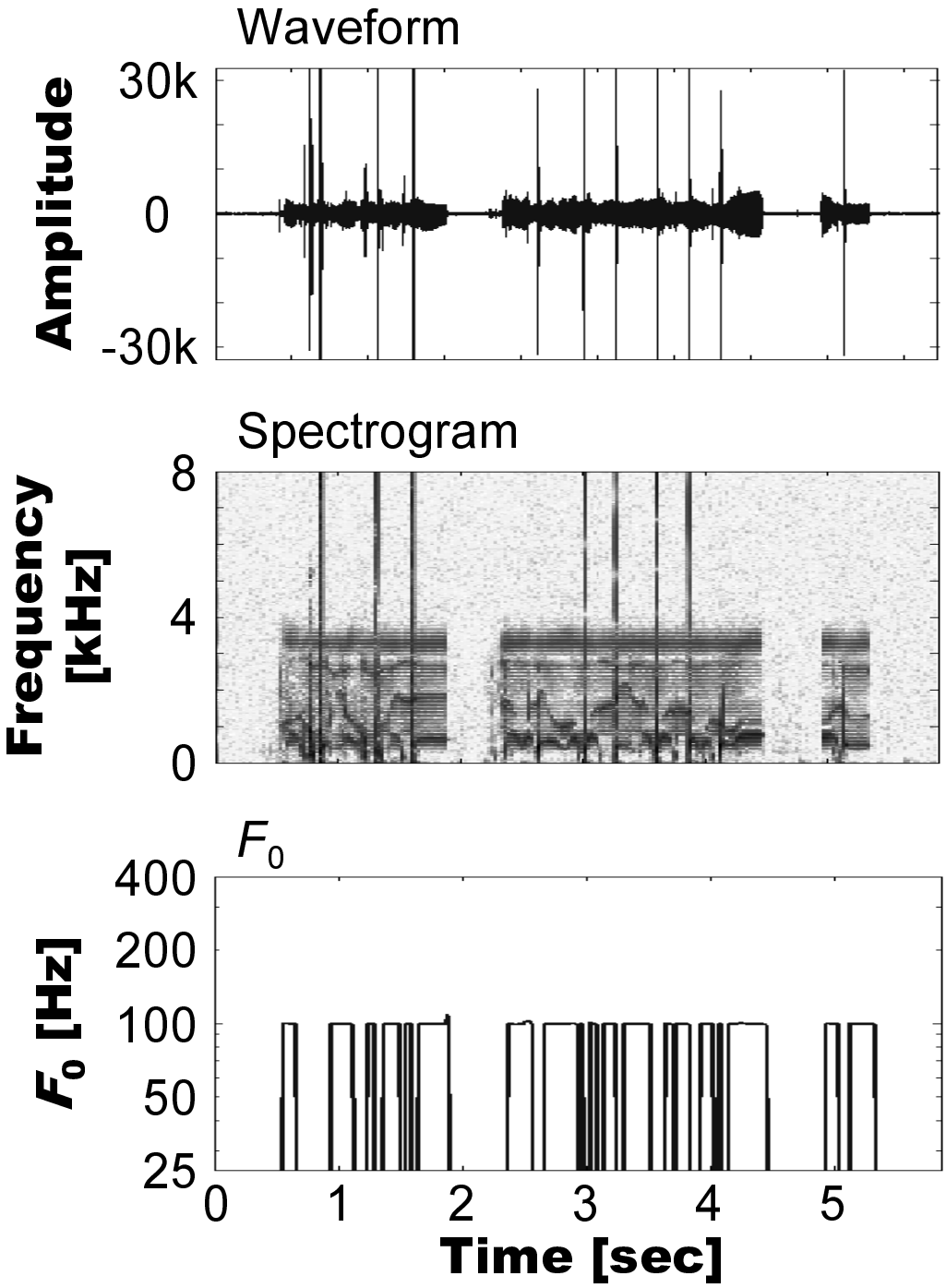

Figure 5.4. Example of waveforms, spectrograms, and F_0 contours using pulse train produced by a non-laryngectomee.

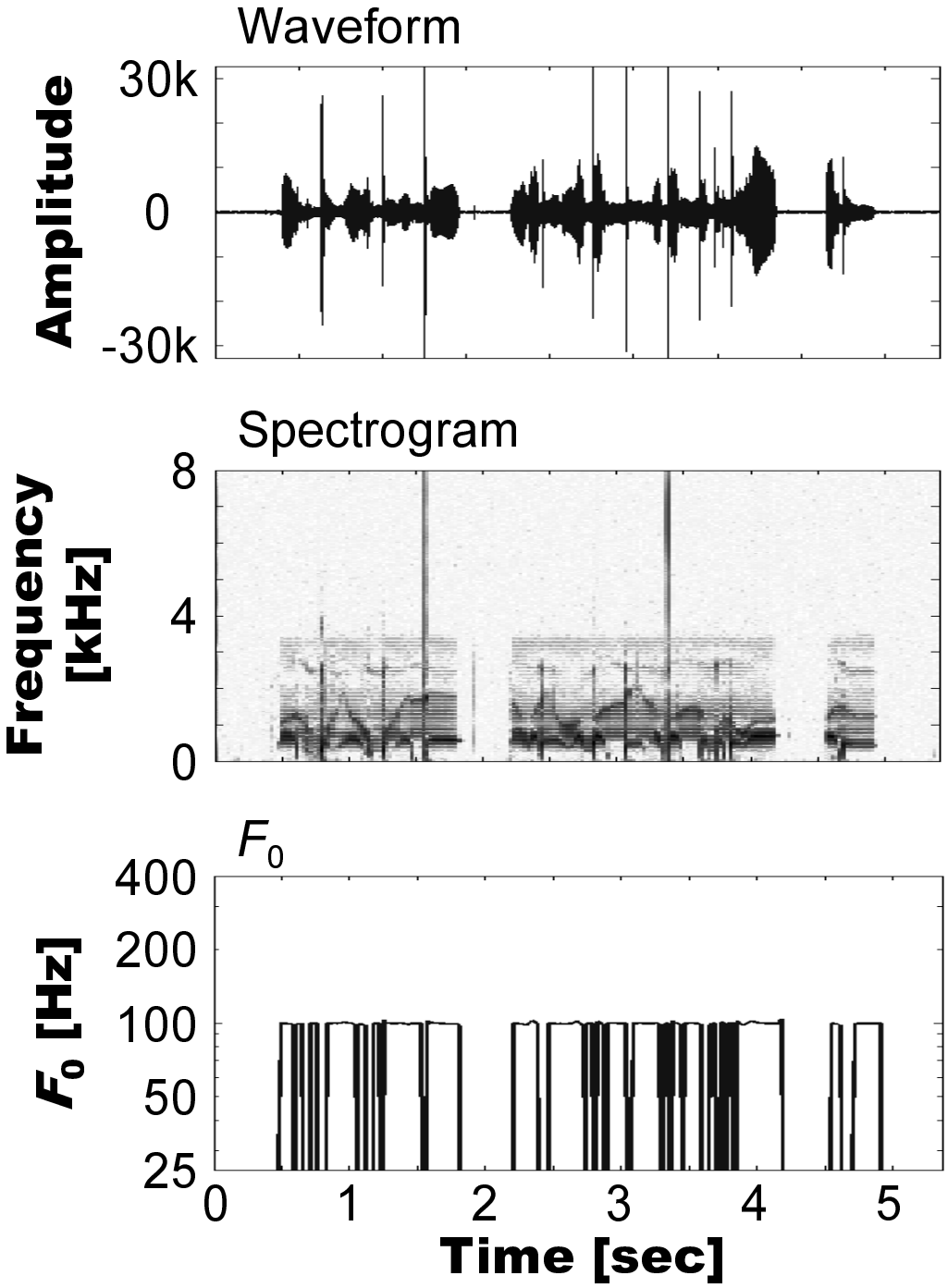

Figure 5.5. Example of waveforms, spectrograms, and F_0 contours using sawtooth waves produced by a non-laryngectomee.

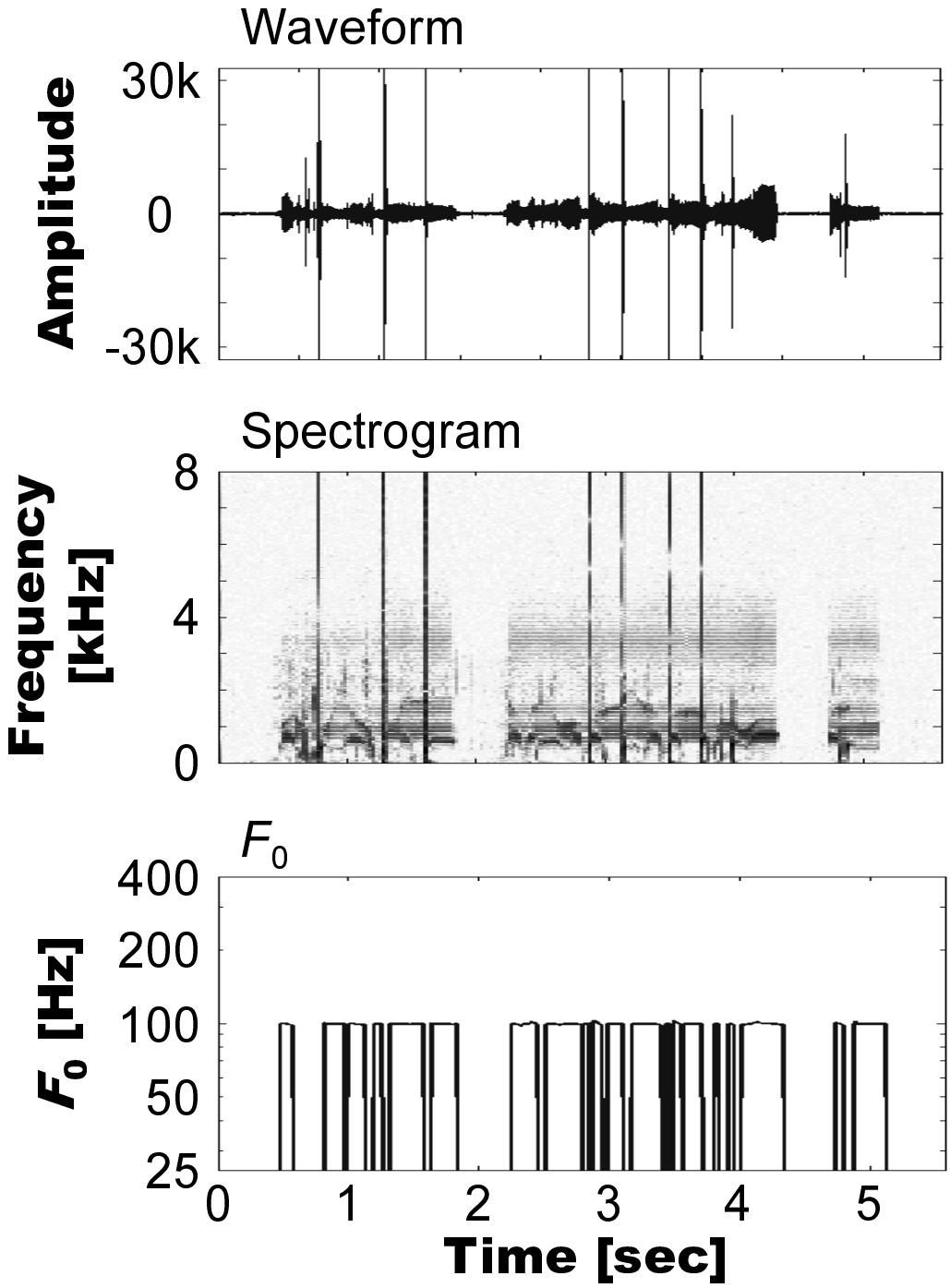

Figure 5.6. Example of waveforms, spectrograms, and F_0 contours using compensation waves into whispering produced by a non-laryngectomee.

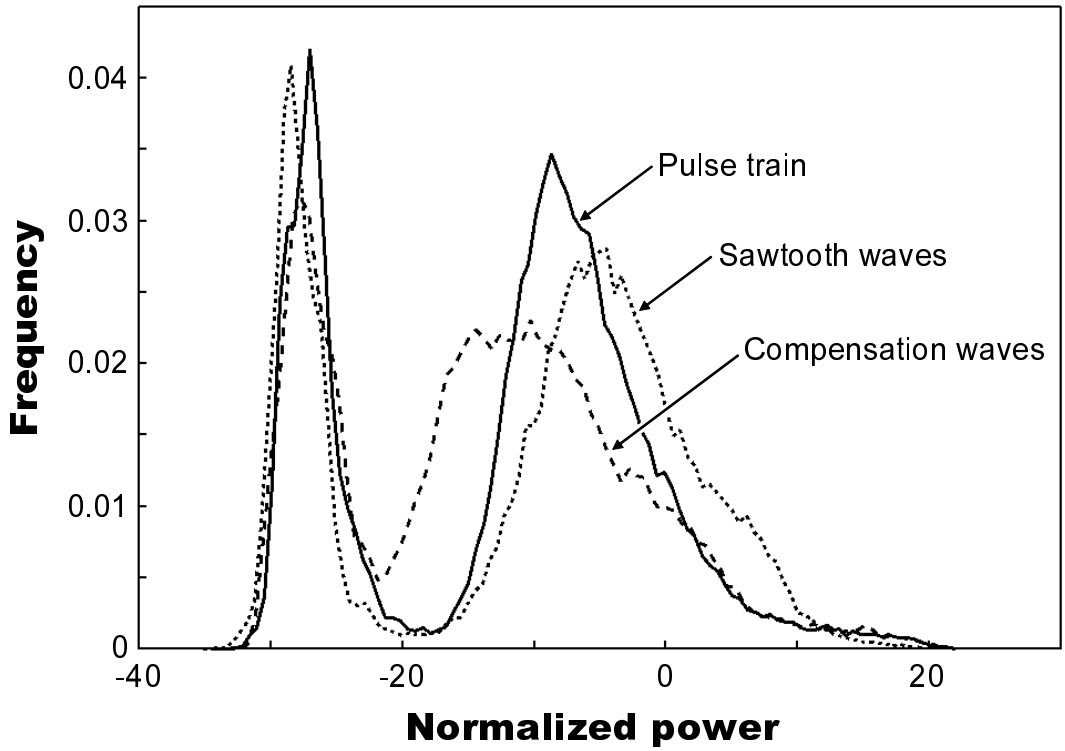

Figure 5.7. Histograms of normalized powers of EL(small) speech using pulse train, sawtooth waves, and compensation waves into whispering.

EL(small) speech using sawtooth waves which power is extremely smaller than the previous one. The other histogram comes from speech without using any sound source signals, in which recording, the speaker only moves the mouth like his/her speaks. This speech is assumed to be the lower limit of the power. As the figure shows, it is difficult to distinguish the speaking and silence parts if the power of the sound source signals is too small.

5.3. Voice Conversion for EL(small) Speech

The basic idea of the VC for EL(small) is the same as other VC for alaryngeal speech described in this thesis. Because the source speech does not have effective F_0 information, EL(small) speech is converted to whispering or normal speech that is EL(small)-to-Whisper or EL(small)-to-Speech in this thesis, respectively.

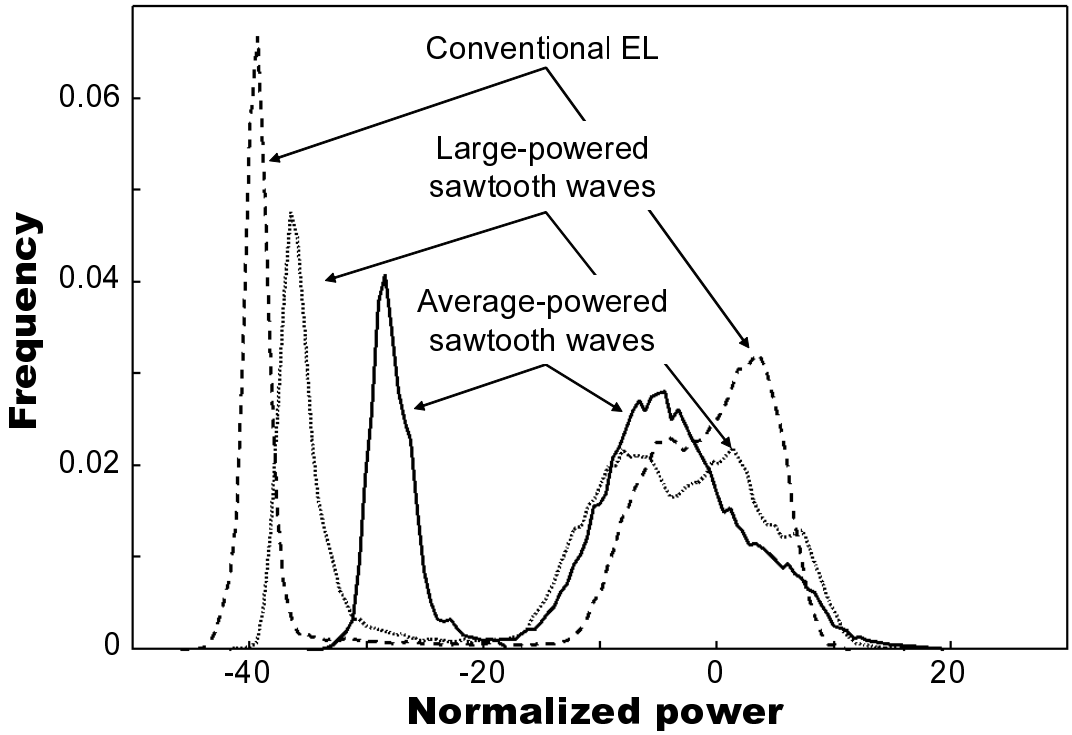

Figure 5.8. Histograms of normalized powers of EL(small) speech using average-powered sawtooth waves, large-powered sawtooth waves, and EL.

The way to constructing the training data of EL(small)-to-Whisper and EL(small)-to-Speech is the same as EL-to-Whisper and EL-to-Speech, respectively. In EL(small)-to-Whisper VC, only a GMM to estimate spectral information is trained, in which procedure joint feature vectors of source segmental feature vectors and target joint feature vectors of static and dynamic feature are modeled. In EL(small)-to-Speech VC, three GMMs each of which independently estimates spectrum, F_0 contours, and aperiodic components are trained. In the training procedure, spectral segmental feature vectors are set to source features of the three GMMs, and joint feature vectors of static and dynamic features of target spectrum, F_0 values, and aperiodic components are used to train the individual GMMs.

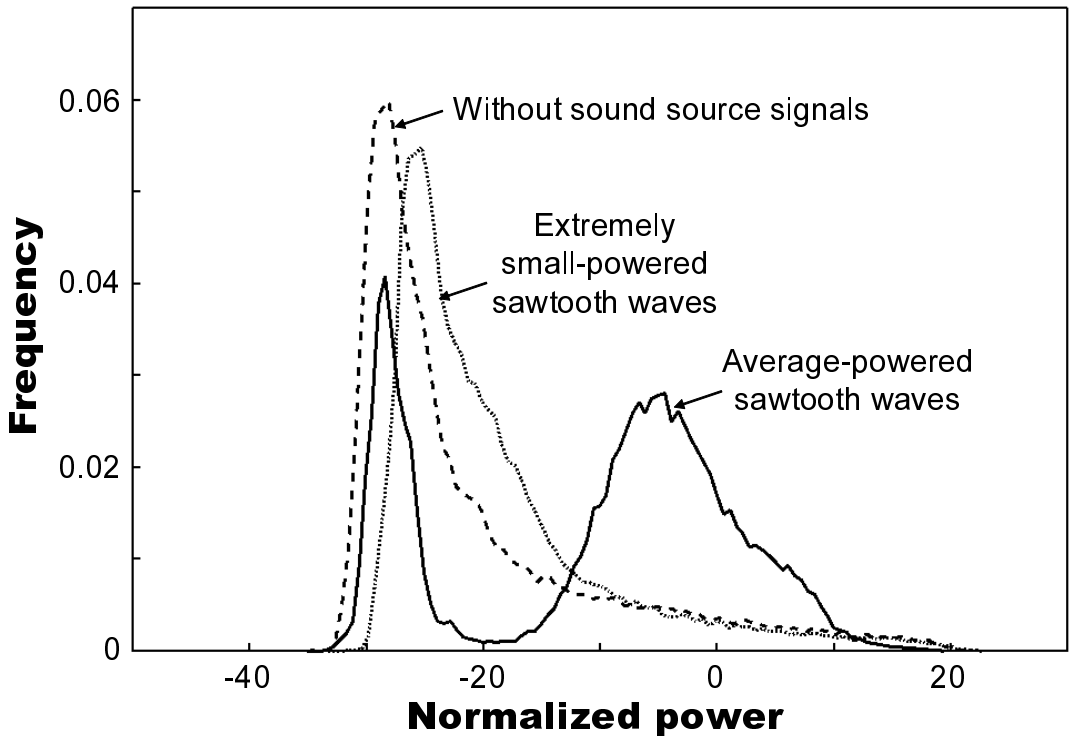

Figure 5.9. Histograms of normalized powers of EL(small) speech using average-powered sawtooth waves, extremely small-powered sawtooth waves, and articulation in which any sound source signals are not used when speaking.

5.4. Preliminary Experimental Evaluation of the Speaking-Aid System for EL(small) Speech

This section describes preliminary experimental evaluations using imitated EL(small) speech produced by a non-laryngectomee. This evaluation is intended to investigate following two items:

- Effectiveness of enhancing auditory feedback
- Impact of using different sound source signals for the VC accuracy

First, the effectiveness of enhancing the speaker's auditory feedback is investigated. And then, the second concern is investigated.

5.4.1 Effectiveness of enhancing auditory feedback

The effectiveness of enhancing auditory feedback is evaluated through the following two steps. The simple method to provide the speaker with amplified auditory feedback is firstly evaluated whether it is suitable to enhance the auditory feedback. Next, the effectiveness and the necessity of enhancing auditory feedback are evaluated by converting the EL(small) speech to normal speech.

Experimental conditions

The speaker was the same one non-laryngectomee of a Japanese male as **Section 4.4**. All speech data were recorded in a sound proof room, and an air-conditioner noise [75] was output as a background noise. As shown in **Figure 5.10**, the user spoke 50 cm away from two loudspeakers. The noise levels, which were 50 dBA, 55 dBA, 60 dBA, 65 dBA respectively, were controlled at the user's location. These situations were expected to simulate indoor noise environments in our daily life. The usual noise level in the recording room was almost 31 dBA. The user articulated using the pulse train described in **Section 5.2** to produce EL(small) speech. When enhancing the auditory feedback, EL(small) speech passed through an amplifier was directly given to the user's one ear with a closed-type ear phone. The volume of the amplified auditory feedback was controlled to each noise level so that the user was the most comfortable to listen. When the auditory feedback was not enhanced, the user got the auditory feedback only from the produced EL(small) speech signals. The user read out nine sentences of newspaper articles for each condition, and totally 90 utterances (nine sentences, five noise levels, and with/without auditory feedback) were recorded that was used to evaluate the simple method to enhance the auditory feedback. The speech data were recorded with 48000 Hz sampling and 24 bit for each sample.

The stability of user's articulation with or without amplified auditory feedback was subjectively evaluated by eight non-laryngectomees. They evaluated recorded all 90 sentences with five scaled opinion score (1: Bad, 2: Poor, 3: Fair, 4: Good, 5: Excellent). The audio format of the stimuli was 16000 Hz sampling and 16 bit for each sample. Stimuli were randomly given to both ears by a headphone in the recording room.

In order to evaluate the effectiveness of the enhancing the auditory feedback, a VC experiment was also conducted. In this evaluation, EL(small) speech using the pulse train was set to the source speech and whispering recorded using a headset microphone was set to the target speech. The same speaker as the above experiment additionally recorded 75 sentences including 49 phoneme-balanced sentences for the training and 26 newspaper articles for the test. The noise level was set to 50 dBA or 55 dBA, respectively. The other recording conditions were the same as described above.

The 0th through 24th mel-cepstral coefficients [71] were extracted from the source and the target speech, respectively. In order to alleviate the degradation of the VC accuracy due to the lack of information because of recording the speech using a NAM microphone, segmental feature vectors were used in the previous study of VC from NAM to normal speech [15]. Therefore, this thesis also used spectral segmental feature vectors, which were constructed by the following frame-by-frame procedures: a vector was prepared by concatenating the extracted static feature vectors at a current $\pm$ eight frames, and then, the dimension of the concatenated feature vector was reduced by principal component analysis (PCA). Finally, 50-dimensional segmental feature vectors were established as the source data. A joint feature vector consisting on the static and the delta of the first-order information was constructed at each frame to be set to the target data. The number of GMM mixture components was set to 32. The mel-cepstral distortion between the target and the converted mel-cepstra, which was given by equation (4.1), was used as an evaluation metric of the voice conversion accuracy.

Experimental result

Figure 5.11 shows the result of the subjective evaluation. In case of not giving the amplified auditory feedback, the stability of the user's articulation is obviously degraded. On the other hand, the user can utter with stable articulation by getting his auditory feedback explicitly. **Figure 5.12** shows a waveform example of utterances with or without amplified auditory feedback. In case of not giving the amplified auditory feedback, many bursts supposed to be appeared in consonants are suppressed. On the other hand, the user can clearly articulate even under the large background noise environment by getting his amplified au-

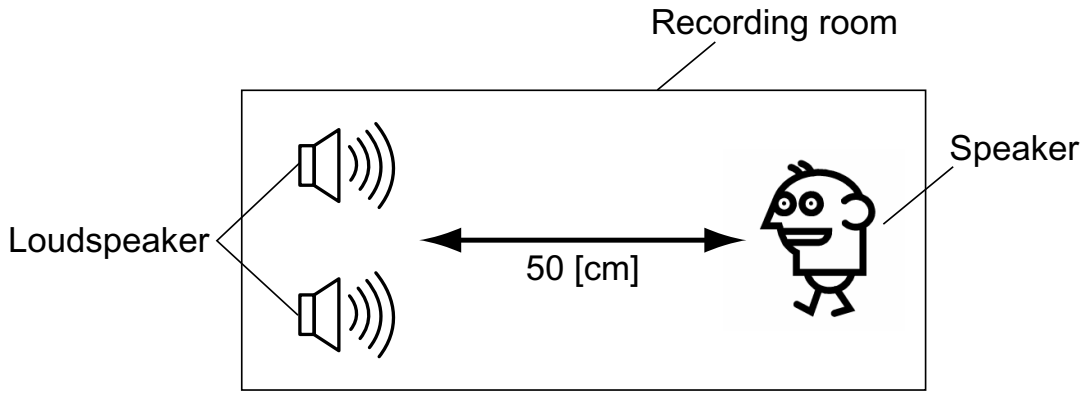

Figure 5.10. Speech recording condition under existing background noises.

ditory feedback with the ear phone. From these result, the user needs to receive one's amplified auditory feedback from the system to make the articulation stable.

Figure 5.13 shows the result of a voice conversion experiment. When the auditory feedback is not enhanced, the voice conversion accuracy is significantly degraded. This degradation is still observed even if the noise level is set to 50 dBA. On the other hand, the auditory feedback enhancement yields significant improvements of the voice conversion accuracy. Therefore, our method of enhancing the auditory feedback is significantly useful to record the EL(small) speech used for training the conversion model.

5.4.2 Robustness of VC for several small-powered sound source signals

Experimental conditions

The speaking-aid system for EL(small) speech was experimentally evaluated as a preliminary test. Employing the result of previous evaluation, all EL(small) speech described in this subsection was produced by giving the speaker amplified auditory feedback.

In order to investigate the robustness of VC for sound source signals, several EL(small) speech was converted to whispering, in which the spectrum and the power of the sound source signals were independently changed. Three different

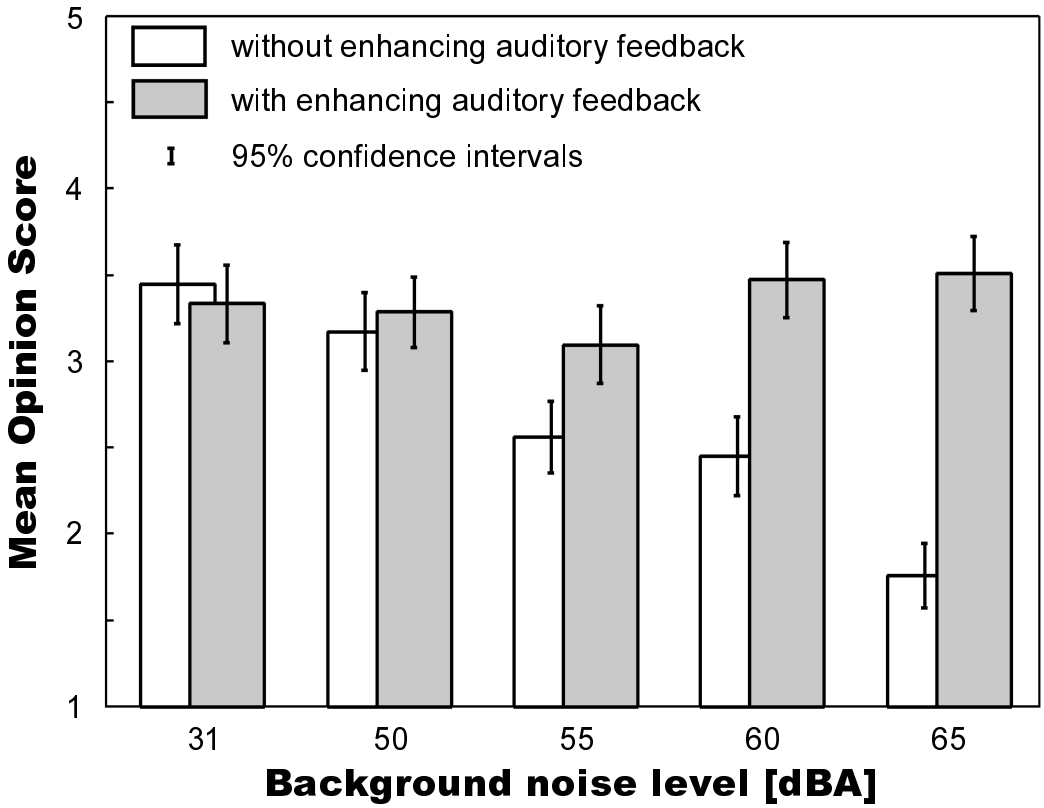

Figure 5.11. Mean opinion score of the stability of articulation under existing background noises.

spectra of pulse train, sawtooth waves, and compensation waves into whispering were evaluated as the influences of the spectrum changes. Six different powers of the sawtooth waves of -27 dB, -18 dB, -9 dB, 0 dB, +9 dB, and +18 dB were evaluated as the other influence of the power changes. 0 dB, which was the basic power of the small-powered sound source signals, means the same averaged power as the pulse train as described in **Section 5.2**. Totally 10 kinds of different imitated alaryngeal speech were evaluated to find out which sound source signals would degrade the VC accuracy. Frequencies of all small excitations were set to 100 Hz. Cross validation in which 50 newspaper articles were for training, and other 20 ones were for evaluation is conducted. Other experimental conditions were the same as evaluations conducted in **Section 4.4**.

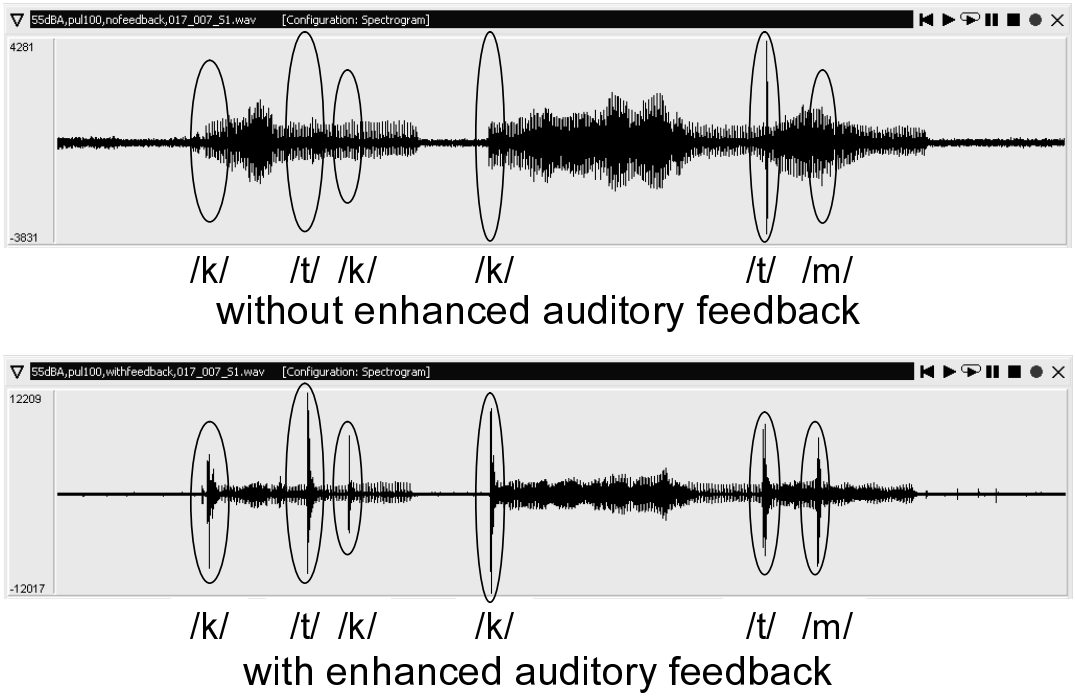

Figure 5.12. Example of EL(small) speech with or without amplified auditory feedback under existing background noises.

Not only objective but also two subjective evaluations were conducted for EL(small)-to-Whisper. One subjective evaluation was to evaluate the system of EL(small)-to-Whisper, and the other was to investigate influences due to using different sound source signals.

Five-scaled opinion score (1: Bad, 2: Poor, 3: Fair, 4: Good, 5: Excellent) was used as the measurement of the subjective evaluation. Stimuli were following seven kinds of converted whispering:

(1) converted whispering from only articulation without using any sound source signals

(2) converted whispering from EL(small) speech using minimum power of the sawtooth waves

(3) converted whispering from EL(small) speech using basic power of the sawtooth waves

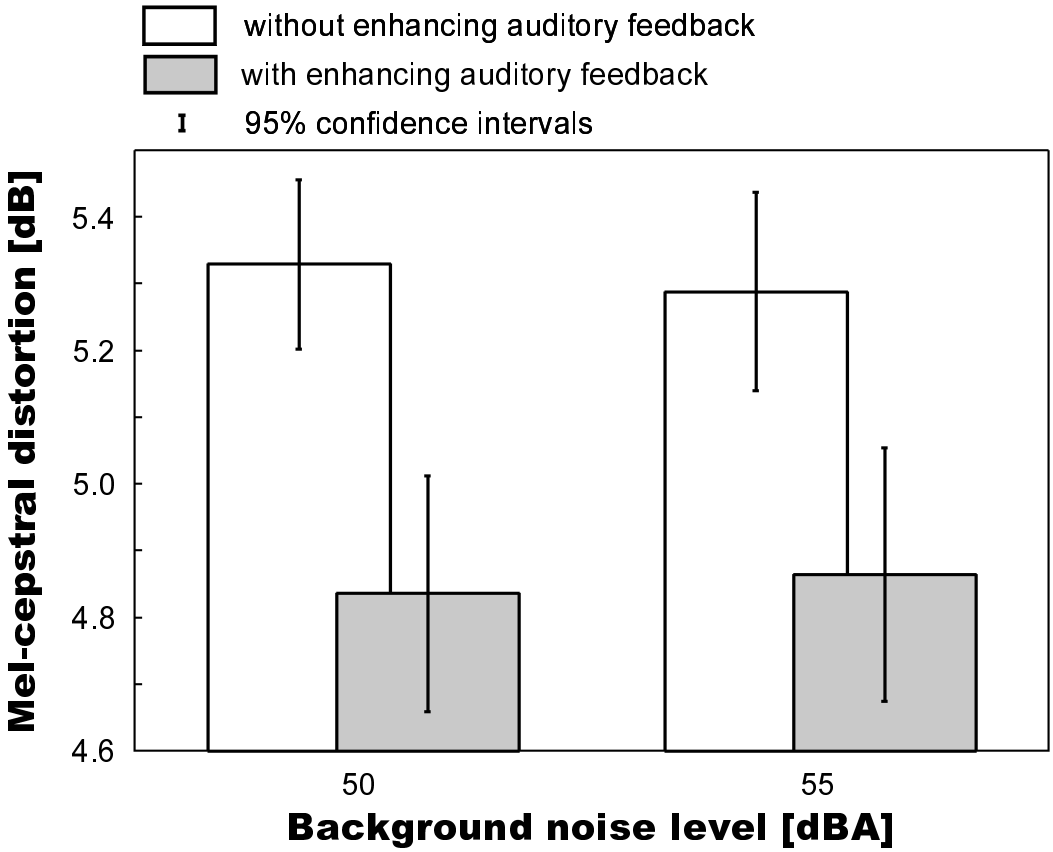

Figure 5.13. Mel-cepstral distortions without power information under existing background noises.

(4) converted whispering from EL(small) speech using the pulse train

(5) converted whispering from EL(small) speech using the compensation waves into whispering

(6) converted whispering from EL(small) speech using maximum power of the sawtooth waves

(7) converted whispering from conventional EL speech

Totally 140 utterances (20 utterances, 7 source signals) were randomly evaluated by five non-laryngectomees with respect to the naturalness.

Experimental result

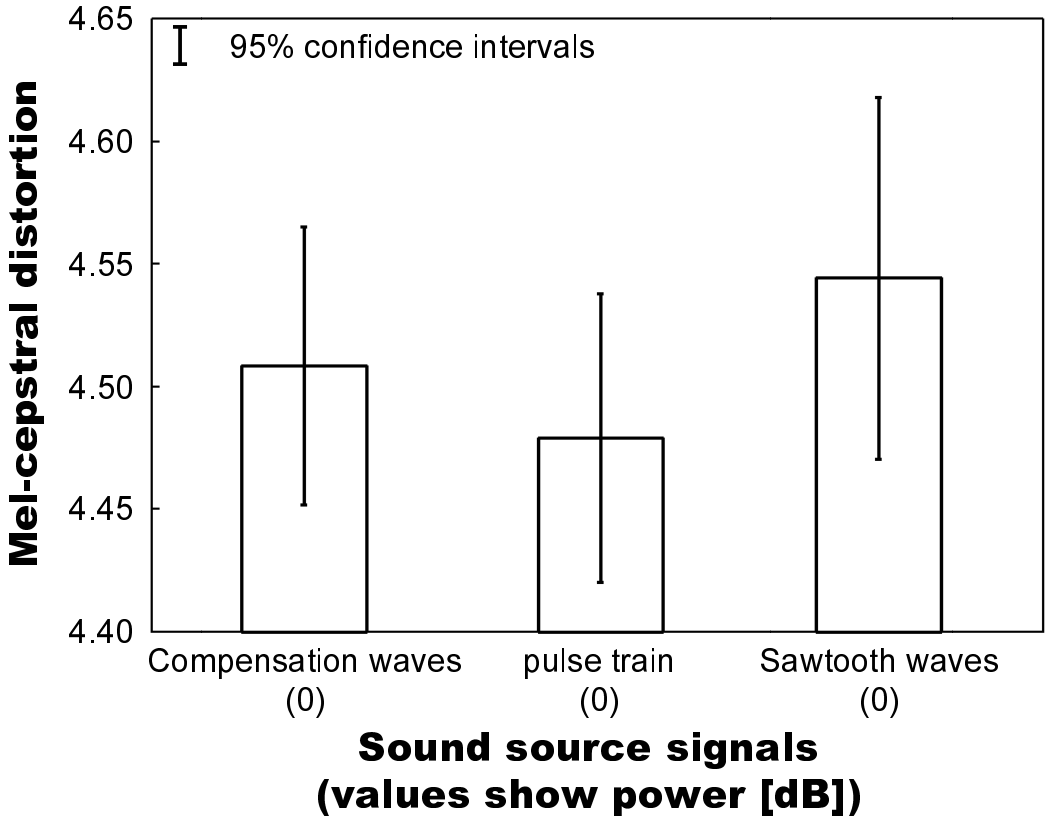

Figure 5.14. Mel-cepstral distortion of EL(small) speech without power information using sound sources with different spectra.

Figure 5.14 shows averaged mel-cepstral distortions utterance by utterance for EL(small)-to-Whisper, in which influences of different spectra are investigated. **Figure 5.15** shows other averaged mel-cepstral distortions, in which influences of different powers are investigated. The results show that the VC accepts different spectra of sound source signals. Changing the power of the sound source signals shows a tendency that the distortion is getting larger as the power going down especially between the articulation and the power of -18 dB. On the other hand, variations of the results cover individual results. Finally, in the objective evaluation, this thesis regards that once sound source signals are given to produce EL speech, VC using sound sources with different powers works in almost the same accuracy.

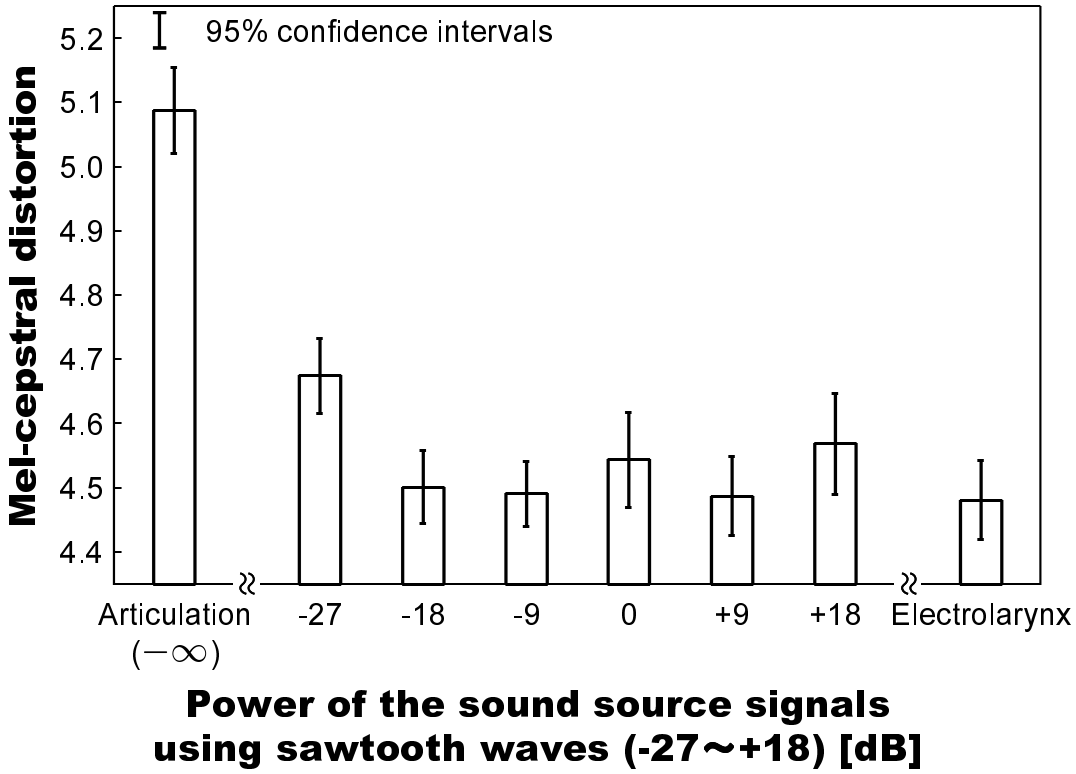

Figure 5.15. Mel-cepstral distortion of EL(small) speech without power information using sound sources with different powers.

Table 5.1. Averaged mel-cepstral distortion for EL-to-Speech in which imitated EL speech produced by a non-laryngectomee is used as the source speech

Source/Converted	With power information [dB]	Without power information [dB]
Source	9.42	8.43
Converted	4.73	3.99

Table 5.1 shows other averaged mel-cepstral distortions utterance by utterance for EL-to-Speech. **Table 5.2** shows the other result about F_0 estimation. Although the correlation coefficient of F_0 contours calculated from only voiced frames of between converted and target normal speech is not high, F_0 values with certain tones are estimated.

Table 5.2. Voiced or unvoiced error rates and correlation coefficients for voiced frames for EL-to-Speech between converted F_0 values estimated from imitated EL speech produced by a non-laryngectomee and target ones. Notations are same as those in **Table 4.3**

Correlation coefficients	0.317 ± 0.105
V → V	41.92 %
U → U	48.96 %
V → U	7.09 %
U → V	2.04 %

Figure 5.16 shows subjective results of using different spectra and powers of the sound source signals. The result for EL(small) speech using different spectral shows the same tendency as the objective result. The VC accuracy is almost the same within the spectral changes. On the other hand, different tendencies from the objective results are seen in case of power changes. The quality is slightly degraded as the power going larger; however, the mean opinion score keeps around three. This thesis regards that this result is acceptable. The quality degradation is seen as the power goes down. Especially using -27 dB of the sawtooth waves, the converted voice quality is degraded. Considering objective and subjective results, this thesis expects that VC accepts large variety of sound source signals with different spectra and powers except the case that the power of the speaking part is almost the same as that of the silence part.

5.5. Conclusion

To suppress the volume of ELs, this thesis used another sound source unit that outputs arbitrary source signals with extremely small power as EL(small). The produced EL(small) speech was preliminarily converted to whispering or normal speech as other speaking-aid systems of EL(small)-to-Whisper or EL(small)-to-Speech, respectively using the imitated EL(small) speech produced by a non-laryngectomee. In EL(small)-to-Whisper, only spectral features were estimated

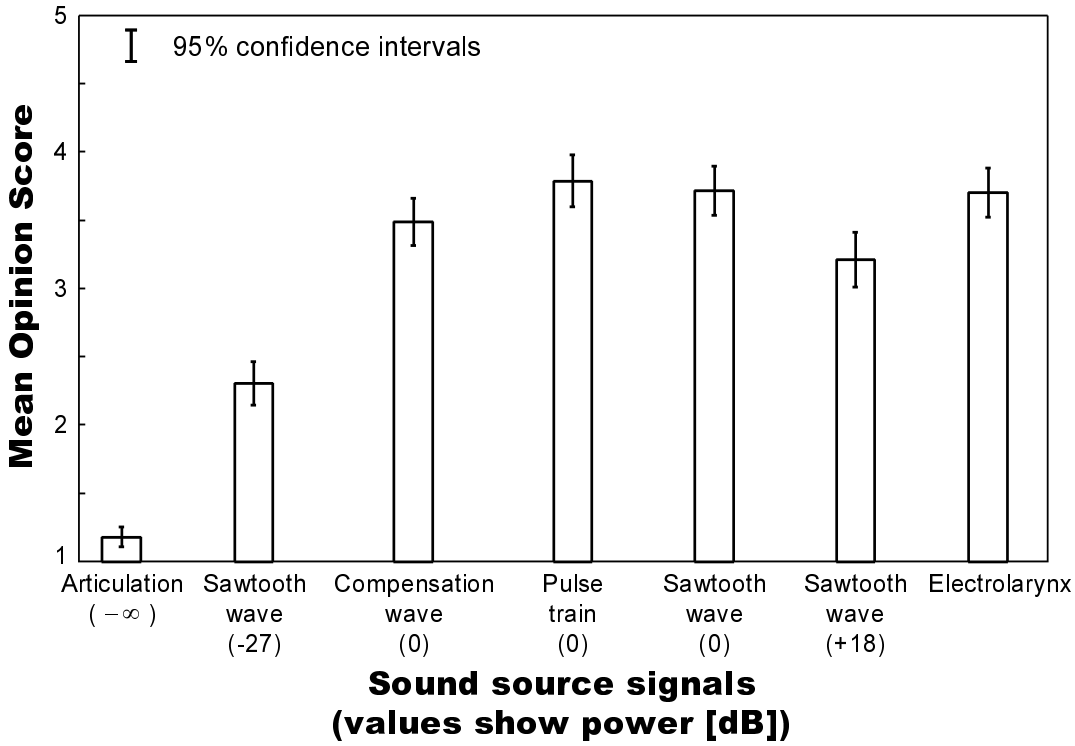

Figure 5.16. Mean opinion score of EL(small) speech using sound sources with different spectra and powers.

from the source spectral features. In EL(small)-to-Speech, spectral, F_0, and aperiodic features were estimated from only the source spectral features. Three sound source signals were; pulse train, sawtooth waves, and compensation waves to make the spectral features of EL(small) close to those of whispering. From the preliminary experimental evaluation, VC accepted all of these signals except the case in which the power of the speaking parts was the almost the same as that of the silent parts.

Chapter 6

Experimental Evaluations

This chapter experimentally evaluates all conversion frameworks of speaking-aid systems proposed in this thesis; EL-to-Whisper, EL-to-Speech, EL(air)-to-Speech, EL(small)-to-Whisper, and EL(small)-to-Speech. The speaker is one male laryngectomee who is proficient at using an external utterance device. Objective and subjective evaluations achieve the enhancement of all kinds of EL speech. Unfortunately, intelligibility is slightly degraded. On the other hand, naturalness of all converted speech is dramatically high compared to that of source speech. Finally, converted speech is preferred to the source speech, which indicates the effectiveness of the VC for alaryngeal speech using an external device. The effectiveness of using an air-pressure sensor is additionally evaluated.

6.1. Introduction

Table 6.1 shows advantages of the all types of speaking-aid systems proposed in this thesis.

In order to evaluate three types of speaking-aid systems using EL, EL(air), and EL(small) speech, respectively, this chapter conducts objective and subjective evaluations. This thesis uses a laryngectomee's data to experimentally investigate following concerns:

(1) Is VC objectively and subjectively effective for EL speech enhancement using different sound source signals?

Table 6.1. Advantages and effective use of individual speaking-aid systems proposed in this thesis

Sound source unit	Output speech	System	Desired advantages
EL	Whispering	EL-to-Whisper	Speaking with natural voice
	Normal speech	EL-to-Speech	
EL(air)	Normal speech	EL(air)-to-Speech	Speaking with more natural voice
EL(small)	Whispering	EL(small)-to-Whisper	Speaking with natural voice without annoying others
	Normal speech	EL(small)-to-Speech	

(2) Is the air-pressure sensor effective in EL(air)-to-Whisper and EL(air)-to-Speech?

(3) Is F_0 information effective in EL(air)-to-Speech?

This chapter is organized as follows. In **Section 6.2**, three proposed speaking-aid systems are objectively evaluated. In **Section 6.3**, these systems are subjectively evaluated. Additional experimental evaluation is conducted in **Section 6.4**. This section is summarized in **Section 6.5**.

6.2. Objective Evaluations of the Speaking-Aid Systems

6.2.1 Experimental conditions

A source speaker was one Japanese male laryngectomee in his 50s. He was undergone the total laryngectomee more than 10 years ago, which meant his physical condition around the neck after the operation was stable. Not only his larynx but also most of his left-side muscles including his sternocleidomastoid were removed, where was a key location to attach NAM microphone. He always used EL in his daily conversations; therefore, he was a proficient to utter using external sound

source unit. One non-laryngectomee was set to a target speaker who was a different from the source speaker. Both of source and target speakers recorded 50 phoneme-balanced balanced sentences for training data and uttered 30 newspaper utterances for test data. Five kinds of source speech signals were converted, which were EL speech, EL(air) speech, EL(small) speech using pulse train, EL(small) speech using sawtooth waves with the same averaged power as the pulse train, and EL(small) speech using sawtooth waves with larger averaged power as the pulse train. These source speech signals were converted to whispering or normal speech in each proposed aid system. Only EL(small) speech signals were recorded by NAM microphone, and all other utterances were recorded by a headset microphone. All speech data were recorded in 48000 Hz sampling with 24 bit for each sample. After down-sampling to 16000 Hz and down-bitrate to 16 bit for each sample, acoustic features were extracted. For target speech data, the 0th through 24th mel-cepstral coefficients, which were extracted by STRAIGHT analysis, were used as the target spectral parameters in which 0th coefficient captures power information. F_0 values and aperiodic components of the target speech were extracted by STRAIGHT analysis. Time-domain Excitation extractor using Minimum Perturbation Operator (TEMPO) [72] analysis was employed for F_0 extraction. For other speech data of target whispering and all kinds of source speech data, the 0th through 24th mel-cepstral coefficients [71] were used as the spectral parameters in which the 0th coefficient captures power information. F_0 values of the EL(air) speech were extracted by Robust Algorithm for Pitch Tracking (RAPT) [76]. Aperiodic components of the EL(air) speech were extracted by STRAIGHT analysis. The number of a GMM component to estimate spectral and aperiodic parameters was set to 32, respectively, and that of another GMM component to estimate F_0 parameters was set to 16, 32, 64, or 128. The segmental feature vector of source spectral data to estimate target spectral data was constructed by the following procedures; first, the current, previous and succeeding eight frames were concatenated into one vector, and then, the dimension of the concatenated vector was compressed by PCA procedures. Finally, 50-dimensional segmental feature vector was constructed frame by frame. For F_0 estimation, the frame length to construct spectral segmental feature vectors was set to 8.

Table 6.2. Averaged mel-cepstral distortions for all kinds of speaking-aid systems proposed in this thesis. Values in front of and behind the slash respectively shows distortions considering and not considering power information (i.e., 0^{th} coefficient). 'Sawtooth waves 1' means averaged power of sawtooth waves is same as that of pulse train, and 'Sawtooth waves 2' means another sawtooth waves including larger power compared to 'Sawtooth waves 1'

System	Source-Target	Converted-Target
EL(small: Pulse train)-to-Whisper	11.61 / 7.85	5.21 / 4.23
EL(small: Pulse train)-to-Speech	17.39 / 10.43	5.27 / 4.37
EL(small: Sawtooth waves 1)-to-Whisper	11.49 / 7.81	5.10 / 4.18
EL(small: Sawtooth waves 1)-to-Speech	17.00 / 10.04	5.07 / 4.21
EL(small: Sawtooth waves 2)-to-Whisper	12.41 / 8.44	5.41 / 4.35
EL(small: Sawtooth waves 2)-to-Speech	17.49 / 9.83	5.23 / 4.38
EL-to-Whisper	9.66 / 7.85	4.96 / 4.12
EL-to-Speech	10.63 / 8.12	4.77 / 4.02
EL(air)-to-Speech	12.28 / 9.05	4.90 / 4.08

The objective measures were the same as described in **Section 4.4.1**; mel-cepstral distortion for spectral evaluation, U/V error rate and correlation coefficient calculated from only voiced frames for the F_0 evaluation.

6.2.2 Experimental results

Table 6.2 shows averaged mel-cepstral distortion for all kinds of speaking-aid systems proposed in this thesis. As the table shows, all kinds of VC work well to reduce the mel-cepstral distortion even though the small-powered sound source signals are used.

Figure 6.1 show U/V errors between target and converted F_0 values, and **Figure 6.2** shows correlation coefficients between target and converted F_0 contours about only voiced frames for both data.

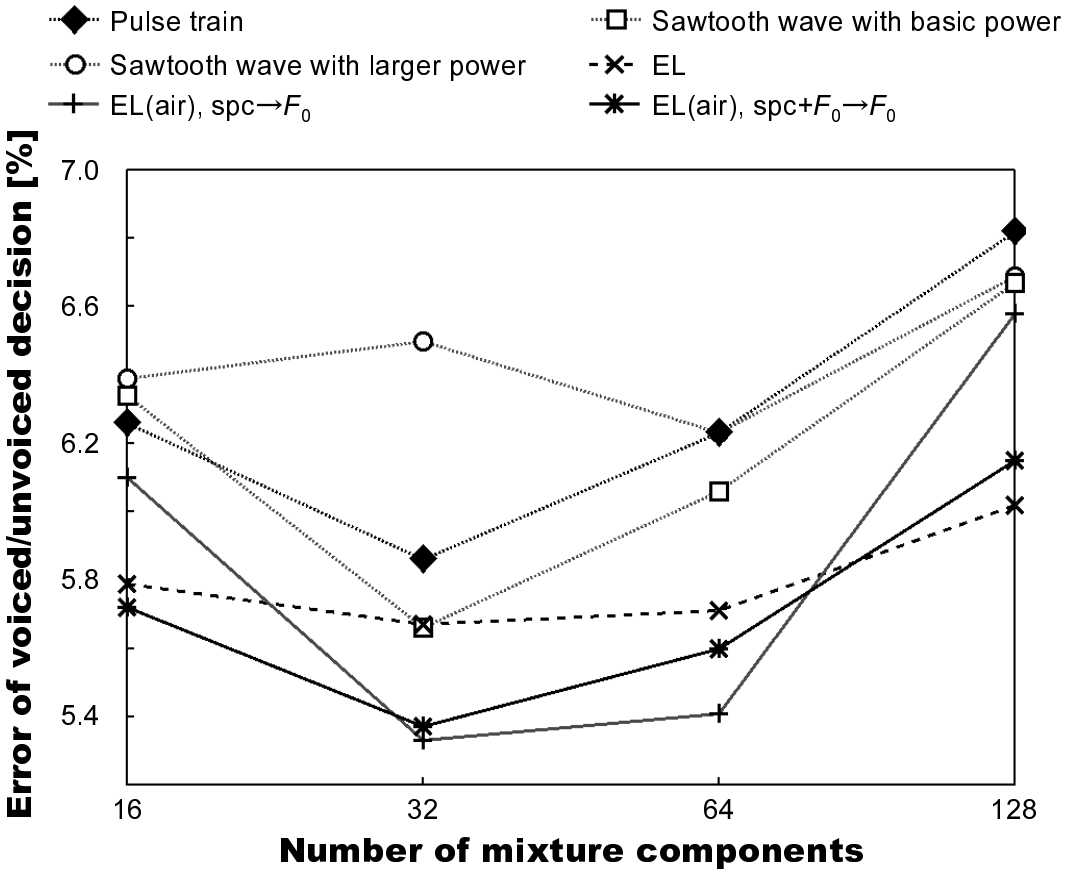

Figure 6.1. Errors of voiced or unvoiced decision for all speaking-aid systems proposed in this thesis. Basic power denotes same power as pulse train.

Figure 6.3, **Figure 6.4**, **Figure 6.5** show examples of waveforms, spectrograms, and F_0 contours of a set of source speech of EL speech, EL(air) speech, and EL(small) speech using pulse train, respectively, and converted normal speech. An example of the target speech is shown in **Figure 6.6**.

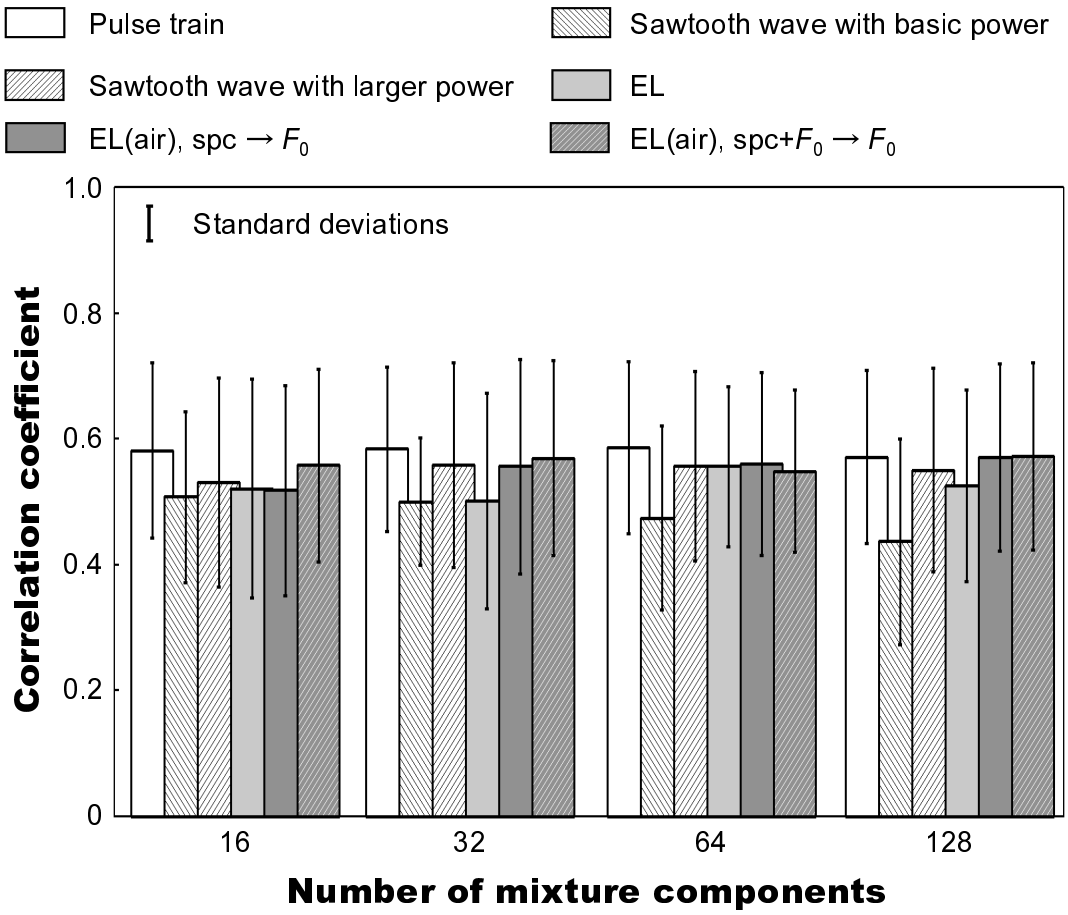

Figure 6.2. Correlation coefficients between converted and target F_0 contours about only voiced frames for both data. 'spc → F_0' denotes target F_0 contours are estimated using source spectral features.

6.3. Subjective Evaluations of the Speaking-Aid Systems

6.3.1 Experimental conditions

10 non-laryngectomees and one laryngectomee who was the source speaker himself evaluated intelligibility, naturalness, and preference with 5-scaled opinion score (1: awful, 2: bad, 3: fair, 4: good, and 5: excellent). Intelligibility was scored as how the contents of the stimuli could be understood by listeners. Naturalness

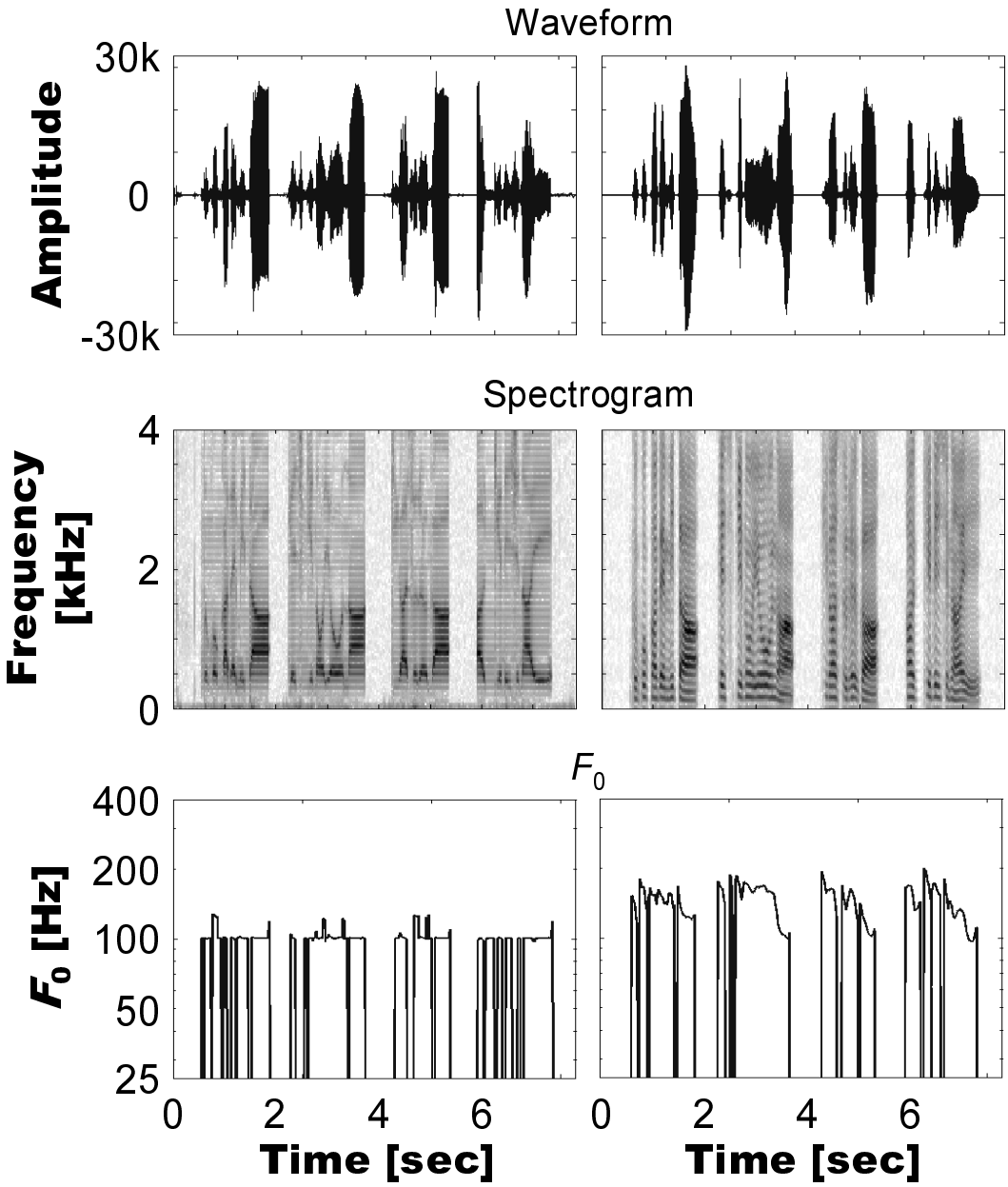

Figure 6.3. Examples of waveforms, spectrograms, and F_0 contours of EL speech produced by a laryngectomee and those of converted normal speech.

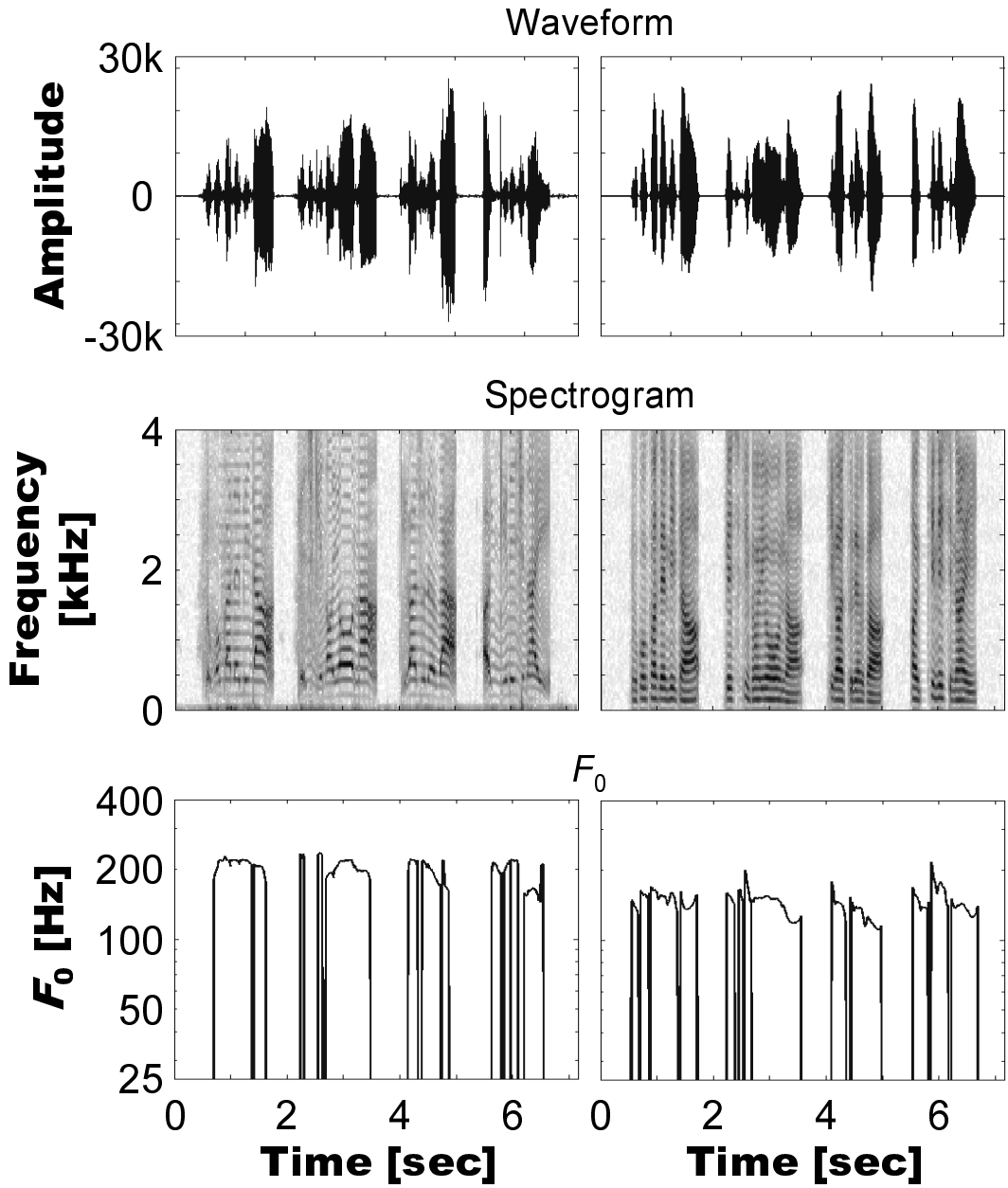

Figure 6.4. Examples of waveforms, spectrograms, and F_0 contours of EL(air) speech produced by a laryngectomee and those of converted normal speech.

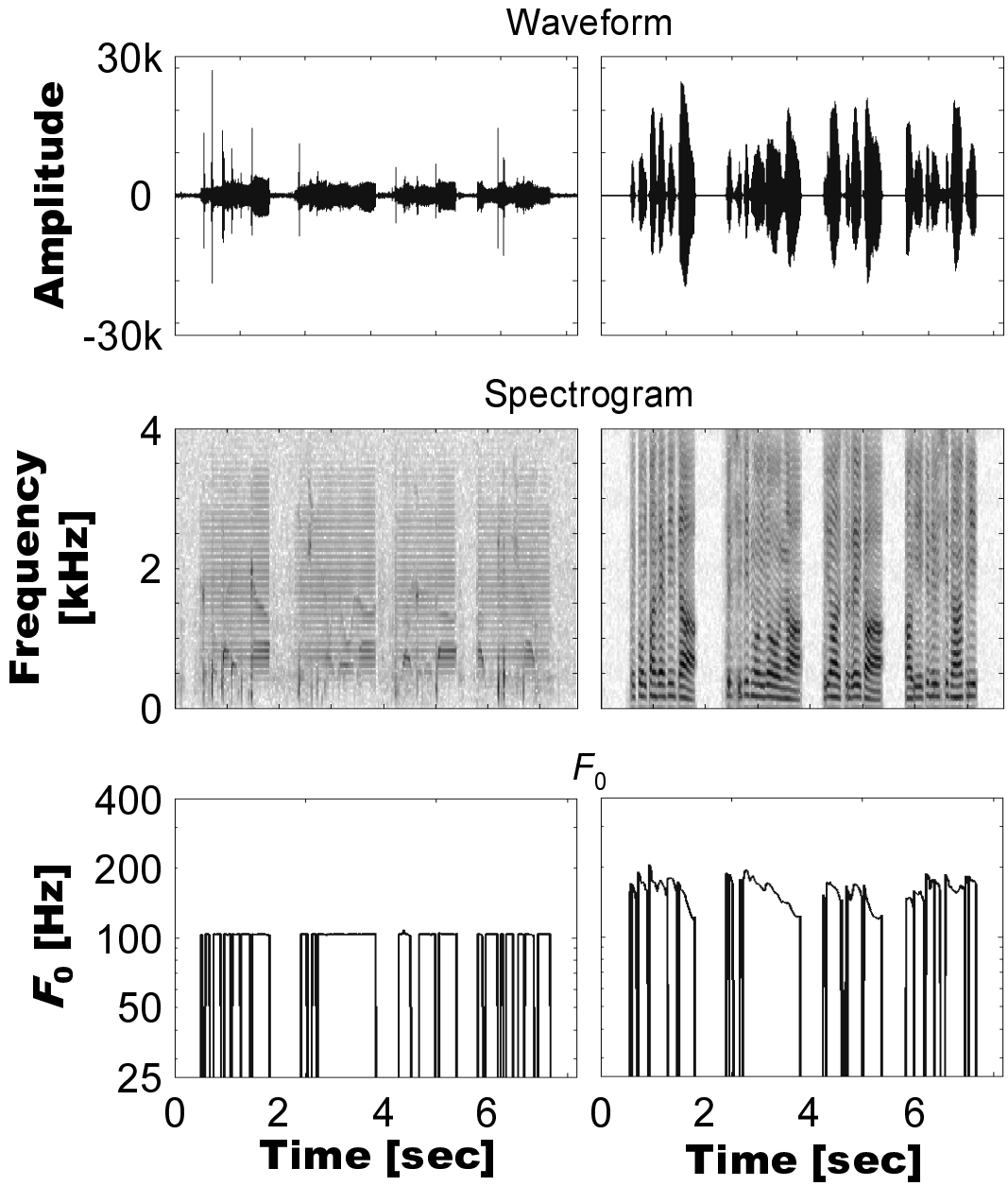

Figure 6.5. Examples of waveforms, spectrograms, and F_0 contours of EL(small) speech using pulse train produced by a laryngectomee and those of converted normal speech.

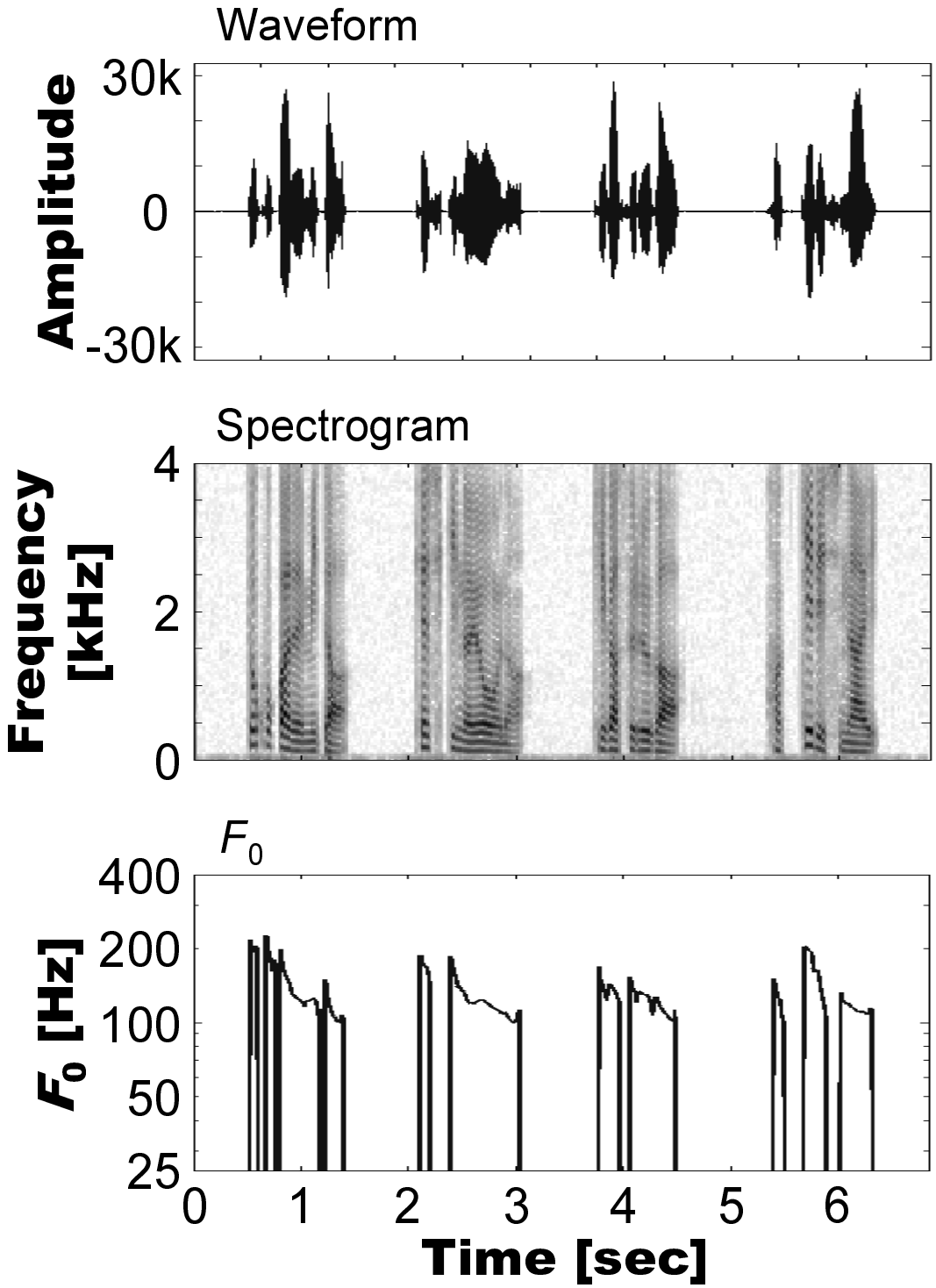

Figure 6.6. Example of waveforms, spectrograms, and F_0 contours of target normal speech.

was scored as how the stimuli were close to the human voices. Preference was scored as how listeners prefer to listen to the stimuli in their conversations in case of non-laryngectomees, and how the listeners preferred to speak with the stimuli in their conversation in case of the laryngectomee, respectively. Stimuli were following 10 speech signals:

(1) source EL speech

(2) converted whispering from the EL speech

(3) converted normal speech from the EL speech

(4) source EL(air) speech

(5) converted normal speech form the EL(air) speech

(6) source EL(small) speech produced using pulse train

(7) converted whispering from the EL(small) speech

(8) converted normal speech from the EL(small) speech

(9) target whispering after analysis-synthetic procedures

(10) target normal speech after analysis-synthetic procedures.

Three speech samples were randomly selected for individual stimuli, speakers, and items of speech qualities. As a result, 30 samples per listener were evaluated for each speech quality. The evaluations were conducted in a sound proof room, and all stimuli were provided to both ears of listeners by a head-phone.

6.3.2 Experimental results

Figure 6.7, **Figure 6.8**, and **Figure 6.9** show the results of intelligibility, naturalness, and preference test, respectively.

From the result of the intelligibility, slight degradation of the intelligibility for converted speech signals is shown. The source speaker has used an EL for more than 10 years, and therefore, he is a proficient to speak with EL speech. This is the reason why the intelligibility of the source EL and EL(air) speech is high

scored. The problem of this intelligibility degradation has not addressed yet, and this is one of future works. Although the intelligibility of converted speech signals from EL or EL(air) speech is degraded, the score is kept around three in mean opinion score. This thesis regards this as acceptable scores. The result of VC from EL(small) is worse than other two systems. One reason of this is thought the difficulty of recording EL(small) speech using current NAM microphone. NAM microphone used in this thesis employs soft silicone as the facial material between the microphone and the skins. Soft silicone does not have adherence property, and therefore, it needs another supplementary device to fix the microphone on the speaker's skins. Using this soft-silicone NAM microphone, special noises are mixed in the recorded signals when the speaker moves one's neck in uttering. Moreover, the source laryngectomee lost not only his larynx but also a part of his muscles. This derives ripple noises that has not been seen in the speech recording of the non-laryngectomee's imitation are observed on waveforms. This thesis thinks that these unnecessary noises unevenly mixed not only in silence parts but also speaking parts derives the result of VC from EL(small) speech. Although the intelligibility of converted EL(small) speech is not achieved to that of others, VC does work to enhance the original intelligibility.

From the result of the naturalness, drastic enhancement is observed in all kinds of converted speech. Although there are no statistical significances between the results of converted whispering and converted normal speech in VC from EL(small) and EL speech, the naturalness of the converted whispering is slightly higher than that of the converted normal speech. This is thought to be the result of avoiding the problem of errors of F_0 estimation, since F_0 is not estimated in the conversions to whispering. The results of VC from EL speech are slightly better than that of VC from EL(small) speech. This thesis thinks that major source of this problem is the reason described in the results of intelligibility. Comparing results between converted normal speech from EL speech and that from EL(air) speech, although the intervals of converted EL(air) speech is slightly smaller than the other one, the result is almost the same scored. The VC is significantly effective to enhance EL speech utterances, on the other hand, the result of the converted speech signals is far from that of target normal speech, and it is necessary for further enhancement.

From the result of the preference, all converted speech signals are higher scored than source speech. There are no statistical significances between results of the converted whispering and converted normal speech in VC from EL(small) and EL speech, converted normal speech is slightly better than converted whispering, which is the inverse trend of the result of naturalness. Whispering is generated by humans and people often speak with whispering; however, Situations for people to communicate with others using whispering is limited. This is thought to be the source problem of the result. There are also no statistical significances between converted speech from EL speech and that from EL(air) speech. As the result of this, the effectiveness of using the air-pressure sensor is not high. Although the laryngectomee had trained the usage of the air-pressure sensor for only one month, the training period of the air-pressure sensor is shorter than that of the EL. The result might be changed as the training efforts of using the air-pressure sensor by the speaker. Moreover, other devices that can control the F_0 values of the EL [8] [77] might derive different results, which is one of future works. The preference score of the converted speech signals is far from that of target normal speech, and it is necessary for further enhancement. On the other hand, converted speech is preferred to source speech in all combinations of VC. This result indicates the effectiveness of the proposed system that enhances EL speech by conversion.

After the subjective evaluation scored by the laryngectomee, he gave important comments about the proposed system. He commented that the modification of the source F_0 contours is no problem since the control of the F_0 contours using breath is not perfect. He also commented that the F_0 contours of converted speech utterances are preferred to those of the source EL(air) speech if the unnatural contours of the source EL(air) speech are removed. The laryngectomee scored almost the same preference score for converted whispering from EL speech and converted normal speech from EL(air) speech in the result of **Figure 6.9**. Therefore, modification of source F_0 contours is no problem for the user.

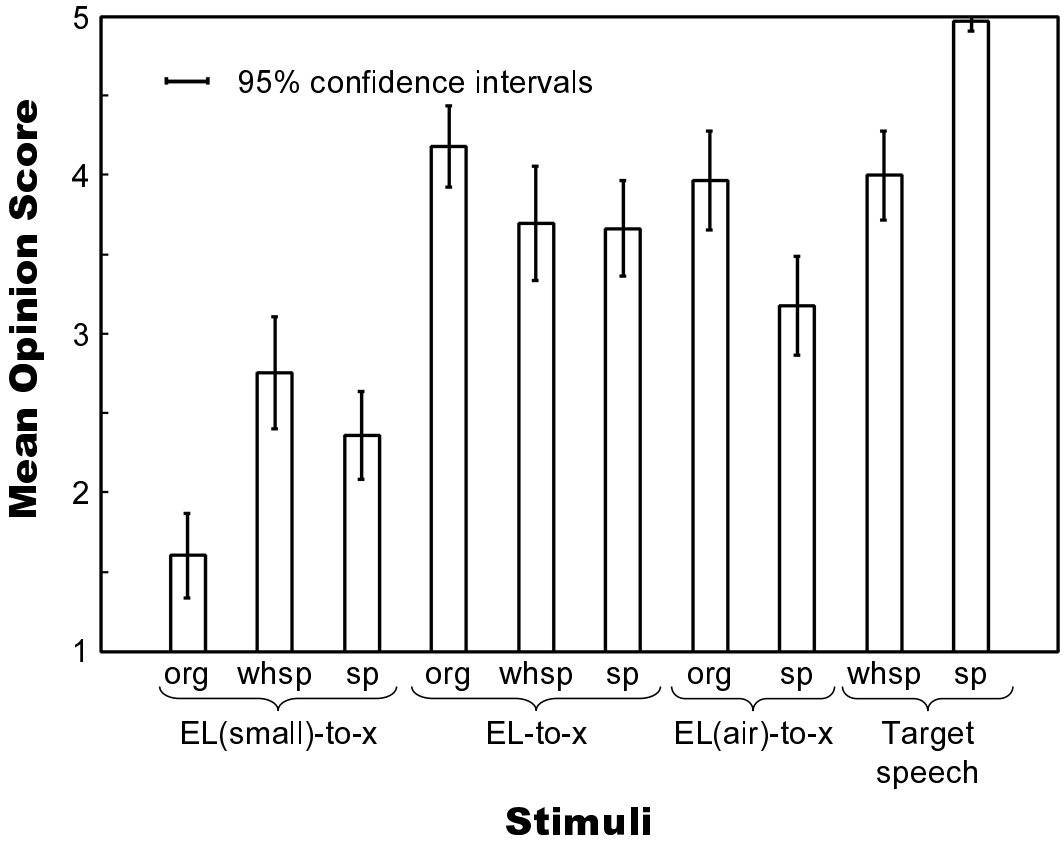

Figure 6.7. Mean opinion score with related to intelligibility for all proposed aid systems.

6.4. Discussion

In order to investigate the effectiveness of the air-pressure sensor, this section conducts additional experimental evaluation for the proposed system of EL(air)-to-Speech.

6.4.1 Experimental conditions

The same laryngectomee as **Section 6.2.1** trained to controll F_0 contours of the EL(air) speech as much as he could for 21 days. High correlation of F_0 contours between the source and the target speech would derive F_0 contours close to the

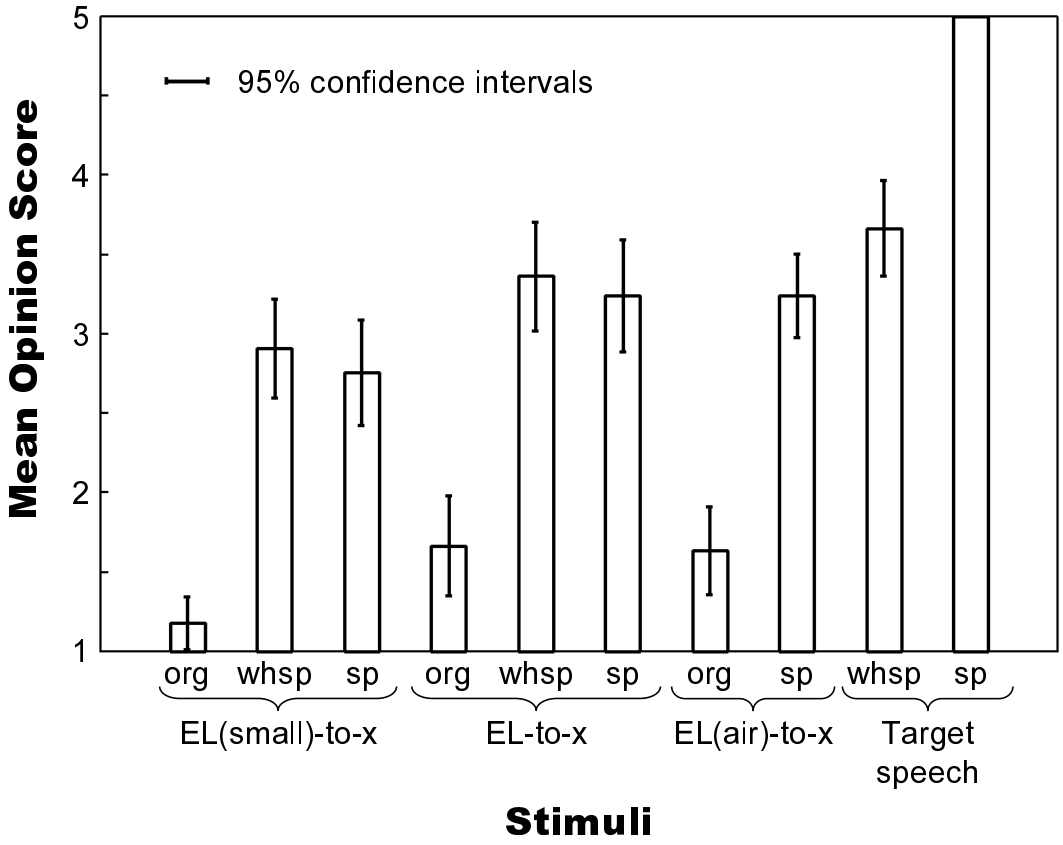

Figure 6.8. Mean opinion score with related to naturalness for all proposed aid systems.

target ones. Therefore, the laryngectomee trained to produce EL(air) speech of which the pitch similarly represents that of the target normal speech used in **Section 6.2.1**.

After the 21-days training, the laryngectomee additionally recorded EL(air) speech with the same contents as **Section 6.2.1** so that the pitch of the EL(air) speech similarly represented that of the target normal speech as much as he could. Although the laryngectomee trained the usage of the air-pressure sensor, the control of the F_0 contours using the air was not perfect. Therefore, the same target speaker as **Section 6.2.1** also additionally recorded the target normal speech of which F_0 contours similarly represented those of the additionally recorded EL(air) speech. This additional recording of the normal speech was aimed to

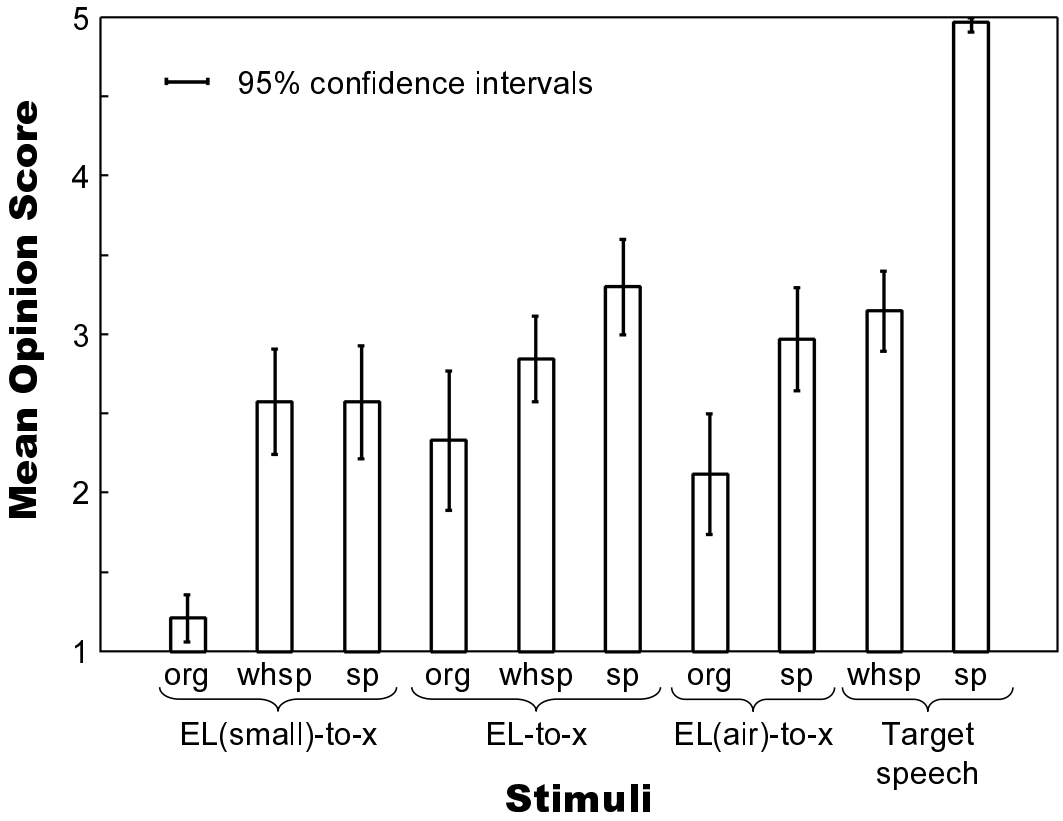

Figure 6.9. Mean opinion score with related to preference for all proposed aid systems.

make the correlation of the F_0 contours between the source EL(air) speech and the target normal speech high. Two VC experiments were conducted from additionally recorded EL(air) speech into normal speech used in **Section 6.2.1** or into another normal speech additionally recorded in this section. Other source electrolaryngeal speech including EL speech and EL(small) speech were also converted into additionally recorded normal speech for the comparison. Note that other experimental conditions were the same as **Section 6.2.1**.

6.4.2 Experimental results

Mel-cepstral distortions in the VC from additionally recorded EL(air) speech into the normal speech used in **Section 6.2** are following: 11.08 dB and 8.67 dB

for the results between the target and the source speech with or without power information, respectively, and 4.60 dB and 3.92 dB for the results between the target and the converted speech with and without power information, respectively. U/V error rates for the results of this conversion framework is 5.38 %, 5.10 %, 4.55 %, and 4.93 % for 16, 32, 64, and 128 mixture components, respectively. Correlation coefficients of this conversion are 0.60, 0.58, 0.59, and 0.55 for 16, 32, 64, and 128 mixture components, respectively. Although the speaker trained to control F_0 contours and recorded EL(air) speech so that the F_0 contours of the EL(air) speech similarly represent those of the target normal speech, the trend of these results is almost the same as **Section 6.2**. In other words, the experimental result of the conversion from additionally recorded EL(air) speech to target speech used in the **Section 6.2** is still insufficient, and the effectiveness of the air-pressure sensor is still not cleared.

Mel-cepstral distortions of the VC from the additionally recorded EL(air) speech into the additionally recorded normal speech are shown in **Table 6.3**. U/V error rates and correlation coefficients of this conversion framework are shown in **Figure 6.10** and **Figure 6.11**, respectively. As the result of VC from additionally recorded EL(air) speech into additionally recorded normal speech is better than other results for both U/V errors and correlation. These results show that the use of air-pressure sensor powerfully works to estimate F_0 contours which are close to target ones compared to EL speech or EL(small) speech.

Figure 6.12, **Figure 6.13**, **Figure 6.14**, and **Figure 6.15** show examples of waveforms, spectrograms, and F_0 contours of additionally recorded EL(air) speech, normal speech used in **Section 6.2**, another normal speech additionally recorded in this section, and converted normal speech when the additionally recorded normal speech is set to the target speech, respectively. From these figures, unnatural steps of F_0 contours are observed in EL(air) speech. These steps are usually not observed in natural speech, and therefore, unpleasant results might be output when the source F_0 contours are directly used to synthesize waveforms. It is seen that unnatural F_0 contours are addressed by converting the F_0 contours, and therefore, it is meaningful not only to use the air-pressure sensor but also convert the source F_0 contours to present F_0 contours.

Table 6.3. Averaged mel-cepstral distortions for all proposed aid systems using target normal speech of which F_0 contours similarly represents those of additionally recorded EL(air) speech. Format of this table is same as **Table 6.2**

System	Source-Target	Converted-Target
EL(small: Pulse train)-to-Speech	17.00 / 11.42	5.50 / 4.55
EL(small: Sawtooth waves 1)-to-Speech	16.59 / 11.01	5.30 / 4.41
EL(small: Sawtooth waves 2)-to-Speech	16.97 / 10.88	5.47 / 4.55
EL-to-Speech	10.93 / 8.96	5.09 / 4.25
EL(air)-to-Speech	11.45 / 9.51	4.90 / 4.12

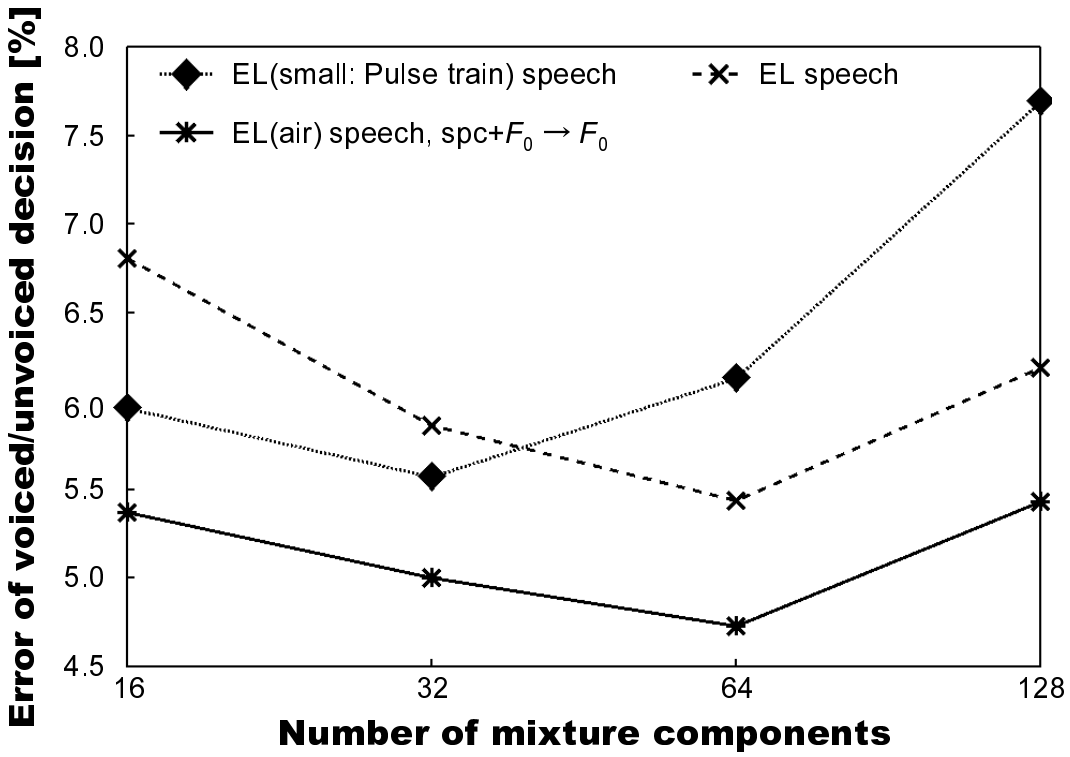

Figure 6.10. Errors of voiced or unvoiced decision using target normal speech of which F_0 contours similarly represents that of EL(air) speech.

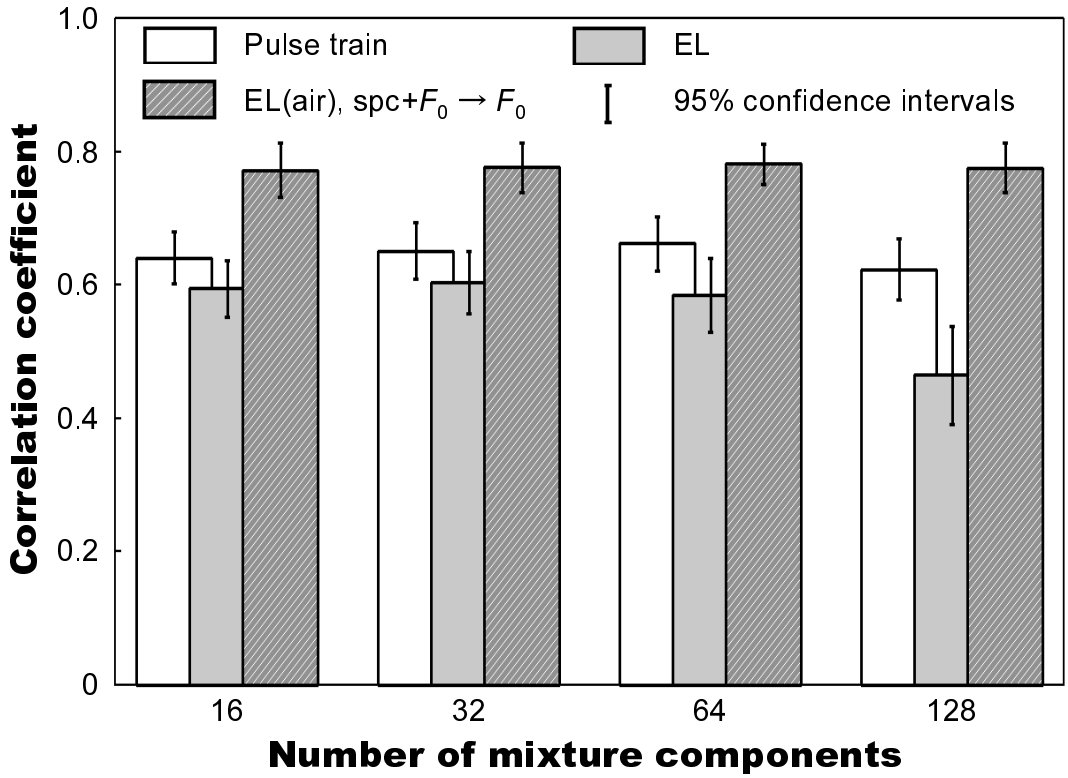

Figure 6.11. Correlation coefficients between converted and additionally recorded target F_0 contours about only voiced frames for both data.

6.5. Conclusion

This chapter investigated three concerns of (1) the effectiveness of VC for all EL speech signals, (2) the effectiveness of using an air-pressure sensor, and (3) the effectiveness of using F_0 information in estimating target F_0 information. From objective and subjective experimental evaluations, VC is significantly effective for all kinds of EL speech used in this thesis. In the estimation of target F_0 information, the conversion accuracy from EL(air) speech is almost the same as that from EL speech in spite of using the air-pressure sensor. As a result, in the case of a short period for the laryngectomee to train to produce EL(air) speech, conventional EL speech is enough to be the input of the speaking-aid system. On the other hand, in the case of a long period to train the sensor, the conversion accuracy was not investigated and this remains for future work. This thesis

employed only an air-pressure sensor to manipulate F_0 patterns. Using different devices [8][77] to manipulate F_0 patterns might give different results, which also remains for future work.

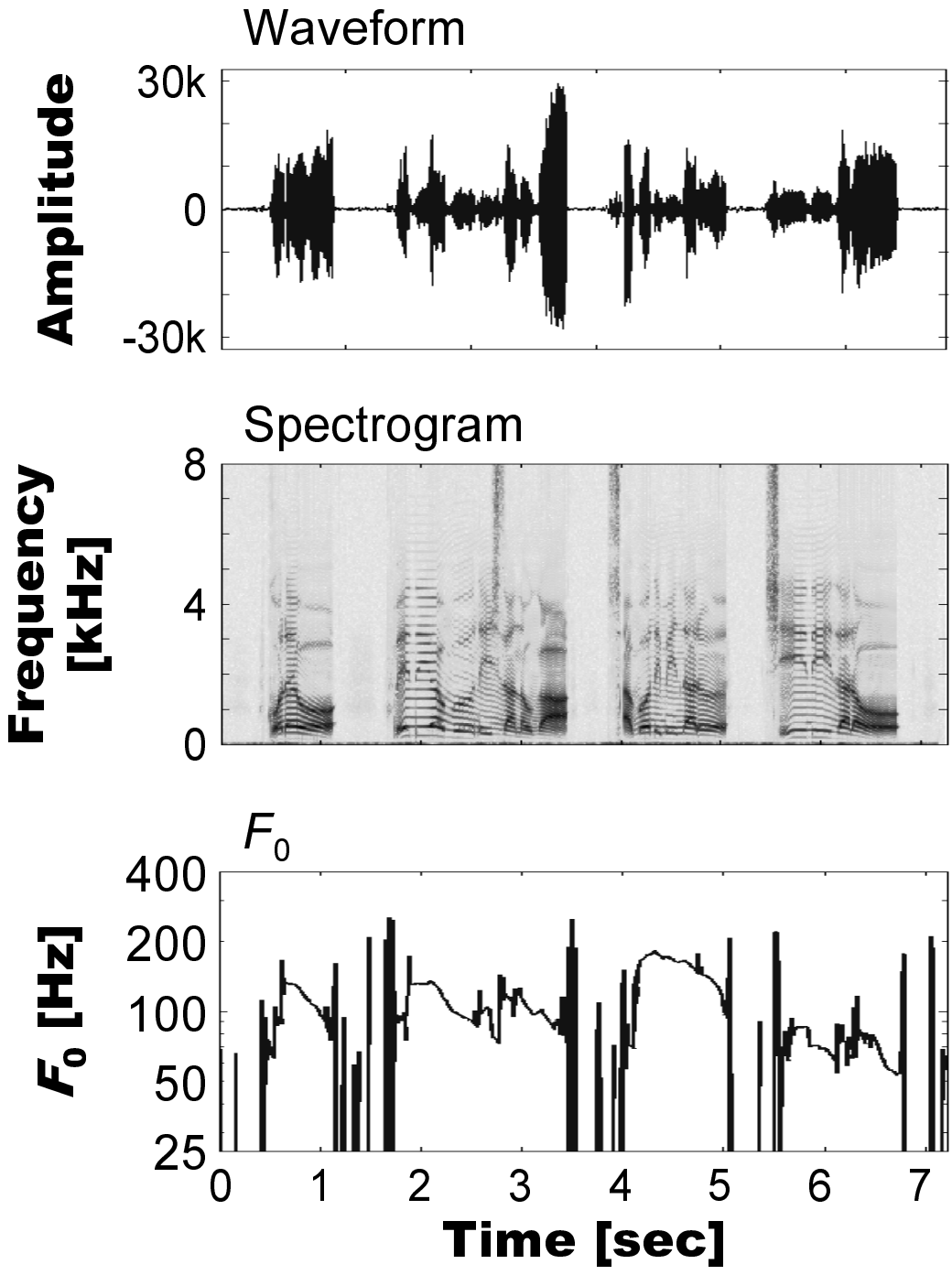

Figure 6.12. Examples of waveforms, spectrograms, and F_0 contours of EL(air) speech of which F_0 contours similarly represents those of normal speech shown in **Figure 6.13**.

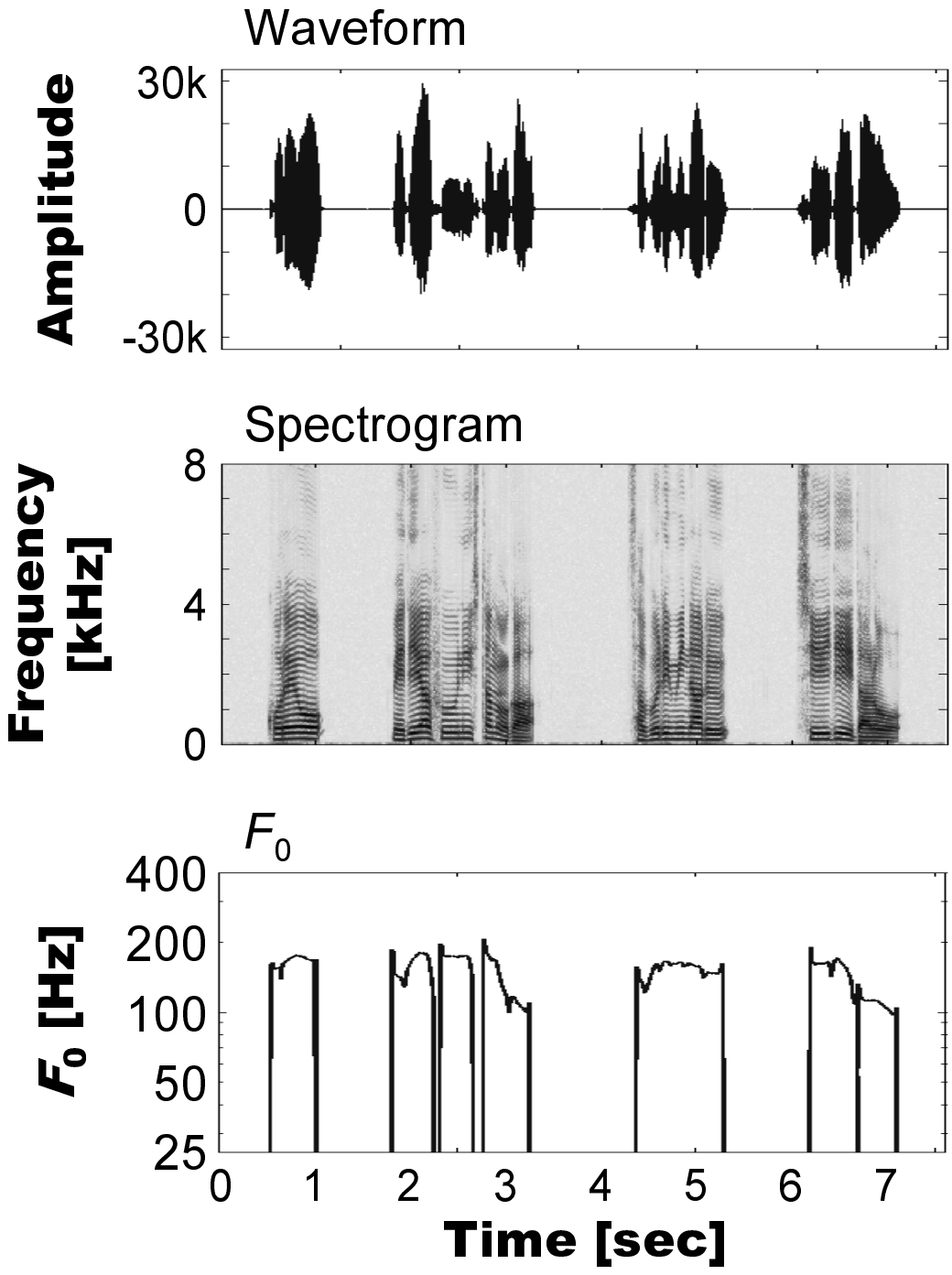

Figure 6.13. Examples of waveforms, spectrograms, and F_0 contours of target normal speech of which acoustic parameters including F_0 contours are same as those of target speech shown in **Section 6.2**.

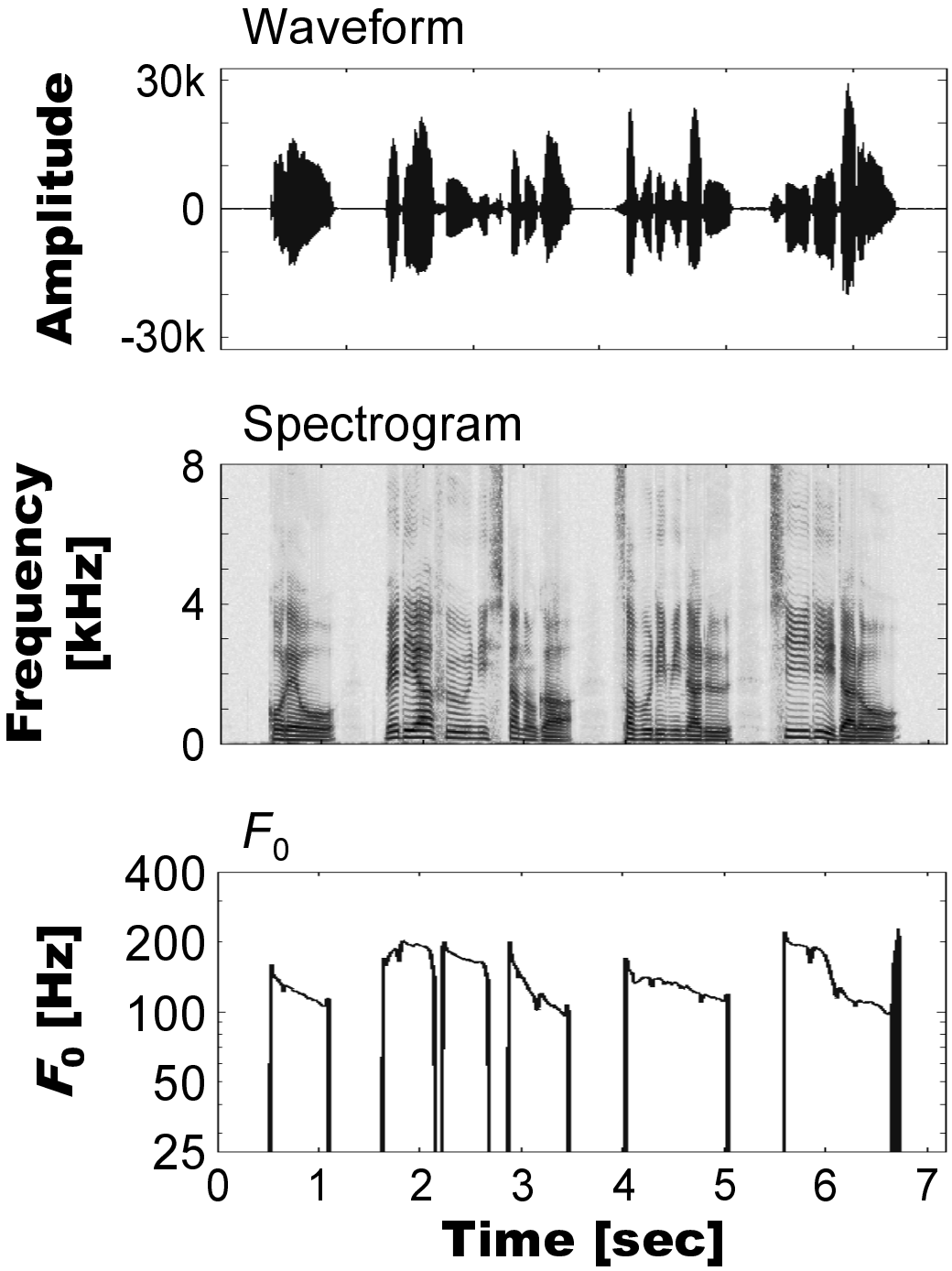

Figure 6.14. Examples of waveforms, spectrograms, and F_0 contours of target normal speech of which F_0 contours are produced to similarly represent those of EL(air) speech shown in **Figure 6.12**.

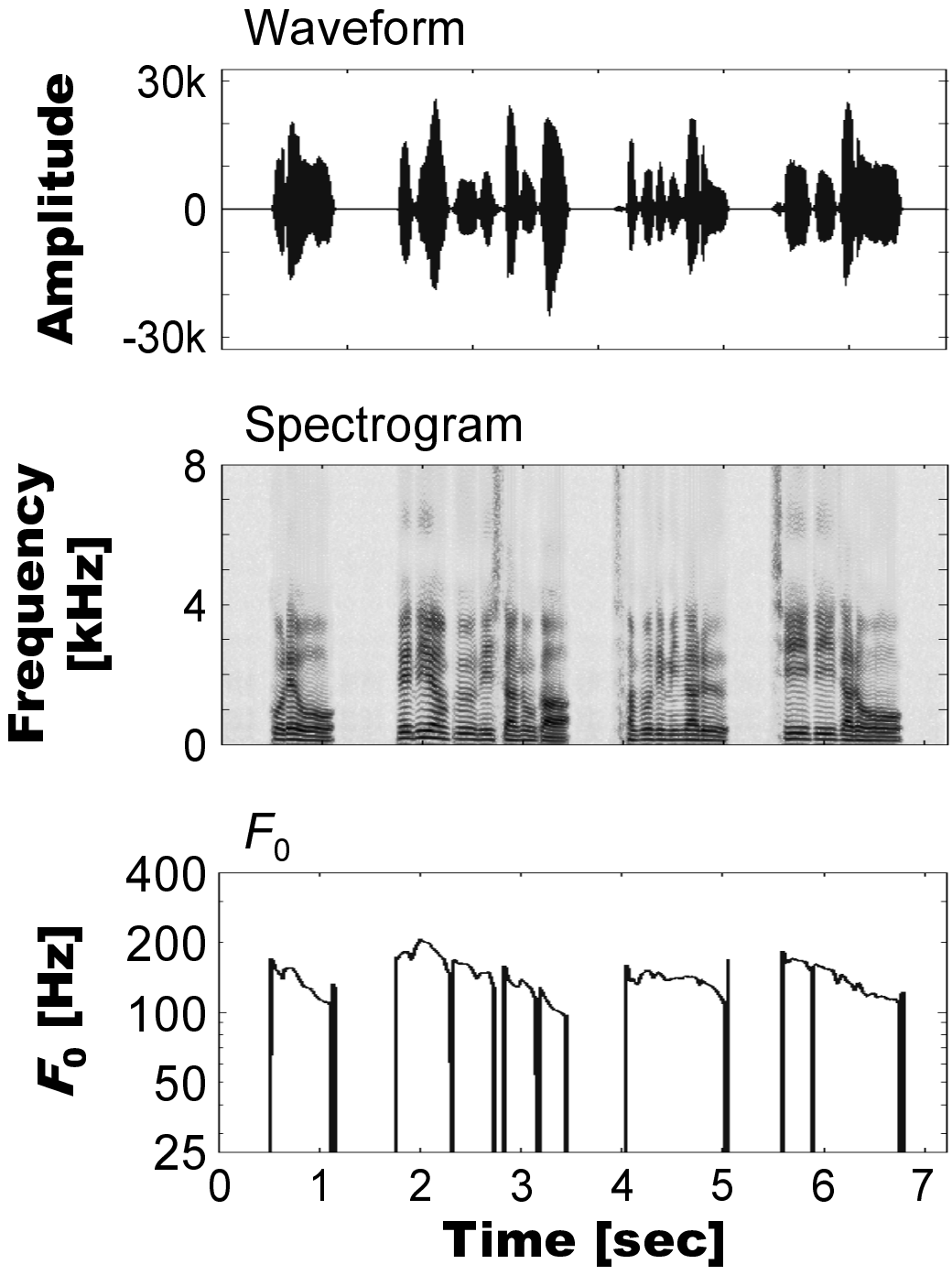

Figure 6.15. Examples of waveforms, spectrograms, and F_0 contours of converted normal speech of which F_0 contours are estimated to represent those of normal speech shown in **Figure 6.14**.

Chapter 7

Conclusion

7.1. Summary of This Thesis

EL is an external medical device to enable a laryngectomee to regain their voices, and the use of the EL is one of the major alternative methods for laryngectomees. Two problems with the current EL have been addressed in this thesis: (1) unnaturalness of the EL speech and (2) noisy radiated sound source signals of the EL itself. In order to address the problem of unnatural EL speech, this thesis introduced a statistical VC technique using GMMs based on the maximum likelihood criterion. In order to address the other problem of the radiated noises, this thesis employed a novel sound source unit that generates extremely small signals that were too faint to be heard by people around the user. This thesis proposed speaking-aid systems using the statistical VC to enhance three kinds of EL speech, which are EL speech, EL(air) speech, and EL(small) speech. This suggestion and the evaluation of the aid systems are the major contribution of this thesis.

In **Chapter 2**, laryngectomees and major alternative speaking methods were described. A novel sound source unit was introduced, which generated extremely small signals so that the people around the speaker were almost could not hear the sound. A NAM microphone was also introduced, which captures extremely small signals through the soft tissues of the head. Moreover, conventional studies were overviewed in this chapter.

In **Chapter 3**, statistical VC used in this thesis was described. VC is a tech-

nique to modify input speech data as if it were uttered by a different speaker while maintaining its linguistic information. This thesis employed statistical VC using GMMs based on the maximum likelihood criterion. This method consisted of the training and conversion parts. Before the training part, two speakers were set to an input speaker (so-called source speaker) and output speaker (so-called target speaker), respectively. Joint probability density function of the source and the target speaker were described by a GMM in the training part. After training the GMM, in the conversion part, target features were estimated based on the conditional probability density function given source features. Three GMMs are usually trained, among which each GMM estimates one acoustic feature which is spectrum, F_0, or aperiodic components, respectively. The VC method employs not only static features but also dynamic features to capture smooth movement of the acoustic features. The VC method also introduces GV to suppress the problem of over-smoothing of the conversion features due to the maximum likelihood criterion.

In **Chapter 4**, one speaking-aid system was proposed, in which conventional EL speech was input and converted natural speech by the statistical VC was output as the user's new voice. Although it is reasonable to set the normal speech to the output speech, it would be difficult to estimate natural F_0 contours from source EL speech because the EL speech does not include effective F_0 information. To avoid this problem, this thesis converts EL speech not only to normal speech as one conversion framework of EL-to-Speech but also to whispering as the other conversion framework of EL-to-Whisper. In EL-to-Whisper, only one GMM that estimates spectral information is trained. In order to capture complicated multivariable factors, this thesis used segmental spectral feature vectors of the source EL speech. Segmental feature vectors are constructed in the following procedures; first, the current $\pm L$ frames are concatenated to one feature vector, and then, the segmental feature vector for the current frame was established by principal component analysis procedure. EL-to-Speech is achieved by estimating spectra, F_0s, and aperiodic components from only the spectral information of the source EL speech using individual GMM. Both EL-to-Whisper and EL-to-Speech were experimentally evaluated using imitated EL speech by a non-laryngectomee as a preliminary evaluation. From objective evaluations, spectral estimation works

well. Correlation coefficient between converted and target F_0 values were almost 0.3 and U/V errors between those features were almost 9 %. Although the correlation was not good, non-predefined F_0 contours were estimated and the proposed system was expected to be effective for laryngectomee's data to enhance EL speech. This thesis additionally introduces an air-pressure sensor with which the speaker can control intonations using their breath. Using this sensor, EL(air) speech includes non-predefined F_0 contours. Then, another speaking-aid system was proposed, which enhances EL(air) speech in another conversion framework of EL(air)-to-Speech.

In **Chapter 5**, another sound source unit generating extremely small sound source signals was introduced to address the problem of noisy radiated noises. EL(small) speech was obtained by recording the small EL speech with a NAM microphone. Then, the other speaking-aid system was proposed, which enhances EL(small) speech to output normal speech as another conversion framework of EL(small)-to-Speech or output whispering as the other conversion framework of EL(small)-to-Whisper. Acoustic features for VC and its procedure were the same as the proposed system for EL speech enhancement. Different sound source signals were designed from the viewpoint of independently changing its spectra or powers. In case of changing its spectra, three sound sources were designed, which were pulse train for all frequency bands, sawtooth waves for lower frequency bands, and compensation waves into target whispering for higher frequency bands. In case of changing its powers, the spectra were fixed to the sawtooth waves that have the largest dynamic ranges among those three sound source signals. EL(small)-to-Whisper and EL(small)-to-Speech were also preliminarily evaluated using the EL(small) speech imitated by the same non-laryngectomee as the evaluation of VC from EL speech. From experimental results, correlation coefficient between converted and target F_0 values were almost 0.3 and U/V errors between those features were almost 9 %. The results of EL(small) speech enhancement were almost the same as those of EL speech enhancement, and therefore, the proposed system for EL(small) speech enhancement was also expected to be effective for laryngectomee's data.

In **Chapter 6**, all proposed speaking-aid systems of EL-to-Whisper, EL-to-Speech, EL(air)-to-Speech, EL(small)-to-Whisper, and EL(small)-to-Speech were

experimentally evaluated using one male laryngectomee's data. From objective evaluations, estimation of spectral information works well. Moreover, another estimation of F_0 information also works well in which correlation coefficient between the estimated and the target F_0 values were more than 0.5 with less U/V errors (less than 7 % for all kinds of source speech). The converted voice quality related to intelligibility, naturalness, and preference was subjectively evaluated by 10 non-laryngectomees and one laryngectomee who was the source speaker. From the results, intelligibility of the converted speech was slightly degraded from that of the source speech. This is a problem that has not been addressed yet, and this remains for future work. On the other hand, the naturalness of the converted speech was dramatically improved from that of the source speech. Moreover, the preference, which was a total evaluation of the voice quality, of the converted voice was higher scored than that of the source speech. As the result of experimental evaluation, the proposed speaking-aid systems addressed the unnaturalness of EL speech by VC technique. The proposed systems also accept several kinds of sound source signals. This advantage is expected to give great versatility to the aid systems when those are used in our daily lives. Moreover, the effectiveness of using an air-pressure sensor was investigated by using additionally recorded target normal speech so that the pitch of the target normal speech is close to that of the source EL(air) speech.

7.2. Future Work

Although EL speech quality has been enhanced by VC technique, a number of problems remain to be addressed.

- **Degradation of intelligibility**

The source speaker in this thesis is proficient at producing alaryngeal speech using an external device, and therefore, the intelligibility of the speaker is high-scored. The intelligibility of the converted speech is, however, lower scored than that of the source EL speech as **Figure 6.7** shows. It is preliminarily investigated that VC accuracy from EL speech produced by another laryngectomee who is a beginner at producing EL speech and was worse than the result in this thesis; and

the converted speech was not intelligible although the source EL speech was also not intelligible. In order to address this problem, other VC frameworks might be necessary. For example, an external device such as an air-pressure sensor might contribute to suppressing the degradation of the intelligibility of the converted speech, or another constraint such as linguistic information might be effective to suppress unnatural transitions between mixture components. Conditional Random Field (CRF) [78] is interesting for this study. In CRF, not a generative model but a discriminative model is introduced to describe features considering many rules called feature functions. A novel VC framework considering such rules might be necessary to suppress the degradation of the intelligibility of converted speech.

- **Controlling speaker individuality**

In the interview of the laryngectomee who was the source speaker, he was concerned that the speaker individuality of the converted speech would be changed because the different speaker from the source speaker was set as the target speaker. In order to address this problem, three methods are concerned. One is to use the original voice of the source speaker before undergoing laryngectomy. If some data before the laryngectomy remains, those data would be able to be set as the target data so that the user speaks with his or her original voice. When there is no such data, the second idea is to set another person whose voice quality is similar to the user as the target speaker. For example, the user would choose one speaker from a speech corpus, and then, the laryngectomee reads the same contents as what the target speaks. As a result, the laryngectomee would obtain natural voices such that speaker individuality is much closer to the original speaker's individuality. When there are no desirable speakers or the quality is not satisfied, the third idea is concerned to introduce Eigenvoice conversion (EVC) [79, 80] that enables users to control speaker individuality with a small amount of parameters.

- **Further evaluation of EL(air)-to-Speech**

As **Section 6.4** shows, the effectiveness of using an air-pressure sensor is investigated. In order to confirm the result, subjective evaluation should be

conducted by non-laryngectomees and the laryngectomee. It is desired to confirm the advantage of using an air-pressure sensor by comparing converted normal speech from EL(air) speech with that from EL speech.

- **Introducing another device to control F_0 contours without giving the speaker stress**

The air-pressure sensor introduced in this thesis is effective to enable the speaker to control the F_0 contours of the EL speech. The laryngectomee, however, gave an important comment that he felt huge stress in producing EL(air) speech because the use of the air-pressure sensor forced the speaker to expose the tracheostoma. Moreover, the speaker needs both hands to produce EL(air) speech, and therefore, the usefulness of the air-pressure sensor is limited. From these issues, introducing another device to enable laryngectomees to control F_0 contours without giving the speaker stress is necessary to establish the proposed aid systems. This thesis used air flowing from the tracheostoma; on the other hand, other methods are available such as myoelectric information, a pressure-sensitive switch, and so on.

- **Optimization of constructing segmental feature vectors for source data**

This thesis used PCA procedure and Tran *et al.* used LDA procedure [53] to construct segmental feature vectors. Other applied research also used PCA procedures; however it was not revealed that PCA or LDA is the best method to reduce the dimension of the vectors. There is another technique of factor analysis to reduce the dimension. In order to exactly represent a multi-dimensional feature vector with small parameters, further investigation is desired. Moreover, acoustic features, namely spectral features, to be concatenated in constructing segmental feature vectors have not been discussed. Mel-cepstral coefficients have so far been used; however, it is not revealed that those coefficients are the best features especially in the conversion from source speech including certain F_0 contours such as EL(air) speech.

- **Evaluation using EL speech produced by other laryngectomees**

This thesis evaluated the proposed system for only one laryngectomee. More users would be necessary in order to evaluate proposed systems for other laryngectomees. The variation of voice humanities of EL speech is smaller than that of normal speech because voice characteristics of EL speech are mainly occupied by the external sound source signals. Therefore, similar results as described in this thesis would be expected for other proficient people to produce alaryngeal speech using an external device.

- **Real-time procedure for VC procedures**

The VC procedures used in this thesis are not specified for real-time procedures. It is essential to update the current VC framework to the real-time VC framework to make the proposed systems useful in our daily lives. In order to achieve the real-time VC, a time-recursive conversion algorithm based on maximum likelihood estimation of spectral parameter trajectory is proposed [81]. This method is inspired by the parameter generation algorithm for HMM-based speech synthesis [55] and the vector quantization algorithm for speech coding [82]. This method is expected to be effective for the proposed aid systems.

Appendix

A. Case Studies of Speech Recognition for Impaired Speech

A.1 Introduction

It is reasonably natural for human beings to be interested in human-machine communications with speech because of the convenience of speech. In human-machine communication, the technique that extracts linguistic information from input speech waveforms is called automatic speech recognition (ASR). It is almost impossible to establish an ASR system that achieves equal performance to the hearing ability of human beings even though the latest scientific techniques are used. Therefore, ASR systems in which conditions such as users, tasks, and the dictionary size are limited are usually established for use in our daily lives. Introducing such limitations derives higher recognition accuracy, and ASR technique is applied to many applications such as a car navigation system, text reader systems, and so on.

The difficulty of the ASR system differs with relation to acceptable users, words, and the size of dictionary. Large vocabulary continuous speech recognition for speaker independent ASR systems is a greatly more difficult task than small vocabulary isolated word speaker recognition for speaker dependent ASR systems. Moreover, speech signals including loud background noises are more difficult to recognize than speech signals without loud noises.

The basic flow chart of a continuous ASR system is shown in **Figure A.1**. The acoustic model holds patterns of acoustic characteristics for each recognition unit (e.g. phoneme) to conduct pattern matching with input parameters. The

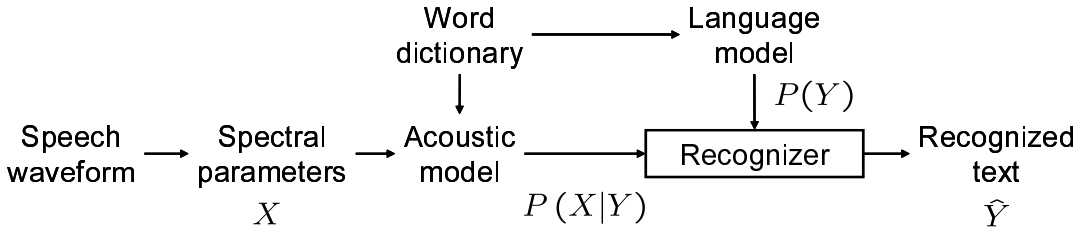

Figure A.1. Overview of speech recognition system.

language model decides connections of words. The dictionary defines the words and those phoneme identities accepted in the system. Only the words defined in the word dictionary are the target of the ASR system. The recognizer decodes the input speech parameters to the output texts using the acoustic and the language model.

This section is organized as follows. Speech recognition using hidden Markov models (HMMs) and a major adaptation technique for unseen speech data are overviewed in **Section A.2**. In **Section A.3**, some acoustic models for impaired-speech data are experimentally evaluated. Finally, this section is concluded in **Section A.4**.

A.2 Speech Recognition Using Hidden Markov Models

From the definition of ASR, the estimation of the word sequence $\hat{Y}$ given input speech parameters X is represented as

$$\hat{Y} = \arg\max P(Y|X) \tag{A.1}$$

$$= \arg\max \frac{P(Y,X)}{P(X)}. \tag{A.2}$$

Although the Eqn. (A.2) is the definition of ASR, the direct calculation of $P(Y|X)$ is difficult. Therefore, using Bayes' rule, ASR is rewritten as

$$\hat{Y} = \arg\max \frac{P(X|Y)P(Y)}{P(X)}. \tag{A.3}$$

The denominator of the Eqn. (A.3) is independent for the decision of Y, and therefore, the solution of ASR, namely the maximization of the posterior proba-

bility for Y given X, is also rewritten as

$$\hat{Y} = \arg\max P(X|Y)P(Y). \quad \text{(A.4)}$$

$P(X|Y)$ in the Eqn. (A.4) is calculated by comparing the input patterns with the templates defined in the acoustic model. HMM is the major model to describe $P(X|Y)$ in these days [83], and HMM enables users to effectively and flexibly process time-sequence data.

Figure A.2 shows an example of left-to-right without skip nor back loop HMM, which is one of typical HMM structure in ASR system in these days in which s_i is the state index and π_i is the initial state probability of the state i. $a_{ij} = P(s_{t+1} = j|s_t = i)$ is the transition probability in which the current frame is in the state i at the time t and it is in the state j at the time $t + 1$. b_i is the probability distribution function of the state i. Each state in **Figure A.2** has the continuous output distribution such as Gaussian distribution or GMM). **Figure A.2** effectively expresses characteristic of human voices. The existence of self-loop means that human voices are regarded as quasi-periodic signals. The undefined loop of skip transition expresses that the vocal tract parameters smoothly moves, and another undefined loop of back loop is seen that the speech signals does not go back to the past.

When the recognition unit is phoneme (and this would be the most popular unit), each phoneme is modeled by a HMM. This type of acoustic model is called monophone model. Speech signals are continuously uttered, and therefore, the acoustic characteristic of the current phoneme differ with related to the previous and succeeding phonemes. In order to express this transition, another acoustic model called triphone prepares different templates with related to the previous and succeeding phonemes, which is the most popular acoustic model. The triphone HMM reasonably expresses characteristic of human voices; on the other hand, the number of parameters to be estimated of the triphone HMM is dramatically larger than those of the monophone HMM. As the result, many utterances are required to train triphone HMM.

Maximum likelihood parameter estimation includes two essential problems to be addressed: Over training for the training data and robustness for the unseen data in the training data. Establishing a robust HMM acoustic model is an important topic and it is achieved by sharing parameters. This thesis employs

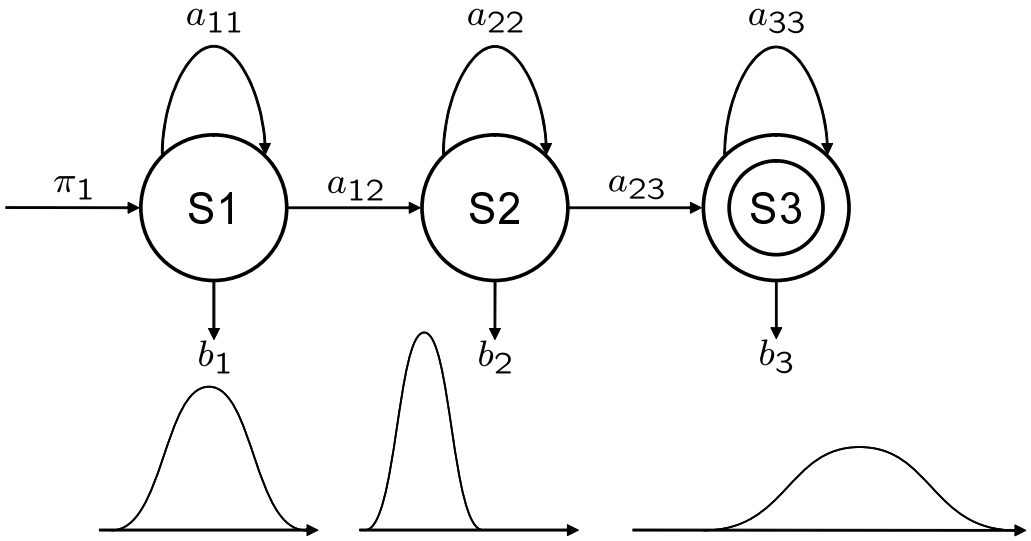

Figure A.2. Example of left-to-right HMM.

Phonetically Tied-Mixture (PTM) [84] acoustic model that shares a set of mixture components among states of the same central phoneme of the state-shared triphone. This state-based PTM models Gaussian distributions more accurately by having independent mixture components on triphones with different central phonemes compared to other tying methods [85][86].

Speaker independent (SI) acoustic model accepts arbitrary speakers as the input speakers. SI model is useful for systems in which switching of the input speaker frequently occurs such as book searching system in book stores or libraries. On the other hand, SI model is not suitable for specified speaker. Speaker dependent (SD) acoustic model, on the other hand, accepts only the specified speaker defined in advance. The SD model is often powerful model when the target of the system is known.

In order to straightforwardly obtain a SD triphone HMM for the new user, environment, or speech style, hundreds of utterances are required, which is almost too difficult for users to record those data especially for speaking-impaired people. Model adaptation technique plays the important roll to have the acoustic model well represent unknown input data by transforming the model parameters. The issue of the model adaptation is that how the transformations are estimated

using small amount of speech data of the new acoustic environment, so-called adaptation data, so that it maximizes the likelihood for the adaptation data given the current model parameter set.

In the speech recognition of speech-impaired people, the usual acoustic model that was suitable for non-laryngectomees was not appropriate because the acoustic features of the impaired people were extremely different from those of the non-impaired people. Moreover, it is rare to be able to obtain huge amount of speech data of impaired people.

Maximum likelihood linear regression (MLLR) [87] is one of the most powerful and popular adaptation techniques using only small amount of the adaptation data. Here, consider the case of continuous density HMM with Gaussian output distributions. Given a speech parameter vector at frame t,$\boldsymbol{o}_t$, the probability density of that vector generated from a particular distribution s, $b_s(\boldsymbol{x}_t)$, is written as

$$
\begin{aligned}
b_s(\boldsymbol{x}_t) &= \mathcal{N}\left(\boldsymbol{x}_t; \boldsymbol{\mu}_s, \boldsymbol{\Sigma}_s\right) && \text{(A.5)} \\
&= \frac{1}{(2\pi)^{\frac{n}{2}} \left|\boldsymbol{\Sigma}_s\right|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}\left(\boldsymbol{x}_t - \boldsymbol{\mu}_s\right)^\top \boldsymbol{\Sigma}_s^{-1}\left(\boldsymbol{x}_t - \boldsymbol{\mu}_s\right)\right\}. && \text{(A.6)}
\end{aligned}
$$

Model-space linear transformation updates the mean vector $\boldsymbol{\mu}_s$ to $\hat{\boldsymbol{\mu}}_s$ written as

$$
\begin{aligned}
\hat{\boldsymbol{\mu}}_s &= \boldsymbol{A}_s\boldsymbol{\mu}_s + \boldsymbol{b}_s = \boldsymbol{W}_s\boldsymbol{\xi}_s, && \text{(A.7)} \\
& && \text{(A.8)}
\end{aligned}
$$

where $\boldsymbol{W}_s$ is the d-by-$(d+1)$ transformation matrix to maximizes the likelihood of the HMM to the adaptation data, and $\boldsymbol{\xi}_s$ is the extended mean vector defined as

$$
\begin{aligned}
\boldsymbol{\xi}_s &= [1, \boldsymbol{\mu}_s^\top]^\top && \text{(A.9)} \\
&= [1, \mu_s(1), \cdots, \mu_s(d), \cdots, \mu_s(D)]^\top, && \text{(A.10)}
\end{aligned}
$$

where $\mu_s(d)$ is the dth dimensional mean value. The current issue is to find the transformation $\boldsymbol{W}_s$. This thesis also updates covariance matrices as follows [88]:

$$
\begin{aligned}
\hat{\boldsymbol{\Sigma}_m} &= \boldsymbol{B}_m^\top \boldsymbol{H}_m \boldsymbol{B}_m, && \text{(A.11)} \\
\boldsymbol{B}_m &= \boldsymbol{C}_m^{-1}, && \text{(A.12)} \\
\boldsymbol{\Sigma}_m^{-1} &= \boldsymbol{C}_m \boldsymbol{C}_m^\top. && \text{(A.13)}
\end{aligned}
$$

Let $\boldsymbol{X}$ be a series of T observations:

$$\boldsymbol{X} \quad = \quad \boldsymbol{x}_1, \cdots, \boldsymbol{x}_T. \tag{A.14}$$

The total likelihood of the model set generating the objective sequence is represented as

$$\mathcal{L}(\boldsymbol{X}|\lambda) \quad = \quad \sum_{\boldsymbol{\theta} \in \Theta} \mathcal{L}(\boldsymbol{X}, \boldsymbol{\theta}|\lambda), \tag{A.15}$$

where $\mathcal{L}(\boldsymbol{X}|\lambda)$ is the likelihood of generating $\boldsymbol{X}$ using the state sequence $\boldsymbol{\theta}$ given the model parameter set λ. Θ denotes a set of all state sequences. Then, in order to maximize the objective function, a following auxiliary function is defined [89]:

$$Q(\lambda, \hat{\lambda}) \quad = \quad \sum_{\theta \in \Theta} \mathcal{L}(\boldsymbol{X}, \boldsymbol{\theta}|\lambda) \log \left(\mathcal{L}(\boldsymbol{X}, \boldsymbol{\theta}|\hat{\lambda}) \right). \tag{A.16}$$

These updates take two stages [88]. First, the transform for the mean vector is found given the current covariance matrix. Next, another transform for the covariance matrix is found given the current mean vector. Finally, the whole process would iteratively update the model parameters so that it maximizes the likelihood of the given adaptation data.

A.3 Experimental Evaluation

Experimental conditions

This thesis used two kinds of speech data sets of (1) EL speech utterances of the laryngectomee described in this thesis and (2) other types of speaking-impaired people.

One Japanese male laryngectomee recorded 50 phoneme-balanced balanced sentences for the adaptation data and recorded 30 newspaper utterances for the test data. He recorded those 80 utterances of EL speech, EL(air) speech, and EL(small) speech using three kinds of sound source signals, which were pulse train, sawtooth waves with the same power as the pulse train, and another sawtooth wave with larger powers than the pulse train. Only EL(small) speech utterances were recorded by NAM microphone, and all other utterances were recorded by a

head-set microphone. Data format were 16000 Hz sampling with 16 bit for each sample.

As other types of speaking-impaired people, ten subjects of speaking-impaired people recorded speech data. Two of them had a disorder due to cerebral palsy. One another subject was a congenital hearing-impaired person. One another subject had a problem due to measles encephalitis. Other three subjects were hearing-impaired or hearing loss patients due to taking the streptomycin. The remaining three subjects were acquired hearing-impaired patients. The contents were Japanese syllables, words expected to be used in their daily lives, several numbers with one through four digits, telephone numbers also expected to be used in their daily lives, and short sentences of "north winds and the sun" and "the ants and the grasshopper" which were often used in clinics. All contents including the meaning and the theoretical phonemes were listed as follows:

1 Japanese syllables

/a/, /i/, /u/, /e/, /o/, /ka/, /ki/, /ku/, /ke/, /ko/, /sa/, /shi/, /su/, /se/, /so/, /ta/, /chi/, /tsu/, /te/, /to/, /na/, /ni/, /nu/, /ne/, /no/, /ha/, /hi/, /fu/, /he/, /ho/, /ma/, /mi/, /mu/, /me/, /mo/, /ya/, /yu/, /yo/, /ra/, /ri/, /ru/, /re/, /ro/, /wa/, /o/, /N/

Note that the second o from the tail was the same pronunciation as the other /o/ of vowel (fifth from the head in the above list), although the Japanese character was different.

2 Words

/ka: sa n/(mother), /to: sa N/(father), /o ba: cha N/(grand mother), /o ji: cha N/(grand father), /go ha N/(rice), /o ka zu/(side dish), /ha na/(flower), /ma do/(window), /o ha yo:/(good morning), /ko N ni chi wa/(hello), /ko N ba N wa/(good evening), /o ya su mi/(good night), /ko za ka na/(little fish), /pa so ko N/(personal computer), /ma u su/(mouse), /a i/(love), /de N sha/(train), /hi ko: ki/(plain), /sa ka na/ or /to to/(fish), /ya ma ya ma/(mountains), /hyo: ta N/(gourd), /sha N de ri a/(chandelier),

/ku ri su ma su/(Christmas), /ro: a ky: ka i/(association of the deaf),
/go ho: bi/(reward), /fu ku o ka)(Fukuoka), /he ri ko pu ta:/(helicopter),
/hi de ri/(dry weather), /ka bi N/(vase), /fu na de/(departure of vessel)

Alphabets surrounded by the slash denoted the theoretical phoneme labels to be recognized, and words surrounded by parentheses denoted the meaning of individual words. The label of /to to/ was a Japanese ancient word representing fish. Although these words were individually uttered as isolated words, some of those were recorded in one file. Six speakers uttered all words in one file, and the other subjects uttered several (two, three, five, or six) words in one file. When several words were recorded in one file, not individual words but the individual files were recognized to make the size of dictionary small. Note that a part of the above words were used in evaluations.

3 Numbers with one through four digits

/ze ro/(0), /i chi/(1), /ni/(2), /sa N/(3), /shi/ or /yo N/(4), /go/(5),
/ro ku/(6), /shi chi/ or /na na/(7), /ha chi/(8), /kyu:/ or /ku/(9), /ju:/(10),
/ni ju: i chi/(21), /sa N ju: ro ku/(36), /ju: na na/(17), /yo N ju: ro ku/(46),
/go ju: kyu:/ or /go ju: ku/(59), /ro ku ju: na na/(67), /ha chi ju: yo N/(84),
/kyu: ju: go/(95), /kyu: hya ku ha chi ju: i chi/(981),
/na na hya ku sa N ju: ro ku/(736), /ha q pya ku ha chi ju: ha chi/(888),
/ro q pya ku ju: ni/(612), /go hya ku sa N ju: ky u:)(539),
/yo N hya ku go ju: kyu:/(459), /se N sa N bya ku sa N ju: i chi/(1331),
/ni se N ky u: hy a ku ha chi ju: ni/(2982),
/sa N ze N sa N bya ku kyu: ju: ha chi/(3398),
/yo N se N na na hya ku na na ju: na na/(4777),
/na na se N go hya ku ro ku ju: ha chi/(7568)

Note that one subject read three digit and four digit numbers as the connection of one digit; not /na na hya ku sa N ju: ro ku/ for 736 but /na na sa N ro ku/ just like 7-3-6.

4 Phone numbers

It was almost the combination one digit pronunciation. The contents were hidden to protect the personal information.

5 The ants and the grasshopper

/a ri to ki ri gi ri su/(The ants and the grasshopper),
/a tsu i a tsu i na tsu no hi no ko to/(In a field one hot summer's day),
/a ri sa N ta chi wa sa mu i fu yu ni so na e te i q sho: ke N me: ha ta ra i te i ma shi ta/ (the ants were working hard for cold the winter.),
/mi N na de chi ka ra o a wa se te ta be mo no o ha ko bu a ri sa N/ (Some ants united their strength to carry foods.),
/to N ne ru o ga N ba q te ho ri su su me ru a ri sa N/ (Other ants hard continued digging a tunnel.),
/mi N na mi N na so re wa i q sho: ke N me: ha ta ra i te i ma shi ta/ (All of them were very hard working.).

Note that although the almost all speakers uttered all the sentences of each work into one file, sentences were split into individual sentences in the adaptation and the test.

6 The north wind and the sun

/ki ta ka ze to ta i yo:/(The north wind and the sun),
/a ru hi no ko to ki ta ka ze ga ta i yo: ni chi ka ra ji ma N o shi te i ma su/ (One day, the north wind boasted of great strength to the sun.),
/ki ta ka ze ga i: ma shi ta/(The north wind said),
/bo ku wa do N na mo no de mo ka N ta N ni fu ki to ba su ko to ga de ki ru yo/ ("I can easily blow out anything."),
/se ka i de ichi ba N no tsu yo i no wa ya q pa ri bo ku da ne/ ("It is me that am the one with the greatest strength in the world."),
/su ru to ta i yo: ga i: ma shi ta/(Then, the sun said),
/fu fu N, ta shi ka ni ki mi wa chi ka ra mo chi da/ ("Well, you surely have the great power."),
/de mo i chi ba N q te i u no wa do: ka na/ ("But I doubt that you are the top.").

Note that although the almost all speakers uttered all the sentences of each work into one file, sentences were split into individual sentences in the adaptation and the test.

Table A.1 and **Table A.2** showed the information of disorder, adaptation, and test data for individual speakers.

Table A.1. Information of speaking-impairment, adaptation data and test data for individual speakers. Note that the number in the parenthesis after the words shows that the number of the represented words are recorded in one file. For example, "words(3)" and "words(all)" show 3 words and all words are recorded in one file, respectively. NS and AG notes "The North Winds and the Sun" and "The Ants and the grasshopper", respectively

Speaker index (speaking-impairment)	adaptation data	test data
YT060910a (cerebral palsy)	41 utterances including syllables, words (3), NS, AG, 1-4 digits numbers, and phone numbers	16 utterances including syllables, words (3), and one digit numbers and one digit numbers
KK061005c (acquired hearing loss)	18 utterances including syllables, words (all), AG, and one digit numbers	18 utterances including syllables, words (all), AG, and one digit numbers
KS061005c (acquired hearing loss)	18 utterances including syllables, words (all), AG, and one digit numbers	18 utterances including syllables, words (all), AG, and 1 digit numbers
NT061004b (hearing loss due to streptomycin)	17 utterances including syllables, words (all), AG, and one digit numbers	16 utterances including syllables, words (all), AG, and one digit numbers
TF061005b (hearing loss due to streptomycin)	18 utterances including syllables, words (all), AG, and one digit numbers	18 utterances including syllables, words (all), AG, and one digit numbers

Mel-frequency cepstral coefficients (MFCCs) [90] were employed to calculate spectral parameters of all speech utterances including adaptation and test utterances. MFCC parameters were calculated by filter bank analysis using triangular windows that were put equally spaced on the mel frequency domain that was

Table A.2. Information of speaking-impairment, adaptation data and test data for the remaining individual speakers. Notations are same as **Table A.1**

Speaker index (speaking-impairment)	adaptation data	test data
TT060910c (acquired hearing loss)	17 utterances including syllables, words (all), AG, and one digit numbers	17 utterances including syllables, words (all), AG, and numbers (1 digit)
YS060910b (hearing-impaired due to streptomycin)	16 utterances including syllables, words (all), AG, and one digit numbers	16 utterances including syllables, words (all), AG, and one digit numbers
YI060910a (cerebral palsy)	40 utterances including syllables, words (3), NW, three sentences out of AG, 1-4 digits numbers, and phone numbers	15 utterances including syllables and words (3)
NS061213d (congenital hearing loss)	26 utterances including syllables, words (5 or 6), phone numbers, and 2-4 digits numbers	17 utterances including syllables and words (5 or 6)
TY070130e (dysarthria due to measles encephalitis)	23 utterances including syllables, words (2 or 3), 1-3 digits numbers, phone numbers, and four utterances out of NW	17 utterances including syllables, words (2 or 3), and one digit numbers

calculated from linear frequency domain as follows:

$$Mel(f) = 2595 \log_{10} \left(1 + \frac{f}{700}\right), \tag{A.17}$$

where the unit of f is Hz. In other words, the data in the lower frequency components were analyzed in detain using triangular windows that have higher resolutions, and the data in the higher frequency components were roughly an-

alyzed using triangular windows that had lower resolutions. The mel-scale filter bank was analyzed by weighting sum of amplitude spectrum $|S(k)|$ to output the power of the frequency band corresponding to the window range as follows:

$$m(l) = \sum_{k=lo}^{hi} W(k;l)|S(k)|(l = 1, \cdots, L), \tag{A.18}$$

$$W(k;l) = \begin{cases} \frac{k-k_{lo}(l)}{k_c(l)-k_{lo}(l)} & (k_{lo}(l) \leq k \leq k_c(l)) \\ \frac{k_{hi}-k(l)}{k_{hi}(l)-k_c(l)} & (k_c(l) \leq k \leq k_{hi}(l)) \end{cases}, \tag{A.19}$$

where L was the number of filter bank, $k_{lo}(l)$, $k_c(l)$, and $k_{hi}(l)$ were lower, center, and higher frequency bins of lth filter, respectively. These three indexes satisfied the following relationships between neighboring filters:

$$k_c(l) = k_{hi}(l-1) = k_{lo}(l+1). \tag{A.20}$$

Finally, dth MFCC, $MFCC(d)$, was calculated by applying logarithm transform and discrete cosine transform to the L powers derived from the filter bank analysis in the mel-frequency domain:

$$MFCC(d) = \sqrt{\frac{2}{L}} \sum_{l=1}^{L} \left\{ \log\left(m(l)\right) \cos\left(\left(l - \frac{1}{2}\right)\frac{d\pi}{L}\right)\right\}. \tag{A.21}$$

All speech utterances were parameterized for ASR evaluation in which frame length was set to 25 msec and the frame shift was set to 10 msec. In each frame, 26 dimensional acoustic parameter vector to recognize the impaired speech was constructed, in which 12 of those parameters are MFCCs that were top 12 MFCCs extracted from 24 ones, another one was log energy of the static feature, other 12 dimensions were delta coefficients, and the other was the delta energy.

Phoneme identities were manually given to all the recorded utterances. This thesis basically defined phoneme identities as roman letters. The differences from the public notation were following:

: representing long vowel

N representing syllabic nasal (different from syllables beginning with /n/)

q representing chocked sound

sp representing silence in a sentence

silB representing another silence at a beginning of a sentence

silE representing the other silence at a end of a sentence

Adding one more notice to the above, phoneme labels were given as the labeler heard. For example, in a case of labeling 'mother', Japanese written characters of 'mother' is /o ka a sa N/. If the labeler heard as /o ka: sa N/, then that sample is labeled as /o ka: sa N/.

Julius [91] was used for the recognition decoder. Word accuracy was calculated as the recognition rate for the input speech, which was written as follows:

$$Accuracy\ [\%] = \frac{N - D - S - I}{N} \times 100, \tag{A.22}$$

where N denoted the number of total words, D, S, I denoted number of error words due to deletion, substitution, and insertion, respectively.

Experimental results

Table A.3 shows the result of word accuracy for impaired speech uttered by individual speakers in the recording first group. The averaged word accuracy of impaired speech signals due to the disorder in the brain including YT060910a, YI061119a, TY070130e is 39.66 %, on the other hand, that of impaired speech due to hearing-impairment including KK061005c, KS061005c, NT061004b, TF061005b, TT061005c, and YS061005b is 83.02 %. If subjects have a disorder in their brain, the articulations are often distorted. As a result, the impaired speech is difficult to be recognized even though the recognition task is extremely simple and strict. In the recognition of impaired speech including distorted articulation, not only simple transformations of linear regressions but also addressing many problems such as considering acoustic features [62], modeling specific fillers that occur before and after the utterance and so on. Although the task is difficult, the possibility of the ASR system for such people remains by setting strictly defined task such as simple yes or no question. On the other hand, the result shows that acquired hearing loss of impaired speech is well recognized. This is because the problem of those impaired speech utterances are mainly unstable speech volumes

and the articulation is comparatively stable. Although it might be difficult for them to use difficult task such as dictations, acquired hearing loss or impaired people would be able to use ASR system in a simple task in which the size of the vocabulary is not large.

Figure A.3, **Figure A.4**, and **Figure A.5** show the other results of word accuracy for EL speech, EL(air) speech, and EL(small) speech, respectively. The adaptation technique dramatically improves the recognition accuracy even the initial model is normal speech of non-laryngectomees. The merit of using converted normal speech as the adaptation data is the recognition accuracy using the initial SI model (0th iteration in the figures) is better than the original electrolaryngeal speech. It is notable that the converted normal speech from EL speech is recognized more than 70 % even though the initial SI model is used. In the other cases of converted speech from EL(air) and EL(small) speech, the word accuracy is around 50 % that is not high; however, it would be effective by introducing constraints for the recognition tasks. In the case of adaptation from the converted normal speech, the over-training is occurred since the recognition accuracy is degraded after a few adaptation data. Although the recognition accuracy of the converted speech using SI model is better than that of the source EL speech, the performance of the source speech goes above that of the converted speech as the adaptation goes. It is thought that source EL speech keeps clear phoneme boundaries and identities in the acoustic feature space; on the other hand, such phoneme environment would be distorted by the conversion procedure. Since the model parameters are adapted in maximum likelihood criterion, more suitable adaptation might be applied by setting the input acoustic features more close to the maximum likelihood values. The results, however, show the contrast trend. Although the differences are little, the word accuracy considering GV is better than the word accuracy not considering GV. This fact would be seen that GV works as the penalty factor for the conversion errors and it is interesting fact as one of the GV tricks. From these results, high recognition accuracies are seen in the significantly different task of the dictation. These results are worthwhile since the possibility of using the ASR system for laryngectomees who usually speaks with EL speech is confirmed.

Table A.3. Word accuracy of various kinds of speaking-impaired people. Acoustic model after 10 iterations of MLLR is used for the SD-AM

Speaker index	Acc. using SI-AM [%]	Acc. using SD-AM [%]
YT060910a	6.25	43.75
KK061005c	30.56	77.78
KS061005c	16.67	69.45
NT061004b	15.63	75.00
TF061005b	30.55	94.44
TT060910c	23.53	97.06
YS060910b	31.25	84.38
YI060910a	26.67	46.67
NS061213d	5.88	26.47
TY070130e	10.72	28.57

A.4 Conclusion

This appendix experimentally evaluated ASR for impaired-speech data including disorders of the brain, hearing impairment, and laryngectomy as case studies.

In the ASR for impaired speech of one group including the disorders of the brain and hearing impairment, although the experimental conditions were not completely the same, a total of 10 subjects uttered Japanese syllables, words, short sentences, and so on, in which some utterances were recorded twice. Since the vocabulary size was extremely small, strictly designed network grammar was introduced in which all input utterances were recognized as one word even though it would be sentences, and a few words. MLLR was introduced as a powerful speaker adaptation technique to make the model well represent the acoustic feature spaces of impaired speech. SI acoustic model of normal speech uttered by non-impaired people were used as an initial model. Almost 23 utterances were used for the adaptation, and almost 17 utterances were used for the test. As a result, although the impaired speech due to the impairment of the brain was more difficult to recognize since the articulation was significantly distorted, other impaired speech due to the hearing impairment was well recognized (more than

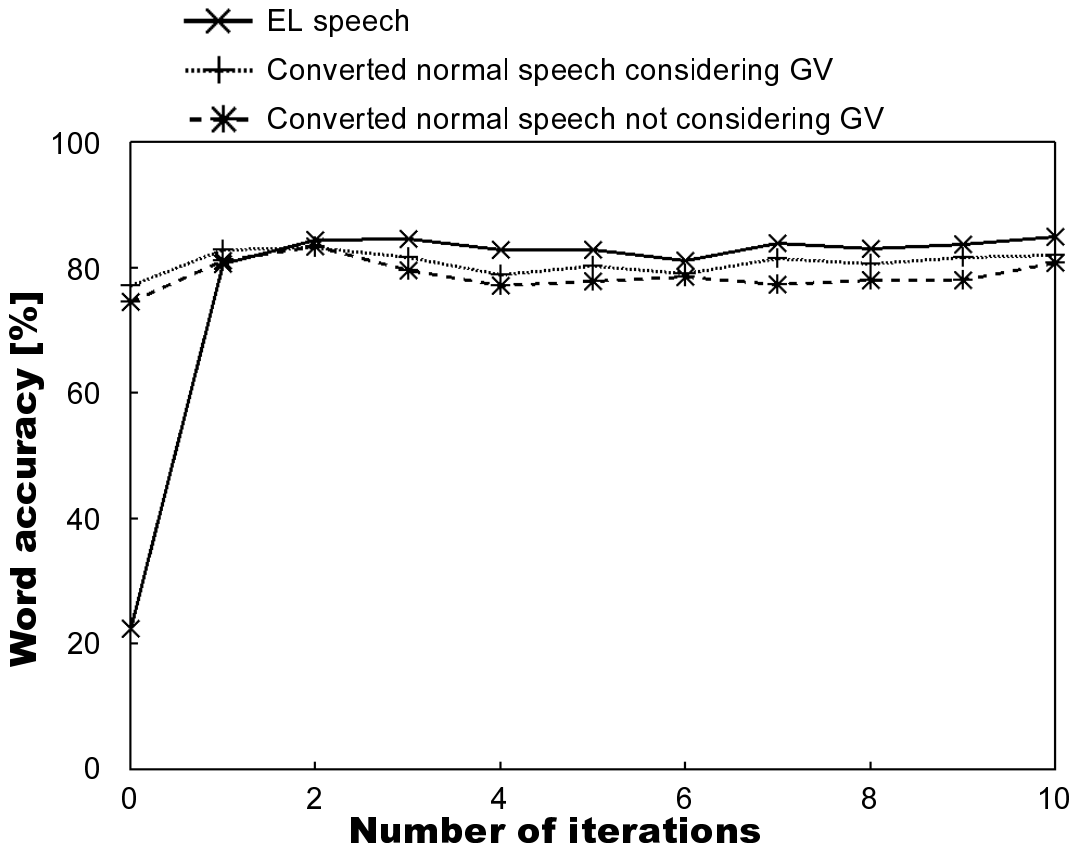

Figure A.3. Word accuracy of EL speech and converted normal speech signals.

80 %) in this simple task.

In the other group of impaired speech due to laryngectomy, which was the source speaker of VC in this thesis, acoustically more stable adaptation data were obtained. 50 phoneme-balanced sentences were used for the speaker adaptation, and 30 newspaper utterances were used for the test. The same SI model as the previous group was used was used as the initial model for speaker adaptation. Three kinds of EL speech were recognized, which were conventional EL speech, EL(air) speech, and EL(small) speech using sawtooth waves, whose power was the same as the pulse train. Moreover, three kinds of speech signals in each EL speech were recognized, which were the original EL speech, the converted normal speech considering GV, and the converted normal speech not considering GV.

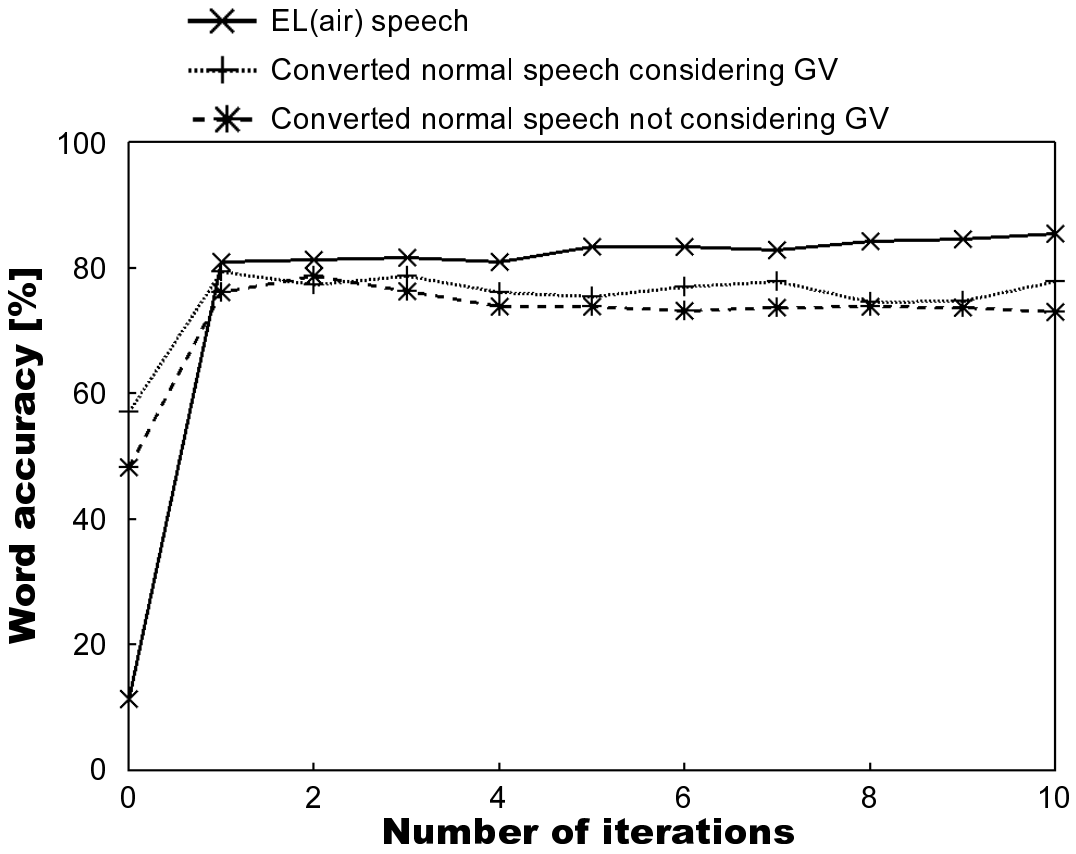

Figure A.4. Word accuracy of EL(air) speech and converted normal speech signals.

Converted normal speech signals were the result of VC conducted in **Chapter 6**. As the result of experiments, the recognition accuracy using the converted normal speech was dramatically higher than that using the original EL speech. Especially in the converted normal speech from EL speech, the recognition accuracy using the initial model was almost 80 %. After the first speaker adaptation, recognition accuracy of EL speech and EL(air) speech was dramatically improved so that it was comparable to the accuracy of converted speech. In the case of EL(small) speech, it needed more iteration times so that the accuracy of the EL(small) speech was comparable to that of the converted speech.

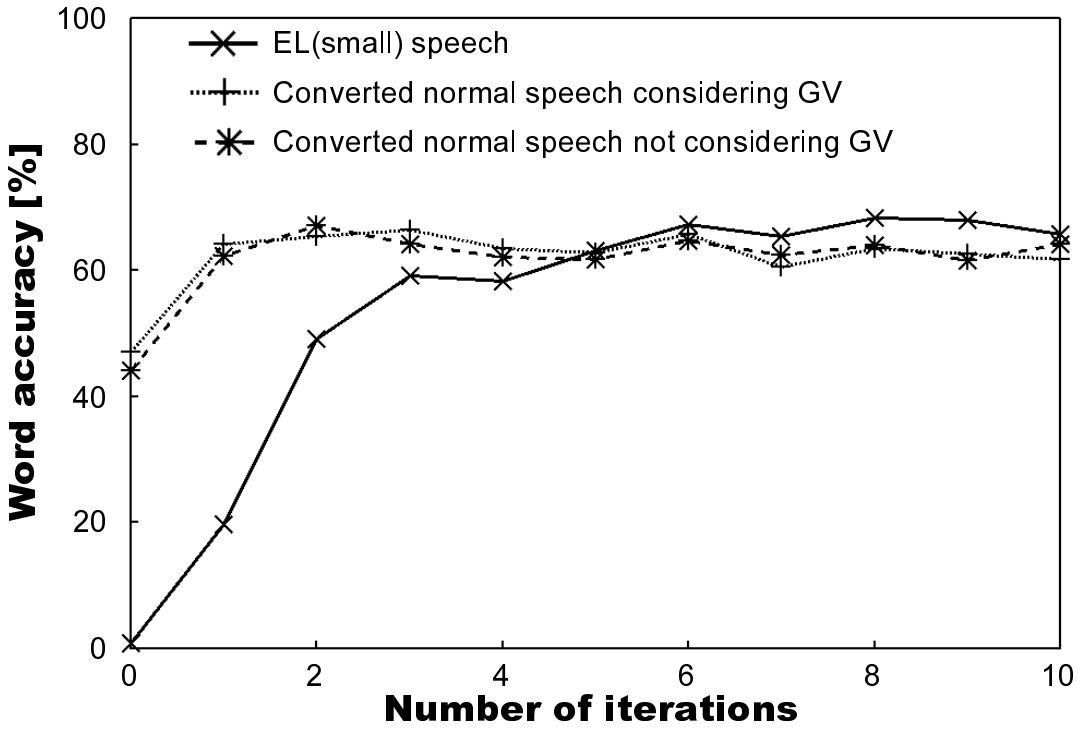

Figure A.5. Word accuracy of EL(small) speech using the sawtooth waves and converted normal speech signals.

References

[1] P. Enderby and R. Philipp, "Speech and Language Handicap: Towards Knowing the Size of the Problem," International Journal of Language and Communication Disorders, Vol. 21, No. 2, pp. 151–165, 1986.

[2] S. Ebihara, "Alaryngeal Speech after Total Laryngectomy," Cancer and Chemotheraphy, Vol. 13, No. 11, pp. 3109–3113, Nov. 1986 (in Japanese).

[3] M. Mohri, M. Amatsu, "Current Situations of Speech Rehabilitation after Total Laryngectomy and Future Works," Journal of Otolaryngology, Head and Neck Surgery, Vol. 18, No. 4, pp. 819–822, April 2002 (in Japanese).

[4] H. Takahashi and K. Yajin, "Voice Rehabilitation after Total Laryngectomy," Journal of the Hiroshima Medical Association, Vol. 56, No. 10, pp. 634–638, Oct. 2003 (in Japanese).

[5] M. I. Singer and E. D. Blom, "An Endoscopic Technique for Restoration of Voice after Laryngectomy," The Analysis of Otology, Rhinology, and Laryngology, Vol. 89, pp. 529–533, 1980.

[6] M. Oitsuki, "Socialization by Laryngectomees," Journal of Otolaryngology, Head and Neck Surgery, Vol. 18, No. 4, pp. 823–826, April 2002 (in Japanese).

[7] M. Hashiba, N. Uemi, M. Oikawa, Y. Yamaguchi, Y. Sugai, and T. Ifukube, "Industrialization of the Electrolarynx with a Pitch Control Function and Its Evaluation," IEICE Transactions D-II Vol.J94-D-II, pp. 1240–1247, Jun 2001 (in Japanese).

[8] Y. Kikuchi and H. Kasuya, "Consideration of F0 Control of Electric Larynx," Technical Report of The Institute of Electronics, Information and Communication Engineers (IEICE), SP2002-106, WIT2002-46, pp. 65–68, Oct. 2002 (in Japanese).

[9] D. Cole, S. Sridharan, M. Moody, and S. Geva, "Application of noise reduction techniques for alaryngeal speech enhancement," Proceedings of IEEE TENCON '97, Vol. 2, pp. 491–494, Dec. 1997. TENCON '97.

[10] P. C. Pandey, S. M. Bhandarkar, G. K. Bachher, and P. K. Lehana, "Enhancement of Alaryngeal Speech Using Spectral Subtraction," Proceedings of 14th International Conference of Digital Signal Processing (DSP 2002), pp. 591–594, Santorini, Greece, 2002.

[11] H. L. Barney, F. E. Haworth, and H. K. Dunn, "An experimental transistorized artificial larynx," Bell system technical Journal, Vol. 38, pp. 1337–1356, July 1959.

[12] M. S. Weiss, G. H. Yeni-Komshian, and J. M. Heinz, "Acoustical and perceptual characteristics of speech produced with an electronic artificial larynx," Journal of Acoustical Society of America, Vol. 65, No. 5, pp. 1298–1308, May 1979.

[13] Y. Nakajima, H. Kashioka, K. Shikano, and N. Campbell, "Remodeling of the Sensor for Non-Audible Murmur (NAM)," Proceedings of Interspeech 2005, pp. 293–296, 2005.

[14] Y. Nakajima, H. Kashioka, N. Campbell, and K. Shikano, "Non-Audible Murmur (NAM) Recognition," IEICE Transactions on Information and Systems, Vol.E89-D, No. 1, pp. 1–8, 2006.

[15] T. Toda and K. Shikano, "NAM-to-Speech Conversion with Gaussian Mixture Models," Proceedings of Interspeech 2005, pp. 1957–1960, September 2005.

[16] M. Nakagiri, T. Toda, H. Kashioka, and K. Shikano, "Improving Body Transmitted Unvoiced Speech with Statistical Voice Conversion," Proceedings of Interspeech 2006, pp. 2270–2273, Pittsburgh, USA, Sep. 2006.

[17] T. Ifukube, "Sound-based Assistive Technology for the Disabled," Corona Publishing Co. Ltd, Tokyo, Japan, 1997 (in Japanese).

[18] J. F. Bosma, Martin W. Donner, E. Tanaka, and D. Robertson, "Anatomy of the Pharynx, Pertinent to Swallowing," Dysphagia, Vol. 1, No. 1, pp. 23–33, March, 1986.

[19] K. Yoshino, "An Epidemiology and Clinico-statistics of Laryngeal Cancer," Journal of Otolaryngology, Head and Neck Surgery, Vol. 18, No. 4, pp. 723–729, April 2002 (in Japanese).

[20] K. Umatani, Y. Tsuruta, K. Yoshino, H. Miyahara, and T. Sato, "Laryngectomee Statistics in Japan," Journal of Japan Bronchoesophagology Society, Vol. 36, No. 3, pp. 261–266, 1985 (in Japanese).

[21] The Research Group for Population-based Cancer Registration in Japan, "Cancer Incidence and Incidence Rates in Japan in 1996: Estimates Based on Data from 10 Population-based Cancer Registries," Japanese Journal of Clinical Oncology, Vol. 31, No. 8, pp. 410–414, 2001.

[22] T. Takafuji, "Current Situations of Alaryngeal Speech by Laryngectomees," Journal of Otolaryngology, Head and Neck Surgery, Vol. 2, No. 5, pp. 527–531, May 1986 (in Japanese).

[23] T. Nakajima, "Latest Knowledge about Sources Causing the Laryngeal Cancer," Journal of Otolaryngology, Head and Neck Surgery, Vol. 18, No. 4, pp. 731–734, April 2002 (in Japanese).

[24] M. Suzuki, "Current Situations and Future Perspective of Medical Examination for Laryngeal Cancer," Journal of Otolaryngology, Head and Neck Surgery, Vol. 18, No. 4, pp. 771–774, April 2002 (in Japanese).

[25] N. Nishiyama, M. Nishio, M. Myojin, K. Shirai, "Improvement of Radiation Therapy for Laryngeal Cancer," Journal of Otolaryngology, Head and Neck Surgery, Vol. 18, No. 4, pp. 781–786, April 2002 (in Japanese).

[26] Y. Hisa, "Laser Surgery in Laryngeal Cancer : Application and Current Situation," Journal of Otolaryngology, Head and Neck Surgery, Vol. 18, No. 4, pp. 788–792, April 2002 (in Japanese).

[27] R. Hayashi and S. Ebihara, "Partial Laryngectomy in Therapy for Laryngeal Cancer," Journal of Otolaryngology, Head and Neck Surgery, Vol. 18, No. 4, pp. 793–796, April 2002 (in Japanese).

[28] K. Nagahara, "Application and Ranges of Supracricoid Laryngectomy with Cricohyoidoepiglottopexy (SCL-CHEP)," Journal of Otolaryngology, Head and Neck Surgery, Vol. 18, No. 4, pp. 798–802, April 2002 (in Japanese).

[29] S. E. Williams and J. B. Watson, "Differences in Speaking Proficiencies in Three Laryngectomee Groups," Arch Otolaryngol Vol. 111, pp. 216–219, April 1985.

[30] Irena Hočevar-Boltežar and Miha Žargi, "Communication after Laryngectomy," Radiation Oncology, pp. 249–254, Vol. 35, No. 4, 2001.

[31] K. Kotake and M. Sato, "The Relationships between Communication Methods for the Patients after Laryngectomy," Journal of Japanese Society of Nursing Research, Vol. 28, No. 1, pp. 109–113, 2005 (in Japanese).

[32] H. F. Nijdam, A. A. Annyas, H. K. Schutte, and H. Leever, "A New Prosthesis for Voice Rehabilitation after Laryngectomy," Archives of Oto-Rhino-Laryngology, Vol.237, pp. 27–33, 1982.

[33] S. E. Chalstrey, N. R. Bleach, D. Cheung, C. A. Van Hasselt, and M. Med, "A Pneumatic Artificial Larynx Popularized in Hong Kong," The Journal of Laryngology and Otology, Vol. 108, pp. 852–854, Oct. 1994.

[34] T. Ifukube, M. Hashiba, J. Matsushima, "A Role of 'Waveform Fluctuation' on the Naturality of Vowels," Acoustical Society of Japan, Vol. 47, No. 12, pp. 903–910, 1991 (in Japanese).

[35] N. Uemi, T. Ifukube, M. Takahashi, and J. Matsushima, "Design of a New Electrolarynx Having a Pitch Control Function," Proceedings of 3rd IEEE

International Workshop of Robot and Human Communication, pp. 198-203, Nagoya, Japan, 1994.

[36] SECOM MYVOICE,
"http://www.secom.co.jp/personal/medical/myvoice.html (in Japanese)".

[37] H. Takahashi, M. Nakano, T. Okusa, Y. Hatamura, Y. Kikuchi, and K. Kaga, "A Voice-generation System Using an Intramough Vibrator," The Japanese Society for Artificial Organs, Vol. 4, pp. 288–294, 2001.

[38] H. Fujisaki and H. Sudo, "A Model for the Generation of Fundamental Frequency Contours of Japanese Word Accent," Acoustical Society of Japan, Vol. 27, No. 9, pp. 445–452, 1971 (in Japanese).

[39] K. Murakami, K. Araki, M. Hiroshige, and K. Tochinai, "A Method for Speech Transform from Electrolaryngeal Speech to Normal Speech," IEICE Transactions D-I Vol. J87-D-I, No. 11 pp. 1030–1040, Nov. 2004 (in Japanese).

[40] S. Hayamizu and R. Oka, "Experimental Studies on the Connected Words Recognition Using Continuous Dynamic Programming," IEICE Transactions Vol. J67-D, No. 6, pp. 677–684, 1984 (in Japanese).

[41] R. L. Norton and R. S. Bernstein, "Improved Laboratory Prototype Electrolarynx (LAPEL): Using Inverse Filtering of the Frequency Response Function of the Human Throat," Annals of Biomedical Engineering, Vol. 21, No. 2, pp. 163–174, March 1993.

[42] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," IEEE Transactions on Acoustics, Speech and Signal Processing, Vol. 27, No. 2, pp. 113–120, April 1979.

[43] H. Liu, Q. Zhao, M. Wan, and S. Wang, "Enhancement of Electrolarynx Speech Based on Auditoy Masking," IEEE Transactions on Biomedical Engineering, Vol. 53, No. 5, pp. 865–874, May 2006.

[44] M. R. Schroeder, B. S. Atal, and J. L. Hall, "Optimizing Digital Speech Coders by Exploiting Masking Properties of the Human Ear," Journal of Acoustical Society of America, Vol. 64, No. S1, pp. S139–S139, Nov. 1978.

[45] M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 208–211, April, 1979.

[46] J. Robbins, H. B. Fisher, E. C. Blom, M. I. Singer, "A Comparative Acoustic Study of Normal, Esophageal, and Tracheoesophageal Speech Production," Journal of Speech and Hearing Disorders, Vol. 49 pp. 202–210, May 1984.

[47] A. Hisada and H. Sawada, "Real-time Clarification of Esophageal Speech Using a Comb Filter," Proceedings of 4th International Conference on Disability, Virtual Reality and Associated Technologies, pp. 39–46, 2002.

[48] K. Matsui, N. Hara, N. Kobayashi, and H. Hirose, "Enhancement of Esophageal Speech Using Formant Synthesis," Acoustical Science and Technology, Vol. 23, No. 2, pp. 69–76, 2002.

[49] D. H. Klatt, "Software for a cascade/parallel formant synthesize," Journal of Acoustical Society of America, Vol. 67, No. 3, pp. 971–995, 1980.

[50] H. Doi, K. Nakamura, T. Toda, H. Saruwatari, and K. Shikano, "Enhancement of Esophageal Speech Using Statistical Voice Conversion," 2009.

[51] E. Hanada, Y. Sobajima, K. Matsumura, S. Tenpaku, K. Kusuhara, T. Umezaki, T. Hara, S. Komiyama, Y. Watanabe, and Y. Nose, "A rapid speech synthesizing software on a PDA for Japanese with speech impairments," Transactions on Neural Systems and Rehabilitation Engineering, Vol. 11, pp. 386–391, Dec. 2003.

[52] Y. Hosoi and T. Sakaguchi, "Silent voice input system without exhalation -theory and applications-," Technical Report of IEICE, SP2003-105, pp. 13–16, 2003.

[53] V. A. Tran, G. Bailly, H. Loevenbruck, C. Jutten, "Improvement to a NAM captured whisper-to-speech system," Proceedings of Interspeech 2008, pp. 1465–1468, Brisbane, Australia, September, 2008.

[54] V. A. Tran, G. Bailly, H. Loevenbruck, and T. Toda, "Multimodal HMM-based NAM-to-speech conversion," Proceedings of Interspeech 2009, pp. 656–659, Brighton, UK, September, 2009.

[55] K. Tokuda, T. Kobayashi, and S. Imai, "Speech parameter generation from HMM using dynamic features," Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 660–663, Detroit, USA, May, 1995.

[56] K. Tokuda, T. Masuko, T. Yamada, T. Kobayashi, and S. Imai, "An algorithm for speech parameter generation from continuous mixture HMMs with dynamic features," Proceedings of Interspeech 1995 (Eurospeech), pp. 757–760, Madrid, Spain, Sep. 1995.

[57] A. Kunikoshi, Y. Qiao, N. Minematsu, and K. Hirose, "Speech Generation from Hand Gestures Based on Space Mapping," Proceedings of Interspeech 2009, pp. 308-311, Brighton, UK, September, 2009.

[58] A. Kain and J. V. Santen, "Using Speech Transformation to Increase Speech Intelligibility for The Hearing- and Speaking-Impaired," Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 3605–3608, Taipei, Taiwan, April, 2009.

[59] Cabinet Office, "The 2009 annual report," http://www8.cao.go.jp/shougai/whitepaper/index-w.html.

[60] S. Young, "A Review of Large-vocabulary Continuous-speech Recognition," IEEE Signal Processing Magazine, Vol. 13, No. 5, pp. 45–57, 1996.

[61] H. Matsumasa, K. Tanaka, T. Takiguchi, Y. Ariki, I. Li, and T. Nakabayashi, "Evaluation of Speech Recognition by a Person with Articulation Disorder in Operation for Home Information Appliances," Technical Report of IEICE, Vol. 107, No. 61, pp. 33–38, May 2007.

[62] H. Matsumasa, T. Takiguchi, Y. Ariki, I. Li, and T. Nakabayashi, "PCA-Based Feature Extraction for Fluctuation in Speaking Style of Articulation Disorders," Proceedings of Interspech 2007 (Eurospeech), pp. 1150–1153, Antwerp, Belgium, 2007.

[63] A. L. Kotler and N. T. Stonel, "Effects of speech training on the accuracy of speech recognition for an individual with a speech impairment," Augmentative and Alternative Communication, Vol. 13, No. 2, pp. 71–80, 1997.

[64] M. Inoue, H. Egashira, Y. Okazaki, K. Watanabe, and H. Kondo, "Input Support System via Voice for a Physically Handicapped Person with an Utterance Obstacle," IEICE Technical Report in Education Technology, Vol. 102, pp. 29–34, 2003 (in Japanese).

[65] A. Kain and M. W. Macon, " Spectral voice conversion for text-to-speech synthesis, " Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 285–288, Seattle, USA, May 1998.

[66] A. P. Dempster, N. M. Laird, D. B. Rubin, "Maximum Likelihood from Incomplete Data via the EM algorithm," Journal of the Royal Statistical Society. Series B (Methodological), Vol. 39, No. 1, pp. 1–38, 1977.

[67] Y. Stylianou, O. Cappe, and E. Moulines, "Continuous probabilistic transform for voice conversion," IEEE Transaction on Speech and Audio Processing (SAP), Vol. 6, No. 2, pp. 131–142, 1998.

[68] T. Toda, A. W. Black, and K. Tokuda, "Voice Conversion Based on Maximum Likelihood Estimation of Spectral Parameter Trajectory," IEEE Transactions on Audio, Speech and Language Processing, Vol. 15, No. 8, pp. 2222–2235, Nov. 2007.

[69] H. Kawahara, J. Estill, and O. Fujimura, "Aperiodicity extraction and control using mixed mode excitation and group delay manipulation for a high quality speech analysis, modificationand synthesis system STRAIGHT, " 2nd Models and Analysis of Vocal Emissions for Biomedical Applications (MAVEBA), Firenze, Italy, Sept. 2001.

[70] Y. Ohtani, T. Toda, H. Saruwatari, and K. Shikano, "Maximum Likelihood Voice Conversion Based on GMM with STRAIGHT Mixed Excitation," the 9th International Conference on Spoken Language Processing (Interspeech 2006 - ICSLP), pp. 2266–2269, Sept. 2006.

[71] T. Fukada, K. Tokuda, T. Kobayashi, and S. Imai, "An adaptive algorithm for mel-cepstral analysis of speech," Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP), Vol. 1, pp. 137–140, San Francisco, USA, March 1992.

[72] 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: Possible role of a repetitive structure in sounds," Speech Communication, Vol. 27, No. 3-4, pp. 187–207, Apr. 1999.

[73] Peter Howell, "Effects of delayed auditory feedback and frequency-shifted feedback on speech control and some potentials for future development of prosthetic aids for stammering," Stammering Resolution, Vol. 1, No. 1, pp. 31–46, April 2004.

[74] A. J. Yates, "Delayed Auditory Feedback," Psychological Bulletin, pp. 213–232, Vol. 60, No. 3, May 1963.

[75] Japan Electronic Industry Development Association, Noise Database, NOS-9601, Mar. 1996

[76] D. Talkin, "A Robust Algorithm for Pitch Tracking (RAPT)," "Speech Coding and Synthesis," 0-444-82169-4, pp. 495–518 (Chapter 14), 2003.

[77] Y. Saikachi, K. N. Stevens, R. E. Hillman, "Development and Perceptual Evaluation of Amplitude-Based F0 Control in Electrolarynx Speech," Journal of Speech, Language, and Hearing Research, Vol. 52, pp. 1360–1369, Oct. 2009.

[78] J. Lafferty, A. McCallum, and F. Pereira, "Conditional random fields: probabilistic models for segmenting and labeling sequence data, " Proceedings of the International Conference on Machine Learning (ICML-2001) ,2001 .

[79] R. Kuhn, J.-C. Junqua, P. Nguyen, and N. Niedzielski, "Rapid speaker adaptation in eigenvoice space," IEEE Transaction on Speech and Audio Processing (SAP), Vol. 8, No. 6, pp. 695–707, 2007.

[80] T. Toda, Y. Ohtani, and K. Shikano, "One-to-Many and Many-to-One Voice Conversion Based on Eigenvoices," International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Vol. 4, pp. 1249–1252, Apr. 2007.

[81] T. Muramatsu, Y. Ohtani, T. Toda, H. Saruwatari, and K. Shikano, "Low-Delay Voice Conversion based on Maximum Likelihood Estimation of Spectral Parameter Trajectory," Proceedings of Interspeech 2008, pp. 1076–1079, Brisbane, Australia, September, 2008.

[82] K. Koishida, K. Tokuda, T. Masko, and T. Kobayashi, "Vector Quantization of Speech Spectral Parameters Using Statistics of Static and Dynamic Features," IEICE Transactions on Information and Systems, Vol. E84-D, No. 10, pp. 1427–1434, 2001.

[83] J. A. Bilmes, "What HMMs can do," IEICE Transactions on Information and Systems, Vol. E89-D, No. 3, pp. 869–891, 2006.

[84] A. Lee, T. Kawahara, K. Takeda, K. Shikano, "A new phonetic tied-mixture model for efficient decoding," Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP), Vol. 3, pp. 1269–1272, 2000.

[85] J. R. Bellegarda and D. Nahamoo, "Tied Mixture Continuous Parameter Models for Large Vocabularyisolated Speech Recognition," Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 13–16, May, 1989.

[86] A. Sankar, "A New Look at HMM Parameter Tying for Large Vocabulary Speech Recognition," Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 2219–2222, 1998.

[87] M. J. F. Gales, "Maximum Likelihood Linear Transformations for HMM-based Speech Recognition," Computer Speech and Language, vol. 12, pp. 75–98, 1998.

[88] M. J. F. Gales, and P. C. Woodland, "Mean and variance adaptation within the MLLR framework," Computer Speech and Language, Vol. 10, pp. 249–264, 1996.

[89] C. J. Leggetter and P. C. Woodland, "Computer Speech and Language, Vol. 9, pp. 171–185, 1995.

[90] S. B. Davis and P. Mermelstein, "Comparison of Parametric representations for monosyllabic word recognition in continuously spoken sentence," IEEE Transactions on Acoustics, Speech, and Signal Processing, Vol. ASSP-28, No. 4, pp. 357–366, 1980.

[91] A. Lee, T. Kawahara, and K. Shikano, "Julius - an open source real-time large vocabulary recognition engine," European Conference on Speech Communication and Technology, pp. 1691–1694, 2001.

List of Publications

Journal Papers

1. K. Nakamura, T. Toda, H. Saruwatari, and K. Shikano, "A Speech Communication Aid System for Total Laryngectomees Using Voice Conversion of Body Transmitted Artificial Speech," IEICE Transactions on Information and Systems, Vol. J90-D, No. 3, pp. 780–787, 2007 (in Japanese).

2. K. Nakamura, T. Toda, H. Saruwatari, and K. Shikano, "Evaluation of Extremely Small Sound Source Signals Used in Speaking-Aid System with Statistical Voice Conversion," IEICE Transactions on Information and Systems, 2010 (accepted).

3. T. Hirahara, M. Otani, S. Shimizu, T. Toda, K. Nakamura, Y. Nakajima, and K. Shikano, "Silent-speech enhancement using body-conducted vocal-tract resonance signals," Speech Communication, Vol. 52, Issue 4, pp. 301–313, 2010.

International Conferences

1. K. Nakamura, T. Toda, H. Saruwatari, and K. Shikano, "Electrolaryngeal Speech Enhancement Based on Statistical Voice Conversion," Proceedings of Interspeech 2009 - Eurospeech, pp. 1431–1434, Brighton, UK, Sep. 2009.

2. K. Nakamura, T. Toda, Y. Nakajima, H. Saruwatari and K. Shikano, "Evaluation of Speaking-Aid System with Voice Conversion for Laryngectomees

Toward Its Use in Practical Environments," Proceedings of Interspeech 2008 - ICSLP, pp.2209–2212, Brisbane, Australia, Sep., 2008.

3. K. Nakamura, T. Toda, H. Saruwatari, and K. Shikano, "Impact of Various Small Sound Source Signals on Voice Conversion Accuracy in Speech Communication Aid for Laryngectomees," Proceedings of Interspeech 2007 - Eurospeech, pp. 2517–2520, Antwerp, Belguim, Aug. 2007.

4. K. Nakamura, T. Toda, H. Saruwatari, and K. Shikano, "A speech communication aid system for total laryngectomies using voice conversion of body transmitted artificial speech," ASA/ASJ Joint Meeting, Nov. 2006.

5. K. Nakamura, T. Toda, H. Saruwatari, and K. Shikano, "Speaking Aid System for Total Laryngectomees Using Voice Conversion of Body Transmitted Artificial Speech," Proceedings of Interspeech 2006 - ICSLP, pp. 1395–1398, Pittsburgh, USA, Sep. 2006.

6. D. Miyamoto, K. Nakamura, T. Toda, H. Saruwatari, and K. Shikano, "Acoustic Compensation Methods for Body Transmitted Speech Conversion", IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP2009), pp. 3901–3904, Taipei, Taiwan, April 2009.

7. T. Toda, K. Nakamura, H. Sekimoto, and K. Shikano, "Voice Conversion for Various Types of Body Transmitted Speech", IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP2009), pp. 3601–3604, Taipei, Taiwan, April 2009 (Special session, invited).

8. K. Morizane, K. Nakamura, T. Toda, and H. Saruwatari, and K. Shikano, "Emphasized Speech Synthesis Based on Hidden Markov Models," Proceedings of Oriental COCOSDA 2009, pp. O2-4, Urumqi, China, Aug. 2009.

9. T. Toda, K. Nakamura, T. Nagai, T. Kaino, Y. Nakajima, and K. Shikano, "Technologies for Processing Body Conductive Speech Detected with Non-Audible Murmur Microphone," Proceedings of Interspeech 2009 - Eurospeech, pp. 632–635, Brighton, UK, Sep. 2009 (Special session, keynote).

10. H. Doi, K. Nakamura, T. Toda, H. Saruwatari, K. Shikano, "Enhancement of esophageal speech using statistical voice conversion," Proceedings of Asia-Pacific Signal and Information PRocessing Association Annual Summit and Conference (APSIPA ASC 2009), pp. 805–808, Sapporo, Japan, Oct. 2009.

Technical Reports

1. K. Nakamura, T. Toda, H. Saruwatari, and K. Shikano, "An Evaluation of Statistical Voice Conversion in Speaking-Aid System Using External Source Signals," Technical Report of IEICE, SP2009-57, Vol. 109, No. 259, pp. 49–54, Oct. 2009 (in Japanese).

2. K. Nakamura, T. Toda, H. Saruwatari, and K. Shikano, "Comparison of Small Sound Source Signals in the Speech Communication-Aid System for Laryngectomees," Technical Report of IEICE, SP2007-39, Vol. 107, No. 165, pp. 91–96, July 2007 (in Japanese).

3. K. Nakamura, T. Toda, H. Saruwatari, and K. Shikano, "Evaluation of Artificial Speech Conversion System Using a Laryngectomee's Data ," Technical Report of IEICE, WIT2007-21, Vol. 107, No. 179, pp. 31–36, Aug. 2007 (in Japanese).

4. K. Nakamura, T. Toda, H. Saruwatari, and K. Shikano, "Voice Conversion of Body Transmitted Speech Using Extremely Small Source Signal for Laryngectomized Patients," Technical Report of IEICE, WIT2006-12, pp. 65–70, May 2006 (in Japanese).

5. H. Doi, K. Nakamura, T. Toda, H. Saruwatari, and K. Shikano, "Enhancement of Esophageal Speech Using Statistical Voice Conversion," 2009 Information Processing Society of Japan (IPSJ) SIG Technical Report 2009-SLP-77, No. 18, pp. 1–6, July 2009 (in Japanese).

6. T. Nagai, K. Nakamura, T. Toda, H. Saruwatari, and K. Shikano, "Effectiveness of Speaker Adaptive Training in Non-Audible Murmur Recognition Based on Speaker Adaptation for Various Speakers," 2008 Information

Processing Society of Japan (IPSJ) SIG Technical Report 2008-SLP-73, pp. 7–12, Oct. 2008 (in Japanese).

7. D. Miyamoto, K. Nakamura, T. Toda, H. Saruwatari, and K. Shikano, "Acoustic Compensation Algorithms for Body Transmitted Speech Conversion," Technical Report of IEICE, SP2008-132, Vol. 108, No. 422, pp. 27–42, Jan. 2009 (in Japanese).

8. D. Miyamoto, K. Nakamura, T. Toda, H. Saruwatari, and K. Shikano, "Adaptive Approach to Varying Recording Conditions in Body Transmitted Voice Conversion Based on Acoustic Compensation," 2009 Information Processing Society of Japan (IPSJ) SIG Technical Report 2009-SLP-75, pp. 7–12, Feb. 2009 (in Japanese).

9. K. Morizane, K. Nakamura, T. Toda, H. Saruwatari, and K. Shikano, "Emphasized Speech Generation Using HMM-Based Speech Synthesis," 2009 Information Processing Society of Japan (IPSJ) SIG Technical Report 2009-SLP-75, pp. 27–32, Feb. 2009 (in Japanese).

Meetings

1. K. Nakamura, T. Toda, H. Saruwatari, and K. Shikano, "Evaluation of Statistical Voice Conversion in Electrolaryngeal Transformation Systems for Laryngectomees," Proceedings of 12th Meetings, Young Researchers of Kansai-section Acoustical Society of Japan, pp. 12, Dec. 2009 (in Japanese).

2. K. Nakamura, T. Toda, H. Saruwatari, and K. Shikano, "Investigation of Target Speech in Statistical Voice Conversion for Artificial Speech Generated with Electrolarynx," Proceedings of Spring Meeting, Acoustical Society of Japan, 2-P-8, pp. 463–464, March 2009 (in Japanese).

3. K. Nakamura, T. Toda, Y. Nakajima, H. Saruwatari, and K. Shikano, "Evaluation of Voice Conversion from Artificial Speech of a Laryngectomee

Using Small Sound Sources into Whisper of a Normal Speaker," Proceedings of Fall Meeting, Acoustical Society of Japan, 1-4-15, pp. 321–322, Sep. 2007 (in Japanese).

4. K. Nakamura, "Proposition and Current Situation of Artificial Speech Transformation System for Laryngectomees," Proceedings of Kansai-section Joint Convention of Institutes of Electrical Engineering, S10-3, pp. S50, Nov., 2007 (invited) (in Japanese).

5. K. Nakamura, T. Toda, Y. Nakajima, H. Saruwatari, and K. Shikano, "Investigation of Enhancement the Naturalness of Artificial Speech Applying Statistical Voice Conversion," Proceedings of 10th Meetings, Young Researchers of Kansai-section Acoustical Society of Japan, p. 9, Nov. 2007 (in Japanese).

6. K. Nakamura, T. Toda, Y. Nakajima, H. Saruwatari, and K. Shikano, "Speaking Aid System Using External Small Sound Source and NAM Microphone," Proceedings of meeting, Information Processing Society of Japan, 4L-4, Vol. 5, pp. 355–356, Mar. 2008 (in Japanese).

7. K. Nakamura, T. Toda, H. Saruwatari, and K. Shikano, "Voice Conversion with Body-Transmitted Artificial Speech in Speech Communication Aid for Laryngectomees," Proceedings of Fall Meeting, Acoustical Society of Japan, 1-6-9, pp. 171–172, Sep. 2006 (in Japanese).

8. K. Nakamura, T. Toda, H. Saruwatari, and K. Shikano, "Voice Conversion of Body Transmitted Artificial Speech for Speech Communication-Aid for Total Laryngectomees," Proceedings of Kansai-section Joint Convention of Institutes of Electrical Engineering, G16-5, pp. 368, Nov., 2006 (in Japanese).

9. K. Nakamura, T. Toda, H. Saruwatari, and K. Shikano, "Influence of Extremely Small Sound Source Signals on Voice Conversion Accuracy in Speech Communication Aid for Laryngectomees," Proceedings of Spring Meeting, Acoustical Society of Japan, 2-8-2, pp. 331–332, March 2007 (in Japanese).

10. K. Nakamura, N. Tamura, and K. Shikano, "Adaptation and Validation of Normal Acoustic Model for Speech Handicapped Voice," Proceedings of Fall Meeting, Acoustical Society of Japan, 3-7-4, pp. 109–110, Sep. 2005 (in Japanese).

11. K. Nakamura, T. Toda, H. Saruwatari, and K. Shikano, "Adaptation and Evaluation of Acoustic Model for Impaired Speech," Proceedings of 8th Meetings, Young Researchers of Kansai-section Acoustical Society of Japan, pp. 5, Dec. 2005 (in Japanese).

12. M. Hatano, K. Nakamura, T. Toda, H. Saruwatari, and K. Shikano, "Investigating factors causing large inter-speaker variation of NAM recognition rate," Proceedings of Fall Meeting, Acoustical Society of Japan, 3-1-18, pp. 131–132, Sep. 2009 (in Japanese).

13. Y. Kisaki, K. Nakamura, T. Toda, H. Saruwatari, and K. Shikano, "Voice Quality Control Using Cluster Adaptive Training in HMM-based Speech Synthesis," Proceedings of Fall Meeting, Acoustical Society of Japan, 1-2-8, pp. 251–252, Sep. 2009 (in Japanese).

14. H. Doi, K. Nakamura, T. Toda, H. Saruwatari, and K. Shikano, "Enhancement of Esophageal Speech Based on Statistical Voice Conversion," Proceedings of Fall Meeting, Acoustical Society of Japan, 2-2-5, pp. 295–296, Sep. 2009 (in Japanese).

15. T. Nagai, K. Nakamura, T. Toda, H. Saruwatari, and K. Shikano, "Evaluation of Individual Speaker Dependent Models Adapted to Various Speakers in NAM Recognition," Proceedings of Fall Meeting, Acoustical Society of Japan, 1-1-7, pp. 17–18, Sep. 2008 (in Japanese).

16. D. Miyamoto, K. Nakamura, T. Toda, H. Saruwatari, and K. Shikano, "Acoustic Compensation for Body Transmitted Voice Conversion Based on Constrained Maximum Likelihood Linear Regression," Proceedings of Fall Meeting, Acoustical Society of Japan, 3-4-8, pp. 297–298, Sep. 2008 (in Japanese).

17. T. Nagai, K. Nakamura, T. Toda, H. Saruwatari, and K. Shikano, "Introducing Speaker Adaptive Training for Non-Audible Murmur Recognition for Various Speakers," Proceedings of 11th Meetings, Young Researchers of Kansai-section Acoustical Society of Japan, pp. 3, Dec. 2008 (in Japanese).

18. D. Miyamoto, K. Nakamura, T. Toda, H. Saruwatari, and K. Shikano, "Introducing Acoustic Compensation Methods for Body-conductive Voice Transformation," Proceedings of 11th Meetings, Young Researchers of Kansai-section Acoustical Society of Japan, pp. 4, Dec. 2008 (in Japanese).

19. K. Morizane, K. Nakamura, T. Toda, H. Saruwatari, and K. Shikano, "Preliminary Investigation for Emphasized Speech Synthesis Method Based on Hidden Markov Model," Proceedings of 11th Meetings, Young Researchers of Kansai-section Acoustical Society of Japan, pp. 5, Dec. 2008 (in Japanese).

20. T. Nagai, K. Nakamura, T. Toda, H. Saruwatari, and K. Shikano, "Applying Speaker Adaptive Training to NAM Recognition Based on Speaker Adaptation," Proceedings of Spring Meeting, Acoustical Society of Japan, 1-P-31, pp. 197–198, Mar. 2009 (in Japanese).

21. K. Morizane, K. Nakamura, T. Toda, H. Saruwatari, and K. Shikano, "Emphasized Speech Synthesis Based on Hidden Markov Models," Proceedings of Spring Meeting, Acoustical Society of Japan, 1-6-5, pp. 299–300, Mar. 2009 (in Japanese).

22. D. Miyamoto, K. Nakamura, T. Toda, H. Saruwatari, and K. Shikano, "Acoustic Compensation Based on Maximum Likelihood Criterion for Body Transmitted Voice Conversion," Proceedings of Spring Meeting, Acoustical Society of Japan, 1-R-23, pp. 435–436, Mar. 2009 (in Japanese).

Master's Thesis

1. K. Nakamura, "A Speaking-Aid System for Laryngectomees based on Statistical Voice Conversion of Body Transmitted Artificial Speech," Master's thesis, Department of Information Processing, Graduate School of

Information Science, Nara Institute of Science and Technology, NAIST-IS-MT0551094, March 2007 (in Japanese).