NAIST-IS-DD1461005

DOCTORAL DISSERTATION

Statistical electrolaryngeal speech production toward voice restoration

Kou Tanaka

March 14, 2017

Graduate School of Information Science
Nara Institute of Science and Technology

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

Kou Tanaka

Thesis Committee:

Professor Satoshi Nakamura	(Supervisor)
Professor Kenji Sugimoto	(Co-supervisor)
Professor Tomoki Toda	(Nagoya university)
Dr. Hirokazu Kameoka	(NTT communication science laboratories)
Assistant Professor Sakriani Sakti	(Co-supervisor)

Statistical electrolaryngeal speech production toward voice restoration[*]

Kou Tanaka

Abstract

Laryngectomees are people with disabilities as a result of surgery to remove their larynges, including the vocal folds, for various reasons such as injury and laryngeal cancer. An electrolarynx is a medical device to help laryngectomees produce intelligible speech, that is electrolaryngeal (EL) speech, by mechanically generating artificial excitation signals, in place of vocal fold vibrations. Unfortunately, there are three main disadvantages: 1) the resulting sound is characterized as being mechanical and robotic because of the fundamental frequency (F_0) pattern of the monotonic excitation signals, 2) the excitation signals are emitted outside as noise because of the EL speech production mechanism, and 3) one hand is occupied.

In this research, we aim to provide laryngectomees smoothen speech communication and deal with a speaking aid system for EL speech enhancement to recover lost information, such as F_0 patterns, by using speech processing techniques. One technique is A) a speaking aid system capable of modifications at the acoustic level. In this dissertation, we propose a hybrid approach comprising a noise reduction method for enhancing spectral parameters and statistical voice conversion for predicting excitation parameters. Through experimental evaluations, we demonstrate that the proposed hybrid approach effectively addresses not only issue (1) but also (2) by successfully improving the naturalness of EL speech while preserving its intelligibility. However, this approach may not be suitable for face-to-face conversation, because both of the enhanced EL speech

[*]DOCTORAL DISSERTATION, Graduate School of Information Science, Nara Institute of Science and Technology, NAIST-IS-DD1461005, March 14, 2017.

and the original EL speech are presented at the same time. Therefore, we propose another method B), which is a speaking aid system capable of modifications at the physiological level and allows laryngectomees to directly produce the enhanced EL speech from their mouths. In this dissertation, to enable the control of F_0 patterns of the electrolarynx without conscious operation, we propose a method of controlling the F_0 patterns of the electrolarynx by applying a statistical F_0 pattern prediction method to system (B). To address latency issues, we also propose two methods of F_0 pattern modeling. Through experimental evaluations, we demonstrate that the proposed control method incorporating the proposed F_0 pattern modeling effectively addresses issue (1) by successfully improving the naturalness of EL speech.

To recover more natural F_0 patterns, we further propose a new statistical F_0 pattern prediction method, which is applicable to the above-mentioned approaches, considering the speech production mechanism. Since the F_0 patterns are physically constrained by the actual control mechanism of the thyroid cartilage, we incorporate a generative model of the F_0 pattern into the conventional statistical model to predict the F_0 patterns of natural speech. This approach is noteworthy in that it allows the generation of F_0 patterns that are both statistically likely and physically natural. The results of experimental evaluations confirmed that the proposed systems successfully address issue (1) by improving the prediction accuracy of F_0 patterns.

Keywords:

Speaking aid, laryngectomy, electrolarynx, electrolaryngeal speech, statistical F_0 pattern prediction

Acknowledgements (in Japanese)

本論文は，筆者が奈良先端大学大学院情報科学研究科修士課程および博士課程の在学中に，知能コミュニケーション研究室にて行った研究をまとめたものであります．

始めに，修士論文，そしてこの博士論文の完成に至るまで長い間指導して頂いた，恩師であり，かつ，本論文の学位審査の主査である中村哲教授に敬意を込めて感謝します．修士課程および博士課程の在学中に，数多のことを経験できたのは，一重に中村先生のご助力の賜物であります．お忙しい中にも関わらず，学位審査員を引き受けて下さり，中間発表などにおいて数々の有益な助言や指導を頂きました副査の杉本謙二教授に感謝します．また，修士論文，そしてこの博士論文まで長い間指導して頂いた戸田智基教授（名大）に心より感謝します．本研究を進めていく過程で，幾度と無く研究の進むべき方向を正して頂き，数え切れないほどの示唆に富んだご助言を頂きました．当初は長く険しい道のりだと思っていた5年間の大学院生活を楽しく充実して過ごせたのは，日頃の温かい指導のおかげであったように思います．筆者が博士課程からお世話になった亀岡弘和博士（NTT 特別研究員）にも心より感謝いたします．数理的に理路整然と思考する様子には大変感銘を受け，そうした姿を傍らで拝見しながら，大きな刺激を受けるとともに数多くのことを学ぶことができました．知能コミュニケーション研究室のSakriani Sakti助教授とGraham Neubig助教授（CMU）にも感謝します．中村研究室の母である松田真奈美秘書には大変お世話にりました．最後に，いつも暖かく見守ってくれた家族に多大なる感謝をいたします．

Contents

List of Figures

List of Tables

Chapter

1

Introduction

1.1 General background

Information science and technology helps us to break barriers that are insurmountable through our own effort and skill. I believe that research is an action to discover something new of value, realize a better world, and take us to a time of contentment. This dissertation focuses on speech, and the concept of this study is the augmentation of the expressiveness of speech beyond the barriers of speech communication.

Human beings convey information, including their intention to others by various communication media such as speech, letters, gestures, and facial expressions including eye contact. Among them, speech is a natural and principal medium. An advantage of speech is that it is used for not only delivering linguistic information but also conveying paralinguistic information that does not appear in context. Speech is produced by the vocal apparatus and its sound is physically constrained by the conditions of the human body. Speech production is the process by which thoughts are translated into speech as shown in Fig. 1. In the first stage (S-1), how to achieve effects in the listener is determined because communication always starts with the intentions of the speaker. In the second stage (S-2), intention is converted into a suitable word sequence. This stage consists of linguistic study called pragmatics, syntax, and semantics. In the third stage (S-3), related to the studies called phonology, phonetics, and acoustic-phonetics, speech sounds are generated from the word sequence. This stage includes the physiological level. The final stage (S-4) concerns the acoustic level for conveying the speaker's intention.

Speaking disability is a grievous obstacle to communication among human beings. This negative factor leads to communication barriers and deteriorates the quality of life of disabled people. Although it is difficult to know the entire scope of communication disorders [1], in a rough classification, developmental disorders are included in stages S-1 to S-3, language disorders are related to stages S-2 and S-3, and speaking disorders are a part of stage S-3 [2]. The disorders focused on in this dissertation are speaking disorders of total laryngectomees[1], who have undergone surgery to remove their larynges, including the vocal folds,

[1] The term laryngectomee [2] includes partial, total, and supracricoid laryngectomees, in whom the vocal folds have been partially or totally removed.

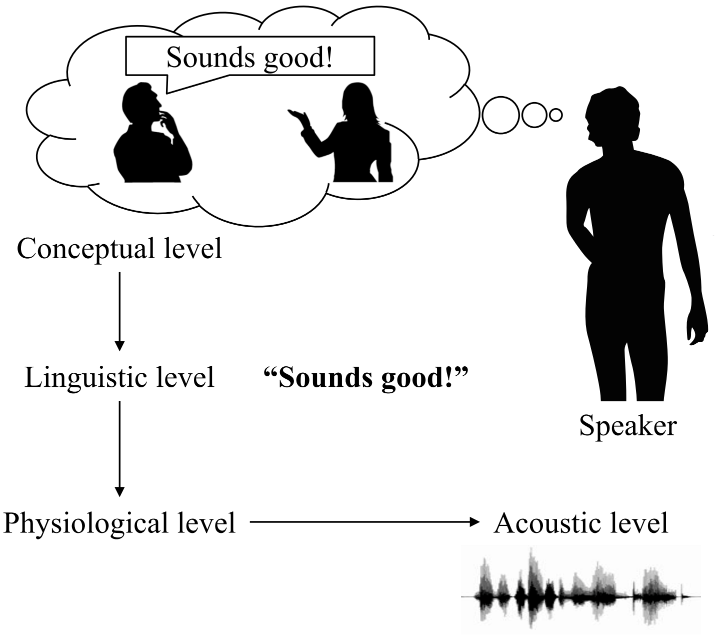


Figure 1. Overview of 4-stage speech production.

for various reasons such as injury and laryngeal cancer. Their ability to generate vocal fold vibrations is severely impaired because they no longer have their vocal folds although their vocal tract for articulating the excitation signals remains.

In more detail, the larynx is an organ located at a position such that it separates the trachea from the esophagus[3]. Laryngeal cancer has the highest incidence among head and neck cancers, although it is a minor type among all cancers [4, 5]. Laryngeal cancer is categorized according to the location of the tumor: glottis cancer, the supra-glottic cancer, and the subglottic cancer. In Japan, although the number of people afflicted with laryngeal cancer in the 70s was less than two thousand, in 1996, it reached almost three thousand [6]. It is said that the number of laryngectomees was estimated to be 20 thousand more than 20 years ago [7], and has recently increased to 30 thousand [8]. Male patients number much more than females, and the disease developed at an advanced age [4]. Some major causes of the disease are smoking and continuous consumption of excess alcohol [9]. It is said that there are almost 600 thousand speech-impaired patients owing to the loss of vocal folds all over the world.

In order to preserve their individual, social, and regular activities, speech-impaired patients need alternative speaking methods that make it possible to generate excitation signals instead of vocal fold vibrations. Speech produced by an alternative speaking method without vocal fold vibrations is called alaryngeal speech [10]. There are various kinds of alaryngeal speech for laryngectomees, such as esophageal (ES) speech [10], tracheoesophageal (TE) shunt speech using a tracheoesophageal puncture (TEP) technique [11], electrolaryngeal (EL) speech using an electrolarynx [12], silent EL speech [13] with a nonaudible murmur (NAM) microphone [14], and speech using a pneumatic artificial (PA) larynx [15].

Many related works have reported the effectiveness of alaryngeal speech. ES speech is accomplished by articulating excitation signals at the beginning of the esophagus by forcing air up from the stomach. Although the ES speaking method allows laryngectomees to speak without the use of any equipment such as a medical device, ES speech has less than 90% intelligibility under ideal conditions [16], and less than that of normal speech uttered by non-laryngectomees under adverse noise conditions [17]. In the TE shunt speaking method, a part of the original vocal apparatus is replaced with a voice prosthesis placed in the TEP created by a surgeon. This artificial larynx is a one-way air valve that allows air to pass from the lungs/trachea to the esophagus when the patient covers the stoma and the resulting sound is similar to a hoarse voice. EL speech is generated using an electrolarynx, which is a major external device to mechanically generate excitation signals by exciting the vibrator. The generated excitation signals are conducted into the speaker's oral cavity through the neck and are articulated to produce the EL speech. Although the electrolarynx allows laryngectomees to generate fairly intelligible speech, the sound is characterized as being mechanical and robotic because of the predefined monotonic excitation signals.

In Japan, most laryngectomees still communicate with others by using the electrolarynx, whereas the current trend in foreign countries is to use the TE shunt for several reasons, such as the need for daily maintenance by themselves and regular maintenance by a doctor. There are three main disadvantages of using the electrolarynx:

1) its sound is characterized as being mechanical and robotic because of the fundamental frequency (F_0) pattern of the monotonic excitation signals,

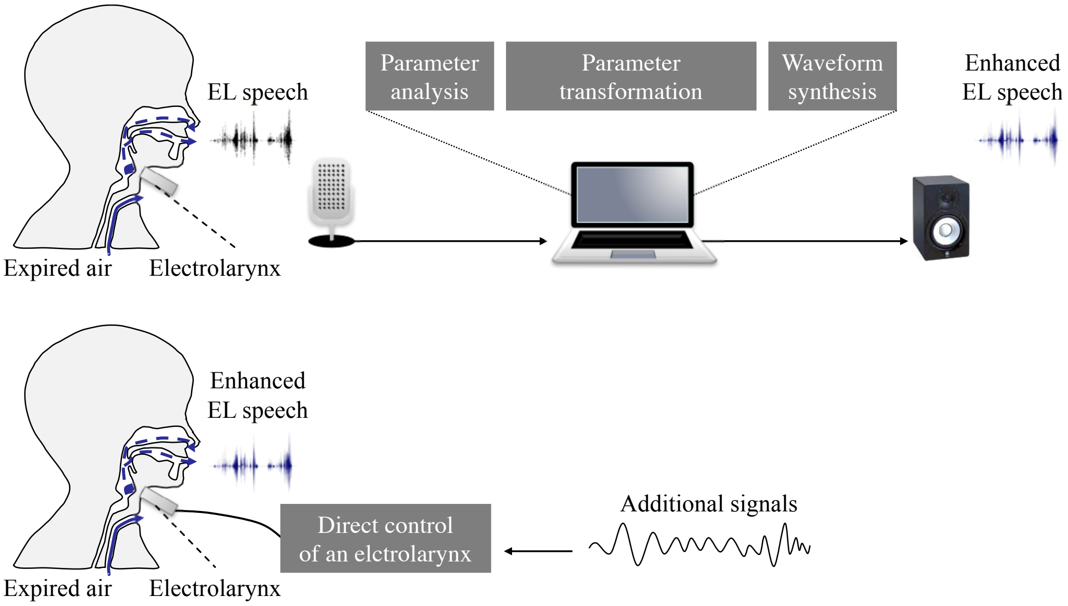


Figure 2. Overview of EL speech enhancement systems at acoustic level (upper side) and physiological level (lower side).

2) the excitation signals are emitted outside as noise because of the EL speech production mechanism,

3) one hand is occupied.

Fortunately, many speech enhancement systems for EL speech have been developed [18, 19, 20, 21]. As shown in the lower illustration of Fig. 2, the concept of these systems is direct control of the F_0 patterns of the vibration of the electrolarynx by using additional signals except for speech signals; then laryngectomees can directly produce more natural sounding EL speech from their mouths. Furthermore, to realize more complex enhancement, which will makes it possible to not only recover F_0 patterns but also enhance spectral properties corresponding to articulation, speech enhancement systems using statistical techniques have also been proposed [16, 22, 23, 24], as shown in the upper illustration of Fig. 2. In the latter systems, the enhanced EL speech is synthesizing on a PC and is presented from a loudness speaker.

However, despite the variety of technologies that have been developed to support laryngectomees, the generated EL speech is still far from achieving a similar quality to natural human speech, in terms of both naturalness and intelligibility. The term naturalness is used as a criterion by the listener to judge the degree to which the generated speech can be perceptually similar to natural human speech, and the term intelligibility is used as the criterion of how well the listener can understand or recognize the content of the speech utterance. Developing a method that can maintain a balance between the naturalness of the speech and the intelligibility of the speech is one of the most important issues that must be resolved. Therefore, in this thesis, we attempt to address those issues. In the next section, we will describe, in detail, the scope of the thesis.

1.2 Research scope

With this research, we aim ultimately to provide laryngectomees with a means of smoothen speech communication. To realize a world in which laryngectomees can communicate like healthy people, we here propose methods that satisfy the following three requirements:

I) improved naturalness of EL speech,

Ⅱ) high intelligibility of EL speech,

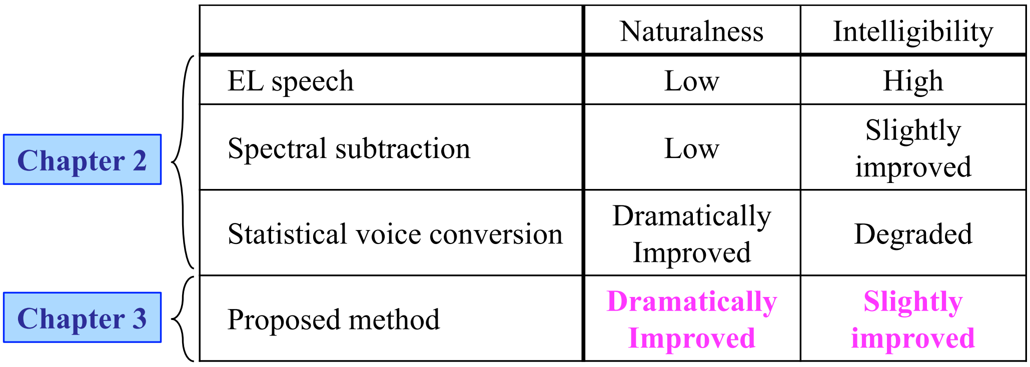
Ⅲ) free laryngectomees from needing to newly learn how to generate the enhanced EL speech.

To realize the above methods, we deal with two types of speaking aid system for EL speech enhancement to recover lost information, such as F_0 patterns, by using speech processing techniques:

A) a speaking aid system capable of modifications at the acoustic level, in which the enhanced speech is presented to the listener using a loudness speaker (related to Fig. 2 upper side),

B) a speaking aid system capable of modifications at the physiological level, allowing laryngectomees to directly produce the enhanced EL speech from their mouths (related to Fig. 2 lower side).

To describe the recovery of F_0 patterns in more detail, we focus on statistical F_0 pattern predictions from spectral parameters corresponding to articulation. As shown in Fig. 1, the shape of F_0 patterns, which are parameters at the acoustic level, is determined at the conceptual level and the linguistic level before the physiological level. In other words, the F_0 patterns are determined from both conceptual parameters, such as the intention and the emotion, and linguistic parameters, such as word sequence. On the other hands, the spectral parameters are also determined by the conceptual parameters and linguistic parameters. Since there is a relationship between the F_0 patterns and the spectral parameters through the conceptual parameters and linguistic parameters, we assume that we can recover the F_0 patterns by using the spectral parameters as a clue. Note that from the point of view of EL speech enhancements at both physiological and acoustic levels, the statistical F_0 pattern prediction has the potential to realize speech production with no apparent differences between the normal speech production and the EL speech production, because there is no requirement, other than to just speaking in a usual manner.

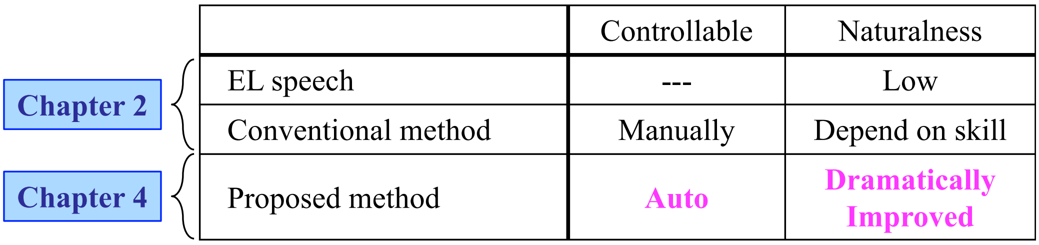


		Naturalness	Intelligibility
Chapter 2	EL speech	Low	High
	Spectral subtraction	Low	Slightly improved
	Statistical voice conversion	Dramatically Improved	Degraded
Chapter 3	Proposed method	**Dramatically Improved**	**Slightly improved**

Figure 3. Problem definition and outline of this dissertation for EL speech enhancement at acoustic level.

For enhancement systems (A), we address the problem arising in conventional studies [16, 22]; the use of statistical techniques in speech enhancement systems to address the unnaturalness of EL speech (see Fig. 3). In the conventional systems using statistical techniques, attempts have been made to address the unnaturalness of EL speech by converting the acoustic features of EL speech into the acoustic features of natural speech while retaining linguistic information. Although these methods dramatically improve the naturalness of EL speech thanks to the recovery of F_0 patterns, the intelligibility degrades for several reasons such as conversion errors and oversmoothing. Since the human ears are sensitive to differences in spectral properties, the intelligibility of speech is severely damaged by spectral conversion errors. Considering the remaining vocal tracts of laryngectomees, we replace the spectral enhancement process using statistical techniques with that using simple signal processing such as spectral subtraction (SS) [25, 26] to reduce the radiated noise, which is emitted excitation signals. The advantage of this replacement is that the data of the natural speech of laryngectomees before surgery is not necessary because the spectral properties of EL speech retain their individuality. The conventional enhancement requires the natural speech data because the speaker individuality of spectral properties should not be changed in this case.

Regarding the F_0 patterns recovered in enhancement systems (A), we modify the original F_0 patterns through accurate modeling with a Gaussian mixture model (GMM). The modified F_0 patterns are obtained by removing rapid movements of F_0 patterns by low-pass filtering after interpolation at unvoiced frames. Thanks to its continuity and less variability, this modification has also realized a accurate modeling in other research of speech processing such as text-to-speech processing [27], in which speech waveforms are synthesized from input text information.



	Controllable	Naturalness
EL speech	---	Low
Conventional method	Manually	Depend on skill
Proposed method	**Auto**	**Dramatically Improved**

Chapter 2
Chapter 4

Figure 4. Problem definition and outline of this dissertation for EL speech enhancement at physiological level.

On the other hand, in the conventional studies of enhancement systems (B), attempts have also been made to address the unnaturalness of EL speech (see

Fig. 4). Some conventional studies have so far striven to make the alternative sound source signals of the electrolarynx close to those of natural vocal fold vibration so that laryngectomees can speak naturally using the electrolarynx. For example, a new electrolarynx [13, 28] and pitch control mechanisms [18, 19, 20, 21] have been proposed. Although these methods allow laryngectomees to directly produce more natural sounding EL speech from their own mouths, the degree of improvement depends on their skills because the control is manual.

To address this issue, in this dissertation, we propose a method of controlling the electrolarynx by predicting F_0 patterns from the acoustic parameters of EL speech by statistical F_0 pattern prediction [16, 22] in real-time [29], which is a part of the statistical voice conversion (VC) technique [30, 31]. Although the proposed approach is capable of reflecting F_0 patterns corresponding to the context information without intentional operation in real-time, the asynchronous problem between articulation and predicted F_0 patterns is caused by the latency occurring in real-time processing. Therefore, in this dissertation, we carefully investigate the relationship between prediction accuracy and latency, and also propose the use of segmental features of smoothed continuous F_0 (CF_0) patterns to make the latency shorter and a forthcoming F_0 pattern prediction method to cancel the asynchronicity.

		Statistically likely	Physically natural	Perceptually natural
Chapter 2	Conventional method	Yes	---	Baseline
Chapter 5	Proposed method	**Yes**	**Yes**	**Improved**

Figure 5. Problem definition and outline of the dissertation for statistical F_0 pattern prediction.

As mentioned, in our proposed methods, the quality of both approaches at the acoustic and physiological levels strongly depends on the prediction accuracy of F_0 patterns. Although the statistical F_0 pattern predictions makes it possible to dramatically improve the naturalness of F_0 patterns, the predicted F_0 patterns are still unnatural compared with those in normal speech. One possible reason is that the predicted F_0 patterns are not necessarily guaranteed to satisfy the physical constraint of the actual control mechanism of the thyroid cartilage, even though they are optimal in a statistical sense (see Fig. 5).

As for the generative process of F_0 patterns, the Fujisaki model [32] has been proposed, and its statistical model [33, 34, 35] has been also proposed to formulate a stochastic counterpart of the Fujisaki model, a well-founded mathematical model representing the control mechanism of vocal fold vibration. These statistical models have made it possible to estimate the underlying parameters of the Fujisaki model that best explain the given F_0 pattern, through the use of powerful statistical inference techniques. To take account of the physical mechanism of vocal phonation, we incorporate these statistical models [33, 34, 35] into the conventional statistical F_0 pattern prediction [16, 22] within a product-of-experts framework [36]. Using this approach, the predicted F_0 patterns are both statistically likely and physically natural.

1.3 Rest of this dissertation

The rest of this dissertation is organized as follows.

Chapter 2: We give anatomical descriptions of the laryngectomees. The way to cure laryngeal cancer is described, and the major alternative speaking methods for laryngectomees are also described. Moreover, we introduce conventional studies of EL speech enhancement systems at both the acoustic and physiological levels. In particular, for a conventional enhancement system at the acoustic level, the statistical F_0 pattern prediction, which is the core technique of the proposed system, is described. The statistical F_0 pattern prediction method used in this dissertation consists of training and conversion parts. Before the training part, the statistical F_0 pattern prediction first defines the source and the target speech to prepare parallel data constructed of time-aligned utterances. Then, a GMM is trained in the training part to model the 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 on the basis of the conditional probability density given the time sequence of the input feature.

Chapter 3: As an EL speech enhancement system at the acoustic level, we present a hybrid approach with a noise reduction method for enhancing spectral parameters and a statistical VC method for predicting excitation parameters. We also propose the modification of trained F_0 patterns to improve prediction accuracy. Moreover, we discuss the proposed method in comparison with the conventional method in terms of naturalness, intelligibility, and listenability and demonstrate that the proposed method yields significant improvements in naturalness compared with EL speech while maintaining sufficiently high intelligibility.

Chapter 4: As an EL speech enhancement system at the physiological level, we propose direct F_0 pattern control of the electrolarynx using real-time statistical F_0 pattern prediction to develop an EL speech enhancement technique also effective for face-to-face conversation. To flexibly investigate the performance of our proposed control method, we also design a simulation method of the EL speech production process using the controlled electrolarynx. Furthermore, we describe the negative impact of latency caused by real-time processing and propose the methods to address the latency issues. By implementing a prototype system and its simulation, we demonstrate that our proposed system successfully addresses the unnaturalness of the electrolarynx and the latency issues.

Chapter 5: To improve the accuracy of statistical F_0 pattern prediction, we propose a statistical F_0 pattern prediction considering the generative process of F_0 patterns within the product-of-experts framework. We introduce the Fujisaki model, which is a well-founded mathematical model representing the control mechanism of vocal fold vibration, and also review its stochastic model. To incorporate the stochastic model for the Fujisaki model into the conventional statistical F_0 pattern prediction, we introduce a latent trajectory model and reformulate the prediction model with a latent trajectory model. Using the constructed model, we derive algorithms for parameter training and F_0 pattern prediction. Through the experimental evaluations, we reveal that the proposed method successfully surpasses the conventional statistical F_0 pattern prediction.

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

Chapter

2

Laryngectomee and speaking-aid system

2.1 Introduction

The larynx is an organ located at the position at where the trachea and the esophagus are separated [3]. 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. Since laryngectomees have undergone to remove their larynges, they cannot produce speech sounds without EL speech speaking method which is a major alternative speaking method. Although EL speech is quite intelligible, there are three major disadvantages: 1) its naturalness is low due to the monotonic excitation signals generated by an electrolarynx, 2) artificial excitation signals to make the produced EL speech sufficiently audible are leaked outside as noise, and 3) one hand is occupied.

To address these issues, several attempts, called EL speech enhancements at physiological or acoustic level, have been developed [18] [19] [20] [21] [37] [13] [22]. Among of them, statistical approaches [22] including statistical F_0 pattern prediction which is a core technique in this dissertation have achieved to dramatically improve the naturalness of EL speech. This data-driven approach is capable of more complicated acoustic modifications to compensate for the large acoustic differences between EL speech and normal speech. As typical conventional methods, the codebook mapping method [38] and a probabilistic conversion method based on GMMs [30] have been applied to alaryngeal speech enhancement [39, 13, 16]. The GMM-based conversion method is one of the most popular voice conversion methods. It is well founded mathematically and its conversion performance is relatively high. It has been reported that the alaryngeal speech enhancement method based on GMM-based voice conversion method is highly effective for improving the naturalness of the different types of alaryngeal speech [39, 13, 16].

This chapter is organized as follows. In **Section 2.2**, we describe the role of the larynx. In **Section 2.3**, we describe laryngeal cancer and the differential between non-disabled person and them. In **Section 2.4**, alternative speaking methods for laryngectomees are explained. In **Section 2.5**, the conventional speaking aid systems at the physiological level are described. In **Section 2.6**, as the conventional speaking aid system at the acoustic level, we review a statistical voice conversion based on GMM.

2.2 Phonation

Phonation is the definition used among those who study laryngeal anatomy and physiology and speech production in general. Phoneticians in other subfields, such as linguistic phonetics, call this process voicing, and use the term phonation to refer to any oscillatory state of any part of the larynx that modifies the airstream, of which voicing is just one example. Voiceless and supra-glottal phonations are included under this definition.

The voicing occurs when air is expelled from the lungs through the glottis, creating a pressure drop across the larynx. When this drop becomes sufficiently large, the vocal folds start to oscillate. The minimum pressure drop required to achieve phonation is called the phonation threshold pressure [40, 41]. The motion of the vocal folds during oscillation is mostly lateral, though there is also some superior component as well. However, there is almost no motion along the length of the vocal folds. The oscillation of the vocal folds serves to modulate the pressure and flow of the air through the larynx, and this modulated airflow is the main component of the sound of most voiced phones.

The sound that the larynx produces is a harmonic series. In other words, it consists of the F_0 accompanied by harmonic overtones, which are multiples of the F_0 [42]. According to the source–filter theory, the resulting sound excites the resonance chamber that is the vocal tract to produce the individual speech sounds.

The vocal folds will not oscillate if they are not sufficiently close to one another, are not under sufficient tension or under too much tension, or if the pressure drop across the larynx is not sufficiently large [43]. In linguistics, a phone is called voiceless if there is no phonation during its occurrence. In speech, voiceless phones are associated with vocal folds that are elongated, highly tensed, and placed laterally when compared to vocal folds during phonation [44].

The F_0 can be varied through a variety of means. Large scale changes are accomplished by increasing the tension in the vocal folds through contraction of the cricothyroid muscle. Smaller changes in tension can be effected by contraction of the thyroarytenoid muscle or changes in the relative position of the thyroid and cricoid cartilages, as may occur when the larynx is lowered or raised, either volitionally or through movement of the tongue to which the larynx is attached via the hyoid bone [44]. In addition to tension changes, F_0 is also affected by the pressure drop across the larynx, which is mostly affected by the pressure in the lungs, and will also vary with the distance between the vocal folds. Variation in F_0 is used linguistically to produce intonation and tone.

2.3 Laryngeal cancer and laryngectomees

Laryngeal cancer is the highest incidence among head and neck cancers although it is a kind of minor disease among all cancers [4, 5]. Although it is a terrible problem for us, early detection of the cancer is comparatively easier than other cancers because in the most cases, some troubles of the neck are observed by the output speech utterances [45]. In these days, the ways to cure the disease are becoming diverse according to the progress of the cancer [46, 47, 48, 49]. 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.

Fig. 6 shows anatomical images of non-laryngectomees and laryngectomees. Larynx works as a valve so that the trachea carries air and the esophagus does

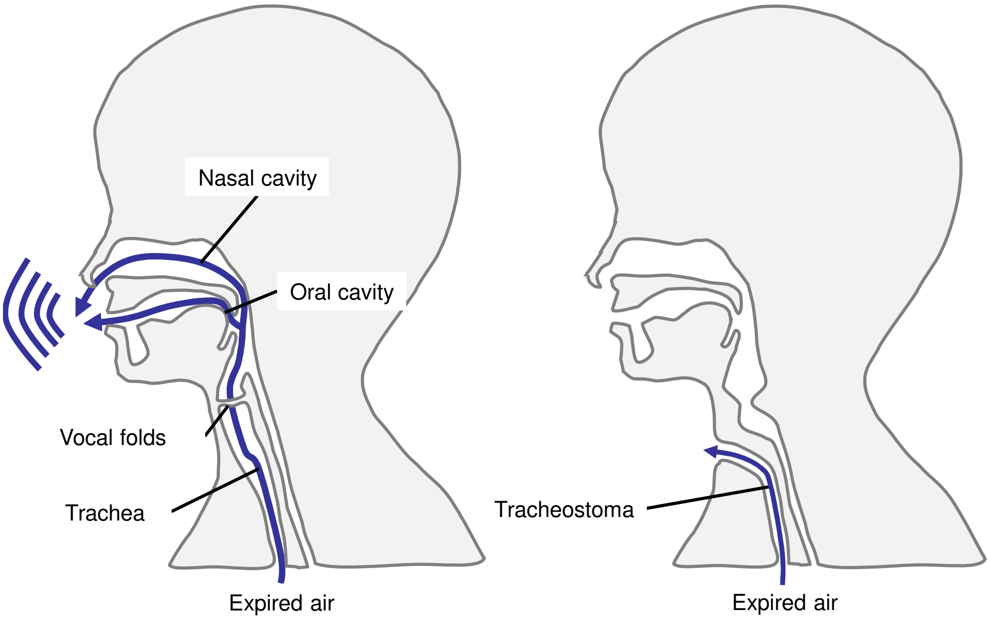


Figure 6. Anatomical image of non-laryngectomees (left side) and total laryngectomees (right side).

foods. 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 to breath. To keep the tracheostoma clean, it is covered by gauze, and certain constraints such as bath is concerned.

2.4 Alaryngeal speech

Laryngectomees mainly have three kinds of alternative speaking methods that are different ways of obtaining the sound sources: 1) esophageal speaking method, 2) tracheoesophageal shunt speaking method, and 3) a method using an external unit [50] such as a pneumatic artificial larynx [51, 52] or an electrolarynx [12]. Fig. 7 shows the route of floating air from the lungs to expiring. The benefits and defects of these methods are shown in Table 1.

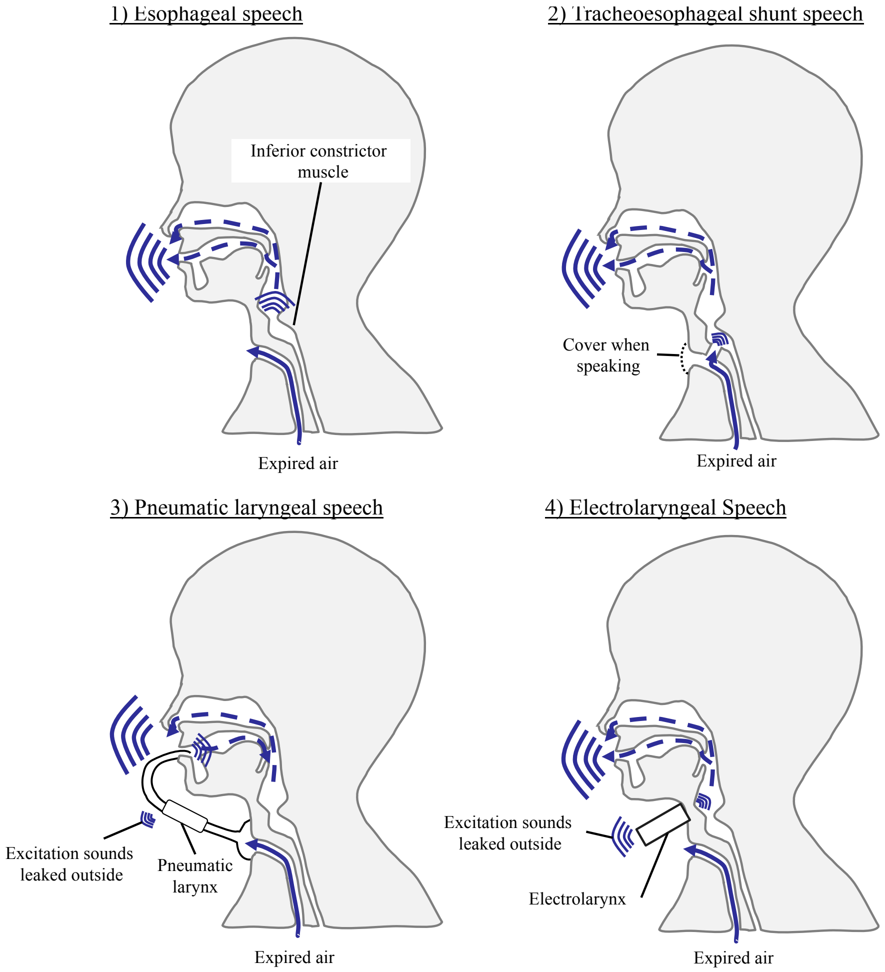

1) Esophageal speech
Inferior constrictor muscle
Expired air
2) Tracheoesophageal shunt speech
Cover when speaking
Expired air
3) Pneumatic laryngeal speech
Excitation sounds leaked outside
Pneumatic larynx
Expired air
4) Electrolaryngeal Speech
Excitation sounds leaked outside
Electrolarynx
Expired air

Figure 7. Major alternative speaking methods for laryngectomees.

2.4.1 Esophageal speech

The ES speaking method is conducted in the following procedure; taking air from the mouth to 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 ES 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 ES speech in Japan. As the result, the ES speaking method is the major alternative speaking method in Japan. On the other hand, the ES speaking methods requiring strength for the speakers, and

Table 1. Benefits and defects of alternative speaking methods for total laryngectomees.

Speech	Naturalness	Intelligibility	Difficulty	Popularity (in Japan)
ES	+	−	Difficult	+
TE shunt	+	±	Easy	−
PA	±	+	Not difficult	+ (in past days)
EL	−	+	Easy	+

therefore, some people are difficult to speak with the ES speaking methods to use another method such as the electrolaryngeal speaking method.

2.4.2 Tracheoesophageal shunt speech

The TE shunt speaking method is similar a speaking method to ES. The only difference from ES is the method of producing the air flow used to vibrate the vocal folds. In TE shunt speaking method, the air flow is delivered from the lungs and trachea into the esophagus through a voice prosthesis, which is a valve inserted between the trachea and esophagus. When speaking, laryngectomees block the tracheostoma to make the air flow to the esophagus through the prosthesis. The air vibrates tissues around the entrance of the esophagus, inducing sound source vibration similarly to in ES speech. It is easier to produce TE shunt speech than ES speech and the resulting speech like a hoarse voice because laryngectomees can use their breath in the same way as non-laryngectomees. Moreover, for the same reason, TE shunt speech has greater power than ES speech. Thus, TE shunt speech sounds more natural than ES speech and other alaryngeal speech. On the other hand, some elderly patients or those with a lung disease cannot undergo the operation required to enable TE shunt speech. Moreover, the voice prosthesis must be maintained every several years. Therefore, it is a less popular method in Japan.

2.4.3 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 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 electrolarynx. An interesting pneumatic artificial larynx was developed [53]; 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.

2.4.4 Electrolaryngeal speech

The other major external medical device is an electrolarynx. The basic structure of an electrolarynx is shown in Fig. 8. An electrolarynx is pushed on the lower jaw in speaking, and the on/off is switched by the button. The defect of the electrolarynx is its fixed F_0 the Fig. 8 shows deriving artificial and mechanical unnatural speech even though human speaks. Moreover, the sound source signals are noisy and may disturb people around the speaker especially in quiet situations. On the other hand, there are mainly two advantages of the electrolarynx: it is easier to produce speech by using the electrolarynx than other types of alaryngeal speech, and EL speech exhibits higher intelligibility than the other types of alaryngeal speech. Therefore, EL speech is the most popular alternative speaking method in Japan.

Figure 9 shows waveforms, F_0 patterns, and spectrograms of EL speech and normal speech. These acoustic features were extracted by STRAIGHT analysis [54]. In this figure, we can find that the extraction of F_0 is not effective for EL speech because F_0 of EL speech is almost constant value. The unnaturalness of EL speech is mainly caused by this monotonic F_0 patterns. Unfortunately, it is inherently difficult to mechanically generate F_0 patterns corresponding to linguistic contents. As for the articulation part, spectrogram of the EL speech is also different from that of the normal speech because the excitation signals generated by the electrolarynx leak out as noise although the vocal tracts of laryngectomees

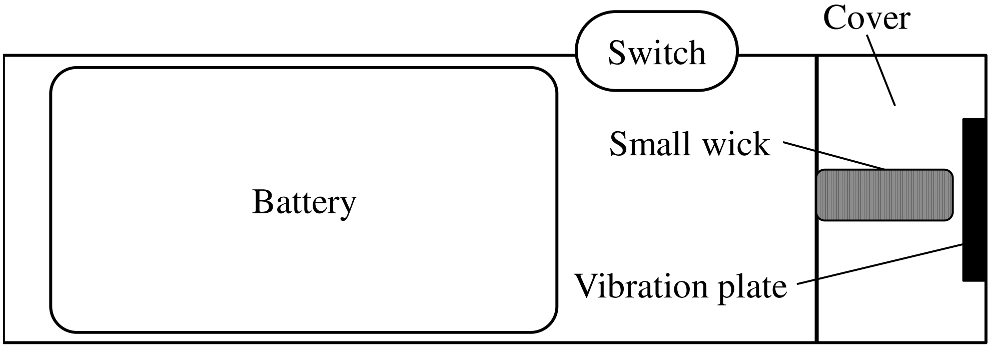


Figure 8. Basic structure of existing electrolarynxes.

are still remaining. Since the excitation signals are monotonic, horizontal stripes are observed in the spectrogram of the EL speech.

2.5 Review on electrolaryngeal speech enhancement system at physiological level

In this section, we introduce the conventional EL speech enhancement systems at physiological level. These frameworks aim to directly control F_0 patterns of excitation signals of the electrolarynx by using additional signals generated by: 1) air-pressure [18], 2) finger movement [19] [20], 4) forearm movement [21], and so on. The benefits and defects of these methods are shown in Table 3.

Table 2. Benefits and defects of conventional EL speech enhancement systems at physiological level.

Electrolarynx	Controllable	Convenience
Original	—	One hand is occupied.
+ air-pressure	Manually	Both hands are occupied.
+ finger movement	Manually	One hand is occupied.
+ forearm movement	Manually	One hand is occupied.

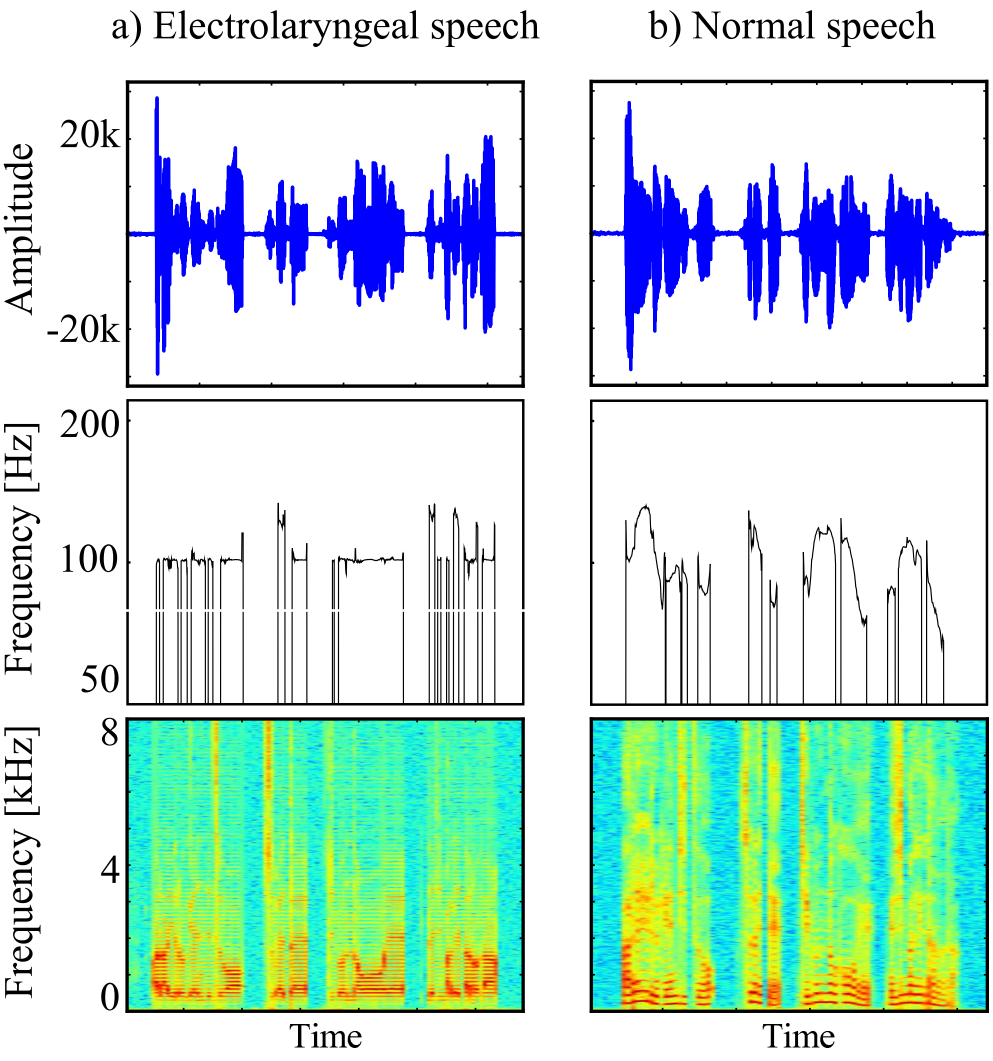


Figure 9. Example of waveforms, F_0 patterns, and spectrograms of a) electrolaryngeal speech and b) normal speech.

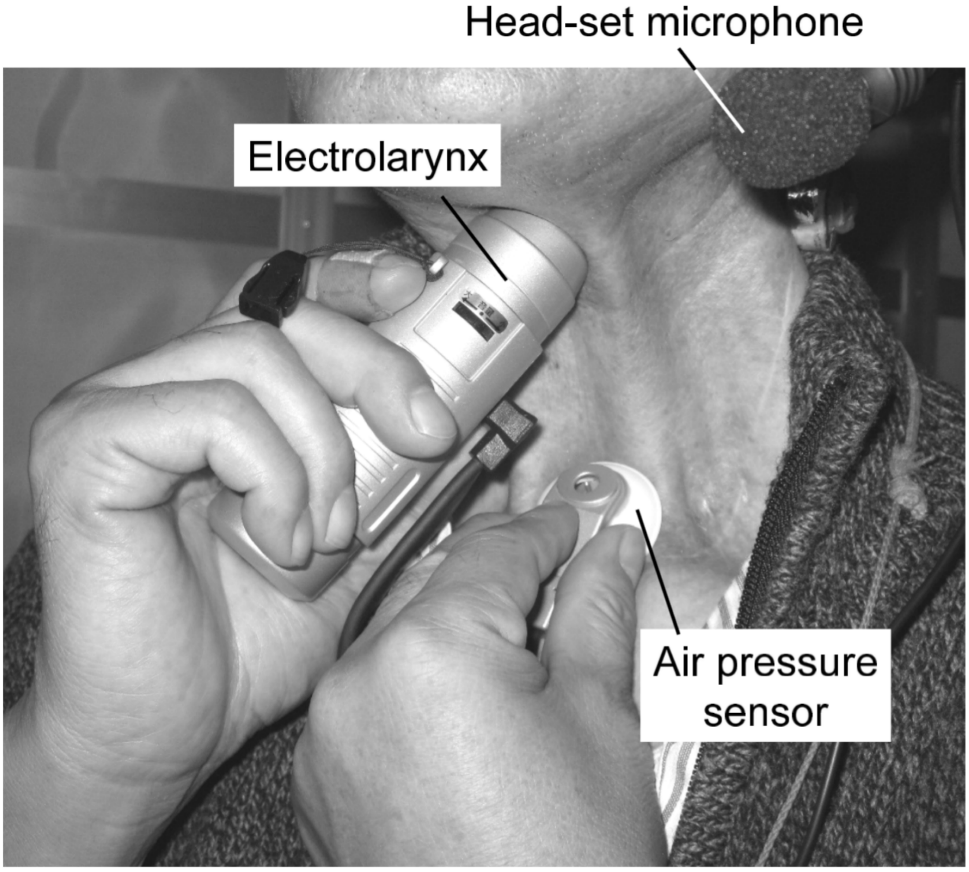


Figure 10. Recording scene of EL speech using air-pressure sensor.

2.5.1 Air-pressure

One electrolarynx 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 [28, 18]. The electrolarynx 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 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 fluctuation appeared in outline shapes of waveforms compared to another fluctuation caused by durations. 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 [18]. A recording scene of the EL speech using the air-pressure sensor (EL(air) speech) is shown in Fig. 10. 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 electrolarynx and the air-pressure sensor. More than one thousand of 'yourtone' is used and it is preferred by buyers. This is the first electrolarynx made in Japan, and its effectiveness is experimentally and practically confirmed.

2.5.2 Finger movement

Finger pressure [19] A pressure sensor is built into a push button of the electrolarynx, and the F_0 patterns are controlled by the finger pressure with which the button is pressed. As this device works by pressure, the displacement amount of the finger for adjusting F_0 patterns is zero in principle. However, it is difficult to control the level of pressure.

Up and down switch As the thumb can move freely up and down or left and right, [20] use the thumb to control F_0 patterns. Among of mentioned movements, the value of F_0 is controlled by the up-down displacement amount of the finger, and ON/OFF of the vibration is controlled by the left-right movement. AS the reported [20], this system using the up-down switch is easier to control F_0 patterns compared with the system using finger pressure [19].

2.5.3 Forearm movement

Hand gestures are usually used in inter-person speech communication. The electrolarynx using forearm movement (shown in Fig. 11) was conducted in order to

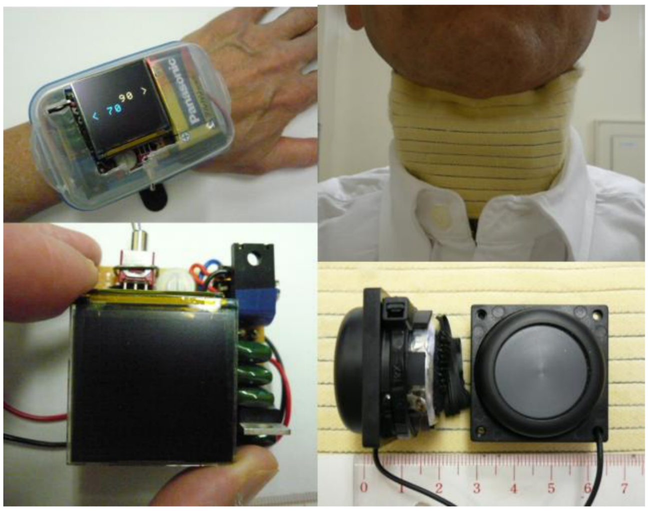


Figure 11. EL Controller and Transducer unit (upper left: EL controller on the wrist, lower left: ARM-PC board, upper right: transducer with neck bandage, lower right: transducer unit) [21].

evaluate feasibility of the control mechanism of F_0 patterns. Fig. 12 shows the forearm tilt and the MEMS output (x-axis) when the controller was placed on the wrist. From the horizontal position (0 °) to the 75 ° upward position is the normal pitch control zone. From the horizontal position to the -25 ° downward position is the fading out zone, where phrase ending pitch pattern is adjusted based on the forearm moving speed. As for the conversion from the MEMS output to the pitch frequency, there are four pitch ranges. Fig.3 shows the relation between the MEMS output and the four ranges of pitch frequency, i.e. high, mid-high, mid-low, and low. Users can select one of the four ranges. It has been demonstrated that the naturalness of EL speech is improved while preserving the intelligibility of EL speech.

2.6 Review on electrolaryngeal speech enhancement system at acoustic level

In this section, we introduce the conventional EL speech enhancement systems at acoustic level. These frameworks aim to enhance acoustic parameters extracted

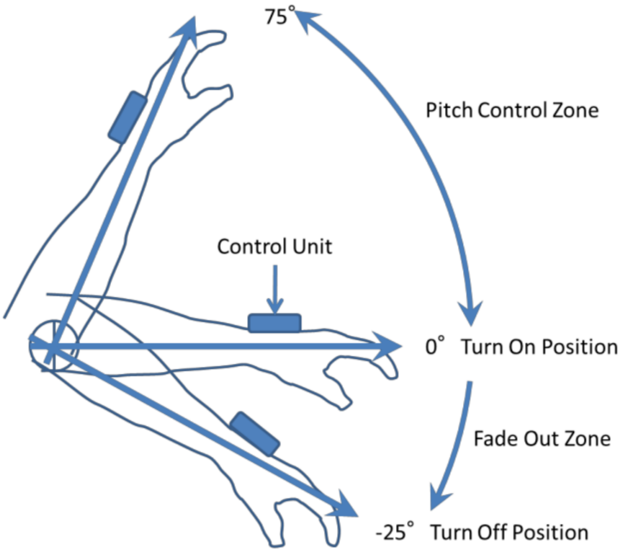


Figure 12. Forearm tilt and Control of F_0 patterns [21].

from recorded speech. After the enhancement, the enhanced speech signals are synthesized from the enhanced acoustic parameters. To enhance acoustic parameters of EL speech, there are two major approach: 1) signal processing, such as spectral subtraction [25, 26], to reduce the excitation signals leaking out as noise and 2) statistical voice conversion [22, 16] to convert acoustic features of EL speech into those of normal speech. The properties of acoustic parameters after using these methods are shown in Table 3.

2.6.1 Spectral subtraction

Spectral subtraction is a method for restoration of the amplitude spectrum of a speech signal that has been observed together with the leaked noise of an electrolarynx. This is done by subtracting an estimate of the amplitude spectrum of the noise from the amplitude spectrum of the noisy speech signal. The noisy speech signal model in the frequency domain is expressed as follows:

$$Y(\omega, t) = S(\omega, t) + L(\omega, t) \tag{1}$$

where $Y(\omega, t)$, $S(\omega, t)$, and $L(\omega, t)$ are respectively components of the noisy speech signal, the clean speech signal, and the additive noise signal at frequency ω and

Table 3. Properties of acoustic parameters enhanced by conventional EL speech enhancement systems at physiological level.

	Parameters	Properties	Naturalness	Intelligibility
Original EL speech	Excitation	Mechanically	Low	High
	Spectral	Noisy		
Spectral subtraction	Excitation	Mechanically	Low	Slightly improved
	Spectral	Less noise		
Statistical voice conversion	Excitation	Normal speech w/ conversion errors	Dramatically improved	Degraded
	Spectral			

time frame t. Assuming that the additive noise signal is stationary, the generalized spectral subtraction scheme [37] is described as follows:

$$|\hat{S}(\omega,t)|^{\gamma} = \begin{cases} |Y(\omega,t)|^{\gamma} - \alpha|\hat{L}(\omega)|^{\gamma} & (\frac{|\hat{L}(\omega)|^{\gamma}}{|Y(\omega,t)|^{\gamma}} < \frac{1}{\alpha+\beta}) \\ \beta|\hat{L}(\omega)|^{\gamma} & (otherwise) \end{cases} \tag{2}$$

where α ($\alpha > 1$) is an over-subtraction parameter, β ($0 \leq \beta \leq 1$) is a spectral flooring parameter, γ is an exponential domain parameter, and $\hat{L}(\omega)$ is an estimate of the averaged amplitude spectrum of the additive noise signal. The enhanced speech signal is generated using the processed amplitude spectrum and the original phase extracted from the noisy speech signal.

For EL speech enhancement, the averaged amplitude spectrum of the additive noise signal is estimated in advance using the excitation signals generated from the electrolarynx as shown in Fig. 13. In order to record only the excitation signals leaked from the electrolarynx as accurately as possible, the excitation signals are recorded with a close-talking microphone while keeping speaker's mouth closed. The excitation signals are generated with the electrolarynx held in the usual manner, as shown in Fig. 7.

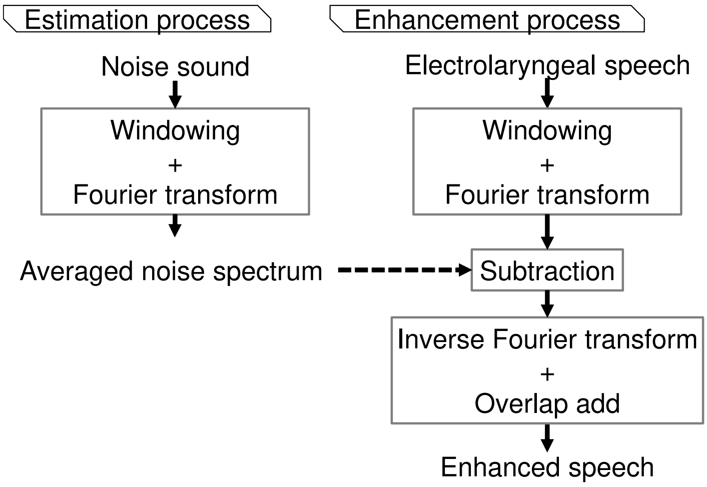


Figure 13. Process flow of spectral subtraction.

2.6.2 Statistical voice conversion

Many statistical approaches to VC have been studied. In an early statistical VC approach, the codebook mapping method based on hard clustering and discrete mapping was proposed [38]. In this method, a converted acoustic feature is determined by quantizing the source speaker's acoustic feature to the nearest centroid feature of the source codebook and substituting it with the corresponding centroid feature of the mapping codebook. For typical statistical VC, two Gaussian mixture model (GMM) based conversion methods as shown in Fig. 15 have been proposed. The GMMs are trained to convert the spectral parameters of the source speaker into those of target speaker, and the aperiodic components of source speaker into those of target speaker, respectively. One is a frame-based conversion method that converts features based on the minimum mean square error (MMSE) [30]. The other is a trajectory-based conversion method [31] that simultaneously converts a feature sequence over an utterance based on maximum likelihood estimation (MLE). Although the former method is capable of real-time conversion because source features at individual frames are converted independently of each other, this method sometimes causes feature discontinuities with inappropriate dynamic characteristics because the inter-frame feature correlation is ignored. On the other hand, the latter method provides converted

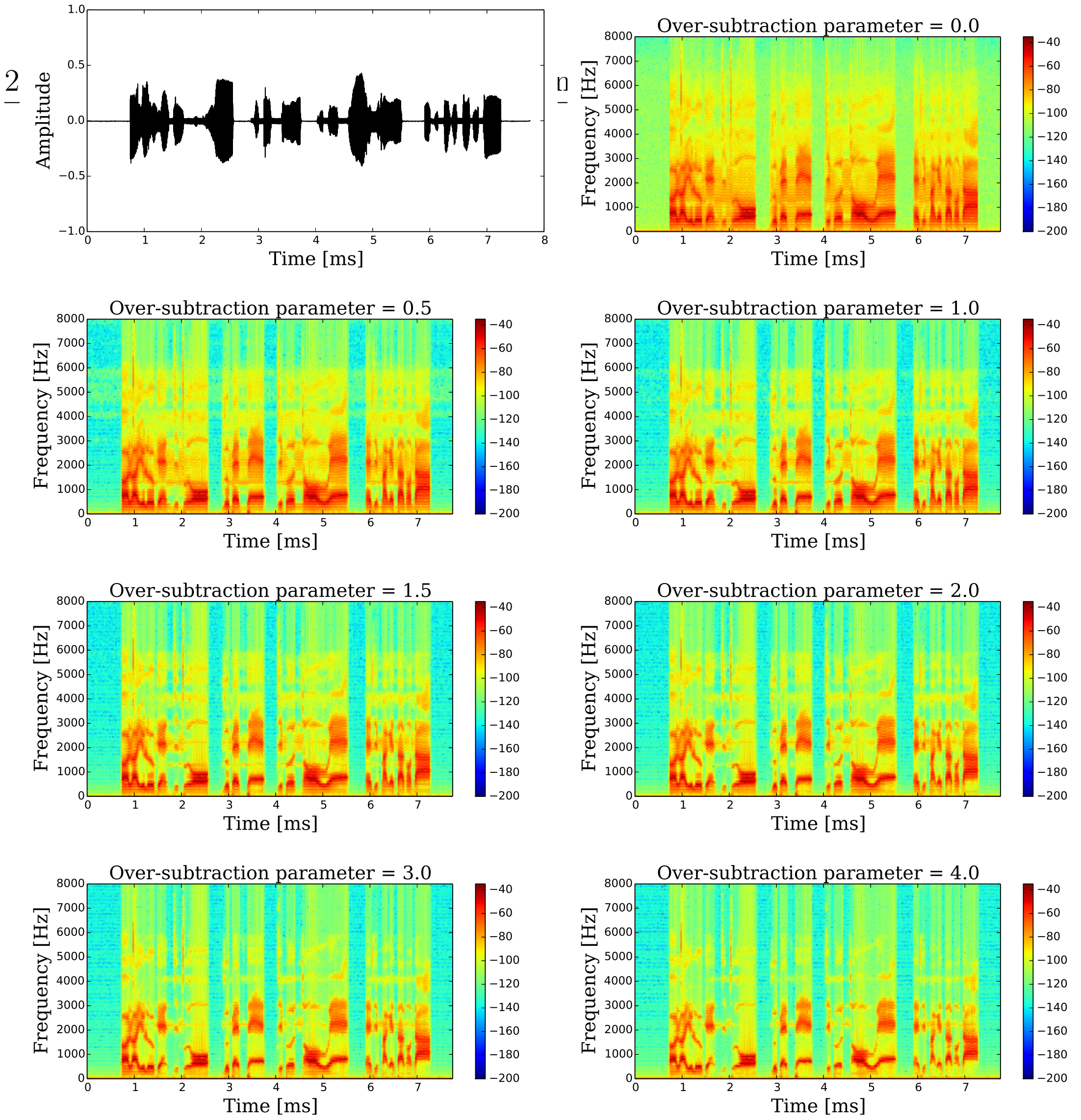


Figure 14. Spectrogram of EL speech enhanced by spectral subtraction with various over-subtraction parameter α in **Section 2.6.1**.

feature sequences exhibiting appropriate dynamic characteristics by considering the dynamic features of the converted speech. Additionally, to alleviate the over-smoothing of the converted features due to statistical modeling, trajectory-based VC considering the global variance (GV), which is the variance of features over a time sequence, has also been proposed [31]. Although the trajectory-based conversion method results in significant quality improvements of the converted speech, it does not work in real time because the source features over an utterance need to be converted simultaneously o consider the inter-frame correlation. To achieve real-time conversion considering the dynamic characteristics of converted features, low-delay VC based on MLE has been proposed [29].

As an EL speech enhancement system at acoustic level, an enhancement method based on statistical VC has been proposed [22]. In this method, three GMMs are trained as shown in Fig. 16 while the basic VC framework has two GMMs. In the training process, the GMMs are trained: a GMM that converts the spectrum of EL speech into a spectrum of normal speech, a GMM that converts the spectrum of EL speech into the F_0 of normal speech, and a GMM that converts the spectrum of EL speech into aperiodic components of normal speech. This is because F_0 and the aperiodic components of EL speech are not informative. In the conversion process, an arbitrary utterance of EL speech is converted into normal speech while keeping the linguistic information unchanged. Although this method significantly improves the naturalness of EL speech, its intelligibility deteriorates owing to inevitable conversion errors caused by the complex conversion process.

Acoustic feature The spectral structure of some phonemes of EL speech are unstably because of the production mechanism of EL speech, such as totally voiced speech. Therefore, it might have a risk of degradations of the conversion accuracy to apply the feature extraction method of natural speech to EL speech. On the other hand, information to be converted is supposed to be included in the current, several previous and succeeding frames. To realize this requirement, the following segment feature $\boldsymbol{X}_t$ [55] are extracted by applying principal component analysis (PCA) [56] to the stacked vector consisting of the mel-cepstra [57, 58] of multiple frames around the current frame t as source feature:

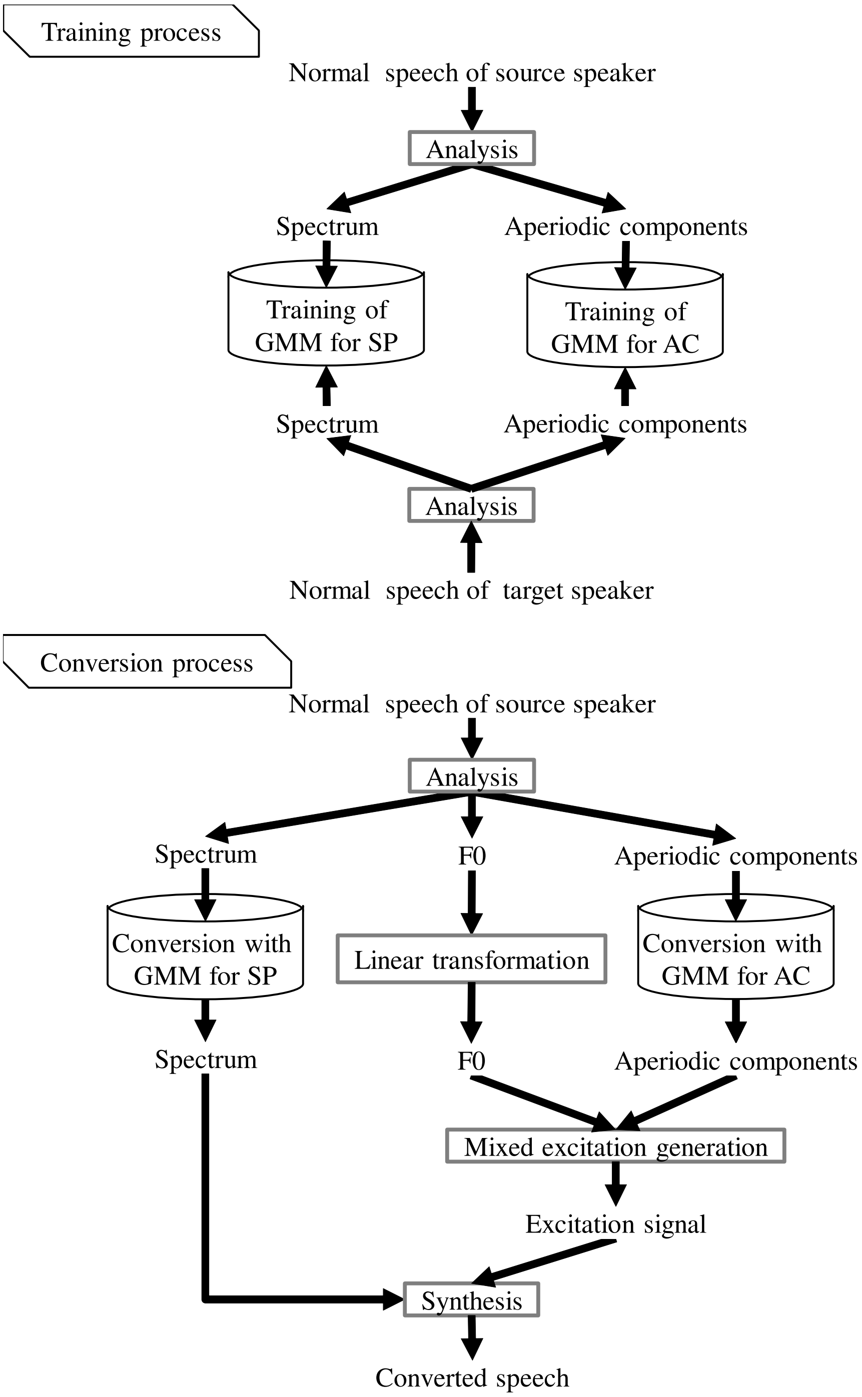


Figure 15. Overview of training and conversion process of basic voice conversion.

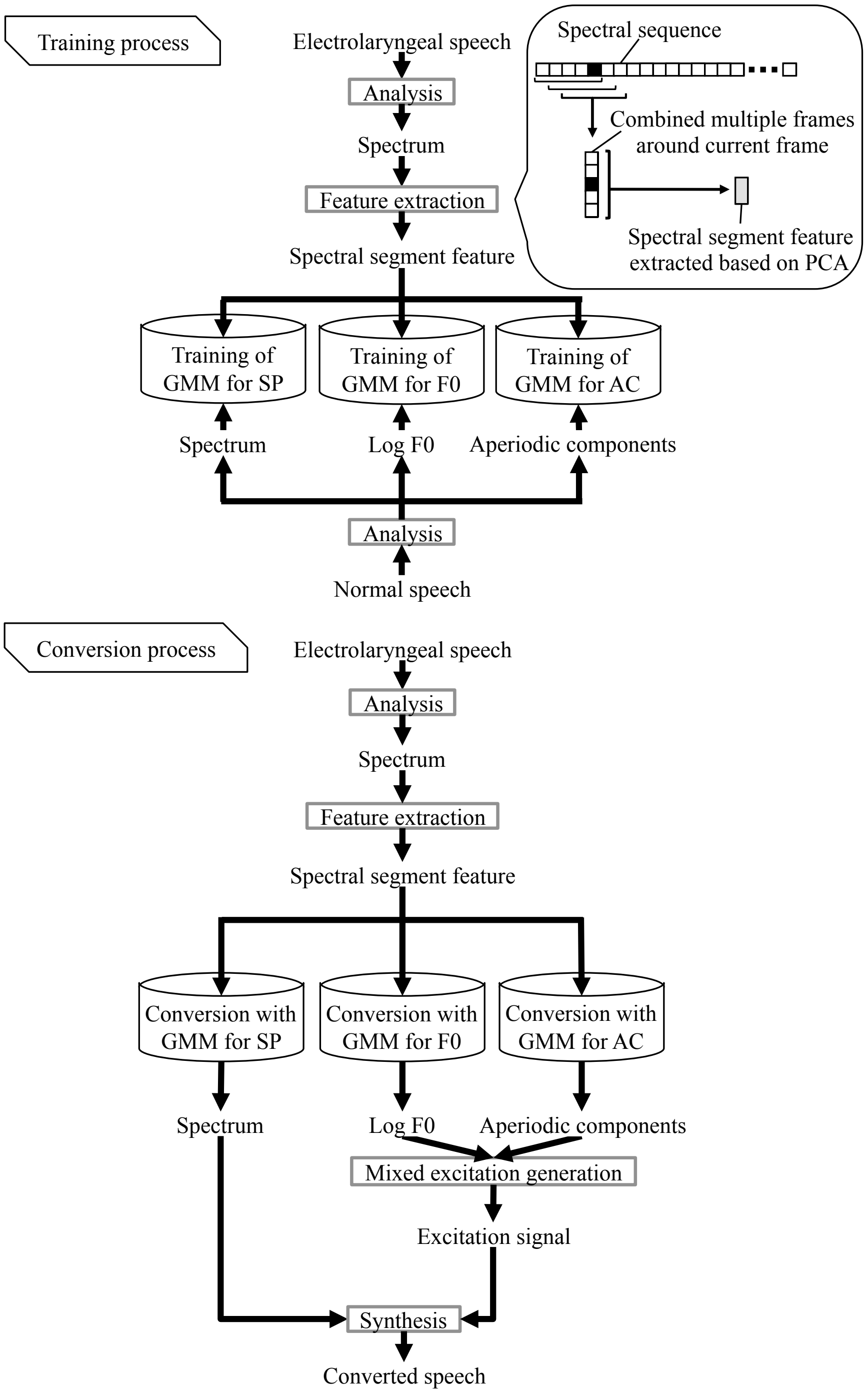


Figure 16. Overview of training and conversion process of the voice conversion for EL speech enhancement.

$$\boldsymbol{X}_t = \boldsymbol{C}\left[\boldsymbol{x}_{t-i}^\top, \cdots, \boldsymbol{x}_t^\top, \cdots, \boldsymbol{x}_{t+i}^\top\right]^\top + \boldsymbol{d}, \tag{3}$$

where $^\top$ is transposition, and $\boldsymbol{C}$ and $\boldsymbol{d}$ are a transformation matrix and a bias vector extracted by PCA, respectively. As target feature, we use $\boldsymbol{Y}_t = [\boldsymbol{y}_t^\top, \Delta\boldsymbol{y}_t^\top]^\top$ consisting of the static and dynamic features of each type of the natural speech: spectral parameters, F_0 patterns, and aperiodic components.

GMM training Let $\boldsymbol{\lambda}_G$ be the parameters of the following joint *p.d.f.* of source and target features defined as a GMM:

$$P\left(\boldsymbol{Z}_t|\boldsymbol{\lambda}_G\right) = \sum_{m=1}^{M} \alpha_m \mathcal{N}\left(\boldsymbol{Z}_t; \boldsymbol{\mu}_m^{(\boldsymbol{Z})}, \boldsymbol{\Sigma}_m^{(\boldsymbol{Z})}\right), \tag{4}$$

$$\boldsymbol{Z}_t = \begin{bmatrix} \boldsymbol{X}_t \\ \boldsymbol{Y}_t \end{bmatrix}, \tag{5}$$

$$\boldsymbol{\mu}_m^{(\boldsymbol{Z})} = \begin{bmatrix} \boldsymbol{\mu}_m^{(\boldsymbol{X})} \\ \boldsymbol{\mu}_m^{(\boldsymbol{Y})} \end{bmatrix}, \tag{6}$$

$$\boldsymbol{\Sigma}_m^{(\boldsymbol{Z})} = \begin{bmatrix} \boldsymbol{\Sigma}_m^{(\boldsymbol{XX})} & \boldsymbol{\Sigma}_m^{(\boldsymbol{YX})} \\ \boldsymbol{\Sigma}_m^{(\boldsymbol{XY})} & \boldsymbol{\Sigma}_m^{(\boldsymbol{YY})} \end{bmatrix}, \tag{7}$$

where α_m is a m-th mixture component weight, and $\mathcal{N}(\cdot; \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m)$ denotes a m-th Gaussian distribution with a mean vector $\boldsymbol{\mu}_m$ and a covariance matrix $\boldsymbol{\Sigma}_m$. The mean vector $\boldsymbol{\mu}_m^{(\boldsymbol{Z})}$ consists of a mean vector $\boldsymbol{\mu}_m^{(\boldsymbol{X})}$ for source features and a mean vector $\boldsymbol{\mu}_m^{(\boldsymbol{Y})}$ for target features. The covariance matrix $\boldsymbol{\Sigma}_m^{(\boldsymbol{Z})}$ consists of source and target covariance matrices $\boldsymbol{\Sigma}_m^{(\boldsymbol{XX})}$ and $\boldsymbol{\Sigma}_m^{(\boldsymbol{YY})}$ and cross-covariance matrices $\boldsymbol{\Sigma}_m^{(\boldsymbol{XY})}$ and $\boldsymbol{\Sigma}_m^{(\boldsymbol{YX})}$. The total number of mixture components is M. The corresponding joint feature vectors can be obtained by performing automatic frame alignment with dynamic time warping (DTW). The model parameters are estimated by expectation-maximization (EM) algorithm [59].

GMM based conversion process Individual speech parameters of the target natural speech are independently estimated from the spectral segment features extracted from the EL speech using each of the trained GMMs as follows:

$$
\begin{aligned}
P\left(\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{\lambda}_G\right) &= \sum_{\text{all}\,\boldsymbol{m}} P\left(\boldsymbol{m}|\boldsymbol{X},\boldsymbol{\lambda}_G\right) P\left(\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{m},\boldsymbol{\lambda}_G\right),\\
&= \prod_{t=1}^{T}\sum_{m=1}^{M} P\left(m|\boldsymbol{X}_t,\boldsymbol{\lambda}_G\right) P\left(\boldsymbol{Y}_t|\boldsymbol{X}_t,m,\boldsymbol{\lambda}_G\right),
\end{aligned} \tag{8}
$$

where $\boldsymbol{m} = [m_1,\cdots,m_t,\cdots,m_T]^\top$, $\boldsymbol{X} = \left[\boldsymbol{X}_1^\top,\cdots,\boldsymbol{X}_t^\top,\cdots,\boldsymbol{X}_T^\top\right]^\top$, and $\boldsymbol{Y} = \left[\boldsymbol{Y}_1^\top,\cdots,\boldsymbol{Y}_t^\top,\cdots,\boldsymbol{Y}_T^\top\right]^\top$ are time sequence vectors of the mixture component, the input spectral segment features, and the output features over an utterance, respectively. The m-th mixture weight and conditional *p.d.f.* are the following form:

$$
P\left(m|\boldsymbol{X}_t,\boldsymbol{\lambda}_G\right) = \frac{\alpha_m,\mathcal{N}\left(\boldsymbol{X}_t;\boldsymbol{\mu}_m^{(\boldsymbol{X})},\boldsymbol{\Sigma}_m^{(\boldsymbol{XX})}\right)}{\sum_{n=1}^{M}\alpha_n,\mathcal{N}\left(\boldsymbol{X}_t;\boldsymbol{\mu}_n^{(\boldsymbol{X})},\boldsymbol{\Sigma}_n^{(\boldsymbol{XX})}\right)}, \tag{9}
$$

$$
P\left(\boldsymbol{Y}_t|\boldsymbol{X}_t,m,\boldsymbol{\lambda}_G\right) = \mathcal{N}\left(\boldsymbol{Y}_t;\boldsymbol{E}_{m,t}^{(\boldsymbol{Y}|\boldsymbol{X})},\boldsymbol{D}_m^{(\boldsymbol{Y}|\boldsymbol{X})}\right), \tag{10}
$$

where $\boldsymbol{E}_{m,t}^{(\boldsymbol{Y}|\boldsymbol{X})}$ is the conditional mean vector of the m-th mixture at frame t, which is given by the mixture-dependent linear transformation of the source feature vector $\boldsymbol{X}_t$, and $\boldsymbol{D}_m^{(\boldsymbol{Y}|\boldsymbol{X})}$ is the conditional covariance matrix depending of the mixture component m_t:

$$
\boldsymbol{E}_{m,t}^{(\boldsymbol{Y}|\boldsymbol{X})} = \boldsymbol{\mu}_m^{(\boldsymbol{Y})} + \boldsymbol{\Sigma}_m^{(\boldsymbol{YX})}\boldsymbol{\Sigma}_m^{(\boldsymbol{XX})^{-1}}\left(\boldsymbol{X}_t - \boldsymbol{\mu}_m^{(\boldsymbol{X})}\right), \tag{11}
$$

$$
\boldsymbol{D}_m^{(\boldsymbol{Y}|\boldsymbol{X})} = \boldsymbol{\Sigma}_m^{(\boldsymbol{YY})} - \boldsymbol{\Sigma}_m^{(\boldsymbol{YX})}\boldsymbol{\Sigma}_m^{(\boldsymbol{XX})^{-1}}\boldsymbol{\Sigma}_m^{(\boldsymbol{XY})}, \tag{12}
$$

The most likely static sequence $\hat{\boldsymbol{y}} = \left[\hat{\boldsymbol{y}}_{t-i}^\top,\cdots,\hat{\boldsymbol{y}}_t^\top,\cdots,\hat{\boldsymbol{y}}_{t+i}^\top\right]^\top$ of the target feature is predicted from given source feature sequence $\boldsymbol{X} = \left[\boldsymbol{X}_1^\top,\cdots,\boldsymbol{X}_t^\top,\cdots,\boldsymbol{X}_T^\top\right]^\top$ as follows:

$$
\hat{\boldsymbol{y}} = \underset{\boldsymbol{y}}{\operatorname{argmax}}\, P\left(\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{\lambda}_G\right) \quad \text{subject to} \quad \boldsymbol{Y} = \boldsymbol{W}\boldsymbol{y}, \tag{13}
$$

where the matrix $\boldsymbol{W}$ is a transform to extend the static feature vector sequence into the joint static and dynamic feature vector sequence [60]. To avoid the complicated formula $\sum_{\boldsymbol{m}}$ in Eq. (8), the sub-optimum mixture component sequence $\hat{\boldsymbol{m}} = [\hat{m}_1, \cdots, \hat{m}_t, \cdots, \hat{m}_T]^\top$ is applied and Eq. (13) is approximated as follows:

$$P(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{\lambda}_G) \simeq P(\boldsymbol{Y}|\boldsymbol{X}, \hat{\boldsymbol{m}}, \boldsymbol{\lambda}_G)\, P(\hat{\boldsymbol{m}}|\boldsymbol{X}, \boldsymbol{\lambda}_G), \tag{14}$$

$$\hat{m}_t = \operatorname*{argmax}_m P(m|\boldsymbol{X}_t, \boldsymbol{\lambda}_G), \tag{15}$$

$$\begin{aligned} P(\boldsymbol{Y}|\boldsymbol{X}, \hat{\boldsymbol{m}}, \boldsymbol{\lambda}_G) &= \mathcal{N}\left(\boldsymbol{Y}; \boldsymbol{E}_{\hat{\boldsymbol{m}}}^{(\boldsymbol{Y}|\boldsymbol{X})}, \boldsymbol{D}_{\hat{\boldsymbol{m}}}^{(\boldsymbol{Y}|\boldsymbol{X})}\right) \\ &= \prod_{t=1}^{T} \mathcal{N}\left(\boldsymbol{Y}_t; \boldsymbol{E}_{\hat{\boldsymbol{m}},t}^{(\boldsymbol{Y}|\boldsymbol{X})}, \boldsymbol{D}_{\hat{\boldsymbol{m}}}^{(\boldsymbol{Y}|\boldsymbol{X})}\right), \end{aligned} \tag{16}$$

The maximum-likelihood estimation of static sequence $\hat{\boldsymbol{y}}$ is analytically determined as follows:

$$\hat{\boldsymbol{y}} = \left(\boldsymbol{W}^\top \boldsymbol{D}_{\hat{\boldsymbol{m}}}^{(\boldsymbol{Y}|\boldsymbol{X})^{-1}} \boldsymbol{W}\right)^{-1} \boldsymbol{W}^\top \boldsymbol{D}_{\hat{\boldsymbol{m}}}^{(\boldsymbol{Y}|\boldsymbol{X})^{-1}} \boldsymbol{E}_{\hat{\boldsymbol{m}}}^{(\boldsymbol{Y}|\boldsymbol{X})}, \tag{17}$$

After estimating time sequences of the converted spectrum, F_0, and aperiodic components, a mixed excitation signal is generated using the converted F_0 and aperiodic components [61]. Finally, the converted speech signal is synthesized by filtering the generated excitation signal with the converted spectral parameters.

Global variance One essential problem in maximum likelihood criterion is that estimated parameters tend to over-smoothed as shown in Fig. 17. In order to address the over-smoothing problem, trajectory based conversion considering GV has been proposed [31]. The GV $\boldsymbol{v}(\boldsymbol{y}) = [v(1), \cdots, v(d_y). \cdots, v(D_y)]^\top$ is defined as the variance of features over one utterance and its *p.d.f.* of the output static feature sequence is written as follows:

$$P\left(\boldsymbol{v}(\boldsymbol{y})|\boldsymbol{\lambda}_{(v)}\right) = \mathcal{N}\left(\boldsymbol{v}(\boldsymbol{y}); \boldsymbol{\mu}^{(\boldsymbol{v})}, \boldsymbol{\Sigma}^{(\boldsymbol{v})}\right), \tag{18}$$

$$v(d_y) = \frac{1}{T}\sum_{t=1}^{T}\left(y_t(d) - \frac{1}{T}\sum_{\tau=1}^{T} y_\tau(d)\right)^2. \tag{19}$$

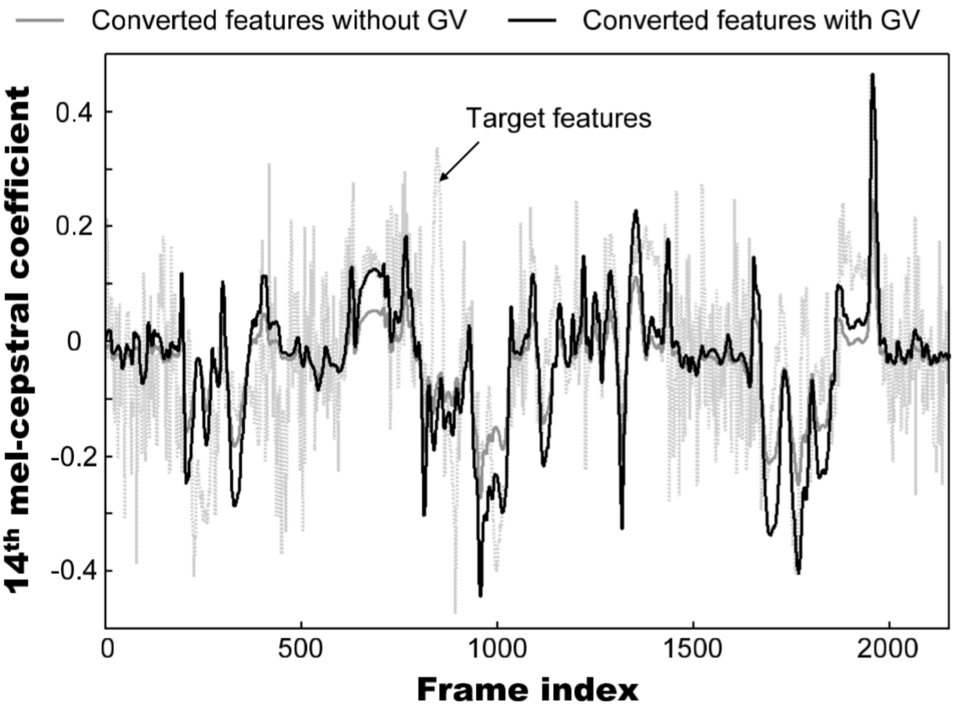


Figure 17. Example of converted features considering GV.

In the conversion process, considering GV, the Eq. (13) is modified as follows:

$$\hat{\boldsymbol{y}} = \underset{\boldsymbol{y}}{\operatorname{argmax}}\, P\left(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{\lambda}_G\right) P\left(\boldsymbol{v}(\boldsymbol{y})|\boldsymbol{\lambda}_{(v)}\right)^{\omega}, \tag{20}$$

The GV likelihood weight is given by ω, which is set to the ratio of the number of dimensions between vector $\boldsymbol{v}(\boldsymbol{y})$ and $\boldsymbol{Y}$, i.e., $\frac{1}{2T}$. The GV likelihood is usually considered only in the spectral estimation, i.e., ω is set to zero in the F_0 estimation and the aperiodic estimation.

Real-time conversion process The real-time prediction process is achieved by using a computationally efficient real-time voice conversion method [29] based on a low-delay conversion algorithm [62]. To approximate the batch-type prediction process with the frame-wise prediction process, we divide the F_0 sequence $\boldsymbol{y}$ into overlapped $(L+1)$-dimensional segment vectors $\boldsymbol{y}_d^{(t)} = [y_{t-L}, \ldots, y_t]\top$ at individual frames. Treating the segment vectors as a latent variable and initializing a $(L+1)$-by-$(L+1)$ state covariance matrix $\boldsymbol{P}_d^{(0)}$ and a state vector $\hat{\boldsymbol{y}}_d^{(t-1)}$ as the zero matrix and the zero vector, the following linear dynamical system can be designed:

$$\boldsymbol{P}_d'^{(t-1)} = \boldsymbol{J}_L \boldsymbol{P}_d^{(t-1)} \boldsymbol{J}_L^\top + \mathrm{diag}\left[\boldsymbol{0}_{1\times L}, \Sigma_{m_t,d}^{(y|X)}\right], \tag{21}$$

$$\hat{\boldsymbol{y}}_d'^{(t-1)} = \boldsymbol{J}_L \hat{\boldsymbol{y}}_d^{(t-1)} + \left[\boldsymbol{0}_{1\times L}, \mu_{m_t,t,d}^{(y|X)}\right]^\top, \tag{22}$$

$$\boldsymbol{P}_d^{(t)} = \left(\boldsymbol{I} - \boldsymbol{k}_d^{(t)} \boldsymbol{w}_L\right) \boldsymbol{P}_d'^{(t-1)}, \tag{23}$$

$$\hat{\boldsymbol{y}}_d^{(t)} = \hat{\boldsymbol{y}}_d'^{(t-1)} + \boldsymbol{k}_d^{(t)} \left(\mu_{m_t,t,d}^{(\Delta y|X)} - \boldsymbol{w}_L \hat{\boldsymbol{y}}_d'^{(t-1)}\right), \tag{24}$$

where the $(L+1)$-dimensional vector $\boldsymbol{k}_d^{(t)}$ is calculated as

$$\boldsymbol{k}_d^{(t)} = \boldsymbol{P}_d^{(t-1)} \boldsymbol{w}_L^\top \left(\Sigma_{m_t,d}^{(\Delta y|X)} + \boldsymbol{w}_L \boldsymbol{P}_d^{(t-1)} \boldsymbol{w}_L^\top\right)^{-1}, \tag{25}$$

and the $(L+1)$-dimensional row vector $\boldsymbol{w}_L$ and the $(L+1)$-by-$(L+1)$ matrix $\boldsymbol{J}_L$ are given by

$$\boldsymbol{w}_L = \left[\boldsymbol{0}_{1\times(L-1)}, -1, 1\right], \quad \boldsymbol{J}_L = \begin{bmatrix} & 0 \;\; 0 \\ \boldsymbol{I}_{L\times L} & \boldsymbol{0}_{1\times L} \end{bmatrix}, \tag{26}$$

respectively. The d-th dimensional static feature components, $\mu_{m,t,d}^{(y|X)}$ and $\Sigma_{m,d}^{(y|X)}$, of the mean vector $\boldsymbol{\mu}_{m,t}^{(Y|X)}$ and the covariance matrix $\boldsymbol{\Sigma}_m^{(Y|X)}$ are used to predict the state covariance matrix and the state vector as shown in Eq. (21) and Eq. (22). Their dynamic feature components, $\mu_{m,t,d}^{(\Delta y|X)}$ and $\Sigma_{m,d}^{(\Delta y|X)}$, are used to optimize the Kalman gain in Eq. (25) and update the state covariance matrix and the state vector as shown in Eq. (23) and Eq. (24). The first component of $\hat{\boldsymbol{y}}_d^{(t)}$ is used as the d-th component of the d converted static feature vector at frame $t-L$, $\hat{y}_{t-L,d}$.

2.7 Summary

This chapter described the process of speech production, laryngectomees, several alaryngeal speech, and conventional speaking aid systems for EL speech. The number of laryngectomees as the target of this dissertation is around 30 thousands person in Japan. They cannot produce speech sounds without alternative speaking methods such as electrolaryngeal and esophageal speaking method. Among of those alaryngeal speech, the EL speech is quite intelligible. However, there are several disadvantages: 1) unnatural owing to the monotonic F_0 patterns of the excitation signals, 2) excitation signals leak out as noise, and 3) one hand is occupied. To address these negative factors, the conventional EL speech enhancement systems at physiological and acoustic level have been proposed. Among of them, statistical F_0 pattern prediction, which a part of technique of statistical voice conversion, has achieved to dramatically improve the naturalness of EL speech.

Chapter

3

Hybrid approach to electrolaryngeal speech enhancement based on spectral enhancement and statistical F_0 pattern prediction

3.1 Introduction

The EL speech enhancement method based on spectral subtraction [25, 26] essentially estimates EL speech produced by the lips while reducing the impact of leaked excitation sounds. Even if the leaked excitation sounds are completely removed, improvements in naturalness yielded by this method will be small because the produced EL speech intrinsically suffers from the lack of naturalness caused by highly artificial F_0 patterns and the mechanical excitation sound quality. On the other hand, this method does not cause any significant degradation in intelligibility of EL speech. In other words, this method may cause small improvements, but very rarely degradations in speech quality. The EL speech enhancement method based on statistical voice conversion [22, 16] has the potential to significantly improve naturalness of EL speech by converting EL speech into normal speech. As the converted speech signal is generated from statistics of normal speech parameters, it does not suffer from the artificial F_0 patterns and mechanical sound quality. However, the conversion process in this method is quite complex, and therefore, errors in conversion are inevitable. These errors tend to cause degradation in intelligibility of converted speech as adverse effects.

In order to develop an EL speech enhancement method that allows for the large improvements of naturalness realizable by the statistical voice conversion while ameliorating its adverse effects, we propose a hybrid approach based on spectral subtraction and statistical voice conversion. Furthermore, we also propose the modification of F_0 patterns for accurate modeling and prediction. To simplify characteristics of the parameter sequence to be modeled, smoothed continuous F_0 (CF_0) patterns are obtained by removing rapid movements [63] with low-pass filtering after interpolating F_0 values at unvoiced frames of the original F_0 patterns. This modification is reasonable because 1) it is difficult to accurately model and reproduce these movements with a GMM, 2) a constant value at the unvoiced frames, clearly different from F_0 values (e.g., 0), disturbs accurate modeling of F_0 trajectory, and 3) it might be not required to model discontinuous F_0 patterns obtained in the natural voice because the EL speech is totally voiced speech.

In the experimental evaluation, we objectively and subjectively compared the performance of the proposed hybrid system with the conventional systems and the effectiveness of the modification of F_0 patterns. The experimental results demonstrate the proposed method yields significant improvements in naturalness compared with EL speech while keeping intelligibility high enough.

The rest of this chapter is organized as follows.In **Section 3.2**, we describe the hybrid approach to realize EL speech enhancement method capable of significantly improving naturalness of EL speech while causing no degradation in its intelligibility In **Section 3.3**, we propose a method to improve prediction accuracy of statistical F_0 prediction method. In **Section 3.4**, we evaluate the performance.

3.2 Enhancement based on a hybrid approach

The proposed EL speech enhancement method is shown in Fig. 18. As laryngectomees have the capability to properly articulate the excitation signals, spectral parameters of EL speech do not have to be changed greatly to generate intelligible speech. Therefore, we use the spectral parameters refined with spectral subtraction without applying voice conversion. On the other hand, it is essentially difficult to generate excitation signals exhibiting natural F_0 patterns in EL speech production. Therefore, we use voice conversion to estimate the excitation parameters: i.e., F_0 and aperiodic components. The proposed hybrid method can be expected to yield much larger improvements in naturalness compared with the enhancement method based on spectral subtraction thanks to the use of more natural excitation signals generated from statistics of normal speech. It also can be expected to alleviate the degradation in intelligibility observed in the conventional enhancement method based on voice conversion by avoiding errors in spectral conversion.

3.3 F_0 pattern modification to improve statistical F_0 pattern prediction

3.3.1 Continuous F_0 patterns

The F_0 patterns extracted from natural voices are discontinuous because the F_0 values are not observed at the unvoiced and silent frames. In the conventional

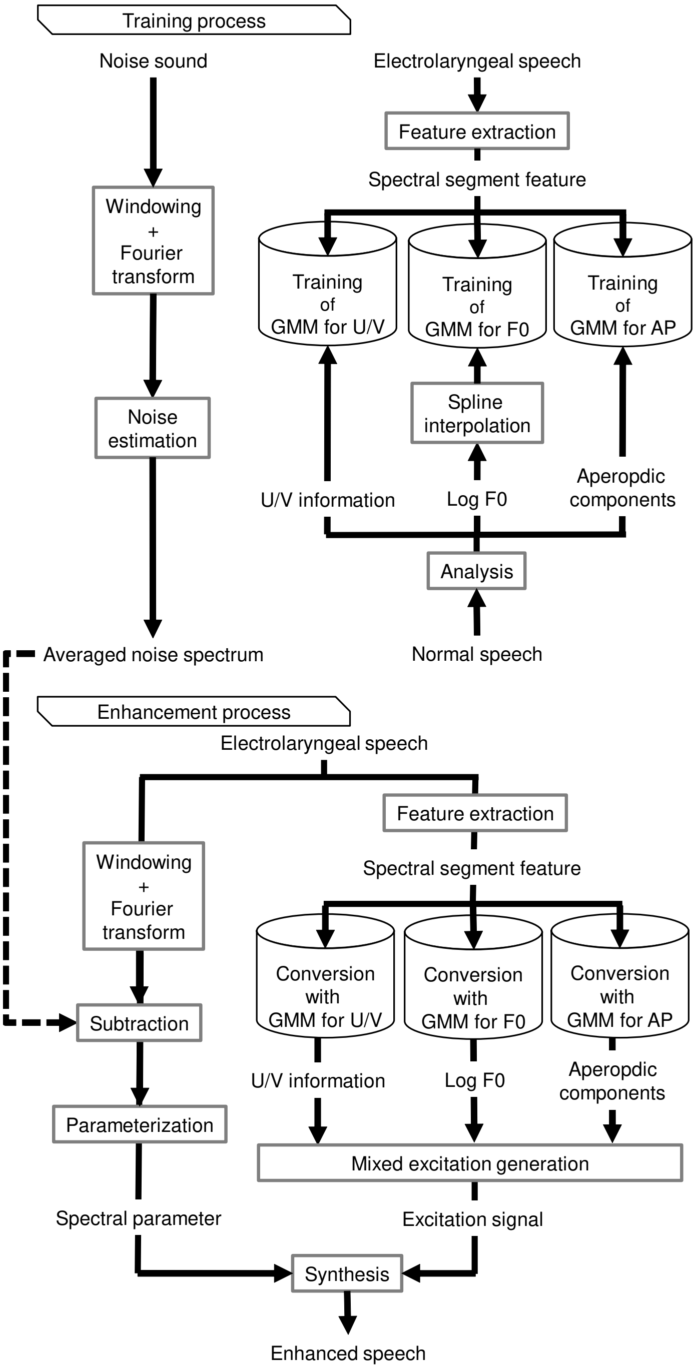


Figure 18. Process flow of EL speech enhancement based on the proposed hybrid approach. The upper side shows estimation of leaking out noise of the electrolarynx and model training to predict F_0 pattern of natural voice. The lower side indicates enhancement process.

enhancement method based on voice conversion, a constant value clearly different from F_0 values (e.g., a value much less than the minimum F_0 value) is used to represent F_0 values at those frames, and both unvoiced/voiced (U/V) prediction and F_0 value estimation are performed with a single GMM in the same manner as described in [55]. However, it is not straightforward to accurately model such a discontinuous F_0 pattern.

To simplify characteristics of the parameter sequence to be modeled, we apply a continuous F_0 pattern to the statistical excitation prediction. In the training process, continuous F_0 patterns of normal speech are generated by using spline interpolation to produce F_0 values at unvoiced frames, as shown in the middle of Fig. 19. The resulting continuous F_0 patterns are used as the target parameter in the GMM training. The conventional discontinuous F_0 patterns are also modeled with another GMM to predict U/V information. In the conversion process, a continuous F_0 pattern and U/V information are separately predicted using the corresponding GMMs. A discontinuous F_0 pattern to be used in synthesis is finally generated by combining them. The effectiveness of using the continuous F_0 pattern has also been reported in the field of statistical parametric speech synthesis [27, 64].

3.3.2 Remove micro-prosody through low-pass filtering

Rapid movements, called micro-prosody, are often observed in F_0 patterns extracted from natural voices. However, it is difficult to accurately model and reproduce these movements with a GMM. Moreover, an impact of micro-prosody on naturalness of synthetic speech is much smaller than that of F_0 patterns corresponding to phrase and accentual components. Therefore, it is helpful to make the GMM focus on modeling only those patterns. To achieve this, we propose the use of a method to smooth the continuous F_0 patterns with low-pass filtering [65] as shown in the middle of Fig. 19. The smoothed continuous F_0 patterns are then modeled with the GMM.

3.3.3 Avoiding unvoiced/voiced information prediction errors

In the excitation parameter prediction, unvoiced/voiced information is also predicted as mentioned above. Errors during this prediction process are also un-

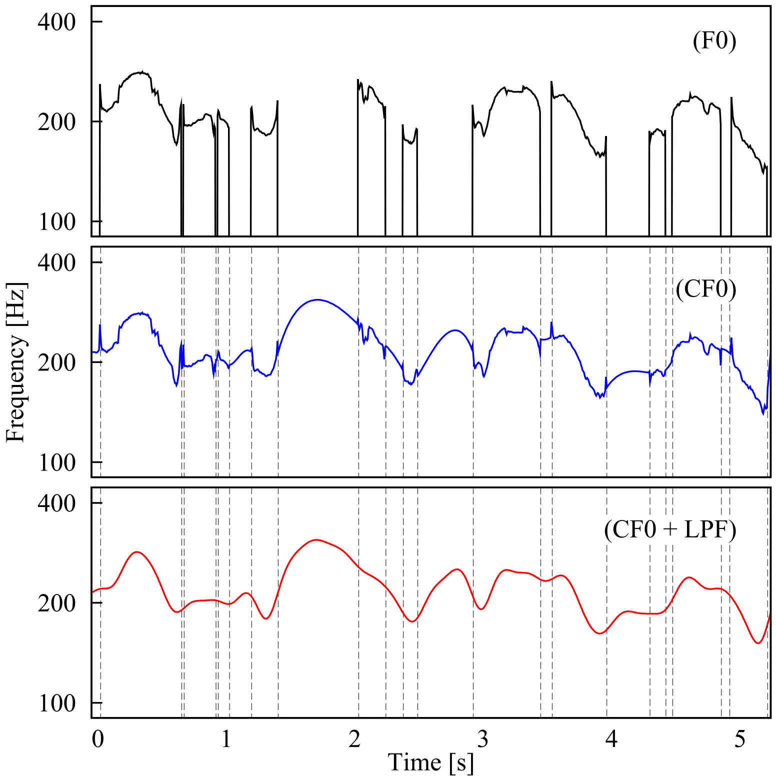


Figure 19. Each type of F_0 patterns. Top figure is target F_0 counters, middle is continuous F_0 patterns using the spline interpolation, and bottom is smoothing continuous F_0 patterns using the low-pass filter.

avoidable and they may cause adverse effects in intelligibility.

As EL speech is totally voiced speech, no degradation is caused even if the converted speech is generated by regarding all speech frames as voiced frames. To further reduce the possibility of degradation in intelligibility caused by U/V prediction errors, we also propose the use of continuous F_0 patterns without any unvoiced frames for speech segments to generate the excitation signals. In the conversion process, continuous F_0 patterns are predicted over all frames. Then, only silence frames are automatically detected using waveform power and unvoiced excitation signals are generated only at those frames. Unvoiced phoneme sounds cannot be generated in this method as in the original EL speech but the converted speech does not suffer from wrongly predicted unvoiced frames.

3.4 Experimental setup and corpora

3.4.1 Experimental setup

We objectively and subjectively compared the performance of the proposed system with the conventional systems. In our experiments, the source speaker was one laryngectomee and the target speaker was one non-disabled speaker. Both speakers recorded 50 sentences in the ATR phonetically balanced sentence set [66]. We conducted a 5-fold cross validation test in which 40 utterance pairs were used for training, and the remaining 10 utterance-pairs were used for evaluation. Sampling frequency was set to 16 kHz. In the EL speech enhancement methods based on voice conversion, the 0th through 24th mel-cepstral coefficients extracted by STRAIGHT analysis [54] were used as the spectral parameters. The shift length was set to 5 ms. STRAIGHT analysis was also used for spectral extraction of EL speech but F_0 values were set to 100 Hz without F_0 extraction, which was equivalent to F_0 values of excitation signals generated by the electrolarynx used in the experiments. For the segment feature extraction, the current ± 4 frames were used to extract a 50-dimensional feature vector. PCA was conducted to determine the transformation matrix and the bias vector using all EL speech samples in the training data. The numbers of mixture components were set to 32 for the spectral and aperiodic estimation, 64 for the F_0 estimation, and 32 for continuous F_0 estimation. In the EL speech enhancement method based on spectral subtraction, the number of FFT points was set to 512 and individual parameters were set to $\alpha = 2.0$, $\beta = 0, 0$, and $\gamma = 1.0$, which were manually determined by listening to the enhanced speech so that its voice quality was improved as much as possible. The cut-off frequency of the low-pass filter was set to 10 Hz. Note that in the EL speech enhancement methods based on voice conversion, mel-cepstral and aperiodic distortion is shown in Table 4.

Table 4. Conversion accuracy in enhancement methods based on voice conversion.

Mel-cepstral distortion without power information	5.09 dB
Aperiodic distortion	3.19 dB

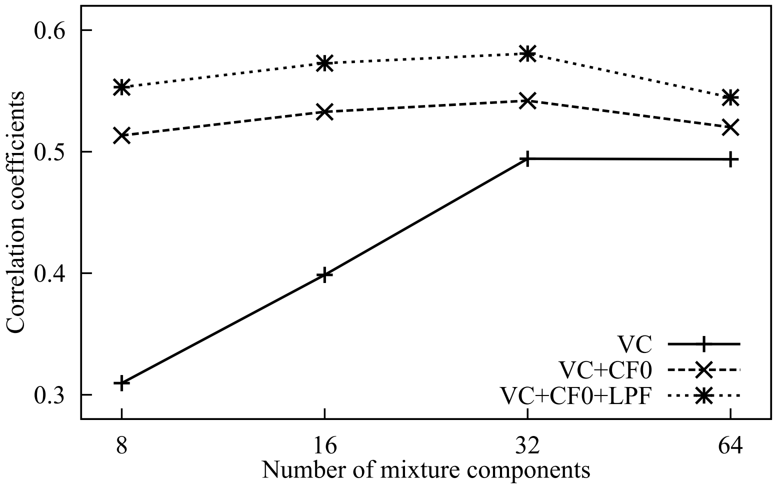


Figure 20. Prediction accuracy for the original F_0 patterns (F0), the continuous F_0 patterns (CF0), and the smoothed continuous F_0 patterns (CF0+LPF).

3.4.2 Objective evaluation

We evaluated the effectiveness of the proposed preprocessing for the training data, including the continuous F_0 estimation method and the low-pass filter. Note that the effectiveness of the continuous F_0 estimation method has already been reported in [27, 64]. As measures to evaluate the prediction accuracy of the excitation features, we used the correlation coefficient and U/V error rate on F_0 components between the converted speech parameters and the natural target speech parameters. As for the evaluation of the F_0 correlation coefficient, we set the number of GMM mixture components to 8, 16, 32, or 64. We evaluate three systems using the normal F_0 patterns extracted from target natural voices, the continuous F_0 patterns interpolated with the normal F_0 patterns using spline interpolation, and the smoothing continuous F_0 patterns extracted from continuous F_0 patterns through the low-pass filter. On the other hand, for the evaluation of U/V error rate, we also set the number of GMM mixture components for VC to 8, 16, 32, or 64.

Fig. 20 shows the result of the evaluation for the F_0 correlation coefficient. In the case of using only the voice conversion method **F0**, the F_0 correlation coefficient depends on the number of GMM mixture components, and is maximized

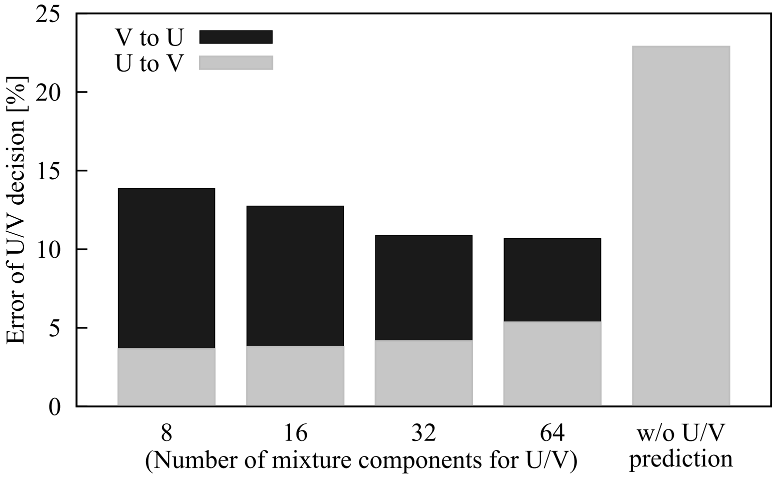


Figure 21. Prediction error rate for unvoiced/voiced information on each setting.

with 64 mixture components. However, as for the proposed method, **CF0** and **CF0+LPF**, those are not depended well. Especially, in the small number of GMM mixture components, a significant degradation is not observed compared with **F0**. Moreover, it can be observed that the F_0 correlation coefficient is improved by the continuous F_0 estimation and also improved by using the low-pass filter. Note that in the case of the use of the continuous F_0 estimation method and the low-pass filter, the F_0 correlation coefficient is maximized with 32 mixture components.

Fig. 21 shows the result of the evaluation in term of the prediction error rate for unvoiced/voiced information. We have found that large errors in the F_0 estimation tend to be observed at short voiced segments that are sometimes generated in only the VC-based enhancement method. This improvement is similar to that yielded by the continuous F_0 modeling in HMM-based speech synthesis [27, 64]. As the number grows larger, voiced-to-unvoiced error rate decreases while unvoiced-to-voiced increases. With 64 mixture components, the unvoiced/voiced information error rate is minimized. On the other hand, without unvoiced/voiced prediction, the unvoiced/voiced error rate is constant. In particular, the V-to-U error rate is practically zero. The voiced-to-unvoiced errors still exist without the continuous F_0 estimation method owing to errors in the automatic silence frame detection with waveform power, but they are almost negligible. However, unvoiced-to-voiced significantly increases owing to the continuous F_0 estimation method. Note that as this increase causes no adverse effect compared with EL speech because EL speech is totally voiced speech.

3.4.3 Subjective evaluation

We conducted two opinion tests of listenability and naturalness and a dictation test on intelligibility. In this dissertation, the term "listenability" is used to indicate a score that was measured by asking the listener to subjectively evaluate how easy it was to understand the utterance. The term "intelligibility" is used to indicate a score that was calculated by asking the listener to write down the content of the utterance, and measuring the accuracy of transcription. The term "naturalness" is used to indicate a score that was measured by asking the listener to subjectively evaluate whether or not the evaluated speech is similar to natural human speech. In the opinion tests, each listener evaluated the naturalness and listenability of the enhanced voices using a 5-scaled opinion score (1: Bad, 2: Poor, 3: Fair, 4: Good, and 5: Excellent). We evaluated the following five types of speech samples:

EL original EL speech

SS enhanced EL speech using spectral enhancement based on only spectral subtraction

VC enhanced EL speech using the enhancement method based on only voice conversion for not only spectral parameters but also excitation features

SS+VC enhanced EL speech using the proposed hybrid enhancement method with discontinuous F_0 pattern prediction and unvoiced/voiced information prediction

SS+VC+CF0 enhanced EL speech using the proposed hybrid enhancement method with smoothed continuous F_0 pattern prediction and without unvoiced/voiced information prediction

On the other hand, in the dictation test, in order to demonstrate the effectiveness of avoiding unvoiced/voiced information prediction errors, we evaluated the following five types of speech samples:

EL original EL speech

SS enhanced EL speech using spectral enhancement based on only spectral subtraction

Hybrid(V) EL speech enhanced by the proposed hybrid enhancement method with smoothed continuous F_0 estimation

Hybrid (U/V) EL speech enhanced by the proposed hybrid enhancement method with the combination of unvoiced/voiced prediction and smoothed continuous F_0 estimation

Hybrid (target U/V) EL speech enhanced by the proposed hybrid enhancement method with the combination of ideal unvoiced/voiced information and smoothed continuous F_0 estimation

As the reference unvoiced/voiced information, we use target unvoiced/voiced information obtained by performing dynamic time warping between the enhanced speech parameters using the voice conversion and the natural target speech parameters. Intelligibility was evaluated using word correct rate and word accuracy, which were calculated as follows:

$$\texttt{word correct rate [\%]} = \frac{C}{S+D+C} \tag{27}$$

$$\texttt{word accuracy [\%]} = \frac{C-I}{S+D+C} \tag{28}$$

where S is the number of substitutions, D is the number of deletions, I is the number of insertions, and C is the number of correct words. Note that the EL speech enhancement method based on voice conversion generally causes a significant degradation in intelligibility (around 3% word recognition rate reduction) compared to EL speech as reported in [16]. All tests were performed by 5 listeners. Each listener evaluated 50 samples, 10 samples per system.

First, in Fig. 22, we show the results of the subjective opinion test on listenability. It can be seen that a slight improvement is yielded by **SS**. On the other hand, **VC** causes significant degradation as reported in [16]. **SS+VC** doesn't cause a degradation compared with **EL** but it still causes a very small degradation compared with **SS**. This adverse effect on listenability is not observed in the

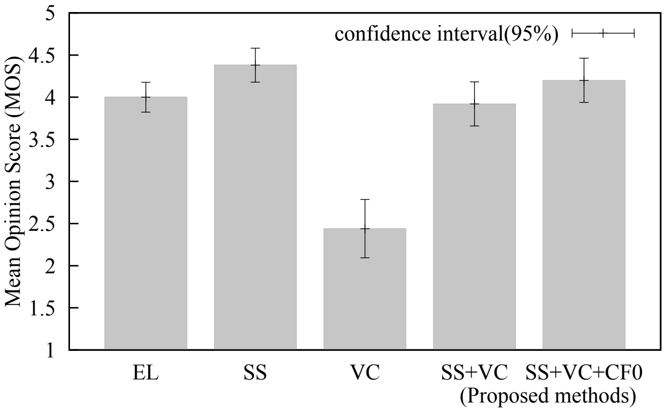


Figure 22. Result of opinion test on listenability.

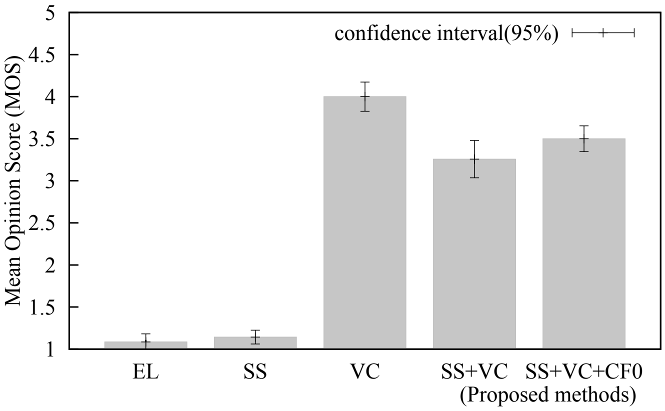


Figure 23. Result of opinion test on naturalness.

proposed hybrid methods (**SS+VC** and **SS+VC+CF0**) thanks to no spectral conversion error.

Fig. 23 shows a result of the opinion test on naturalness. **SS** yields a very small improvement in naturalness. On the other hand, **VC** yields a significantly larger improvement. The proposed hybrid methods (**SS+VC** and **SS+VC+CF0**) also yield significantly larger improvements compared with **SS** as they are capable of generating more natural F_0 patterns. We can also observe that the continuous F_0 estimation is effective for improving naturalness as well.

Fig. 24 shows a result of the dictation test on intelligibility. We found that the hybrid methods do not cause any degradation in intelligibility compared with EL

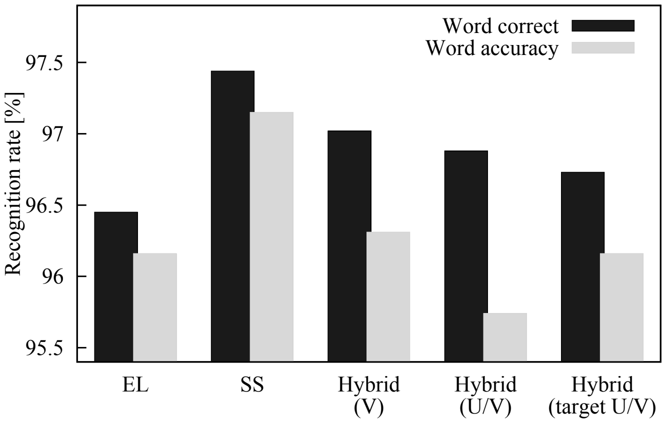


Figure 24. Result of opinion test on intelligibility.

speech. Furthermore, in the hybrid method that avoided unvoiced/voiced prediction by using the continuous F_0 estimation method, the intelligibility is preserved, similarly to the hybrid method using ideal unvoiced/voiced information. Hence, it can be said that unvoiced/voiced prediction is not always required. On the other hand, the hybrid methods tend to degrade intelligibility slightly compared to **SS**, owing to several issues, such as the effect of synthesis by using vocoder and using 24-dimensional mel-cepstral coefficients as spectral features.

These results suggest that the proposed hybrid approach to EL speech enhancement based on the continuous F_0 estimation and using the low-pass filter is effective in significantly improving naturalness of EL speech while avoiding degradation in listenability that is often observed in the conventional VC-based enhancement method.

3.5 Summary

In this chapter, we have proposed a hybrid approach to EL speech enhancement based on spectral subtraction for spectral parameter estimation and statistical voice conversion for excitation parameter prediction. To further improve the excitation parameters estimation, we have also proposed smoothed continuous F_0 pattern prediction as part of the proposed approach. Moreover, we have investigated the importance of U/V information. As a result of an experimental evaluation, it has been demonstrated that the proposed approach is capable of significantly improving naturalness of EL speech while causing no adverse effect such as the degradation in intelligibility. Furthermore, U/V prediction is not always required for EL speech enhancement.

Chapter

4

Direct control of F_0 pattern generated by an electrolarynx based on statistical F_0 pattern prediction

4.1 Introduction

To generate more natural F_0 patterns, we have proposed methods that make it possible to convert acoustic features of EL speech to those of normal-sounding speech by predicting the F_0 pattern from the spectrum sequence of the EL speech based on GMMs as reported in [22, 16] and **Section 3**. With a similar aim, whisper-to-speech conversion [23] and whisper-to-audible speech conversion [24] have been proposed. These methods have successfully shown to improve the naturalness of EL speech [22, 16] while preserving its intelligibility in **Section 3**. However, these frameworks restrict the use situation because the enhanced EL speech is presented with a loudspeaker. For example, these frameworks are effective for telecommunication, but not suitable to face-to-face conversation because the enhanced EL speech presented from a loudspeaker and the original EL speech produced from their own mouthes are presented to a listener at the same time.

The several techniques, which make it possible to directly produce the enhanced EL speech from their own mouthes, have been proposed to control F_0 patterns of the excitation signals generated from an electrolarynx additionally using intentionally controllable signals, such as expiratory air pressure [18], finger movements [19, 20], and forearm movements [21]. Although these methods can change the F_0 patterns, it is inherently difficult to control these signals to generate natural F_0 patterns corresponding to linguistic content of the speech. To make it possible to control without conscious operation, methods using neck surface electromyography (EMG) or intramuscular cricothyroid (CT) EMG have been proposed [67, 68, 69]. Although myoelectricity measured by CT-EMG is strongly correlated with F_0 patterns, the CT muscles are accessible only through invasive needle electrodes. On the other hand, although the surface EMG makes it easy to measure the myoelectricity, the quality of the myoelectricity measured by the surface EMG strongly depends on the position of EMG. Even if we successfully measure the myoelectricity by using the surface EMG, the correlation results in around 0.5.

In this chapter, we propose an EL speech enhancement system, as shown in Fig. 25, effective for any situation, including face-to-face conversation. F_0 patterns of the excitation signals produced by the electrolarynx are directly controlled using statistical F_0 pattern prediction. Namely, an F_0 value at a current frame is predicted in real time from the EL speech produced by the laryngectomee articulating the excitation signals with previously predicted F_0 values. Consequently, the proposed system has the potential to allow laryngectomees to directly produce enhanced EL speech with more natural F_0 patterns than the original EL speech, and present only the enhanced EL speech to the listener.

Furthermore, we investigate latency issues caused by real-time processing. To address the negative impacts caused by latency of the real-time prediction on the F_0 prediction accuracy, we also propose two methods: 1) modeling of segmented continuous F_0 (CF_0) patterns to shorten the required delay time in real-time statistical F_0 prediction and 2) prediction of forthcoming F_0 values to cancel the impact of the processing delay of the prototype system.

Through an actual implementation of our proposed speaking aids system and its simulation, we demonstrate that 1) our proposed speaking aid system successfully achieved to cause no degradation in term of naturalness compared with the use of batch-type prediction algorithm. 2) the delay time required to predict CF_0 patterns in the conventional prediction method can be significantly reduced by using the segmented CF_0 modeling strategy, and 3) the negative impacts of the processing delay can be effectively alleviated by predicting the forthcoming F_0 values.

The rest of this chapter is organized as follows.In **Section 4.2**, we describe the strategy of direct control method of the electrolarynx. In **Section 4.3**, we show the setting of actual implementation of our prototype system. In **Section 4.4**, we also design the simulation experiment for flexible evaluations. In **Section 4.5**, we explain the latent issues caused by real-time processing. In **Section 4.6**, we evaluate the performance of our proposed control method.

4.2 Control strategy

Our proposed speaking aids system, as shown in Fig. 25, to directly and automatically control F_0 patterns of the excitation signals generated from an electrolarynx consists of prediction and articulation processes. In the prediction process, the F_0 value is predicted from EL speech produced by a laryngectomee frame by frame using the real-time prediction algorithm. In the articulation process, to produce the EL speech, the laryngectomee articulates the excitation signals of the elec-

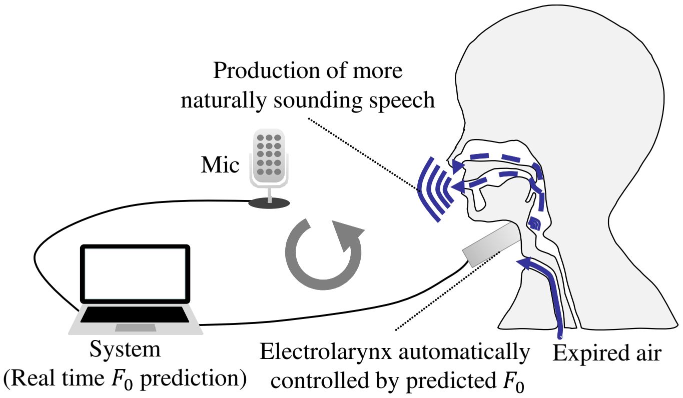

Production of more naturally sounding speech
Mic
System
(Real time F_0 prediction)
Electrolarynx automatically controlled by predicted F_0
Expired air

Figure 25. Proposed system to control F_0 patterns of excitation signals of an electrolarynx using statistical F_0 pattern prediction for laryngectomees.

trolarynx reflecting predicted F_0 values. Therefore, this system allows laryngectomees to directly produce enhanced EL speech with more naturally sounding F_0 patterns corresponding to linguistic contents because the source spectral features of EL speech capture the linguistic contents.

4.3 Implementation detail

A prototype one of our proposed speaking aids system was developed using a microphone, a laptop, and a digital/analog (D/A) converter shown in Table 5. As shown in Fig. 25, EL speech produced from a mouth of a laryngectomee is detected with a usual close-talk microphone. The EL speech signal is recorded on a laptop and F_0 patterns of normal speech are predicted on the fly by using the real-time prediction algorithm. The predicted F_0 values are linearly converted to voltage values to control the F_0 values of the excitation signals. Then, through the D/A converter connected from the laptop to the electrolarynx, an electric signal corresponding to the determined voltage values is generated. Finally, the electrolarynx generates the excitation signals reflecting the predicted F_0 values according to the input electric signal generated from the D/A converter.

As mentioned in the previous section, the F_0 patterns are constantly delayed owing to latency of the real-time prediction process. Moreover, additional latency is caused in our prototype system because of the use of D/A converter.

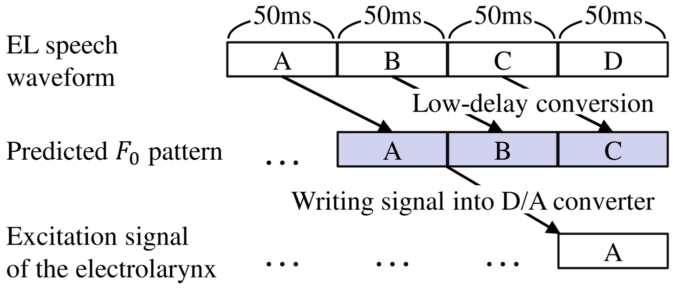


Figure 26. Latency caused by each process of our prototype system.

Fig. 26 shows the latency caused by each process of our prototype system. For the real-time prediction process, 50 msec latency is caused in our conventional implementation [29]. For the D/A part to convey the digital signals, it takes around 50 msec. Consequently, the whole D/A part causes 100 msec latency because the digital signal to be written needs to be determined before starting writing. In total, 150 msec latency is caused in the prototype system. Note that the latency in the D/A part could be addressed by the development of a special device for the electrolarynx. Moreover, we have successfully implemented statistical voice conversion processing on a digital signal processor (DSP) [70]. It is thus expected that all processors could be embedded into the electrolarynx and total latency will be decreased to the 50 msec caused by the real-time statistical F_0 prediction.

4.4 Simulation experiment

To flexibly investigate the performance of our proposed control method, we also design a simulation method of EL speech production process using the con-

Table 5. Electronic devices on the prototype system

Electrolarynx	Yourtone
Microphone	Crown CM-311A
CPU of the laptop	Intel(R) Core(TM) i5-4200U
D/A converter	AIO-160802AY-USB

trolled electrolarynx. The simulated process is shown in the right side of Fig. 27. EL speech signals produced by articulating the excitation signals based on the predicted F_0 values are artificially generated using the STRAIGHT [54] analysis/synthesis method.

At first, 1) we extract spectral envelope parameters and aperiodic components [71] from the original EL speech in advance by using STRAIGHT analysis. These features to approximate the EL speech production process capture acoustic properties determined by articulation and the excitation signals leaking out as noise from the electrolarynx, except for periodicity of the excitation signals. Then, 2) spectral segment features are extracted from EL speech, and F_0 patterns of normal speech are predicted from them based on the real-time F_0 pattern prediction. 3) The predicted F_0 patterns are just delayed to consider the delay time caused by whole process of our prototype system, such as mentioned D/A part. 4) Using the delayed F_0 patterns and the extracted aperiodic components, excitation signals are generated based on the mixed excitation model [61] to replace actual excitation signals of the electrolarynx. 5) Finally, the enhanced EL speech is approximately synthesized by filtering the generated excitation signals with the extracted spectral envelope parameters reflecting the articulation. Note that in our prototype system, the F_0 values of the enhanced EL speech suffer from those of previous time step because the F_0 values are predicted from EL speech reflecting the F_0 values of previous time step. However the above mentioned processing without iterative update, Step 3) to 5), results in the F_0 prediction using the spectral segment features extracted from the original EL speech. To reflect the impact of the predicted F_0 values of previous time step, 6) the spectral segment features are extracted again from the synthesized EL speech and F_0 pattern prediction is also performed again using the extracted spectral segment features. Step 3) to 6) are iteratively repeated until the predicted F_0 patterns converge. If they converge, the proposed system may be expected to work stably because the EL speech produced with the predicted F_0 patterns is consistent with that used in the spectral segment feature extraction.

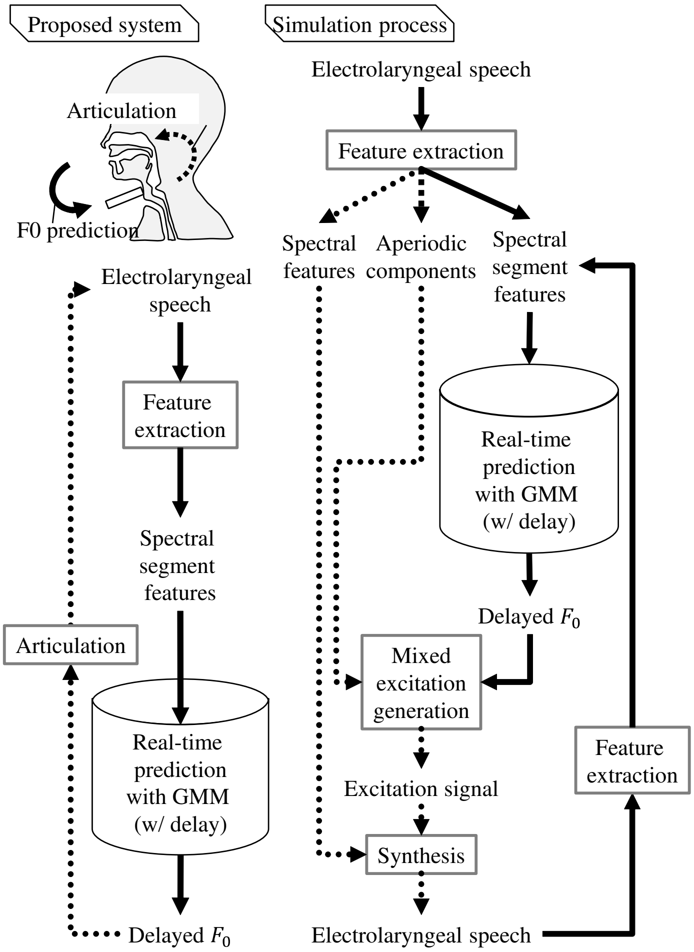


Figure 27. Process flow of the proposed system and its simulation implementation

4.5 Negative impact of latency

Through the use of the prototype system, we confirmed that it yields significant improvements in naturalness of EL speech while preserving its high intelligibility. However, we also found that the naturalness of enhanced EL speech tends to be lower than that yielded by the batch-type prediction.

As mentioned in **Section 2.6**, the latency to predict F_0 patterns is inherent in our proposed speaking aids system. It has been reported in a spectral conversion task [62] that the delay time depending on the segment feature length L in the real-time prediction process requires around 50 to 70 msec to maintain the conversion accuracy of the batch-type prediction process. On the other hand, no previous work has examined the effect of latency for the F_0 prediction accuracy. It is possible that longer delay will be required because F_0 is a suprasegmental feature, which has strong correlation over a wider range compared to segmental features, such as spectral features. Moreover, in our prototype system mentioned in **Section 4.3**, the additional latency is caused by using D/A converter to convey predicted F_0 values to the electrolarynx. These latency on our proposed system leads to asynchronous problem between articulation and F_0 patterns of excitation signals generated by the electrolarynx. To address these issues, we also propose the use of segmented continuous F_0 patterns as trained target features and forthcoming F_0 prediction for reducing the latency caused by the real-time prediction process while preserving F_0 prediction accuracy at the level of the batch-type prediction process.

4.5.1 Segmented continuous F_0 patterns

In the previous CF_0 modeling method, the prediction process given in Eq. (17) is performed utterance by utterance. Because inter-frame correlation over an utterance is considered in this process, a long delay is required in real-time prediction to achieve sufficient prediction accuracy.

To reduce the delay time, we propose a segmented CF_0 pattern modeling method to make the range of which we consider inter-frame correlation shorter than an utterance. Shorter segments are first extracted from each utterance, and then, CF_0 patterns of individual segments (i.e., segmented CF_0 patterns) are modeled and predicted separately. In this dissertation, we determine the individual segments by extracting time frames of which the waveform power is over a pre-determined threshold. An example of the segmented CF_0 patterns is shown in Fig. 28. Note that the segmented CF_0 patterns are still different from the original F_0 pattern, which is segmented by unvoiced frames, in that 1) the segmented CF_0 patterns can also include unvoiced frames, and thus they tend to be longer than segments observed in the original F_0 patterns, and 2) each segmented CF_0 pattern varied more smoothly than the original F_0 patterns.

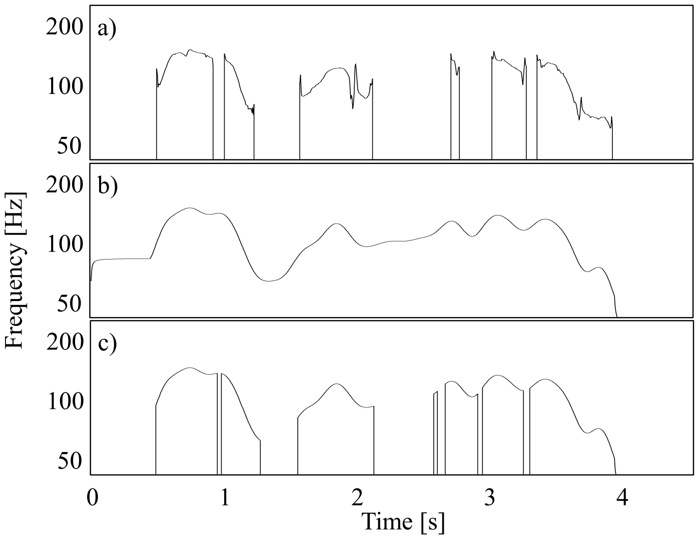


Figure 28. a) F_0 patterns extracted from normal speech, b) smoothed continuous F_0 patterns interpolated at unvoiced frames, and c) segmented CF_0 patterns of (b) extracted by using the power of waveform.

4.5.2 Forthcoming F_0 prediction

In order to cancel the misalignment between articulation and the constantly delayed F_0 patterns predicted in the real-time process, we investigate the possibility of predicting forthcoming F_0 values. We train the GMM for modeling the joint probability density function $P([\boldsymbol{X}_t^\top, \boldsymbol{Y}_{t+F}^\top]^\top | \boldsymbol{\lambda}_G)$ of the source features at time frame t, $\boldsymbol{X}_t$ and the target features at time frame $t+F$, $\boldsymbol{Y}_{t+F}$. The trained GMM is used to predict the F_0 value at F frames ahead. For example, if the latency of the prototype system is set to 200 msec, we train the GMM to predict the F_0 values at 200 msec ahead. Consequently, there is no mismatch between articulation and the predicted F_0 patterns. It is expected that there is a trade-off between the prediction accuracy and the setting of F; i.e., larger F accepts a longer delay time in the real-time prediction process, which makes the real-time prediction accuracy close to the batch-type prediction accuracy; on the other hand, it is obviously more difficult to predict F_0 values at frames far away from the current one than those at closer frames.

4.6 Experimental setup and corpora

4.6.1 Experimental setup

We conducted 5 objective evaluations to examine the performance of the proposed methods, 1 subjective evaluation to investigate the amount of the allowance of the misalignment between articulation and F_0 patterns, and 1 subjective evaluation to examine the naturalness of the proposed methods. The first evaluation of objective evaluations is a comparison of the prediction accuracy among three types of F_0 pattern modeling, F_0 pattern, smoothed continuous F_0 (CF_0) pattern, and the proposed segmented CF_0 pattern, in the batch-type prediction process. The second evaluation is a comparison of the accuracy of batch-type F_0 prediction and real-time F_0 prediction. The third evaluation is for the validity of the proposed simulation experiment to simulate our proposed speaking aids system. The fourth evaluation is conducted to investigate the negative impacts caused by latency on the proposed system and to examine the effectiveness of the proposed segmented CF_0 pattern modeling. The last objective evaluation is conducted to examine the effectiveness of the proposed forthcoming F_0 prediction method. The first evaluation of subjective evaluations is constructed to investigate how much delay is acceptable for the ears of human beings. Since the real-time processing causes the latency problem as mentioned in **Section 4.5**, we investigate the amount of the allowance of the misalignment between articulation and F_0 patterns. Setting L in **Section 2.6.2** to the amount considering the acceptable misalignment, we evaluate the enhanced EL speech in term of the naturalness.

The source speech was EL speech uttered by a male speaker, and the target speech was normal speech uttered by a professional female speaker. Each speaker uttered about 50 sentences in the ATR phonetically balanced sentence set [66]. We conducted a 5-fold cross validation test in which 40 utterance pairs were used for training, and the remaining 10 utterance pairs were used for evaluation. Sampling frequency was set to 16 kHz. We employed FFT analysis with a 25 msec hanning window to extract the mel-cepstra of EL speech as the spectral features. The frame shift length was set to 5 msec. As the source features, the spectral segment features were extracted from the mel-cepstra at the current ± 4 frames. On the other hand, F_0 values of normal speech were extracted with STRAIGHT

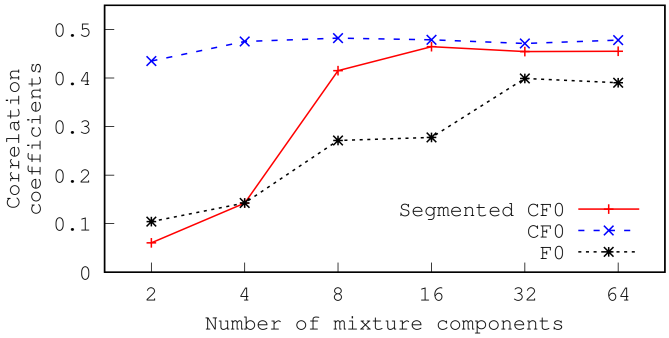


Figure 29. Prediction accuracy on batch-type prediction.

F_0 analysis [54] and CF_0 patterns were generated as the target feature using a low-pass filter with 10 Hz cut-off frequency. Moreover, the target F_0 patterns were shifted so that their mean value was equal to 100 Hz to predict F_0 patterns suitable for the source male speaker.

4.6.2 Comparison of F_0 pattern modeling in batch-type prediction

To choose best setting from a variety number of mixture components for later evaluations, we evaluated the prediction accuracy of each F_0 pattern modeling method in the batch-type process using the correlation coefficient between the predicted F_0 pattern and the target F_0 pattern. As shown in Fig. 29, the best number of mixture components is 32 for F_0, 16 for CF_0, and 16 for segmented CF_0. We found that reducing the variability of F_0 patterns such as rapid movements, we achieved to train F_0 patterns with smaller number of mixture components. Moreover, we also confirmed that CF_0 brings better performance compared with the original F_0 because continuous sequence makes it possible to consider inter-frame correlation over an utterance. The proposed segmented CF_0 preserves such an improvement relatively well while minimizing degradation of the prediction accuracy.

4.6.3 Comparison of batch-type prediction and real-time prediction

As mentioned in **Section 4.5.1**, it is possible in the real-time prediction that the larger delay time is required in the CF_0 pattern than in the F_0 pattern to

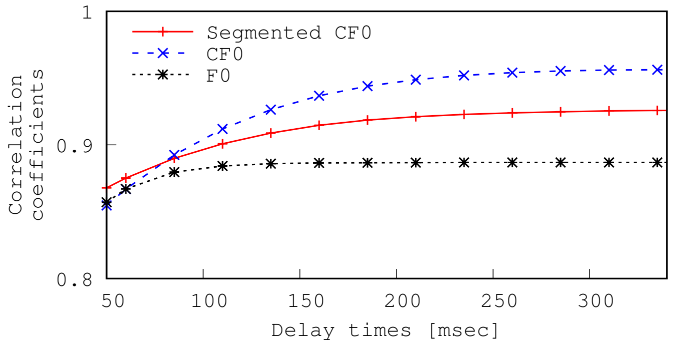


Figure 30. Prediction accuracy on batch-type prediction.

achieve the prediction accuracy comparable to that of the batch-type prediction. To examine this possibility, we calculated a correlation coefficient between the F_0 pattern predicted by the real-time prediction with various settings of the delay time and that by the batch-type prediction.

The result is shown in Fig. 30. As for the F_0 pattern, even if setting the delay time to 85 msec (corresponding to $L = 10$), a quite high correlation coefficient is achieved. On the other hand, as for the CF_0 pattern, the predicted patterns are quite different from those by the batch-type process, showing that the correlation coefficient is similar to the case of F_0 pattern when setting the delay time to less than 85 msec. Moreover, its accuracy convergence is much slower compared to that observed in the F_0 pattern. Consequently, in the CF_0 pattern, the delay time needs to be set to around 250 msec to achieve the prediction accuracy comparable to that of the batch-type prediction. As we expected, the segmented CF_0 modeling converge faster compared with the CF_0 pattern modeling because the number of frames considering inter-frame correlation is limited.

4.6.4 Comparison of prototype system and simulated system

The F_0 patterns predicted by the prototype system strongly correlate to those by the simulated system, with a correlation coefficient higher than 0.9 as shown in Fig. 31. This high correlation demonstrates that the proposed implementation is effective and the simulated system is able to effectively approximate the results of the prototype system. This result allows us to replace the evaluations of our

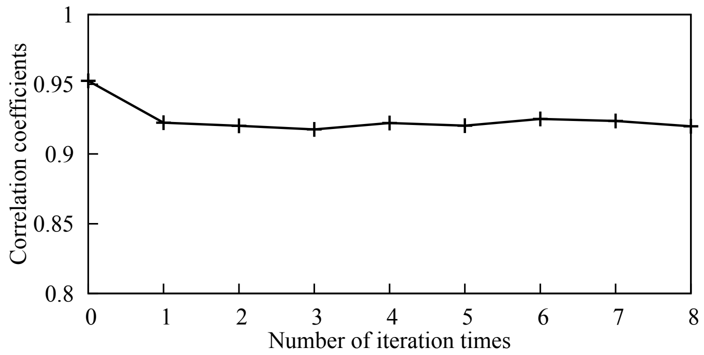


Figure 31. Correlation of predicted F_0 patterns between prototype system and simulated system.

prototype system into those of simulated system.

4.6.5 Negative impacts caused by latency

We evaluated the real-time prediction accuracy of each F_0 modeling method using the correlation coefficient between the predicted F_0 pattern and the target F_0 pattern. To evaluate only the prediction accuracy, we also evaluate predicted F_0 patterns with delay time correction at time of evaluation. As shown in solid lines of Fig. 32, the effect of the misalignment between the predicted and the target F_0 patterns, which is observed on the prototype system, was removed in this evaluation by shifting the predicted F_0 patterns according to the delay time settings in calculation of the correlation coefficient.

The result shows in Fig. 32. As for the solid lines, we confirmed a similar tendency to results in **Section 4.6.3**. as for the F_0 pattern, we found that although the prediction accuracy quickly converges at around 60 msec of the delay time, the resulting correlation coefficient is lower than 0.4 because the prediction accuracy of the batch-type prediction is also low, as shown in Fig. 29. As for the CF_0 pattern, the converged prediction accuracy is significantly higher than that in the F_0 pattern, as also observed in Fig. 29, and its convergence is very slow. To achieve sufficient prediction accuracy, the delay time needs to be set to around 250 msec. On the other hand, the use of the proposed segmented CF_0 patterns makes the convergence faster than that of the CF_0 patterns while

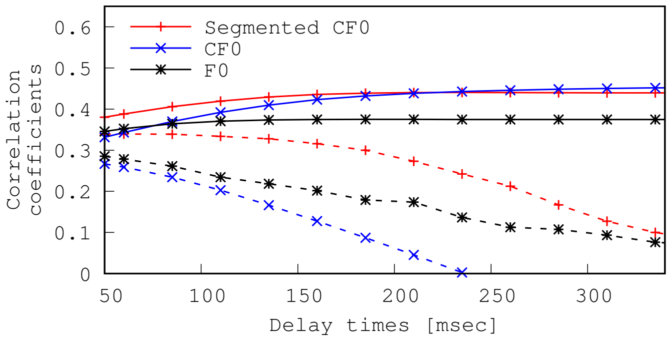


Figure 32. Prediction accuracy on real-time prediction for each F_0 patterns. Solid lines result w/ delay time correction at the time of evaluation, and dash lines result w/o delay time correction.

preserving its prediction accuracy. As for the dash lines, the delay time is set to longer, the prediction accuracy gets lower. However, the segmented CF_0 pattern makes it possible to alleviate the negative impact of latency compared with the other baseline F_0 modelings.

4.6.6 Amount of allowance of misalignment

As mentioned in **Section 4.5**, time asynchronous problems between articulation and F_0 patterns are caused by the real-time processing. Therefore, we investigate that how much asynchronous is unnoticeable for the ears of human beings. This investigation helps laryngectomees to generate more naturally sounding EL speech reflecting predicted F_0 patterns because the longer latency makes the better prediction accuracy of F_0 patterns in real-time statistical F_0 pattern predictions. We use STARIGHT analysis/synthesis framework [54] to generate the speech signals using shifted F_0 patterns. To prepare the shifted F_0 patterns, we interpolate F_0 patterns at unvoiced frames after analysis acoustic parameters. Then we just shift the F_0 patterns, and extract the F_0 patterns at voiced frames of the original speech signals.

The evaluated speech is the normal speech uttered by a female of the professional. For various settings of the delay time, the listener evaluate the modified speech in term of the naturalness by using a 5-scaled opinion score (1: Bad, 2:

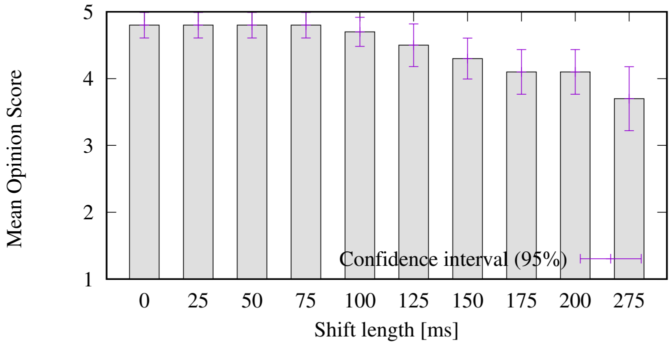


Figure 33. Amount of allowance of misalignment between articulation and F_0 patterns.

Poor, 3: Fair, 4: Good, and 5: Excellent). The term “naturalness” is used to indicate a score that was measured by asking the listener to subjectively evaluate whether the evaluated speech is similar to natural human speech or not. The number of listeners was 5 and each listener evaluate 10 sentences per one system. Hence, each systems are evaluated with 50 sentences.

The result is shown in Fig. 34. We can find that we allow the smaller misalignment less than 100 ms. After that, the larger misalignment degrade the naturalness of speech.

4.6.7 Evaluation of the proposed forthcoming F_0 prediction

We evaluated the real-time prediction accuracy also considering the effect of the misalignment between articulation and the delayed F_0 patterns predicted in the real-time process, which was observed in a practical situation, using the correlation coefficient between the predicted F_0 pattern without any correction of the delay time and the target F_0 pattern.

The proposed forthcoming F_0 prediction method was applied to the CF_0 pattern and proposed segmented CF_0 pattern, and its effectiveness was examined. The result is shown in Fig. 34. If not using the proposed forthcoming F_0 prediction, the delay time is set to longer, the prediction accuracy gets lower. This result shows that the adverse effect of the misalignment on the actual prediction accuracy is significantly large. This issue is well addressed by using the

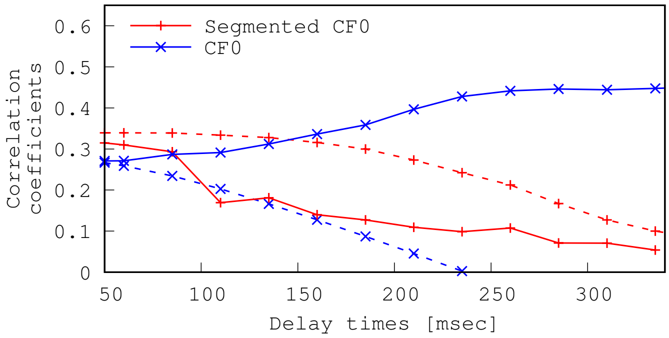


Figure 34. Comparison of basic modeling (dash lines) and forthcoming modeling (solid lines).

proposed forthcoming F_0 prediction for CF_0 pattern modeling. Consequently, by setting the delay time to around 250 msec, the real-time prediction with the proposed forthcoming F_0 prediction method makes it possible to achieve prediction accuracy comparable to that of the batch-type prediction. However, as for the segmented CF_0 patterns, even if we apply the proposed forthcoming F_0 prediction, its prediction accuracy is not improved. This result shows that restricting the unit considering inter-frame correlation makes it difficult to predict F_0 values at frames far away from the current one than those at closer frames.

Considering the result in **Section 4.6.6**, 85 ms of the delay time in segmented CF_0 pattern modeling without the forthcoming F_0 prediction is expected to get the best performance for the prototype system with low speed of CPU clock and small size of memory because of the computational costs in real-time prediction. The update of the Kalman gain $\boldsymbol{k}_d^{(t)}$ in **Section 2.6.2** requires the cost $O(N^3)$. The larger latency to improve the prediction accuracy makes it difficult to realize the real-time processing because the computational costs get larger. On the other hand, 265 ms of the delay time in CF_0 pattern modeling with the forthcoming F_0 prediction is expected to get the best performance if we can use high speed of CPU clock and large size of memory.

4.6.8 Naturalness of predicted F_0 patterns

Through the simulation experiment, we evaluated the naturalness of F_0 patterns predicted by using our proposed F_0 modeling. The term "naturalness" is used to indicate a score that was measured by asking the listener to subjectively evaluate whether the evaluated speech is similar to natural human speech or not. In the opinion tests, 5 listeners evaluated each speech quality using a 5-scaled opinion score (1: Bad, 2: Poor, 3: Fair, 4: Good, and 5: Excellent). The number of listeners was 5 and each listener evaluate 10 sentences per one system. Hence, each systems are evaluated with 50 sentences. Comparison methods are following 4 systems:

EL original EL speech

Batch enhanced EL speech with CF_0 patterns predicted by batch-type prediction algorithm. This is baseline system.

RT enhanced EL speech with real-time prediction algorithm for segmented CF_0 pattern modeling (delay time: 85 msec). This is a simulated system of our proposed system.

Forthcoming enhanced EL speech with forthcoming F_0 prediction on real-time prediction algorithm for CF_0 patterns modeling (delay time: 265 msec). This is also a simulated system.

The result is shown in Fig. 35. We confirmed that **Batch** is significantly improved compared with **EL** by predicting F_0 patterns based on statistical F_0 patterns. For our proposed methods **RT** and **Forthcoming**, we achieved that two proposed systems caused no degradation compared with **Batch**. This results show that our proposed methods successfully overcome the latency issues mentioned in **Section 4.5**.

4.7 Summary

In this chapter, we have proposed a new electrolarynx capable of automatically controlling F_0 patterns of its excitation signals based on statistical F_0 pattern

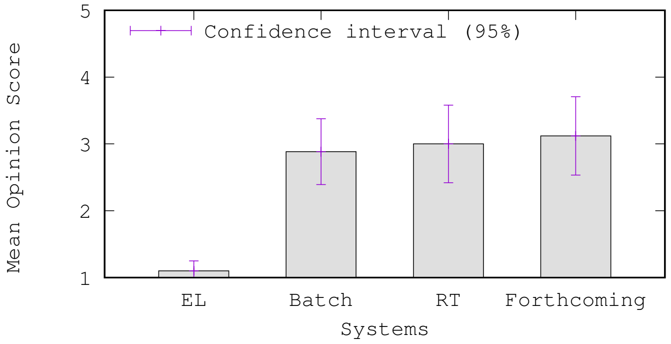


Figure 35. Naturalness of Predicted F_0 patterns.

prediction. Moreover we have also proposed two methods to address the latency issues caused by whole process of our proposed speaking aids system: segmented continuous F_0 patterns modeling and forthcoming F_0 modeling. In additionally, we have also designed the simulation experiment of our proposed speaking aids system to alleviate several construction costs, such as recording to evaluate new proposal. Through implementing a prototype system and its simulation, we have demonstrated that 1) our proposed system is capable of improving the naturalness of EL speech, 2) our ears cannot find smaller misalignment, less than 100 ms, between articulation and F_0 patterns, 3) segmented CF_0 pattern modeling without the forthcoming F_0 prediction is effective in the condition with restricted PC resources such as low speech of CPU clock and small size of memory, and 4) CF_0 pattern modeling with the forthcoming F_0 prediction is effective in the condition without restriction for PC resources.

Chapter

5

Statistical F_0 pattern prediction considering generative process of F_0 pattern

5.1 Introduction

The F_0 patterns predicted using statistical F_0 pattern prediction methods still sounded unnatural compared with that in normal speech. This was because the predicted F_0 patterns were not necessarily guaranteed to satisfy the physical constraint of the actual control mechanism of the thyroid cartilage, even though they were optimal in a statistical sense. In this regard, these methods still had a plenty of room for improvement. One possible solution to improve the naturalness of the F_0 patterns of the converted speech would be to incorporate a generative model of voice F_0 patterns into the statistical F_0 prediction model to take account of the physical mechanism of vocal phonation.

A statistical model of voice F_0 patterns [33, 34, 35], formulated by constructing a stochastic counterpart of the Fujisaki model [32], has been proposed. The Fujisaki model [32], which is a well-founded mathematical model representing the control mechanism of vocal fold vibration, assumes that an F_0 pattern on a logarithmic scale is the superposition of a phrase component, an accent component and a base value. The phrase and accent components are considered to be associated with mutually independent types of movement of the thyroid cartilage with different degrees of freedom and muscular reaction times. The model proposed in [33, 34, 35] has made it possible to estimate the underlying parameters of the Fujisaki model that best explain the given F_0 pattern, by using powerful statistical inference techniques.

In this chapter, we propose a Product-of-Experts (PoE) model [36] combining the above-mentioned two models to incorporate the generative F_0 pattern model into the statistical F_0 prediction framework. Since the PoE model is obtained by multiplying the densities of different models, it usually becomes complicated due to the renormalization term. To avoid this, we introduce a latent trajectory model proposed in [72] to reformulate the prediction model so that it can be smoothly combined with the generative F_0 pattern model.

The rest of this chapter is organized as follows.In **Section 5.2**, we review the generative model of F_0 patterns. In **Section 5.3**, we describe the strategy of integration of the conventional statistical F_0 prediction model and the generative model of F_0 patterns. In **Section 5.4**, we derive the *p.d.f.* of our integration model to train and predict F_0 patterns. In **Section 5.5**, we evaluate the perfor-

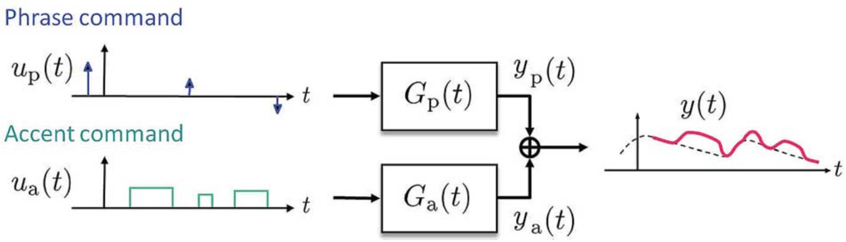


Figure 36. Original Fujisaki model [32].

mance of our integration model.

5.2 Review on Fujisaki model and its stochastic model

5.2.1 Fujisaki model

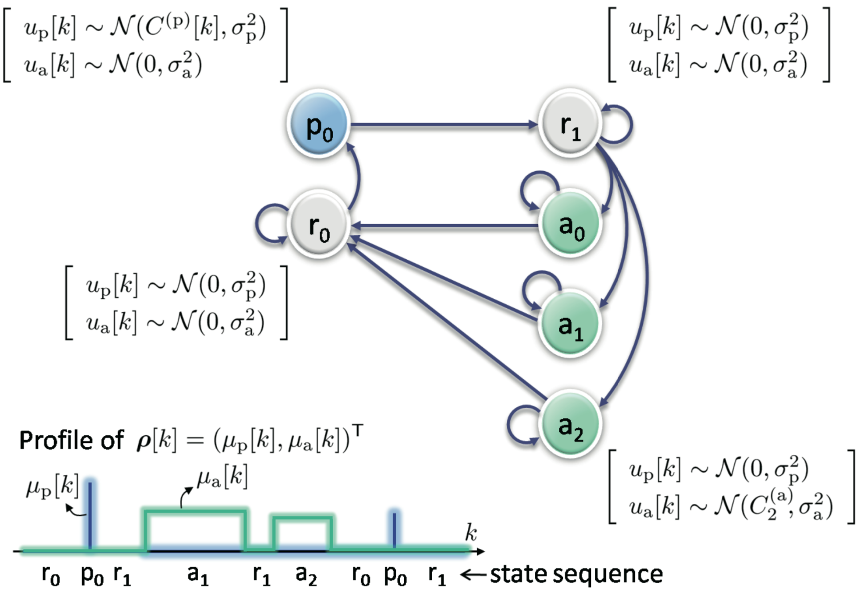
The Fujisaki model, as shown in Fig. 36, assumes that a log-scaled F_0 pattern $y(t)$ is the superposition of a phrase component $y_{\mathrm{p}}(t)$, an accent component $y_{\mathrm{a}}(t)$ and a base value μ_{b}. The phrase and accent components are assumed to be the outputs of different second-order critically damped filters, excited with Dirac deltas $u_{\mathrm{p}}(t)$ (phrase commands) and rectangular pulses $u_{\mathrm{a}}(t)$ (accent commands), respectively. Here, it must be noted that the phrase and accent commands do not usually overlap each other. The base value is a constant value related to the lower bound of the speaker's F_0, below which no regular vocal fold vibration can be maintained. The log F_0 pattern, $y(t)$, is thus expressed as

$$y(t) = y_{\mathrm{p}}(t) + y_{\mathrm{a}}(t) + \mu_b, \tag{29}$$

where

$$\begin{aligned} y_{\mathrm{p}}(t) &= g_{\mathrm{p}}(t) * u_{\mathrm{p}}(t), & (30)\\ y_{\mathrm{a}}(t) &= g_{\mathrm{a}}(t) * u_{\mathrm{a}}(t). & (31) \end{aligned}$$

Here, $*$ denotes convolution over time. $g_{\mathrm{p}}(t)$ and $g_{\mathrm{a}}(t)$ are the impulse responses of the two second-order systems, which are known to be almost constant within an utterance as well as across utterances for a particular speaker.



$u_{\rm p}[k] \sim \mathcal{N}(C^{(\rm p)}[k], \sigma_{\rm p}^2)$
$u_{\rm a}[k] \sim \mathcal{N}(0, \sigma_{\rm a}^2)$
$u_{\rm p}[k] \sim \mathcal{N}(0, \sigma_{\rm p}^2)$
$u_{\rm a}[k] \sim \mathcal{N}(0, \sigma_{\rm a}^2)$
p0
r1
r0
a0
a1
a2
$u_{\rm p}[k] \sim \mathcal{N}(0, \sigma_{\rm p}^2)$
$u_{\rm a}[k] \sim \mathcal{N}(0, \sigma_{\rm a}^2)$
Profile of $\rho[k] = (\mu_{\rm p}[k], \mu_{\rm a}[k])^{\sf T}$
$\mu_{\rm p}[k]$
$\mu_{\rm a}[k]$
k
$u_{\rm p}[k] \sim \mathcal{N}(0, \sigma_{\rm p}^2)$
$u_{\rm a}[k] \sim \mathcal{N}(C_2^{(\rm a)}, \sigma_{\rm a}^2)$
r0 p0 r1 a1 r1 a2 r0 p0 r1 ← state sequence

Figure 37. Command function modeling with HMM.

5.2.2 Stochastic model of generative process of F_0 pattern

The generative model of F_0 patterns proposed in [33, 34, 35] is a stochastic counterpart of a discrete-time version of the Fujisaki model [32].

A key idea of the model proposed in [33, 34, 35] is that the sequence of the phrase and accent command pair (i.e., the underlying parameters of the Fujisaki model) is modeled as a path-restricted hidden Markov model (HMM) with Gaussian emission densities, as shown in Fig. 37, so that estimating the state transition of the HMM directly amounts to estimating the Fujisaki-model parameters.

We hereafter use k to indicate the discrete time index. Given a state sequence $\boldsymbol{s} = (s_1, \ldots, s_K)$ of the above HMM, the conditional distributions of the phrase command sequence $\boldsymbol{u}_{\rm p} = (u_{\rm p}[1], \ldots, u_{\rm p}[K])^{\sf T}$ and the accent command sequence $\boldsymbol{u}_{\rm a} = (u_{\rm a}[1], \ldots, u_{\rm a}[K])^{\sf T}$ are given as

$$p(\boldsymbol{u}_{\rm p}|\boldsymbol{s}, \boldsymbol{\lambda}_F) = \mathcal{N}(\boldsymbol{u}_{\rm p}; \boldsymbol{\mu}_{\rm p}, \boldsymbol{\Sigma}_{\rm p}), \quad (32)$$

$$p(\boldsymbol{u}_{\rm a}|\boldsymbol{s}, \boldsymbol{\lambda}_F) = \mathcal{N}(\boldsymbol{u}_{\rm a}; \boldsymbol{\mu}_{\rm a}, \boldsymbol{\Sigma}_{\rm a}), \quad (33)$$

respectively, where $\boldsymbol{\lambda}_F$ denotes the parameters of the HMM. $\boldsymbol{\mu}_{\rm p}$ and $\boldsymbol{\mu}_{\rm a}$ denote the mean sequences of the state emission densities and $\boldsymbol{\Sigma}_{\rm p}$ and $\boldsymbol{\Sigma}_{\rm a}$ are diagonal matrices whose diagonal elements correspond to the variances of the state emission densities. From Eqs. (30) and (31), the relationships between $\boldsymbol{y}_{\mathrm{p}} = (y_{\mathrm{p}}[1], \ldots, y_{\mathrm{p}}[K])^{\top}$ and $\boldsymbol{u}_{\mathrm{p}}$ and between $\boldsymbol{y}_{\mathrm{a}} = (y_{\mathrm{a}}[1], \ldots, y_{\mathrm{a}}[K])^{\top}$ and $\boldsymbol{u}_{\mathrm{a}}$ can be written as

$$\begin{aligned} \boldsymbol{G}_{\mathrm{p}}\boldsymbol{u}_{\mathrm{p}} &= \boldsymbol{y}_{\mathrm{p}}, & (34) \\ \boldsymbol{G}_{\mathrm{a}}\boldsymbol{u}_{\mathrm{a}} &= \boldsymbol{y}_{\mathrm{a}}, & (35) \end{aligned}$$

where $\boldsymbol{G}_{\mathrm{p}}$ and $\boldsymbol{G}_{\mathrm{a}}$ are Toeplitz matrices where each row is a shifted copy of the convolution kernels $g_{\mathrm{p}}[1], \ldots, g_{\mathrm{p}}[K]$ and $g_{\mathrm{a}}[1], \ldots, g_{\mathrm{a}}[K]$. By using $\boldsymbol{u}_{\mathrm{b}}$ to denote the baseline component, the log F_0 sequence $\boldsymbol{y}$ is given as $\boldsymbol{y} = \boldsymbol{y}_{\mathrm{p}} + \boldsymbol{y}_{\mathrm{a}} + \boldsymbol{u}_{\mathrm{b}} + \boldsymbol{n}$ where $\boldsymbol{n}$ is an additive noise component corresponding to micro prosody. If we assume that $\boldsymbol{n}$ follows a Gaussian distribution with mean $\mathbf{0}$ and covariance $\boldsymbol{\Gamma}$, the conditional distribution of $\boldsymbol{y}$ given $\boldsymbol{u} = (\boldsymbol{u}_{\mathrm{p}}^{\top}, \boldsymbol{u}_{\mathrm{a}}^{\top}, \boldsymbol{u}_{\mathrm{b}}^{\top})^{\top}$ is defined as

$$p(\boldsymbol{y}|\boldsymbol{u}) = \mathcal{N}(\boldsymbol{y}; \boldsymbol{G}_{\mathrm{p}}\boldsymbol{u}_{\mathrm{p}} + \boldsymbol{G}_{\mathrm{a}}\boldsymbol{u}_{\mathrm{a}} + \boldsymbol{u}_{\mathrm{b}}, \boldsymbol{\Gamma}). \tag{36}$$

We further assume that $\boldsymbol{u}_{\mathrm{b}}$ follows a Gaussian distribution with mean $\boldsymbol{\mu}_{\mathrm{b}} = [\mu_{\mathrm{b}}, \cdots, \mu_{\mathrm{b}}]^{\top}$ and covariance $\boldsymbol{\Sigma}_{\mathrm{b}}$. Then, from Eq. (32), Eq. (33) and Eq. (36), the conditional distribution of $\boldsymbol{y}$ given $\boldsymbol{s}$ is given as

$$\begin{aligned} p(\boldsymbol{y}|\boldsymbol{s}, \boldsymbol{\Lambda}_F) &= \int p(\boldsymbol{y}|\boldsymbol{u}) p(\boldsymbol{u}|\boldsymbol{s}, \boldsymbol{\lambda}_F) \mathrm{d}\boldsymbol{u} \\ &= \mathcal{N}(\boldsymbol{y}; \boldsymbol{\mu}_F, \boldsymbol{\Sigma}_F), & (37) \end{aligned}$$

where $\boldsymbol{\mu}_F = \boldsymbol{G}_{\mathrm{p}}\boldsymbol{\mu}_{\mathrm{p}} + \boldsymbol{G}_{\mathrm{a}}\boldsymbol{\mu}_{\mathrm{a}} + \boldsymbol{\mu}_{\mathrm{b}}$ and $\boldsymbol{\Sigma}_F = \boldsymbol{G}_{\mathrm{p}}\boldsymbol{\Sigma}_{\mathrm{p}}\boldsymbol{G}_{\mathrm{p}}^{\top} + \boldsymbol{G}_{\mathrm{a}}\boldsymbol{\Sigma}_{\mathrm{a}}\boldsymbol{G}_{\mathrm{a}}^{\top} + \boldsymbol{\Sigma}_{\mathrm{b}} + \boldsymbol{\Gamma}$.

5.3 Incorporate Fujisaki model into statistical F_0 pattern prediction

5.3.1 Review on Product-of-experts

Product-of-experts [36] is related to a mixture model, where several probability distributions are combined via an ”or” operation, which is a weighted sum of their density functions. A well-known example of mixture approach is a mixture of Gaussians in which each simple model is a Gaussian, and the combination rule consists of taking a weighted arithmetic mean of the individual distributions.

This is equivalent to assuming an overall generative model in which each data vector is generated by first choosing one of the individual generative models and then allowing that individual model to generate the data vector. Combining models by forming a mixture is attractive for several reasons. It is easy to fit mixtures of tractable models to data using expectation-maximization (EM) or gradient ascent, and mixtures are usually considerably more powerful than their individual components. However, mixture models are very inefficient in high-dimensional spaces.

Product-of-experts is also models a probability distribution by combining the output from several simpler distributions, but a very different way of combining distributions is to multiply them together and re-normalize. The core idea is to combine several probability distributions (”experts”) by multiplying their density functions—making the Product-of-experts classification similar to an ”and” operation. Therefore, Product-of-experts can produce much sharper distributions than the individual expert models. This allows each expert to make decisions on the basis of a few dimensions without having to cover the full dimensionality of a problem. By contrast, training a Product-of-experts by maximizing the likelihood of the data is difficult because it is hard even to approximate the derivatives of the renormalization term in the combination rule. To avoid this issue, in generally speaking, a Product-of-experts can be trained using a different objective function called ”contrastive divergence (CD)” whose derivatives with regard to the parameters can be approximated accurately and efficiently.

5.3.2 Strategy

Product-of-experts [36] is a general technique to model a complicated distribution of data by combining relatively simpler distributions (experts). Since the distribution is obtained by multiplying the densities of the experts, the way the experts are combined is somewhat similar to an “and” operation. In this section, we construct a Product-of-experts model by treating the two models introduced in the previous sections as the experts.

Training a Product-of-experts model by maximizing the likelihood of the data usually becomes difficult since it is hard even to approximate the derivatives of the renormalization term. By contrast, we propose an elegant formulation that allows

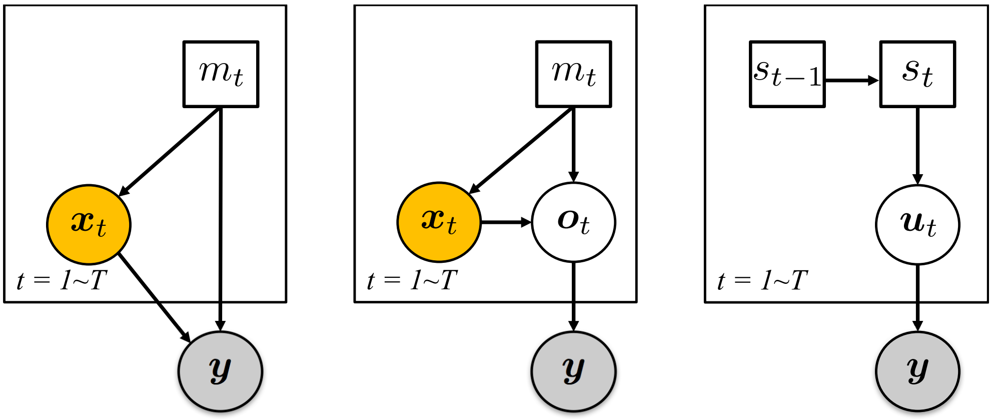


Figure 38. Graphical model of GMM-based statistical F_0 prediction model in **Section 2.6** (left), GMM-based statistical F_0 prediction model reformulated by employing latent trajectory in **Section 5.3.3** (middle), and stochastic model of generative process of F0 pattern in **Section 5.2.2** (right).

the use of the EM algorithm for both parameter training and F_0 prediction. To do so, we first introduce a latent trajectory model proposed in [72] to reformulate the GMM-based statistical F_0 prediction model, which plays a key role in making this possible.

5.3.3 Reformulate by employing latent trajectory

As mentioned in **Section 2.6**, the conventional GMM-based statistical F_0 prediction method adopts the trajectory conversion algorithm (see Eq. (17)). Therefore, if we take a conditional distribution with respect to $\boldsymbol{y}$, the normalization term to calculate the probability is terribly complicated because of the necessity to consider all combinations $\Sigma_{\boldsymbol{m}}$.

To avoid this difficulty, we reformulate the GMM-based statistical F_0 prediction model presented in **Section 2.6** by employing the idea proposed in [72]. Instead of treating $\boldsymbol{o}$ as a function of $\boldsymbol{y}$, we treat $\boldsymbol{o}$ as a latent variable to be marginalized out, that is related to $\boldsymbol{y}$ through a soft constraint $\boldsymbol{o} \simeq \boldsymbol{W}\boldsymbol{y}$. The relationship $\boldsymbol{o} \simeq \boldsymbol{W}\boldsymbol{y}$ can be expressed through the conditional distribution $p(\boldsymbol{y}|\boldsymbol{o})$

$$p(\boldsymbol{y}|\boldsymbol{o}) \propto \exp\left\{-\tfrac{1}{2}(\boldsymbol{W}\boldsymbol{y}-\boldsymbol{o})^\top \boldsymbol{\Lambda}(\boldsymbol{W}\boldsymbol{y}-\boldsymbol{o})\right\} \tag{38}$$

$$= \mathcal{N}(\boldsymbol{y}; \boldsymbol{H}\boldsymbol{o}, \boldsymbol{V}), \tag{39}$$

$$\boldsymbol{H} = (\boldsymbol{W}^\top \boldsymbol{\Lambda} \boldsymbol{W})^{-1} \boldsymbol{W}^\top \boldsymbol{\Lambda}, \tag{40}$$

$$\boldsymbol{V} = (\boldsymbol{W}^\top \boldsymbol{\Lambda} \boldsymbol{W})^{-1}, \tag{41}$$

where $\boldsymbol{\Lambda}$ is a constant positive definite matrix that can be set arbitrarily. As with **Section 2.6**, the joint distribution $p(\boldsymbol{x}, \boldsymbol{o}|\boldsymbol{\lambda}_G)$ is modeled as a GMM. Namely, given mixture indices $\boldsymbol{m}$, the conditional distribution $p(\boldsymbol{x}, \boldsymbol{o}|\boldsymbol{m}, \boldsymbol{\Lambda}_G)$ is defined as a Gaussian distribution. Thus, the joint distribution of $\boldsymbol{y}$, $\boldsymbol{x}$, $\boldsymbol{o}$, and $\boldsymbol{m}$ can be described using the distributions defined above

$$p(\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{o}, \boldsymbol{m}|\boldsymbol{\lambda}_G) = p(\boldsymbol{y}|\boldsymbol{o})p(\boldsymbol{x}, \boldsymbol{o}|\boldsymbol{m}, \boldsymbol{\lambda}_G)p(\boldsymbol{m}|\boldsymbol{\lambda}_G), \tag{42}$$

where $p(\boldsymbol{m}|\boldsymbol{\lambda}_G)$ is the product of mixture weights of the GMM (see Fig. 38). By marginalizing $\boldsymbol{o}$ and $\boldsymbol{m}$ out, we can readily obtain the joint distribution $p(\boldsymbol{y}, \boldsymbol{x}|\boldsymbol{\lambda}_G)$, which can be used as a criterion to train $\boldsymbol{\lambda}_G$ and predict optimal $\boldsymbol{y}$ in a consistent manner, unlike the method presented in **Section 2.6**. We use this model to construct our Product-of-experts model in the next subsection.

5.3.4 Design a probability density function

In the same way as Eq. (42), we write the model presented in **Section 5.2.2** in the form of a joint distribution

$$p(\boldsymbol{y}, \boldsymbol{u}, \boldsymbol{s}|\boldsymbol{\lambda}_F) = p(\boldsymbol{y}|\boldsymbol{u})p(\boldsymbol{u}|\boldsymbol{s}, \boldsymbol{\lambda}_F)p(\boldsymbol{s}|\boldsymbol{\lambda}_F), \tag{43}$$

where $p(\boldsymbol{u}|\boldsymbol{s}, \boldsymbol{\lambda}_F)$ is given as the product of state emission densities and $p(\boldsymbol{s}|\boldsymbol{\lambda}_F)$ the product of the state transition probabilities given a state sequence $\boldsymbol{s}$ (see Fig. 38). We consider constructing a Product-of-experts model by combining Eq. (42) and Eq. (43) followed by marginalization, rather than by simply combining the marginal distributions $p(\boldsymbol{y}, \boldsymbol{x}|\boldsymbol{\lambda}_G)$ and $p(\boldsymbol{y}|\boldsymbol{\lambda}_F)$, which makes the parameter training and F_0 prediction problems excessively hard. To do so, we first combine the densities of $p(\boldsymbol{y}|\boldsymbol{o})$ and $p(\boldsymbol{y}|\boldsymbol{u})$ to obtain $p(\boldsymbol{y}|\boldsymbol{o},\boldsymbol{u})$ (see Fig. 39). Since both of these distributions are Gaussians, the product of their distributions can be easily obtained by completing the square of the exponent

$$\begin{aligned} p(\boldsymbol{y}|\boldsymbol{o},\boldsymbol{u}) &\propto \mathcal{N}(\boldsymbol{y};\boldsymbol{H}\boldsymbol{o},\boldsymbol{V})\cdot\mathcal{N}(\boldsymbol{y};\boldsymbol{G}\boldsymbol{u},\boldsymbol{\Gamma}) \\ &= \mathcal{N}(\boldsymbol{y};\boldsymbol{\mu}_{y|o,u},\boldsymbol{\Sigma}_{y|o,u}), & (44) \\ \boldsymbol{\mu}_{y|o,u} &= (\boldsymbol{V}^{-1}+\boldsymbol{\Gamma}^{-1})^{-1}(\boldsymbol{V}^{-1}\boldsymbol{H}\boldsymbol{o}+\boldsymbol{\Gamma}^{-1}\boldsymbol{G}\boldsymbol{u}), & (45) \\ \boldsymbol{\Sigma}_{y|o,u} &= (\boldsymbol{V}^{-1}+\boldsymbol{\Gamma}^{-1})^{-1}, & (46) \end{aligned}$$

where $\boldsymbol{G}=[\boldsymbol{G}_{\mathrm{p}}\boldsymbol{G}_{\mathrm{a}}\boldsymbol{I}]$ and $\boldsymbol{u}=(\boldsymbol{u}_{\mathrm{p}}^{\top},\boldsymbol{u}_{\mathrm{a}}^{\top},\boldsymbol{u}_{\mathrm{b}}^{\top})^{\top}$. From Eq. (42), Eq. (43) and Eq. (44), the joint distribution of $\boldsymbol{y}$, $\boldsymbol{x}$, $\boldsymbol{o}$, $\boldsymbol{u}$, $\boldsymbol{m}$ and $\boldsymbol{s}$ can be constructed as

$$\begin{aligned} p\quad &(\boldsymbol{y},\boldsymbol{x},\boldsymbol{o},\boldsymbol{u},\boldsymbol{m},\boldsymbol{s}|\boldsymbol{\lambda}_G,\boldsymbol{\lambda}_F) & (47) \\ = \quad &\underbrace{p(\boldsymbol{y}|\boldsymbol{o},\boldsymbol{u})p(\boldsymbol{x},\boldsymbol{o}|\boldsymbol{m},\boldsymbol{\lambda}_G)p(\boldsymbol{u}|\boldsymbol{s},\boldsymbol{\lambda}_F)}_{p(\boldsymbol{y},\boldsymbol{x},\boldsymbol{o},\boldsymbol{u}|\boldsymbol{m},\boldsymbol{s},\boldsymbol{\lambda}_G,\boldsymbol{\lambda}_F)}p(\boldsymbol{m}|\boldsymbol{\lambda}_G)p(\boldsymbol{s}|\boldsymbol{\lambda}_F). \end{aligned}$$

This can be used as the complete data likelihood for parameter training and F_0 prediction as explained later. By marginalizing $\boldsymbol{o}$ and $\boldsymbol{u}$ out, we can readily obtain the joint distribution $p(\boldsymbol{y},\boldsymbol{x},\boldsymbol{m},\boldsymbol{s}|\boldsymbol{\lambda}_G,\boldsymbol{\lambda}_F)$, which can be used as a criterion to train $\boldsymbol{\lambda}_G$ and $\boldsymbol{\lambda}_F$ and predict optimal $\boldsymbol{y}$ in a consistent manner.

Since both $p(\boldsymbol{x},\boldsymbol{o}|\boldsymbol{m},\boldsymbol{\lambda}_G)$ and $p(\boldsymbol{u}|\boldsymbol{s},\boldsymbol{\lambda}_F)$ are Gaussians, let us write them as

$$\begin{aligned} p(\boldsymbol{x},\boldsymbol{o}|\boldsymbol{m},\boldsymbol{\lambda}_G) &= \mathcal{N}\left(\begin{bmatrix}\boldsymbol{x}\\ \boldsymbol{o}\end{bmatrix};\begin{bmatrix}\boldsymbol{\mu}_x\\ \boldsymbol{\mu}_o\end{bmatrix},\begin{bmatrix}\boldsymbol{P}_{xx} & \boldsymbol{P}_{xo}\\ \boldsymbol{P}_{ox} & \boldsymbol{P}_{oo}\end{bmatrix}^{-1}\right), & (48) \\ p(\boldsymbol{u}|\boldsymbol{s},\boldsymbol{\lambda}_F) &= \mathcal{N}(\boldsymbol{u};\boldsymbol{\mu}_u,\boldsymbol{P}_u^{-1}). & (49) \end{aligned}$$

Then, from Eq. (44), Eq. (48) and Eq. (49), it can be shown that $p(\boldsymbol{y},\boldsymbol{x},\boldsymbol{o},\boldsymbol{u}|\boldsymbol{m},\boldsymbol{s},\boldsymbol{\lambda}_G,\boldsymbol{\lambda}_F)$ is given as

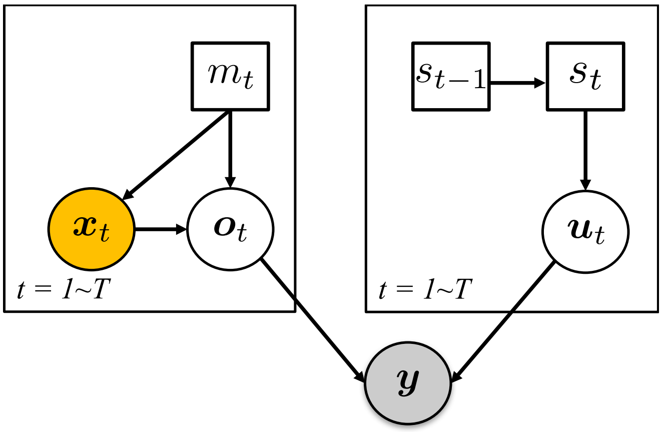


Figure 39. Graphical model of proposed method.

$$\begin{aligned} p & \quad (\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{o}, \boldsymbol{u} | \boldsymbol{m}, \boldsymbol{s}, \boldsymbol{\lambda}_G, \boldsymbol{\lambda}_F) \\ = & \ \mathcal{N}\left(\left[\begin{array}{c} \boldsymbol{y} \\ \boldsymbol{x} \\ \hline \boldsymbol{o} \\ \boldsymbol{u} \end{array} \right]; \left[\begin{array}{c} \boldsymbol{A}_{11}\boldsymbol{b}_1 + \boldsymbol{A}_{12}\boldsymbol{b}_2 \\ \hline \boldsymbol{A}_{21}\boldsymbol{b}_1 + \boldsymbol{A}_{22}\boldsymbol{b}_2 \end{array} \right], \left[\begin{array}{c|c} \boldsymbol{A}_{11} & \boldsymbol{A}_{12} \\ \hline \boldsymbol{A}_{21} & \boldsymbol{A}_{22} \end{array} \right] \right), \end{aligned} \tag{50}$$

where

$$\begin{bmatrix} \boldsymbol{A}_{11} & \boldsymbol{A}_{12} \\ \boldsymbol{A}_{21} & \boldsymbol{A}_{22} \end{bmatrix} = \left[\begin{array}{cc|cc} \boldsymbol{\Sigma}_{y|o,u}^{-1} & \boldsymbol{O} & -\boldsymbol{V}^{-1}\boldsymbol{H} & -\boldsymbol{\Gamma}^{-1}\boldsymbol{G} \\ \boldsymbol{O} & \boldsymbol{P}_{xx} & \boldsymbol{P}_{xo} & \boldsymbol{O} \\ \hline -\boldsymbol{H}^\top \boldsymbol{V}^{-\top} & \boldsymbol{P}_{ox} & \boldsymbol{B}_{11} & \boldsymbol{B}_{12} \\ -\boldsymbol{G}^\top \boldsymbol{\Gamma}^{-\top} & \boldsymbol{O} & \boldsymbol{B}_{21} & \boldsymbol{B}_{22} \end{array} \right]^{-1}, \tag{51}$$

$$\begin{bmatrix} \boldsymbol{B}_{11} & \boldsymbol{B}_{12} \\ \boldsymbol{B}_{21} & \boldsymbol{B}_{22} \end{bmatrix} = \begin{bmatrix} \boldsymbol{P}_{oo} + \boldsymbol{H}^\top \boldsymbol{V}^{-\top} \boldsymbol{\Sigma}_{y|o,u} \boldsymbol{V}^{-1} \boldsymbol{H} & \boldsymbol{H}^\top \boldsymbol{V}^{-\top} \boldsymbol{\Sigma}_{y|o,u} \boldsymbol{\Gamma}^{-1} \boldsymbol{G} \\ \boldsymbol{G}^\top \boldsymbol{\Gamma}^{-\top} \boldsymbol{\Sigma}_{y|o,u} \boldsymbol{V}^{-1} \boldsymbol{H} & \boldsymbol{P}_u + \boldsymbol{G}^\top \boldsymbol{\Gamma}^{-\top} \boldsymbol{\Sigma}_{y|o,u} \boldsymbol{\Gamma}^{-1} \boldsymbol{G} \end{bmatrix},$$

$$\begin{bmatrix} \boldsymbol{b}_1 \\ \boldsymbol{b}_2 \end{bmatrix} = \left[\begin{array}{c} \boldsymbol{O} \\ \boldsymbol{P}_{xx}\boldsymbol{\mu}_x + \boldsymbol{P}_{xo}\boldsymbol{\mu}_o \\ \hline \boldsymbol{P}_{oo}\boldsymbol{\mu}_o + \boldsymbol{P}_{ox}\boldsymbol{\mu}_x \\ \boldsymbol{P}_u \boldsymbol{\mu}_u \end{array} \right], \tag{52}$$

by completing the square of the exponent.

5.4 Parameter training and F_0 pattern prediction

The problems of parameter training and F_0 prediction can be formulated as the following optimization problems:

$$\begin{aligned}
\{\hat{\boldsymbol{\lambda}}_G, \hat{\boldsymbol{\lambda}}_F, \hat{\boldsymbol{m}}, \hat{\boldsymbol{s}}\} &= \underset{\boldsymbol{\lambda}_G, \boldsymbol{\lambda}_F, \boldsymbol{m}, \boldsymbol{s}}{\text{argmax}} \log p(\tilde{\boldsymbol{y}}, \tilde{\boldsymbol{x}}, \boldsymbol{m}, \boldsymbol{s} | \boldsymbol{\lambda}_G, \boldsymbol{\lambda}_F), && (53)\\
\{\hat{\boldsymbol{y}}, \hat{\boldsymbol{m}}, \hat{\boldsymbol{s}}\} &= \underset{\boldsymbol{y}, \boldsymbol{m}, \boldsymbol{s}}{\text{argmax}} \log p(\boldsymbol{y}, \tilde{\boldsymbol{x}}, \boldsymbol{m}, \boldsymbol{s} | \hat{\boldsymbol{\lambda}}_G, \hat{\boldsymbol{\lambda}}_F). && (54)
\end{aligned}$$

where $\tilde{\boldsymbol{y}}$ and $\tilde{\boldsymbol{x}}$ denote the observed F_0 pattern extracted from normal speech and the observed spectral sequence extracted from non-larynx speech. Both of these problems can be solved using the EM algorithm by treating $\boldsymbol{o}$ and $\boldsymbol{u}$ as latent variables. Owing to space limitations, here we only derive an algorithm for solving Eq. (54).

The likelihood of $\boldsymbol{y}$, $\boldsymbol{m}$ and $\boldsymbol{s}$ given the complete data $\{\tilde{\boldsymbol{x}}, \boldsymbol{o}, \boldsymbol{u}\}$ is given by Eq. (47). By taking the conditional expectation of $\log p(\boldsymbol{y}, \tilde{\boldsymbol{x}}, \boldsymbol{o}, \boldsymbol{u} | \boldsymbol{m}, \boldsymbol{s}, \hat{\boldsymbol{\lambda}}_G, \hat{\boldsymbol{\lambda}}_F)$ with respect to $\boldsymbol{o}$ and $\boldsymbol{u}$ given $\tilde{\boldsymbol{x}}$, $\boldsymbol{y} = \boldsymbol{y}'$, $\boldsymbol{m} = \boldsymbol{m}'$ and $\boldsymbol{s} = \boldsymbol{s}'$ and then adding $\log p(\boldsymbol{m}|\hat{\boldsymbol{\lambda}}_G) p(\boldsymbol{s}|\hat{\boldsymbol{\lambda}}_F)$, we obtain an auxiliary function

$$\begin{aligned}
Q(\theta, \theta') &= \mathbb{E}_{\boldsymbol{o},\boldsymbol{u}|\tilde{\boldsymbol{x}},\boldsymbol{y}',\boldsymbol{m}',\boldsymbol{s}'}\left[\log p(\boldsymbol{y}, \tilde{\boldsymbol{x}}, \boldsymbol{o}, \boldsymbol{u} | \boldsymbol{m}, \boldsymbol{s}, \hat{\boldsymbol{\lambda}}_G, \hat{\boldsymbol{\lambda}}_F)\right]\\
&+ \log p(\boldsymbol{m}|\hat{\boldsymbol{\lambda}}_G) + \log p(\boldsymbol{s}|\hat{\boldsymbol{\lambda}}_F), && (55)
\end{aligned}$$

where

$$\theta = \{\boldsymbol{y}, \boldsymbol{m}, \boldsymbol{s}\}. \qquad (56)$$

From Eq. (50), we obtain

$$\begin{aligned}
\mathbb{E}\left[\begin{bmatrix}\boldsymbol{o}\\ \boldsymbol{u}\end{bmatrix} \middle| \begin{bmatrix}\boldsymbol{y}'\\ \tilde{\boldsymbol{x}}\end{bmatrix}, \boldsymbol{m}', \boldsymbol{q}'\right] &= \boldsymbol{A}_{21}\boldsymbol{b}_1 + \boldsymbol{A}_{22}\boldsymbol{b}_2\\
&\quad + \boldsymbol{A}_{21}\boldsymbol{A}_{11}^{-1}\left(\begin{bmatrix}\boldsymbol{y}'\\ \tilde{\boldsymbol{x}}\end{bmatrix} - \boldsymbol{A}_{11}\boldsymbol{b}_1 - \boldsymbol{A}_{12}\boldsymbol{b}_2\right)\\
&=: \begin{bmatrix}\bar{\boldsymbol{o}}\\ \bar{\boldsymbol{u}}\end{bmatrix}, && (57)\\
\mathbb{E}\left[\begin{bmatrix}\boldsymbol{o}\\ \boldsymbol{u}\end{bmatrix}\begin{bmatrix}\boldsymbol{o}\\ \boldsymbol{u}\end{bmatrix}^{\top} \middle| \begin{bmatrix}\boldsymbol{y}'\\ \tilde{\boldsymbol{x}}\end{bmatrix}, \boldsymbol{m}', \boldsymbol{q}'\right] &= \boldsymbol{A}_{22} - \boldsymbol{A}_{21}\boldsymbol{A}_{11}^{-1}\boldsymbol{A}_{12} + \begin{bmatrix}\bar{\boldsymbol{o}}\\ \bar{\boldsymbol{u}}\end{bmatrix}\begin{bmatrix}\bar{\boldsymbol{o}}\\ \bar{\boldsymbol{u}}\end{bmatrix}^{\top} && (58)
\end{aligned}$$

which are the values to be computed at the "E-step" by substituting θ into θ'. At the "M-step", we compute

$$\{\boldsymbol{y}, \boldsymbol{m}, \boldsymbol{s}\} \leftarrow \underset{\boldsymbol{y},\boldsymbol{m},\boldsymbol{s}}{\operatorname{argmax}}\, Q(\theta, \theta'). \tag{59}$$

Given $\bar{\boldsymbol{o}}$ and $\bar{\boldsymbol{u}}$, $\boldsymbol{y}$ is obviously a mean vector of $p(\boldsymbol{y}|\boldsymbol{o}, \boldsymbol{u})$ by maximizing Eq. (44) with respect to $\boldsymbol{y}$. After that, we can update $\boldsymbol{m}$ and $\boldsymbol{s}$ by maximizing Eq. (42) and Eq. (43) with respect to $\boldsymbol{m}$ and $\boldsymbol{s}$, respectively,

$$\hat{m}_t = \underset{m}{\operatorname{argmax}} \left\{ \log p(\tilde{\boldsymbol{x}}, \boldsymbol{W}\boldsymbol{y}|\boldsymbol{m}, \hat{\boldsymbol{\lambda}}_G) + \log p(\boldsymbol{m}|\hat{\boldsymbol{\lambda}}_G) \right\}, \tag{60}$$

$$\hat{\boldsymbol{s}} = \underset{\boldsymbol{s}}{\operatorname{argmax}} \left\{ \log p(\boldsymbol{s}|\hat{\boldsymbol{\lambda}}_F) + \int p(\boldsymbol{u}|\boldsymbol{y}, \boldsymbol{s}_0, \hat{\boldsymbol{\lambda}}_F) \log p(\boldsymbol{u}|\boldsymbol{s}, \hat{\boldsymbol{\lambda}}_F) \mathrm{d}\boldsymbol{u} \right\}, \tag{61}$$

where $\boldsymbol{s}_0$ is an initial state sequence obtained by initial command sequence which was obtained with the method described in method [73].

5.5 Experimental setup and corpora

5.5.1 Experimental setup

We conducted objective and subjective evaluation experiments to evaluate the performance of the proposed method. For the objective evaluation, we evaluated the F_0 correlation coefficients between the predicted and target F_0 contours. We also subjectively evaluated the naturalness of the F_0 contour of converted speech.

The source speech was EL speech uttered by one male laryngectomee, and the target speech was normal speech uttered by a professional female speaker. Each speaker uttered about 50 sentences in the ATR phonetically balanced sentence set [66]. We conducted a 5-fold cross validation test in which 40 utterance pairs were used for training, and the remaining 10 utterance pairs were used for evaluation. The sampling frequency was set at 16 kHz. The settings of HMM in the generative F_0 contour model were the same as reported in [33, 34, 35].

For initialization of the mixture index sequence $\boldsymbol{m}$ and state sequence $\boldsymbol{s}$, we performed the conventional GMM-based and Fujisaki-model-based methods. Note that the Fujisaki-model-based method refers to a post-processing method that consists of first applying the GMM-based method and then fitting the Fujisaki model to the predicted F_0 contour using the method of [34, 35], which is similar to a post-processing method for HMM-based speech synthesis [74]. Note that

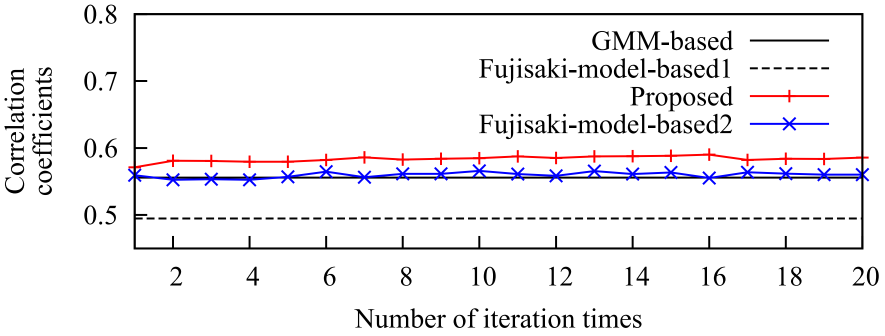


Figure 40. F_0 correlation coefficients between predicted F_0 patterns from EL speech and extracted F_0 patterns from normal speech.

in this experiment, we implemented a simplified and approximated version of the proposed method, in which the E-step procedure is replaced with the conventional Fujisaki-model-based and GMM-based methods. Therefore, the convergence of the algorithm implemented for the current evaluation is not strictly guaranteed. The speech used for evaluation were synthesized using STRAIGHT [54] given the mel-cepstrum sequence and F_0 contour. The methods selected for comparison were:

GMM-based : Predict F_0 contours with the GMM-based method.

Fujisaki-model-based1 : Fit the Fujisaki model to the predicted F_0 contours obtained with the **GMM-based** method.

Proposed : Predict F_0 contours with an approximated version of the proposed method, in which the E-step is replaced with the **GMM-based** and **Fujisaki-model-based2** methods.

Fujisaki-model-based2 : Fit the Fujisaki model to the predicted F_0 contours obtained with **Proposed**.

5.5.2 Objective evaluation

As Fig. 40 shows, **Proposed** obtained the highest prediction accuracy because of Eq. (44) meaning an "and" operation for Eq. (36) and Eq. (39). In additionally, we can also find the “and” operation from F_0 patterns samples in Fig. 41. Therefore, we found that it is effectiveness to construct a PoE model. Furthermore, since **Fujisaki-model-based2** has higher correlation coefficients than **Fujisaki-model-based1**, we found that the predicted F_0 contours by **Proposed** were given good influences by considering not only the GMM-based method but also the Fujisaki model. Note that we used the predicted F_0 contours obtained with the **GMM-based** method as the input for the **Fujisaki-model-based1** method in this experiment. To make a more fair comparison, it would be necessary to modify the Fujisaki-model-based method so as not to depend on the GMM-based method. In additional, as for proposed, there is no large difference of correlation coefficients in each iteration. To make exactly evaluation, we have to construct the PoE model not replaced E-step with the conventional methods.

5.5.3 Subjective evaluation

As Fig. 42 shows, **Proposed** outperformed the conventional methods, **GMM-based** and *Fujisaki-model-based1.* This result is reasonable since **Proposed** obtained the highest prediction accuracy as in Fig. 40. Therefore, it is demonstrated to construct a PoE model to make it possible to predict F_0 patterns, which is statistically likely and physically natural. Note that as shown in Eq. (57), the calculation of $\bar{\boldsymbol{o}}$ and $\bar{\boldsymbol{u}}$ requires quite high computational cost $O(N^3)$ where N denotes the number of frames over an utterance. Therefore, this proposed PoE model is effective for EL speech enhancement at acoustic level described in **Section 3** while it is difficult to apply applications using real-time processing such as **Section 4**.

5.6 Summary

In this chapter, to improve F_0 prediction performance in electrolaryngeal speech enhancement, we proposed a Product-of-Experts model that combined two conventional methods, a statistical F_0 prediction method and a statistical F_0 pattern modeling method based on its generative process. Experimental results revealed that the proposed method successfully outperformed our previously proposed method in terms of the naturalness of the predicted F_0 patterns.

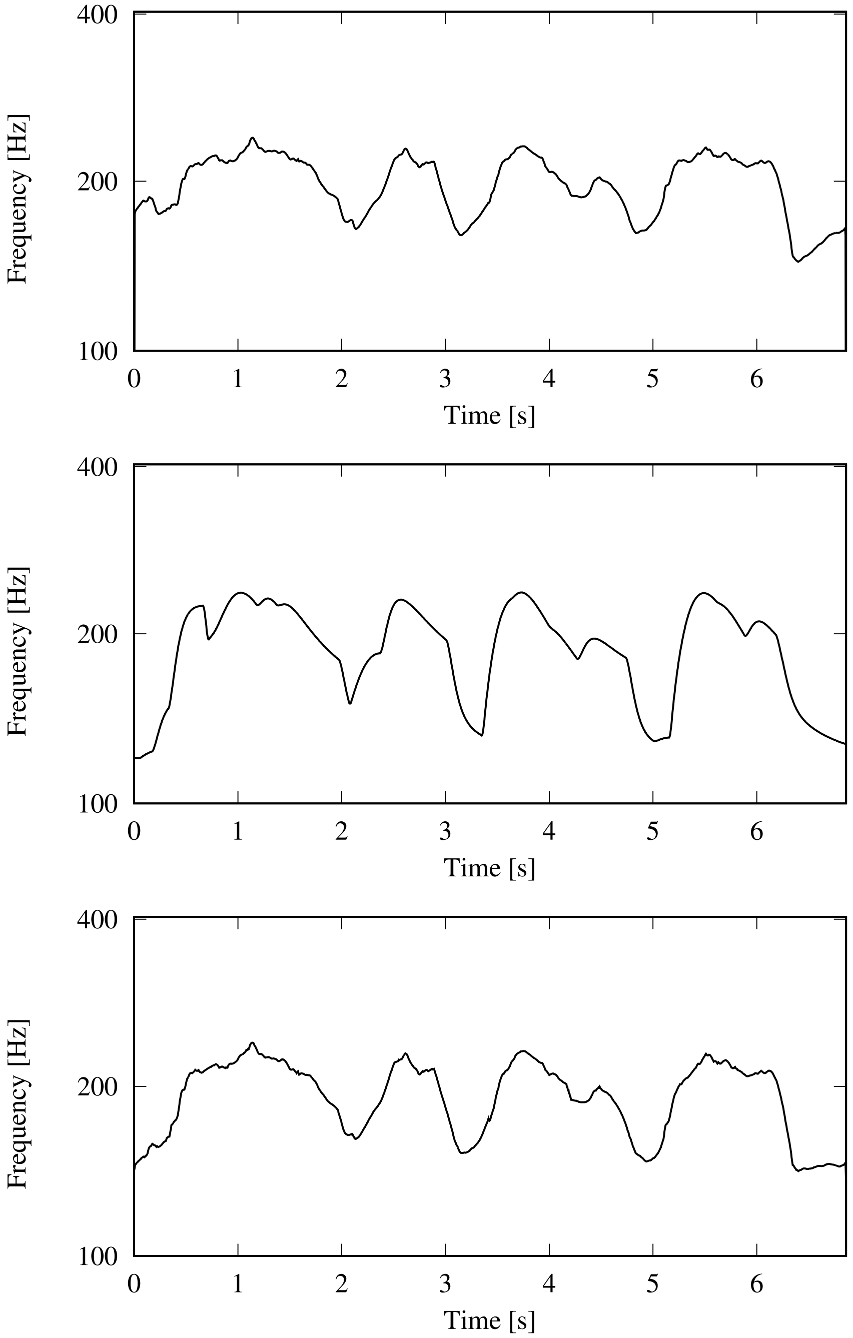


Figure 41. F_0 patterns predicted by **GMM-based** (top), **Fujisaki-model-based1** (middle), and **Proposed** (bottom).

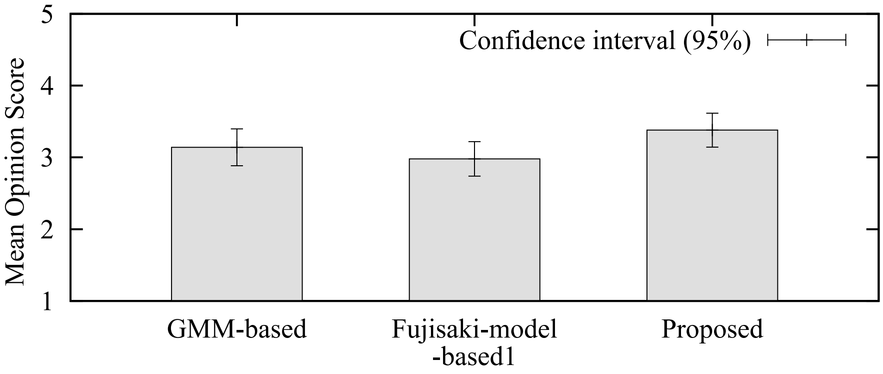


Figure 42. Result of opinion test on naturalness.

Chapter

6

Conclusion

6.1 Summary of dissertation

In this dissertation, toward the realization of a world in which laryngectomees can communicate like healthy people, we addressed two approaches, that is, systems at the physiological and acoustic levels, as statistical EL speech production methods satisfying the following three requirements: I) improved the naturalness of EL speech, Ⅱ) preservation of the high intelligibility of EL speech, and Ⅲ free laryngectomees from the requirement of having to newly learn how to generate the enhanced EL speech.

In chapter 2, we describe the problems of EL speech and the conventional studies addressing the problems of EL speech. EL speech is one type of alaryngeal speech and is produced using an electrolarynx, which allows laryngectomees to produce speech sounds. Although the intelligibility of EL speech is quite high, there are three main disadvantages of using the electrolarynx: 1) its sound is characterized as being mechanical and robotic because of the fundamental frequency (F_0) pattern of the monotonic excitation signals, 2) the excitation signals are emitted outside as noise because of the EL speech production mechanism, and 3) one hand is occupied. To address these issues, two types of speaking aid system for EL speech enhancement were proposed: A) a speaking aid system capable of modifications at the acoustic level, with a loudness speaker, through which the enhanced speech is presented to the listener, and B) a speaking aid system capable of modifications at the physiological level, whereby laryngectomees can directly produce the enhanced EL speech from their mouths.

Despite the variety of technologies that have been developed to address the above issues, the enhanced EL speech is still far from achieving a similar quality to natural human speech, in terms of both naturalness and intelligibility. Developing a method that can maintain a balance between obtaining the naturalness quality of speech while preserving the intelligibility of the speech content is one of the most important problems that must be overcome. Therefore, in chapter 3, we described the proposed EL speech enhancement system at the acoustic level, that is, the hybrid approach with a noise reduction method for enhancing spectral parameters and a statistical voice conversion method for predicting the excitation parameters. The conventional enhancement systems at the acoustic level satisfy either requirement (I) or (Ⅱ). Although using spectral subtraction for noise reduction retains the intelligibility of EL speech, the monotonic excitation signals of EL speech are unimproved. On the other hand, statistical voice conversion dramatically improve the naturalness of EL speech at the expense of the intelligibility of EL speech. Therefore, we propose a combination of these methods to satisfy both requirements (I) and (Ⅱ). Considering that the naturalness of the enhanced EL speech strongly depends on the prediction accuracy of F_0 patterns, we also propose the modification of trained F_0 patterns to improve the prediction accuracy. Moreover, we discuss the proposed method in comparison with the conventional method in term of naturalness, intelligibility, and listenability and demonstrate that the proposed method yields significant improvements in naturalness compared with EL speech while retaining high intelligibility.

Chapter 4 is on tha proposed EL speech enhancement system at the physiological level, that is, direct F_0 pattern control of the electrolarynx using real-time statistical F_0 pattern prediction to develop an EL speech enhancement technique that is also effective for face-to-face conversation. The conventional enhancement systems at the physiological level do not satisfy requirement (Ⅲ) but satisfy both requirements (I) and (Ⅱ). Therefore, the improvements in the naturalness of EL speech on using the conventional enhancement systems depend on the skill of the user. Focusing on the fact that the use of the statistical F_0 pattern prediction satisfies requirement (Ⅲ) by predicting F_0 patterns without manual operations, we proposed the use of the statistical F_0 pattern prediction to control the F_0 patterns of the excitation signals. To flexibly investigate the performance of our proposed control method, we also designed a simulation method of the EL speech production process using the controlled electrolarynx. Furthermore, we have described the negative impact of latency caused by real-time processing and propose methods to address latency issues. By implementing a prototype system and its simulation, we demonstrated that our proposed system is capable of successfully addressing the unnaturalness of the electrolarynx and the latency issues.

Chapter 5 is on improving the accuracy of statistical F_0 pattern prediction (related to requirement (I)). We proposed a statistical F_0 prediction method considering the generative process of F_0 patterns within the product-of-experts framework. We introduced the Fujisaki model, which is a well-founded mathematical model representing the control mechanism of vocal fold vibration, and also reviewed its stochastic model. To incorporate the stochastic model for the Fujisaki model into the conventional statistical F_0 pattern prediction, we introduced a latent trajectory model and reformulated the prediction model with the latent trajectory model. Using the constructed model, we derived algorithms for parameter training and F_0 prediction. Through experimental evaluations, we revealed that the proposed method successfully surpasses the conventional statistical F_0 pattern prediction.

6.2 Direction of future works

We successfully proposed methods satisfying all of requirements (I), (Ⅱ), and (Ⅲ). However, the EL speech still has three problems compared with normal speech: 1) the naturalness of the enhanced EL speech strongly depends on the prediction accuracy of the statistical F_0 pattern prediction, 2) the excitation signals are emitted outside as noise because of the EL speech production mechanism, and 3) one hand is occupied. Consequently, there are still many issues to be resolved in fully statistical EL speech productions.

6.2.1 Where is the limitation of statistical F_0 pattern prediction?

In this dissertation, we realized significant improvements in EL speech in terms of naturalness by using statistical F_0 pattern prediction. Therefore, the naturalness of the enhanced EL speech obviously depends on the prediction accuracy of the statistical F_0 pattern prediction. Even if we set the best experimental conditions in the prediction of F_0 patterns of normal speech, the prediction accuracy is still insufficient and the correlation coefficient between target F_0 patterns and predicted F_0 patterns is around 0.6. In the later paragraphs, we suggest methods that might improve the prediction accuracy: a-1) the modeling of a feature in not individual time frames, but suprasegmental units, a-2) the use of better models to predict F_0 patterns, and a-3) the use of multimodal features as the input of the speaking aid systems that predict F_0 patterns.

For (a-1), one possible factor of insufficient prediction accuracy is a acoustic feature trained in statistical F_0 pattern prediction. In the statistical F_0 pattern prediction mentioned in **Section 2.6.2**, we basically train the *p.d.f.* of $\boldsymbol{Z}_t$, which is a feature in individual time frames. In contrast, F_0 patterns are described as a series of suprasegmental units, such as syllables, stress groups, and intonational units, from a linguistic point of view. Therefore, it might be better to use the technique of training the *p.d.f.* of $\boldsymbol{Z}_{\mathrm{unit}}$, which is associated with the suprasegmental units.

For (a-2), there are several statistical models that might improve the prediction accuracy, e.g., deep neural networks (DNN) [75], long short-term memory (LSTM) [76] which is a recurrent neural network (RNN) architecture, and WaveNet [77]. In particular, the use of the RNN architecture is also related to (a-1) because of the modeling of the time series, $P(\boldsymbol{Z}_t|\boldsymbol{Z}_{t-1}, \cdots, \boldsymbol{Z}_1)$. Note that from the point of view of PC resources, these methods are still unsuitable for the applications using real-time processing, such as the proposed method described in **Section 2.5**, because these methods are known to incur high computational costs. On the other hand, for the GMM-based framework, several techniques to improve the quality of generated speech have also been proposed [78].

For (a-3), we can also use biological signals such as myopotential to recover F_0 patterns. With aid systems for other disabilities, the original abilities have been successfully recovered by using the myopotential obtained at the surface of the chest [79]. One advantage of using these signals is that it is possible to generate truly desired F_0 patterns that reflect the user's intention.

6.2.2 Which F_0 patterns are really desired by laryngectomees in practical use?

In this dissertation, we modeled F_0 patterns of only normal speech. However, in real speech communication such as a conversation, the speech uttered by a speaker differs from normal speech. As mentioned in **Section 1**, the F_0 patterns are determined by not only the word sequence but also several other elements such as b-1) emotion and b-2) intention of the speaker. Therefore, it is necessary to reflect these elements in the F_0 patterns predicted by the proposed speaking aid systems. Note that in the practical use of our implemented prototype system described in **Section 2.5**, a laryngectomee suggested that we realize systems capable of reflecting (b-1) and (b-2).

Considering the laryngectomee's comments about (b-1), an "emotion button" for switching the emotional level of F_0 patterns is required. This is reasonable because we change our speaking style in accordance with our emotion. Therefore, we must model the relationship between the variation of F_0 patterns and emotion, such as anger and sadness.

Regarding (b-2), the laryngectomee said that in real speech communication, he encounters situations when he wants to raise F_0 values for just a moment. One possible way of resolving this request is to implement an ”accent button”, which is associated with the accent command in the Fujisaki model described in **Section 5.2**. While this ”accent button” is activated, the accent command is enabled and the F_0 patterns are changed in accordance with the accent components corresponding to the generated accent command while maintaining the natural.

6.2.3 What is a suitable assistive device?

The electrolarynx allow laryngectomees to produce speech sounds again. However, there are still three essential drawbacks: c-1) all generated speech is totally voiced speech, c-2) the excitation signals are emitted outside as noise, and c-3) one hand is occupied.

For the first issue (c-1), it is necessary to make it possible to produce unvoiced speech because natural speech consists of both unvoiced and voiced speech. For the second issue (c-2), it might be possible to alleviate the emitted noise and convert the noise into sounds suitable for speech by covering the electrolarynx with materials consisting of both soft tissue such as skin and muscle, and the hard tissue such as bone. For the third issue (c-3), the electrolarynx can be fixed against the neck, as reported in [21], or it might be possible to embed it in the body.

Publication, Reference, and Appendix

Publication

Journal papers

[J1] Kou Tanaka, Tomoki Toda, Graham Neubig, Sakriani Sakti and Satoshi Nakamura,
"A Hybrid Approach to Electrolaryngeal Speech Enhancement Based on Noise Reduction and Statistical Excitation Generation,"
IEICE Transactions, Vol. E97-D, No. 6, pp. 1429-1437, Jun. 2014.

International conferences

[I1] Kou Tanaka, Tomoki Toda, Graham Neubig and Satoshi Nakamura,
" Real-time vibration control of an electrolarynx based on statistical F_0 contour prediction,"
Proc. EUSIPCO, pp. 1333–1337, Budapest, Hungary, Aug. 2016.

[I2] Kou Tanaka, Hirokazu Kameoka, Tomoki Toda and Satoshi Nakamura,
"Statistical F_0 prediction for electrolaryngeal speech enhancement considering generative process of F_0 contours within product of experts framework,"
Proc. ICASSP, pp. 5665–5669, Shanghai, China, Mar. 2016.

[I3] Kou Tanaka, Tomoki Toda, Graham Neubig, Sakriani Sakti, Satoshi Nakamura,
" An enhanced electrolarynx with automatic fundamental frequency control based on statistical prediction,"
Proc. ASSETS, Demonstration paper, pp. 435–436, Lisbon, Portugal, Oct. 2015.

[I4] Shinnosuke Takamichi, Kazuhiro Kobayashi, Kou Tanaka, Tomoki Toda, Satoshi Nakamura,
" The NAIST text-to-speech system for the Blizzard Challenge 2015,"
Proc. Blizzard Challenge 2015 Workshop, 4 pages, Berlin, Germany, Sep. 2015.

[I5] Yusuke Tajiri, Kou Tanaka, Tomoki Toda, Graham Neubig, Sakriani Sakti, Satoshi Nakamura,
" Non-audible murmur enhancement based on statistical conversion using air- and body-conductive microphones in noisy environments,"
Proc. INTERSPEECH, pp. 2769–2773, Dresden, Germany, Sep. 2015.

[I6] Kou Tanaka, Tomoki Toda, Graham Neubig, Sakriani Sakti, Satoshi Nakamura,
" An Inter-Speaker Evaluation through Simulation of Electrolarynx Control based on Statistical F_0 Prediction,"
Proc. APSIPA ASC, 4 pages, Siem Reap, Cambodia, Dec. 2014.

[I7] Sakura Tsuruta, Kou Tanaka, Tomoki Toda, Graham Neubig, Sakriani Sakti, Satoshi Nakamura,

“ An Evaluation of Target Speech for a Nonaudible Murmur Enhancement System in Noisy Environments,”
Proc. APSIPA ASC, 4 pages, Siem Reap, Cambodia, Dec. 2014.

[I8] Kou Tanaka, Tomoki Toda, Graham Neubig, Sakriani Sakti, Satoshi Nakamura,
“Direct F_0 Control of an Electrolarynx based on Statistical Excitation Feature Prediction and its Evaluation through Simulation,”
Proc. INTERSPEECH, pp. 31–35, MAX Atria, Singapore, Sep. 2014.

[I9] Kou Tanaka, Tomoki Toda, Graham Neubig, Sakriani Sakti, Satoshi Nakamura,
“ An Evaluation of Excitation Feature Prediction in A Hybrid Approach to Electrolaryngeal Speech Enhancement,”
Proc. ICASSP, pp. 4521–4525, Florence, Italy, May 2014.

[I10] Kou Tanaka, Tomoki Toda, Graham Neubig, Sakriani Sakti, Satoshi Nakamura,
“Hybrid approach to electrolaryngeal speech enhancement based on spectral subtraction and statistical voice conversion,”
Proc. INTERSPEECH, pp. 3067–3071, Lyon, France, Aug. 2013.

Technical reports

[T1] Kou Tanaka, Hirokazu Kameoka, Tomoki Toda, Satoshi Nakamura,
“ Product-of-Experts Approach to Integration of F_0 Generative Process Model to Statistical F_0 Prediction for Electrolaryngeal Speech Enhancement ,”
IPSJ SIG Tech. Rep., Vol. 115, No. 523, SP2015-161, pp. 373–377, Mar. 2016 (in Japanese).

[T2] Kou Tanaka, Tomoki Toda, Graham Neubig, Sakriani Sakti, Satoshi Nakamura,
“Implementation of Direct F_0 Control of an Electrolarynx based on Real-time Excitation Prediction,”
IPSJ SIG Tech. Rep., Vol. 115, No. 99, SP2015-9, pp. 47–52, June 2015 (in Japanese).

[T3] Yusuke Tajiri, Kou Tanaka, Tomoki Toda, Graham Neubig, Sakriani Sakti, Satoshi Nakamura,
“ Non-Audible Murmur Enhancement Method using Air- and Body-Conductive Microphones in Noisy Environments and its Evaluation,”
IPSJ SIG Tech. Rep., Vol. 115, No. 99, SP2015-11, pp. 59–64, June 2015 (in Japanese).

[T4] Sakura Tsuruta, Kou Tanaka, Tomoki Toda, Graham Neubig, Sakriani sakti, Satoshi Nakamura,
“ An Evaluation of Target Speech for Nonaudible Murmur Enhancement Focusing on Intelligibility under Noisy Environments,”
IPSJ SIG Tech. Rep., Vol. 114, No. 303, SP2014-102, pp. 71–76, Nov. 2014 (in Japanese).

[T5] Kou Tanaka, Tomoki Toda, Graham Neubig, Sakriani Sakti, Satoshi Nakamura,
“ An Evaluation through Simulation for Direct F_0 Control of an Electrolarynx based on Statistical Excitation Feature Prediction,”
IPSJ SIG Tech. Rep., Vol. 114, No. 91, SP2014-52, pp. 33–38, June 2014 (in Japanese).

[T6] Kou Tanaka, Tomoki Toda, Graham Neubig, Sakriani Sakti, Satoshi Nakamura,
“[Reseach introduction] An Evaluation of a Hybrid Approach to Electrolaryngeal Speech Enhancement Based on Noise Reduction and Statistical Excitation Prediction,”
IPSJ SIG Tech. Rep., Vol. 114, No. 52, SP2014-32, pp. 331–336, May 2014 (in Japanese).

[T7] Kou Tanaka, Tomoki Toda, Graham Neubig, Sakriani Sakti, Satoshi Nakamura,
“Evaluation of Excitation Feature Prediction in a Hybrid Approach to Electrolaryngeal Speech Enhancement,”
IPSJ SIG Tech. Rep., Vol. 113, No. 308, SP2013-71, pp. 7–12, Nov. 2013 (in Japanese).

[T8] Kou Tanaka, Tomoki Toda, Graham Neubig, Sakriani Sakti, Satoshi Nakamura,
“ A Hybrid Approach to Electrolaryngeal Speech Enhancement Based on Spectral Compensation and Statistical Source Excitation Generation,”
IPSJ SIG Tech. Rep., Vol. 113, No. 76, SP2013-37, pp. 37–42, June 2013 (in Japanese).

Domestic conferences

[D1] Kou Tanaka, Hirokazu Kameoka, Tomoki Toda, Satoshi Nakamura,
“Statistical F_0 Prediction For Electrolaryngeal Speech Enhancement Based on Product-of-Experts Frame work Considering F_0 Generative Process,”
Proc. of Spring Meeting, Acoust. Soc. Jpn., 2-2-13, pp. 259–260, Mar. 2016 (in Japanese).

[D2] Kou Tanaka, Tomoki Toda, Graham Neubig, Sakriani Sakti, Satoshi Nakamura,
“ An Investigation of Latency and Prediction Accuracy in Real-Time Control of Electrolarynx based on Sta tistical Excitation Prediction,”
Proc. of Autumn Meeting, Acoust. Soc. Jpn., 3-1-8, pp. 245–246, Sep. 2015 (in Japanese).

[D3] Kou Tanaka, Tomoki Toda，Graham Neubig, Sakriani Sakti, Satoshi Nakamura,
“ A Real-Time Control of Electrolarynx Based on Statistical Excitation Prediction,”
Proc. of Spring Meeting, Acoust. Soc. Jpn., 1-2-7, pp. 239–240, Mar. 2015 (in Japanese).

[D4] Sakura Tsuruta, Kou Tanaka, Tomoki Toda, Graham Neubig, Sakriani Sakti, Satoshi Nakamura,
“Investigation on Intelligibility Improvements by Nonaudible Murmur Enhancement under Noisy Environ ments,”

Proc. of Spring Meeting, Acoust. Soc. Jpn., 3-2-4, pp. 283–284, Mar. 2015 (in Japanese).

[D5] Yusuke Tajiri, Sakura Tsuruta, Kou Tanaka, Tomoki Toda, Graham Neubig, Sakriani Sakti, Satoshi Nakamura,
“NAM enhancement using air- and body-conductive microphones in noisy environments. by TAJIRI, Yusuke,”
Proc. of Spring Meeting, Acoust. Soc. Jpn., 3-2-5, pp. 285–286, Mar. 2015 (in Japanese).

[D6] Kou Tanaka, Tomoki Toda, Graham Neubig, Sakriani Sakti, Satoshi Nakamura,
“ Evaluation of F_0 control of electrolarynx based on statistical excitation feature prediction,”
Proc. of Autumn Meeting, Acoust. Soc. Jpn., 1-7-17, pp. 227–228, Sep. 2014 (in Japanese).

[D7] Sakura Tsuruta, Kou Tanaka, Tomoki Toda, Graham Neubig, Sakriani Sakti, Satoshi Nakamura,
“ Voicing Effects in Statistical NAM Enhancement on Intelligibility of Converted Speech in Noisy Environ ments,”
Proc. of Autumn Meeting, Acoust. Soc. Jpn., 2-7-10, pp. 253–254, Sep. 2014 (in Japanese).

[D8] Sakura Tsuruta, Kou Tanaka, Tomoki Toda, Neubig Graham, Sakti Sakriani, Satoshi Nakamura,
“ Evaluation of target speech for a nonaudible murmur enhancement system in noisy environments,”
Proc. of Spring Meeting, Acoust. Soc. Jpn., 3-6-5, pp. 331–332, Mar. 2014 (in Japanese).

[D9] Kou Tanaka, Tomoki Toda, Neubig Graham, Sakti Sakriani, Satoshi Nakamura,
“ Control of electrolarynx based on statistical excitation feature prediction,”
Proc. of Spring Meeting, Acoust. Soc. Jpn., 3-6-20, pp. 373–374, Mar. 2014 (in Japanese).

[D10] Kou Tanaka, Tomoki Toda, Graham Neubig, Sakriani Sakti, Satoshi Nakamura,
“ Excitation feature prediction in a hybrid approach to electrolaryngeal speech enhancement,”
Proc. of Autumn Meeting, Acoust. Soc. Jpn., 3-7-7, pp. 1477–1478, Sep. 2013 (in Japanese).

[D11] Kou Tanaka, Tomoki Toda, Graham Neubig, Sakriani Sakti, Satoshi Nakamura,
“ Investigation of converted acoustic features in statistical electrolaryngeal speech conversion,”

Proc. of Spring Meeting, Acoust. Soc. Jpn., 2-7-8, pp. 331–332, Mar. 2013 (in Japanese).

Awards

[A1] The 12th Best Student Presentation Award of ASJ.

[A2] Best Student Paper Award Finalist at 15th Annual Conference of the International Speech Communication Association (INTERSPEECH 2014).

[A3] The 2014 Best Student of Nara Institute of Science and Technology.

Articles

1. Kou Tanaka,
“ My dream to speaking aids,”
Rehabilitation Engineering Society of Japan, Vol. 30, No. 2, pp 39, May 2015.
(Invited article in Japanese)

Master’s thesis

1. Kou Tanaka,
”A Hybrid Approach to Electrolaryngeal Speech Enhancement Based on Spectral Compensation and Statistical Source Excitation Generation,”
Master’s thesis, Graduate School of Information Science, Nara Institute of Science and Technology, Mar. 2014.
(in Japanese)

References

[1] P. Enderby and R. Philipp, “Speech and language handicap: Towards knowing the size of the problem,” *British Journal of Disorders of Communication*, vol. 21, no. 2, pp. 151–165, 1986.

[2] World Health Organization, *The ICD-10 classification of mental and behavioural disorders: clinical descriptions and diagnostic guidelines.* Geneva: World Health Organization, 1992.

[3] J. F. Bosma, M. W. Donner, E. Tanaka, and D. Robertson, “Anatomy of the pharynx, pertinent to swallowing,” *Dysphagia*, vol. 1, no. 1, pp. 23–33, 1986.

[4] K. Yoshino, “An epidemiology and clinico-statistics of laryngeal cancer,” *Otolaryngology, Head and Neck Surgery*, vol. 18, no. 4, pp. 723–729, 2002.

[5] K. Umatani, Y. Tsuruta, K. Yoshino, H. Miyahara, and T. Sato, “Laryngectomee statistics in japan,” *Japan Bronchoesophagology Society*, vol. 36, no. 3, pp. 261–266, 1985.

[6] Research Group for Population-based Cancer Registration in Japan and others, “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.

[7] T. Takafuji, “Current situations of alaryngeal speech by laryngectomees,” *Otolaryngology, Head and Neck Surgery*, vol. 2, no. 5, pp. 527–531, 1986.

[8] Japan Federation of Laryngectomees Associations, *https://www.nikkouren.org*.

[9] T. Nakajima, “Latest knowledge about sources causing the laryngeal cancer,” *Otolaryngology, Head and Neck Surgery*, vol. 18, no. 4, pp. 731–734, 2002.

[10] W. M. Diedrich and K. A. Youngstrom, *Alaryngeal speech.* Charles C. Thomas Publisher, 1966.

[11] M. I. Singer and E. D. Blom, “An endoscopic technique for restoration of voice after laryngectomy,” *Annals of Otology, Rhinology & Laryngology*, vol. 89, no. 6, pp. 529–533, 1980.

[12] H. Barney, F. Haworth, and H. Dunn, “An experimental transistorized artificial larynx,” *Bell system technical Journal*, vol. 38, no. 6, pp. 1337–1356, 1959.

[13] K. Nakamura, T. Tomoki, 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*, vol. 93, no. 7, pp. 1909–1917, 2010.

[14] Y. Nakajima, H. Kashioka, N. Campbell, and K. Shikano, “Non-audible murmur (NAM) recognition,” *IEICE TRANSACTIONS on Information and Systems*, vol. 89, no. 1, pp. 1–8, 2006.

[15] R. L. Goode, "Artificial laryngeal devices in post-laryngectomy rehabilitation," *The Laryngoscope*, vol. 85, no. 4, pp. 677–689, 1975.

[16] H. Doi, T. Toda, K. Nakamura, H. Saruwatari, and K. Shikano, "Alaryngeal speech enhancement based on one-to-many eigenvoice conversion," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 1, pp. 172–183, 2014.

[17] Y. Horii and B. Weinberg, "Intelligibility characteristics of superior esophageal speech presented under various levels of masking noise," *Journal of Speech, Language, and Hearing Research*, vol. 18, no. 3, pp. 413–419, 1975.

[18] N. Uemi, T. Ifukube, M. Takahashi, and J. Matsushima, "Design of a new electrolarynx having a pitch control function," in *Robot and Human Communication, 1994. RO-MAN'94 Nagoya, Proceedings., 3rd IEEE International Workshop on.* IEEE, 1994, pp. 198–203.

[19] Griffin laboratories, "Trutone users guide," *http://www.griffinlab.com/*.

[20] Y. Kikuchi and H. Kasuya, "Development and evaluation of pitch adjustable electrolarynx," in *Speech Prosody 2004, International Conference*, 2004.

[21] K. Matsui, K. Kimura, Y. Nakatoh, and Y. O. Kato, "Development of electrolarynx with hands-free prosody control," *The Proc. of the 8th ISCA*, pp. 273–277, 2013.

[22] K. Nakamura, T. Toda, H. Saruwatari, and K. Shikano, "Speaking-aid systems using gmm-based voice conversion for electrolaryngeal speech," *Speech Communication*, vol. 54, no. 1, pp. 134–146, 2012.

[23] J.-j. Li, I. V. McLoughlin, L.-R. Dai, and Z.-h. Ling, "Whisper-to-speech conversion using restricted boltzmann machine arrays," *Electronics Letters*, vol. 50, no. 24, pp. 1781–1782, 2014.

[24] M. Janke, M. Wand, T. Heistermann, T. Schultz, and K. Prahallad, "Fundamental frequency generation for whisper-to-audible speech conversion," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* IEEE, 2014, pp. 2579–2583.

[25] D. Cole, S. Sridharan, M. Moody, and S. Geva, "Application of noise reduction techniques for alaryngeal speech enhancement," in *TENCON'97. IEEE Region 10 Annual Conference. Speech and Image Technologies for Computing and Telecommunications., Proceedings of IEEE*, vol. 2. IEEE, 1997, pp. 491–494.

[26] P. C. Pandey, S. M. Bhandarkar, G. K. Bachher, and P. K. Lehana, "Enhancement of alaryngeal speech using spectral subtraction," in *Digital Signal Processing, 2002. DSP 2002. 2002 14th International Conference on*, vol. 2. IEEE, 2002, pp. 591–594.

[27] K. Yu and S. Young, "Continuous f_0 modeling for HMM based statistical parametric speech synthesis," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 5, pp. 1071–1079, 2011.

[28] 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 Trans. Inf. and Syst.(Japanese Edition)*, vol. 94, pp. 1240–1247, 2001.

[29] T. Toda, T. Muramatsu, and H. Banno, "Implementation of computationally efficient real-time voice conversion." in *INTERSPEECH.* Citeseer, 2012, pp. 94–97.

[30] Y. Stylianou, O. Cappé, and E. Moulines, "Continuous probabilistic transform for voice conversion," *IEEE Transactions on Speech and Audio Processing*, vol. 6, no. 2, pp. 131–142, 1998.

[31] 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, 2007.

[32] H. Fujisaki, "A note on the physiological and physical basis for the phrase and accent components in the voice fundamental frequency contour," *Vocal Fold Physiology: Voice Production, Mechanisms and Functions*, pp. 347–355, 1998.

[33] H. Kameoka, J. Le Roux, and Y. Ohishi, "A statistical model of speech f_0 contours." in *SAPA@ INTERSPEECH*, 2010, pp. 43–48.

[34] K. Yoshizato, H. Kameoka, D. Saito, and S. Sagayama, "Statistical approach to fujisaki-model parameter estimation from speech signals and its quantitative evaluation," *Proc. Speech Prosody 2012*, pp. 175–178, 2012.

[35] H. Kameoka, K. Yoshizato, T. Ishihara, K. Kadowaki, Y. Ohishi, and K. Kashino, "Generative modeling of voice fundamental frequency contours," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 6, pp. 1042–1053, 2015.

[36] G. E. Hinton, "Training products of experts by minimizing contrastive divergence," *Neural computation*, vol. 14, no. 8, pp. 1771–1800, 2002.

[37] B. L. Sim, Y. C. Tong, J. S. Chang, and C. T. Tan, "A parametric formulation of the generalized spectral subtraction method," *IEEE Transactions on Speech and Audio Processing*, vol. 6, no. 4, pp. 328–337, 1998.

[38] M. Abe, S. Nakamura, K. Shikano, and H. Kuwabara, "Voice conversion through vector quantization," in *Acoustics, Speech, and Signal Processing, 1988. ICASSP-88., 1988 International Conference on.* IEEE, 1988, pp. 655–658.

[39] H. R. Sharifzadeh, I. V. McLoughlin, and F. Ahmadi, "Reconstruction of normal sounding speech for laryngectomy patients through a modified CELP codec," *IEEE Transactions on Biomedical Engineering*, vol. 57, no. 10, pp. 2448–2458, 2010.

[40] I. R. Titze, "The physics of small-amplitude oscillation of the vocal folds," *The Journal of the Acoustical Society of America*, vol. 83, no. 4, pp. 1536–1552, 1988.

[41] J. C. Lucero, “The minimum lung pressure to sustain vocal fold oscillation,” *The Journal of the Acoustical Society of America*, vol. 98, no. 2, pp. 779–784, 1995.

[42] I. R. Titze, *Principles of voice production.* National Center for Voice and Speech, 2000.

[43] J. C. Lucero, “Optimal glottal configuration for ease of phonation,” *Journal of Voice*, vol. 12, no. 2, pp. 151–158, 1998.

[44] W. R. Zemlin, “Speech and hearing science, anatomy and physiology.” 1968.

[45] M. Suzuki, “Current situations and future perspective of medical examination for laryngeal cancer,” *Otolaryngology, Head and Neck Surgery*, vol. 18, no. 4, pp. 771–774, 2002.

[46] N. Nishiyama, M. Nishio, M. Myojin, and K. Shirai, “Improvement of radiation therapy for laryngeal cancer,” *Otolaryngology, Head and Neck Surgery*, vol. 18, no. 4, pp. 781–786, 2002.

[47] Y. Hisa, “Laser surgery in laryngeal cancer : Application and current situation,” *Otolaryngology, Head and Neck Surgery*, vol. 18, no. 4, pp. 788–792, 2002.

[48] R. Hayashi and S. Ebihara, “Partial laryngectomy in therapy for laryngeal cancer,” *Otolaryngology, Head and Neck Surgery*, vol. 18, no. 4, pp. 793–796, 2002.

[49] K. Nagahara, “Application and ranges of supracricoid laryngectomy with cricohyoidoepiglottopexy (scl-chep),” *Otolaryngology, Head and Neck Surgery*, vol. 18, no. 4, pp. 798–802, 2002.

[50] S. E. Williams and J. B. Watson, “Differences in speaking proficiencies in three laryngectomee groups,” *Archives of Otolaryngology*, vol. 111, no. 4, pp. 216–219, 1985.

[51] I. Hočevar-Boltežar and M. Žargi, “Communication after laryngectomy,” *Radiology and Oncology*, vol. 35, no. 4, pp. 249–254, 2001.

[52] 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.

[53] S. Chalstrey, N. Bleach, D. Cheung, and C. Van Hasselt, “A pneumatic artificial larynx popularized in hong kong,” *The Journal of Laryngology & Otology*, vol. 108, no. 10, pp. 852–854, 1994.

[54] H. Kawahara, I. Masuda-Katsuse, and A. De Cheveigne, “Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based f_0 extraction: Possible role of a repetitive structure in sounds,” *Speech communication*, vol. 27, no. 3, pp. 187–207, 1999.

[55] T. Toda, M. Nakagiri, and K. Shikano, “Statistical voice conversion techniques for body-conducted unvoiced speech enhancement,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 9, pp. 2505–2517, 2012.

[56] K. Pearson, “Liii. on lines and planes of closest fit to systems of points in space,” *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, vol. 2, no. 11, pp. 559–572, 1901.

[57] T. Fukada, K. Tokuda, T. Kobayashi, and S. Imai, “An adaptive algorithm for mel-cepstral analysis of speech,” in *Acoustics, Speech, and Signal Processing, 1992. ICASSP-92., 1992 IEEE International Conference on*, vol. 1. IEEE, 1992, pp. 137–140.

[58] T. Kobayashi, S. Imai, and Y. Fukuda, “Mel-generalized log spectral approximation filter,” *IECE transaction*, vol. J68-A.

[59] A. P. Dempster, N. M. Laird, and D. B. Rubin, “Maximum likelihood from incomplete data via the EM algorithm,” *Journal of the royal statistical society. Series B (methodological)*, pp. 1–38, 1977.

[60] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, “Speech parameter generation algorithms for HMM-based speech synthesis,” in *Acoustics, Speech, and Signal Processing, 2000. ICASSP’00. Proceedings. 2000 IEEE International Conference on*, vol. 3. IEEE, 2000, pp. 1315–1318.

[61] Y. Ohtani, T. Toda, H. Saruwatari, and K. Shikano, “Maximum likelihood voice conversion based on GMM with STRAIGHT mixed excitation,” 2006.

[62] T. Muramatsu, Y. Ohtani, T. Toda, H. Saruwatari, and K. Shikano, “Low-delay voice conversion based on maximum likelihood estimation of spectral parameter trajectory,” in *INTERSPEECH*, 2008, pp. 1076–1079.

[63] K. J. Kohler, “Macro and micro f_0 in the synthesis of intonation,” *Papers in Laboratory Phonology I: Between the Grammar and Physics of Speech*, pp. 115–138, 1990.

[64] J. Latorre, M. J. Gales, S. Buchholz, K. Knill, M. Tamura, Y. Ohtani, and M. Akamine, “Continuous f_0 in the source-excitation generation for HMM-based TTS: Do we need voiced/unvoiced classification,” in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2011, pp. 4724–4727.

[65] A. Sakurai and K. Hirose, “Detection of phrase boundaries in japanese by low-pass filtering of fundamental frequency contours,” in *Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on*, vol. 2. IEEE, 1996, pp. 817–820.

[66] M. Abe, Y. Sagisaka, T. Umeda, and H. Kuwabara, “Speech database user’s manual,” *ATR Interpreting Telephony Research Laboratories Technical Report, TR-I-0166, Japan*, 1990.

[67] B. Roubeau, C. Chevrie-Muller, and J. L. Saint Guily, “Electromyographic activity of strap and cricothyroid muscles in pitch change,” *Acta oto-laryngologica*, vol. 117, no. 3, pp. 459–464, 1997.

[68] T. Shipp, E. T. Doherty, and P. Morrissey, "Predicting vocal frequency from selected physiologic measures," *The Journal of the Acoustical Society of America*, vol. 66, no. 3, 1979.

[69] K. Nakamura, M. Janke, M. Wand, and T. Schultz, "Estimation of fundamental frequency from surface electromyographic data: EMG-to-f_0," in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2011, pp. 573–576.

[70] T. Moriguchi, T. Toda, M. Sano, H. Sato, G. Neubig, S. Sakti, and S. Nakamura, "A digital signal processor implementation of silent/electrolaryngeal speech enhancement based on real-time statistical voice conversion." in *INTERSPEECH*, 2013, pp. 3072–3076.

[71] H. Kawahara, J. Estill, and O. Fujimura, "Aperiodicity extraction and control using mixed mode excitation and group delay manipulation for a high quality speech analysis, modification and synthesis system STRAIGHT." in *MAVEBA*, 2001, pp. 59–64.

[72] H. Kameoka, "Modeling speech parameter sequences with latent trajectory hidden markov model," in *2015 IEEE 25th International Workshop on Machine Learning for Signal Processing (MLSP)*. IEEE, 2015, pp. 1–6.

[73] S. Narusawa, N. Minematsu, K. Hirose, and H. Fujisaki, "A method for automatic extraction of model parameters from fundamental frequency contours of speech," in *Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on*, vol. 1. IEEE, 2002, pp. I–509.

[74] T. Matsuda, K. Hirose, and N. Minematsu, "Applying generation process model constraint to fundamental frequency contours generated by hidden-markov-model-based speech synthesis," in *Acoustical Science and Technology*, vol. 33, no. 4, 2012, pp. 221–228.

[75] H. Ze, A. Senior, and M. Schuster, "Statistical parametric speech synthesis using deep neural networks," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013, pp. 7962–7966.

[76] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[77] A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," *arXiv preprint arXiv:1609.03499*, 2016.

[78] S. Takamichi, T. Toda, A. W. Black, G. Neubig, S. Sakti, and S. Nakamura, "Postfilters to modify the modulation spectrum for statistical parametric speech synthesis," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 4, pp. 755–767, 2016.

[79] The Johns Hopkins University Applied Physics Laboratory LLC, "Amputee makes history with APL's modular prosthetic limb," *http://www.jhuapl.edu/newscenter/pressreleases/2014/141216.asp*, 2014.