Doctoral Dissertation

A study on effects of glottal source and vocal tract cues on perception of emotional vowels

Yongwei Li

Supervisor: Professor Masato Akagi

School of Information Science
Japan Advanced Institute of Science and Technology

December, 2018

Abstract

In human-human communication, speech is the most direct way for communication. Speech contains a lot of information of speaker, such as emotion, gender, age, native and level of education. Emotions play a vital role in speech understanding. By using appropriate emotions, the same textual information can be used to convey different meanings. Emotions in speech can be used not only to express intentions, but also to understand intended information by combining potential emotions, voice information, and other linguistic factors.

Although acoustic features of emotional speech have been investigated, it is still difficult to model emotions by using only these acoustic features. Many studies have shown that the speech production organs features, such as glottal waveforms and vocal tract shapes are important. However, the properties of glottal source and vocal tract to expressive emotions via acoustic features have not been investigated deeply yet. Thus, this study focuses on investigating the effects of glottal source and vocal tract cues on emotional speech perception, particularly for vowel, since vowel plays more important role in emotional speech.

The research aims to investigate the effects of glottal source and vocal tract cues on perception of emotional vowels, especially after removing known effects of dominant prosody features (e.g., pitch, intensity and duration). Thus, (1) an analysis-by-synthesis method is firstly developed for estimating glottal source waveform and vocal tract shape of emotional vowel. (2) the glottal source waveform and vocal tract shape are estimated from Japanese vowel /a/, and the spectral tilt and character of vocal tract shape are consistent with previous results. (3) F_0/pitch (fundamental frequency), intensity (E_e-related), duration of source related features (prosody features), spectral tilt of glottal source waveform, and F_1 (first formant frequency) are discussed, which in a controlled way of modifying the estimated glottal source waveform and vocal tract shape and utilized for establishing an analysis-by-synthesis method for resynthesizing the emotional vowels. Then, Japanese natives with normal hearing participate in the evaluation of perceptual

rating emotions in the valence and arousal space. The results show that the glottal source information plays an essential role in perception of emotions in vowels, whereas the vocal tract information contributed to the valence and arousal perception after neutralizing the F_0, intensity, and duration cues effects.

This study investigates emotional vowels from the point of view of speech production. The results contribute to further understanding the emotional speech production mechanism, also can enlighten many emotional speech fields, such as emotional speech recognition, synthesis and conversion. Moreover, an accurate estimation method of glottal source waveform and vocal tract shape is proposed for vowels in this study. It can be used in many speech signal processing fields, for example, speech analysis, speech synthesis, voice pathology detection, speaker recognition, and speech recognition.

Keywords: Emotional vowel production, emotional vowel perception, glottal source waveform, vocal tract, ARX-LF model, valence and arousal

Acknowledgments

First of all, I would like to express my deepest gratitude to my supervisor, Professor Masato Akagi, for his guidance, encouragement and support during my master and Ph.D. studies. Without his guidance, this thesis could not have been finished. He also guides me on technical writing and English writing, and makes a lot of publications in my master and Ph.D. course. Studying under his guidance that is my best experiences.

I would like to thank my associate supervisor, Professor Masashi Unoki, for his comments and advises in my master and Ph.D. studies. Especially in the laboratory meetings, he helps me to improve my knowledge, presentation skill and Powerpoing skill.

I would like to thank Professor Jianwu Dang for his discussions, suggestions and comments both on my studies and my life.

I also would like to thank my minor research supervisor, Associate Professor Ken-Ichi Sakakibara, from Health Sciences University of Hokkaido, for his comments and solutions during my Ph.D. study. Especially, he came to Japan Advanced Institute of Science and Technology (JAIST) from Hokkaido many times to help me.

I thank all members of acoustic information science laboratory at JAIST. In particular, thanks to my touter, Dr. Rieko Kubo, for her help in my life from the first day in JAIST to now.

In addition, I would like to express my gratitude to Professor Donna Erickson and Professor Aijun Li for their valuable comments on my study.

Finally, I would like to express to my gratitude to the financial support of my Ph.D study scholarship by China Scholarship Council (CSC). I also thank the Grant-in-Aid for Scientific Research and A3 Foresight program that support me to join a lot of conferences.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Human emotions

In the last few decades, there are a lot of studies on human emotions. Some researchers suggested that emotion is a kind of motivational system of human beings [1], and other researchers pointed out that emotion plays a very important role on sustaining behavior and organizing [2]. Some researchers said that emotion is the essential function, and emotional activity is governed by the nervous system [3–5].

On the other hand, psychiatrists and clinical psychologists have pointed out that one kind of the psychopathology problems is actually an emotional problem [6]. Moreover, psychopathology problem can be alleviated by controlling emotions and suppressing inappropriate emotional responses.

In a word, emotion is very important for human beings. Emotion is a complex phenomenon, and it appears at times when something unexpected happens. Emotion is not stable, and constantly changing with the stimulus, because it affected by many factors. Figure 1-1 shows a simple emotional processing that a person's emotion was generated by the stimulus from outside, and then was delivered to other person using speech, face and eyes, etc. For example, if your teacher suddenly tells you that you got full score in the exam (current stimuli), you will be very happy (psychological state), and you will respond to the teacher using a happy voice with a smile (show a perceivable emotion).

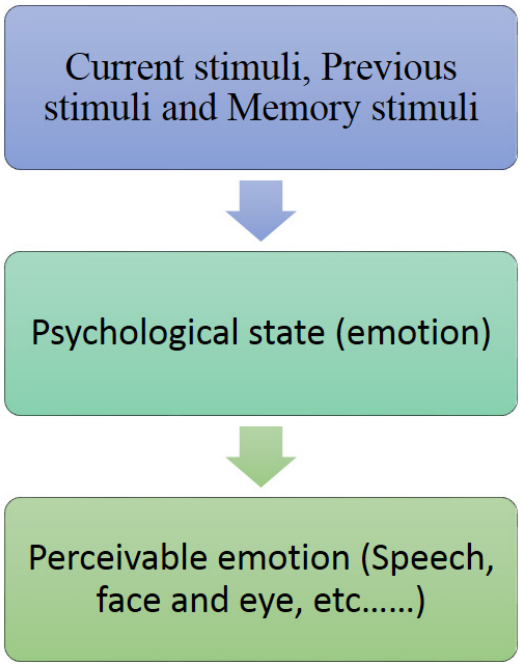

Figure 1-1: A simple emotional processing by human beings

1.2 History of research on emotion in speech

Fujisaki [7] defined that speech contains three kinds of information as shown in Figure 1-2.

- **Linguistic information**: discrete information performing by written language or inferring from context.
- **Paralinguistic information**: continuous and discrete information that is modified by the speaker or supplement the linguistic information.
- **Nonlinguistic information**: continuous and discrete information that cannot be easily controlled by the speaker, like emotion, age, gender, and body size of the speaker, etc.

Emotions in speech can be divided into two groups: (1) spontaneous or acted emotion which is controllable is defined as Para-linguistic information, and (2) spontaneous emotion which is uncontrollable is defined as Non-linguistic information in the thesis. In this study, we focus on acted emotional speech.

Speech is the most natural and direct way for human communication. From the view of psychology and phonology, emotions in speech are the most common way to express emotional state and intentions of speaker, and act a pivotal part in speech understand-

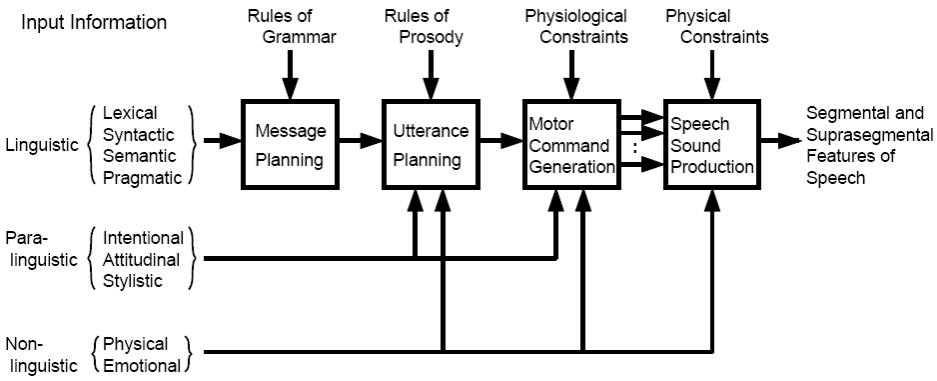

Figure 1-2: Processing of various types of information in speech [7].

ing [8]. By combining with different emotions, the same textual information can be used to convey different meanings. Moreover, the intended information can be understood from underlying emotions as well as paralinguistic information and other linguistic information.

Because the importance of emotion in speech and its huge impact on human communication, studies on emotion in speech started in the early 20th century. The first research was conducted by psychiatrists using electro-acoustic analysis method to diagnose emotional problem [9]. Since 1960s, a number of studies proved that speaker's emotional state, age, gender, individuality, and intelligence could be recognized only by speech [10], and the speech could be used to diagnose emotional states by psychiatrists [11, 12].

With the development of telephone, computer, and other communication technologies, phoneticians found the importance of emotion in speech for speech communication [13]. Therefore, more attention has been drawn to emotions in speech for speech signals processing by speech researchers and engineers. In recent years, many researchers attempted to process emotions in speech to speech technology applications to make it more intelligent for human-computer interactions. Around 2000s, emotional speech topic was developed in leaps and bounds, and there were many conferences and publications on emotional speech [14, 15].

1.2.1 Description of emotions

To describe emotion perceptions in speech, categorical or dimensional approach are frequently used. For the categorical approach, the emotions in speech are generally regarded as discrete states, such as anger, joy, sadness, disgust, fear, and surprise, e.g., [2], cannot

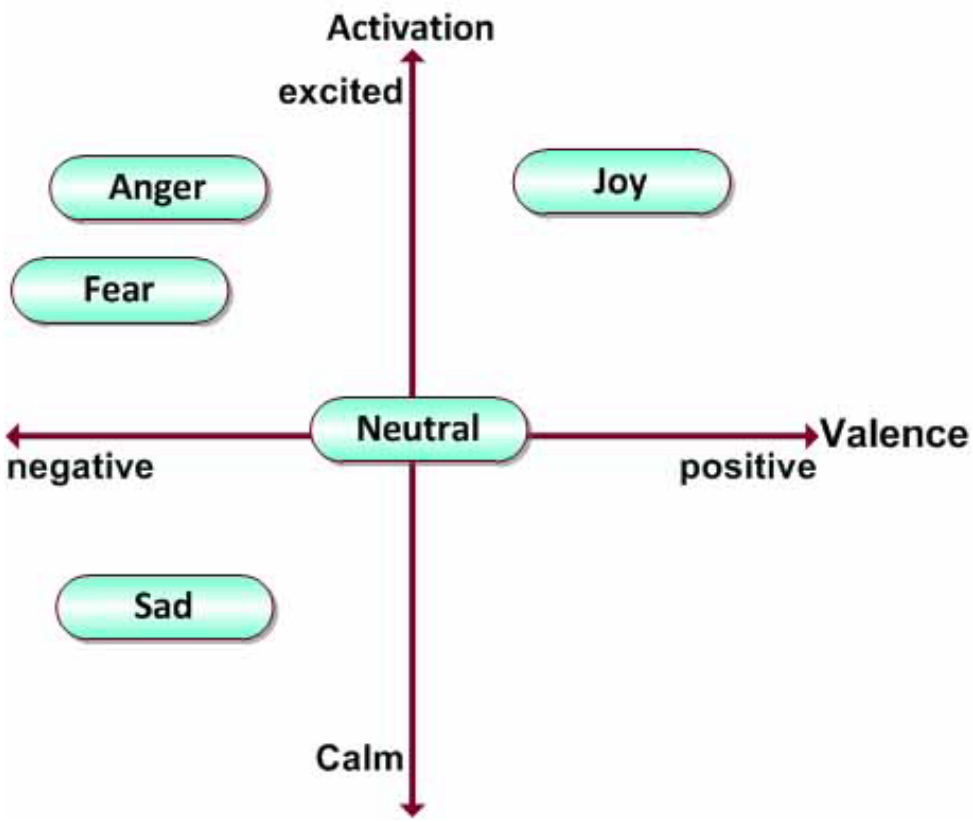

Figure 1-3: Locations of different categories of emotions in valence and activation (arousal) space.

capture the perceptual complexity of emotions [15]. In contrast, the multi-dimensional space approach, which is more flexible and continuous, has been widely utilized to describe emotions by representing emotions with different degrees on different position in the dimensional space [16]. For example, some researchers described emotional states by two dimensional space which includes valence and arousal/activation [16, 17]. Scherer adopted valence and strength in two dimensional space [18]. Grimm *et al.* adopted valence, activation/arousal and dominance in three dimensional space [19].

Among these multi-dimensional space methods, valence and arousal/activation (V-A) space are extensively used to analyze and recognize the emotions in speech in the last ten years. Description of V-A space are as follows:

• Valence (appraisal or evaluation): the degree that the listener perceives the emotion on a scale of negative emotion to a positive one, or from unpleasantness to pleasantness.

• Arousal (activation): the degree that the listener perceives the emotion on a scale from calm to excited.

Therefore, in the present study, valence and arousal emotional space is used to describe

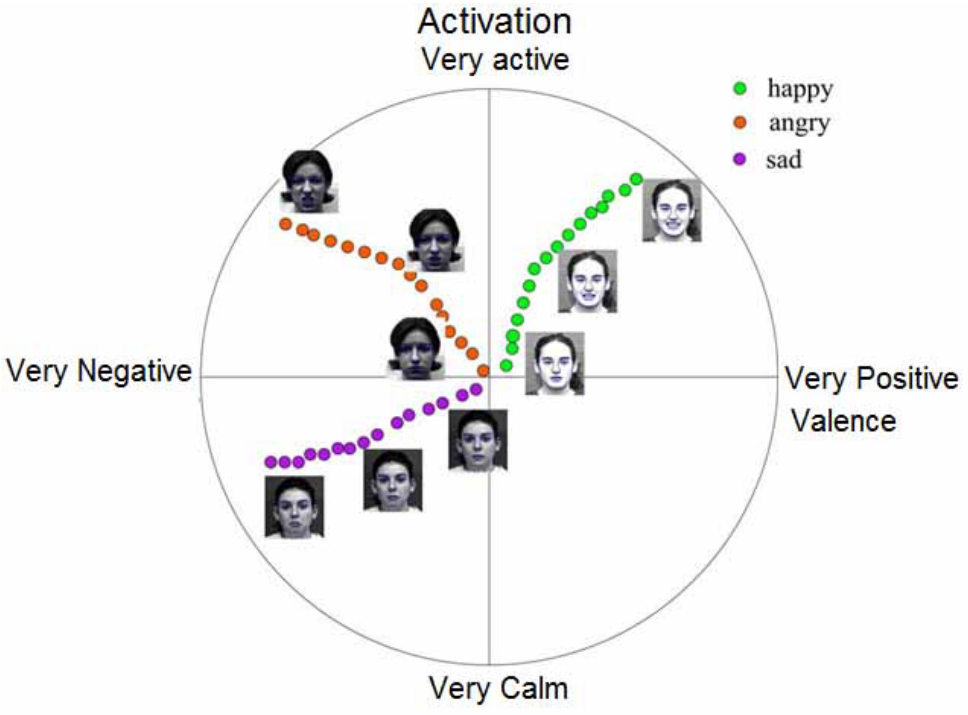

Figure 1-4: Degrees of emotions in valence and activation (arousal) space [20]

emotions, e.g., joy locates in the position of positive valence and high arousal; anger locates in the position of negative valence and high arousal; sadness locates in the negative valence and low arousal as shown in Figs 1-3 and 1-4.

1.2.2 Brunswik's lens model for emotional speech communication

As Fujisaki reported, speech communication includes linguistic, paralinguistic, as well as nonlinguistic information. A speech communication model needs to have ability to represent Fujisaki's description.

Deep learning is very popular for speech signal processing, and it can get some achievements. However, the process of deep learning is a black box. If we want to improve the performance, the only way is to increase the training data. In other words, the deep learning method only works well when large data could be provided, and it cannot help us to understand the mechanisms of human vocal communication. Therefore, many researchers want to open the black box to see that the inside the box form the point of view of speech production and perception. From the point of the engineering, the deep learning method is more practical and the results are better than other method. Our study is to discuss connections/relations from the speech production to perception via

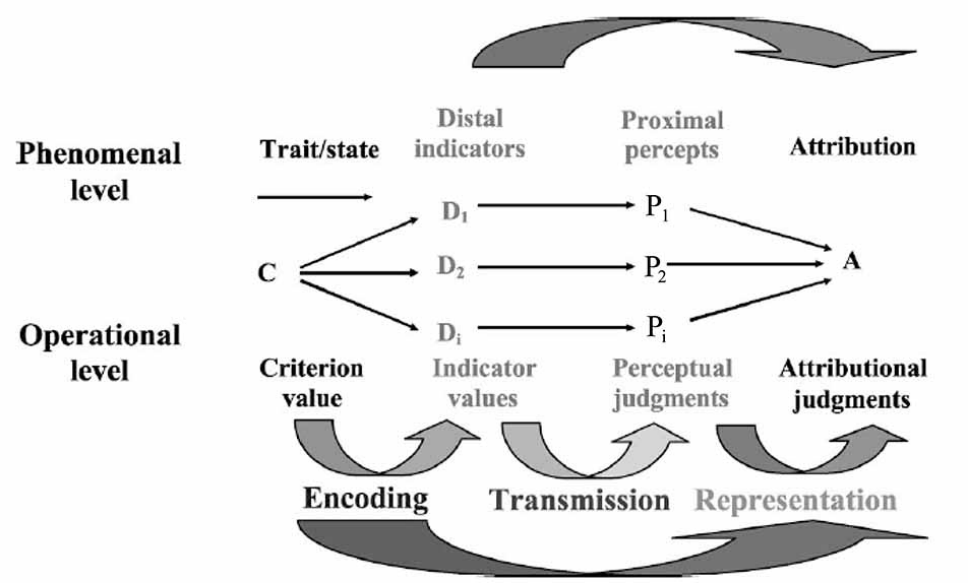

Figure 1-5: A Brunswikian lens model of the vocal communication of emotion [15].

acoustic features, and this study may helps us to understand contributions of the speech production for perception of emotional voices.

Lens model proposed by Brunswik that was originally used in multiple research fields of visual perception [21], and this model is close to both the Fujisaki's point of view and the speech chain. Thus, this model has been widely used in the concept of speech communication. Scherer extended this model to emotional speech communication [15, 22] as shown in Figure 1-5.

The extended Lens model described encoding and decoding processes of emotional speech communication. This process starts at the encoding end, where the speaker's emotional state are encoded in speech signals with the change of speaker's emotional state, and speech production system produces some specific acoustic features, which are called distal indicators, and they are transmitted to the listener through multiple channels. At the decoding end, listener's ear acts as a speech decoder to receive perceived cues (proximal percepts) of emotional speech via the auditory perceptual system.

The extended Lens model is a typical model, guiding researchers to understand the emotional speech communication process and further exploration of emotional speech. For example, Huang and Akagi [23] proposed a three-layer perceptual model of emotional speech based on Lens model which is shown in Figure 1-6.

Process of the three-layer perceptual model is: the acoustic features (bottom layer) are

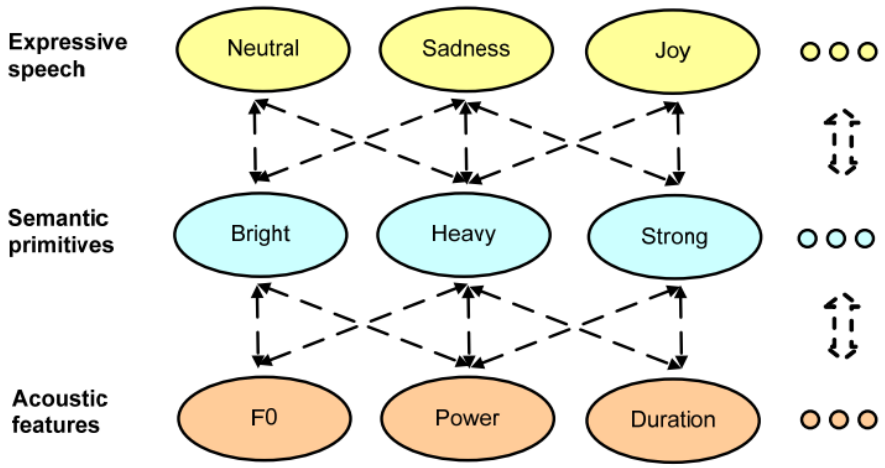

Figure 1-6: A Three-layer perceptual model of emotional speech [23]

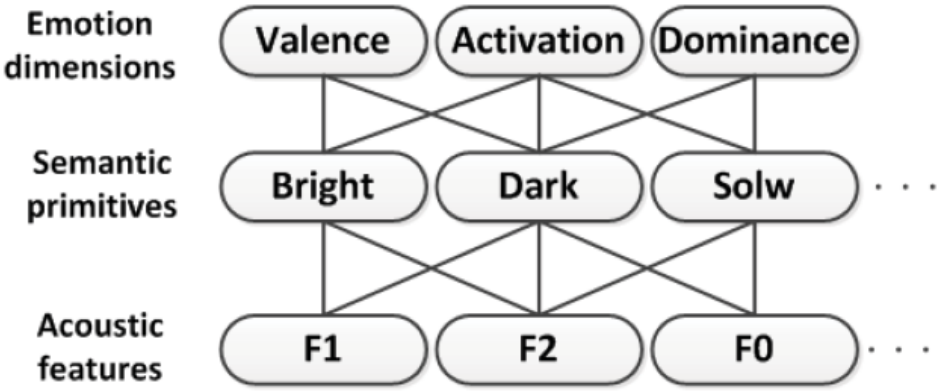

Figure 1-7: Elbarougy's three-layer model [24]

delivered to semantic primitives (middle layer) by related adjectives, and emotions (e.g., neutral, joy and sad) of speaker's speech (top layer) are described by semantic primitives. The essence of the three-layer model can be considered as the Lens model, and at encoding end, the three-layer model starts from acoustic features. At the decoding end of the three-layer model, the attribution (emotions of speaker's speech, which are in the top layer of three-layer model) are perceived via the proximal cues (semantic primitives).

Elbarougy and Akagi [24] further revised Huang's three-layer model which shown in Figure 1-7. In Elbarougy's three-layer model, emotions in the top layer of Huang's three-layer model were replaced by high-level adjectives (Valence, Activation, and Dominance). They adopted this model that each layer were connected by using a fuzzy inference system to recognize emotions form speech signals [19]. Following Elbarougy's idea, this model was further adopted to synthesize emotional speech by Xue, Hamada and Akagi [25].

In a word, most of the studies of emotional speech are still based on the Lens model, including the researches of emotional speech analysis, recognition (decoding end of the lens model) and synthesis (encoding end of the lens model).

1.3 Review of emotional speech literature

Most investigations of emotional speech from the distal cues of the Lens model, in which acoustic features are frequently used for the distal cues, such as fundamental frequency (F_0), intensity, speech rate, envelope amplitude, voice quality, and formant frequency.

In the early studies, Williams *et al.* processed the acoustic features (F_0, duration and amplitude information) of original emotional speech, and concluded that F_0, duration and amplitude information contribute to emotional speech [26]. They also found that same emotion in speech were not always delivered by same acoustic features among different speakers.

Murry *et al.* [27] summarized the acoustic features that correlate to speech emotions (see Table 1.1). Erickson [28] summarized a lot of studies on the emotional speech (acted) as listed in Table 1.2.

These acoustic features in all of previous studies can be divided into six groups:

- F_0-related features;
- Duration features;
- Formant features;
- Voice quality features;
- Spectrum-related features;
- Power-related features.

There are many studies on the relationship between these acoustic features and V-A space, which is the crucial acoustic feature that contributes perception of emotion is a contentious problem among these studies. For example, Yanushevskaya [35] suggested that there may be no one-to-one correspondence between acoustic features and emotional speech, and a particular emotion in the speech is affected by many acoustic features. Mozziconacc [36] suggested that there may be many ways to produce emotional speech by the same acoustic features.

Murry *et al.* [27] found that F_0 (pitch) information (F_0 range, F_0 average and F_0

Table 1.1: Summary of correlated acoustic features with emotional speech [27].

	Anger	Happiness	Sadness	Fear	Disgust
Speech rate	Slightly faster	Faster or slower	Slightly slower	Much faster	Very much slower
Pitch average	Very much higher	Much higher	Slightly slower	Very much higher	Very much slower
Pitch range	Much wider	Much wider	Slightly narrower	Much wider	Slightly wider
Intensity	Higher	Higher	Lower	Normal	Lower
Voice quality	Breathy, chest tone	Breathy, blaring	Resonant	irregular voicing	Grumbled, chest tone
Pitch changes	abrupt, on stressed syllables	smooth, upward inflections	Downward inflections	Normal	Wide, downward terminal inflections
Articulation	Tense	Normal	Slurring	Precise	Normal

contour) and voice quality are important features for emotional speech. Juslin [37] reported that F_0 is the essential cue for emotional speech. A higher average F_0 is usually correspond with a higher level of arousal, while duration and spectral cues are generally utilized for emotion valence prediction. Banziger and Scherer [38] further suggested that F_0 is important feature to perceive emotional speech, in particular, F_0 average and F_0 range play an important role on the perception of arousal, but not F_0 contour shape.

While, Audibert [39] reported that F_0 contour provides more information on the perception of arousal. Mori and Kasuya [40] suggested that first formant frequency(F_1) and second formant frequency (F_2) are essential to perceive the emotional speech valence, and that the higher F_1 and F_2 correlated to the more positive valence perception of emotion.

Schuller and colleagues [41, 42]introduced a lot of studies for emotional speech and summarized a lot of acoustic features that related to emotions, such as pitch, jitter, formant frequency, shimmer, loudness, and spectral parameters (HNR, spectral slope, and formant relative energy). Their studies mainly focus on the acoustic features and mapping to perception (emotional state or emotional dimensional space).

In a word, most of these studies tried to analyze acoustic features of emotional speech,

and find these relationships. These studies are partly investigated and discussed when considering the Lens model.

There is increasing evidence suggesting that the characteristics of the glottal source waveform need to be paid more attention for modeling voice types and expressive speech synthesis [43]. Scherer [44] pointed out that essential point for vocal differentiation of discrete emotions seems to be voice quality, and humans can perceive voice quality independent of F_0. One definition of voice quality [45] is that the quality of two sound can be perceived differences by listener, although the loudness and pitch are same. There are many redundancies contained when expressing emotional speech. Even if many vocal cues are eliminated, some emotional states still can be perceived. Thus, the investigation is necessary from the speech production (Glottal source and vocal tract).

1.4 Motivations and Research goal

To further analyze emotions in speech, in this study, we review the Lens model. In the encoding end, we revisit speech production system. The glottal source and the acoustics of the vocal tract (the filter) are usually used to characterize the speech production process. We focus on investigating the relations of the speech production (glottal source and vocal tract shape) acoustic features and perception of emotions, and we try to answer how glottal source waveform and vocal tract shape affect perception of emotions via related acoustic features. According to the Brunswik's lens model, our motivated process is shown in Fig. 1-8.

- Speech production: glottal source waveform and vocal tract shape.
- Acoustic features: glottal source and vocal tract-related features.
- Speech perception: Valence and arousal.

Emotion in speech is a complex phenonemon, which is affected by many factors, such as linguistic content, culture, and language. Since vowel perception can provide an important insight into speech perception [34, 46–48], the emotion perception in Japanese vowels was examined in this study.

In this thesis, proposed method investigates the effects of glottal source and vocal tract cues on perception of emotional vowels for understanding the emotional speech from the point of view of speech production. Although acoustic features of emotional speech have

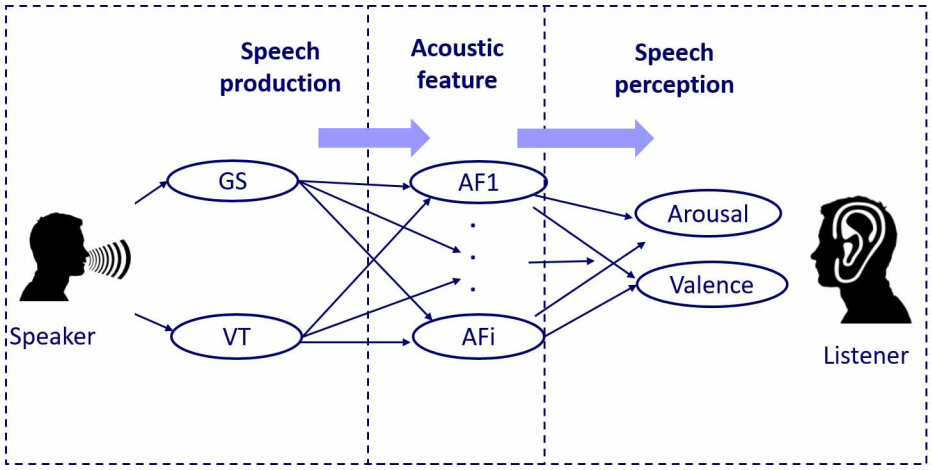

Figure 1-8: The process of this study

been investigated and used for emotional speech recognition and synthesis, the properties of speech production (glottal source and vocal tract) for emotional speech have not been investigated deeply yet. To further analyze emotional speech and understanding emotional speech, this study follows the process of human speech communication system. The investigation consists of three parts: speech production, acoustic feature and perception of emotional speech as shown in Fig. 1-8.

The present study contains three steps. (1) the measurement of vocal folds vibration and vocal tract shape simultaneously and directly during uttering emotional speech, such as using magnetic resonance imaging (MRI), electromagnetic articulography (EMA), and high-speed videoendoscopy (HSV) which is precise but costly, (2) accurately estimating the glottal source and vocal tract from emotional speech by traditional source-filter model is still challenging. In the first step, we developed an analysis-by-synthesis method to estimate glottal source waveform and vocal tract shape from emotional speech signals based on source-filter model [49]. In the second step, source and filter-related features were investigated, and relationships between source-filter and these-related features were discussed. In the final step, the glottal source waveform and vocal tract shape were modified in a controlled way and then used for resynthesizing emotional vowels by applying a recently developed analysis-by-synthesis method, The resynthesized emotional vowels were presented to native Japanese listeners with normal hearing for perceptually rating emotions in valence and arousal dimensions.

1.5 Structure of the Dissertation

The remaining parts of this thesis are organized as follows:

- **Chapter 2** describes speech production system that is often characterized by the glottal source and acoustics of the vocal tract (the filter). Thus, we first reviewed literature on measurement and estimation of the glottal source and vocal tract. Then, source-filter theory is introduced, and source-filter models are reviewed, especially the Auto-Regressive eXogenous with Lilijencrant-Fant (ARX-LF) model.

- **Chapter 3** introduces proposed schemes for estimating glottal source wave and vocal tract shape of emotional speech based on the ARX-LF model. Since, it is difficult to optimize multiple parameters of the LF mdoel, two solutions/scheme were introduced in this chapter: the first method needs the helps of electro-glottograph (EGG) signal to estimate the initial values. In the second method, instead of using EGG signal, the initial values of the LF model are obtained using an inverse filter. The implement of two methods are introduces in details, and both methods are evaluated on synthesized vowels and real vowels. The emotional vowels (joy, neutral, sadness, and anger) with four speakers are analyzed by the method with EGG signal. An analysis-by-synthesis system is constructed base on the method without EGG signal in the final part of this chapter.

- **Chapter 4** focus on discussion of the glottal source and vocal tract related features based on the results of in previous chapter. The source related features of intensity (E_e of the LF model related), F_0 (T_0 of the LF model related) and spectral tilt of glottal source waveform are discussed. The filter related feature of F_1 is mainly discussed in this chapter.

- **Chapter 5** describes the effects of these source and filter related feature to the emotional V-A space. The perceptual scores of synthesized vowels using the ARX-LF model and original vowels are first confirmed on emotional V-A space. The effects of prosody features (pitch/F_0, intensity and duration) associated with glottal source to the emotional V-A space are first discussed. The effects of spectra features

associated with glottal source waveform and first formant frequency associated with vocal tract on emotional V-A space are then discussed.

- **Chapter 6** summarizes all of this work and its contributions to this research filed and other research fields. Moreover, future works are discussed.

Table 1.2: Summary of studies on emotional speech by Erickson [28].

Research	Emotional speech	Language	Speakers	Findings
Paeschke [29]	happiness, anger, anxiety, sadness, disgust, boredom	German	actors, sentences	gradient linear regression of global F_0 trend (final F_0 movement)
Scherer *et al.* [30]	anger, sadness, joy, fear, disgust	German	professional radio actors	F_0, intensity, speech rate, spectral
Hashizawa *et al.* [31]	anger, joy, sadness	Japanese	4 professional NHK announcers	speech rate and F_0
Ehrette *et al.* [32]	speaking styles (normal, warm, dynamic, reassuring, smiling)	French	20 professional female speakers	voice quality attribute (spectral centroide) best correlated parameter for perceptual attributes
Huang and Akagi [23]	joy, sadness, hot anger, cold anger, neutral	Japanese	professional actress (Fujitsu Lab database)	differences in F_0, power and duration
Maekawa [33]	admiration, disappointment, suspicion, indifference, focused, neutral	Japanese	3 non-professional speakers, 10 repetitions, "acted," 3 short phrase	changes in duration, pitch, vowel formants, vowel spectrum, voice quality (laryngealization for suspicion)
Takeda *et al.* [34]	degrees of anger	Japanese	4 non-professional speakers	changes in intensity, speech rate, and F_0

Chapter 2

Speech production

This chapter describes human speech production mechanism. First, three main components in speech production system including lungs, larynx, and vocal tract are briefly introduced. Speech production model, the source-filter theory and source-filter model are then described, in which glottal excitation and vocal tract are introduced as two main functions in source-filter model. Finally, existing source-filter models are reviewed, especially, widely used source filter model are described in detail.

2.1 Speech production system

The speech production process mainly includes three organs: lungs, larynx and vocal tract. The lungs are motor of human respiratory system, which provides airflow to the trachea, and the airflow in the trachea go to the larynx. The larynx is a house for vocal folds, and there is a pair of elastic vocal folds in the larynx. At the larynx, the airflow makes vocal folds to vibrate, which are source excitation to the vocal tract for voiced speech. The vocal tract contains oral cavity and nasal cavity, the sound is colored by the vocal tract, the vocal tract shape determines various voice sound. For example, vowel /a/ has its distinctive vocal tract shape. The spectrum of source excitation/waveform are filtered by vocal tract shape. The speech waveform are finally radiated by the lips and the deliver to the listeners. The speech production system is shown in Figure 2-1.

There are two main types of the speech sound: voiced and unvoiced. The source of voiced sounds is the vibration of vocal folds, which mainly generates periodic waveform

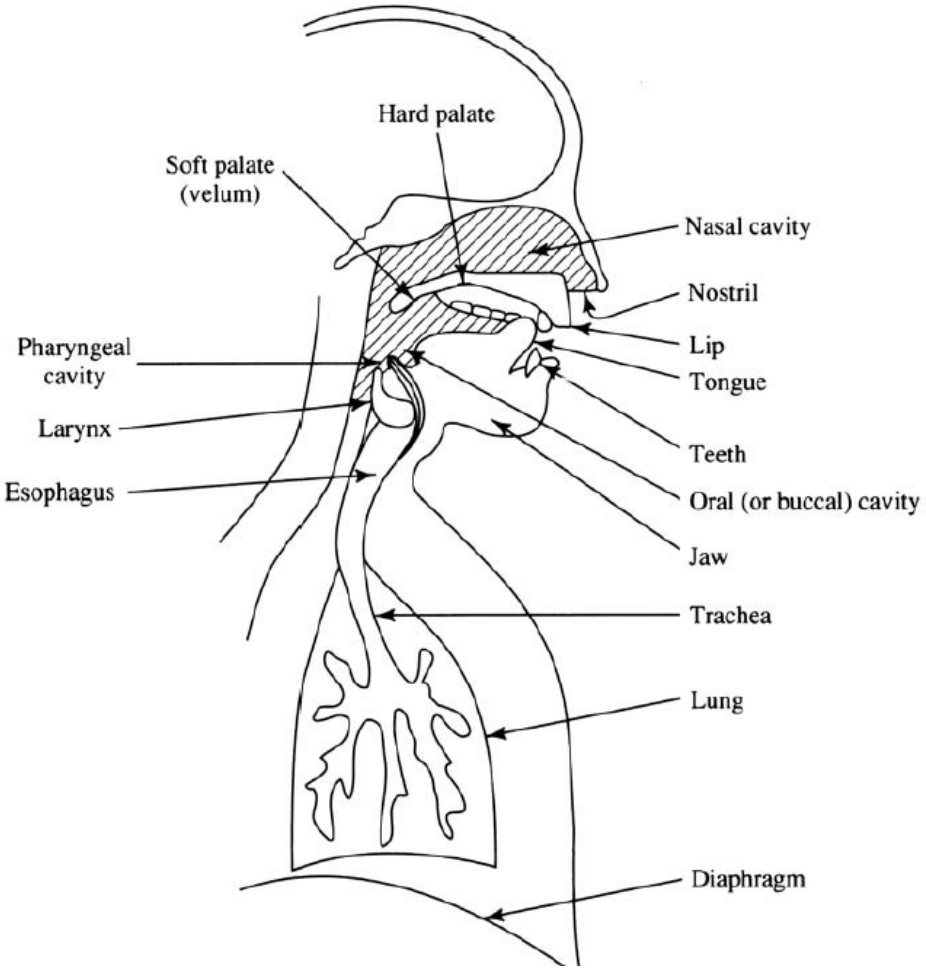

Figure 2-1: Human speech production system [50]

with rich harmonics to the vocal tract. For example, to produced vowels, the vocal folds often periodically vibrates. For unvoiced sounds, the source is a turbulent noise and there is no harmonics, and it is constructed by the non-periodic airflow in the vocal tract. Since this paper focuses on the study of Japanese emotional vowels, the source of voiced sounds is further introduced in next section, whereas there is no further description for unvoiced sounds.

2.2 Vocal folds vibration

The respiratory system of human is a reservoir of air, and the inhalation and exhalation are controlled by lungs. In normal breathing conditions (no speaking), the vocal folds are open, and air can pass freely through the vocal fold and vocal tract, producing no sound. When humans want to produce voiced sounds, two pieces of the vocal folds move toward

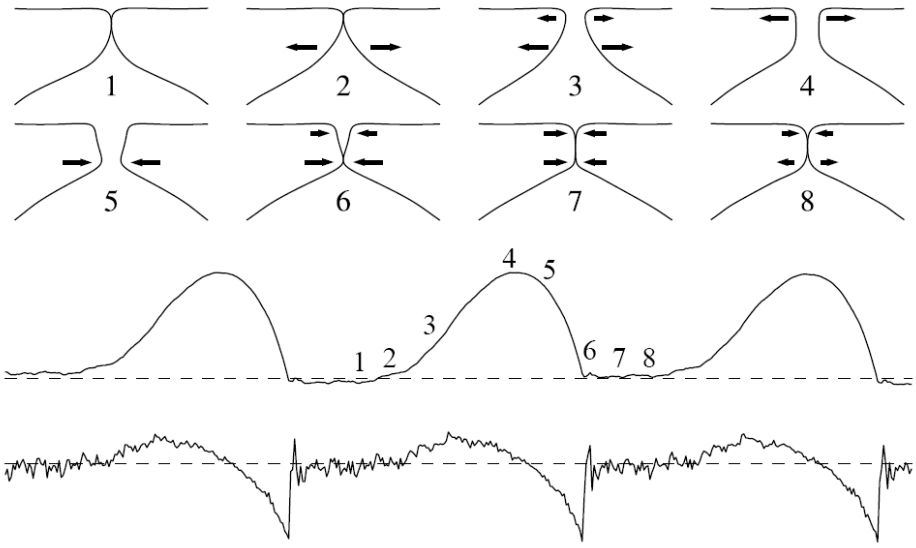

Figure 2-2: Movements of the vocal folds in speaking. The different phases of the vocal folds vibration (top), the corresponding glottal source waveform (middle) and its derivative (bottom) [51].

to contact using the musculature. The vocal folds have vibratory movement due to the effect of the airflow pressure of the respiratory system.

The movements of vocal folds vibration in one period are plotted in Figure 2-2, which contains eight states. From aerodynamic point of view, first, the airflow comes out of the lungs and arrives the closed vocal folds, and the subglottal pressure is increases with the airflow increasing at vocal folds. When the subglottal pressure increases high enough, the vocal folds is opened by high pressure in the subglottal, which corresponds the states 2, 3 and 4 in Figure 2-2, and the state 2 is called glottal opening instant (GOI). Then, the vocal folds closed together due to changes in the elasticity of the vocal folds, which corresponds to the states 5 and 6 in Figure 2-2, and state 6 is called glottal closure instant (GCI), which locates the maximum negative peak in the glottal source waveform derivative as shown in bottom of the Figure 2-2. The states from 1 to 6 are called glottal open phase. After GCI, the subglottal pressure starts to increase, and go to the next vocal folds vibration period, which corresponds the states 7 and 8, and they are called glottal closed phase.

Moreover, humans can freely and flexibly control the tension of the vocal folds to adjust the phonation types and the fundamental frequency (F_0), which is the frequency of the vocal folds vibration. It is not hard to imagine that human producing hight level

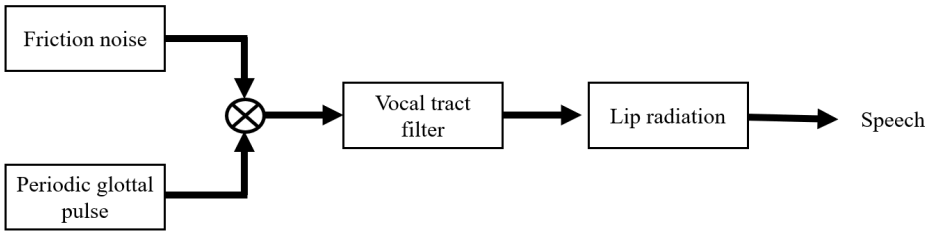

Figure 2-3: Source-filter model of speech production

arousal speech, may have higher subglottal pressure and result a higher F_0. Thus, there are a lot of studies reporting that F_0 is important information on emotional speech as mentioned in previous section. In normal speaking style, the average F_0 is about 132 Hz (male), 223 Hz (female), and 264 Hz (children). The F_0 of Soprano singer can exceed 1300 Hz [51].

2.3 Source-filter theory

According to the physical speech production mechanism, Fant proposed the source-filter theory for speech production [52], where the glottal source excitation and the vocal tract filter are considered independent contributing to speech. The source-filter theory claims that the vocal tract is assumed to be a constant filter and there is no interaction with glottal source excitation in a short-time interval of speech. The source-filter theory guides most researchers who study on speech analysis and speech synthesis.

2.3.1 Source-filter model

Based on source-filter theory, the source-filter model is widely used for studies on speech production, which is shown in Figure 2-3. Speech signal can be independently expressed as acoustical characteristics of source and filter. For voiced speech, the glottal source excitation is mainly generated by periodic vibration of the vocal folds, and it is called the glottal source waveform. There are many harmonics in the glottal source waveform, which causes energy to decay with increased frequency. For the unvoiced speech, the glottal source excitation is assumed as white noise, which is existed at somewhere of vocal tract.

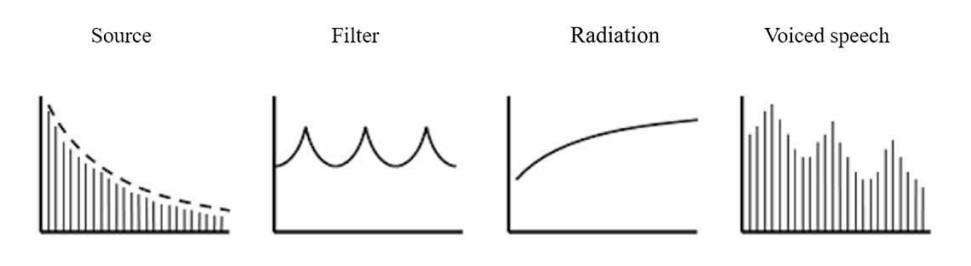

Figure 2-4: Source-filter model in frequency domain

As shown in Fig. 2-3, in the source-filter model for voiced speech, there are three components that are assumed to be linear and independent.

- **Glottal source**: two parts are considered in glottal source, which are glottal source waveform that generated by periodic vibration of the vocal folds, and combined with a aperiodic noise.

- **Vocal tract filter**: the vocal tract contains the pharyngeal cavity, oral cavity and nasal cavity. The vocal tract worked as an all pole filter for voiced speech, and the vocal tract formant frequencies are presented by these poles. The peaks on the spectrum are called the formants, and the peak located at the frequency that called formant frequency.

- **Lip radiation**: sound is delivered from vocal tract to the lips, and it seems that sound is delivered from a sphere with a smaller exit to the outside. This process is approximated as a differencing filter.

The frequency domain of source-filter model is shown in Figure 2-4. For voiced speech with normal speaking style, the spectral slop of glottal source waveform is decreasing about -12 dB/oct, while the vocal tract spectral slope is approximated flat, and the spectral slope of lip radiation increases about +6dB/oct. As a result, the spectral slope of voiced speech is decreasing approximately -6dB/oct.

Since the glottal source, vocal tract filter, and lip radiation are linearly independent, and short-time constant, the lip radiation and glottal source waveform can be combined as a derivative glottal source waveform. The source-filter model for voiced speech (Figure 2-3) can be simplified as Figure 2-5.

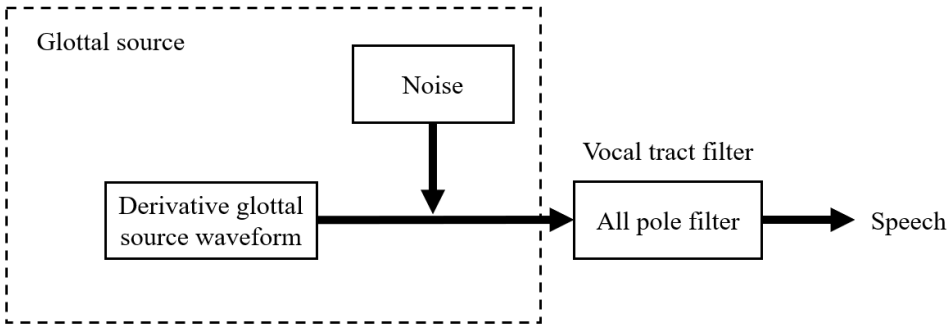

Figure 2-5: Source-filter model for voiced speech

Table 2.1: Definitions of the parameters describing the glottal flow derivative in the LF model.

T_0	one period of glottal flow
T_p	instant of the maximum glottal flow
T_e	instant of the maximum negative differentiated glottal flow
T_a	duration of the return phase
T_c	instant at the complete glottal closure
E_e	amplitude at the glottal closure instant

2.3.2 Glottal source model

According to source-filter model, many glottal source models have been proposed to simulate glottal source waveform or derivative glottal source waveform. In this part, we briefly introduce the well-known glottal source models.

- **LF model**

  LF model was proposed by Fant *et al.* [49]. The glottal flow derivative is formulated in the LF model by 6 parameters, where 5 parameters are related to time and one parameter is related to amplitude. The definitions of these parameters are listed in Table 2.1.

  A typical LF glottal flow derivative is plotted in Fig. 3-1. The explicit expression

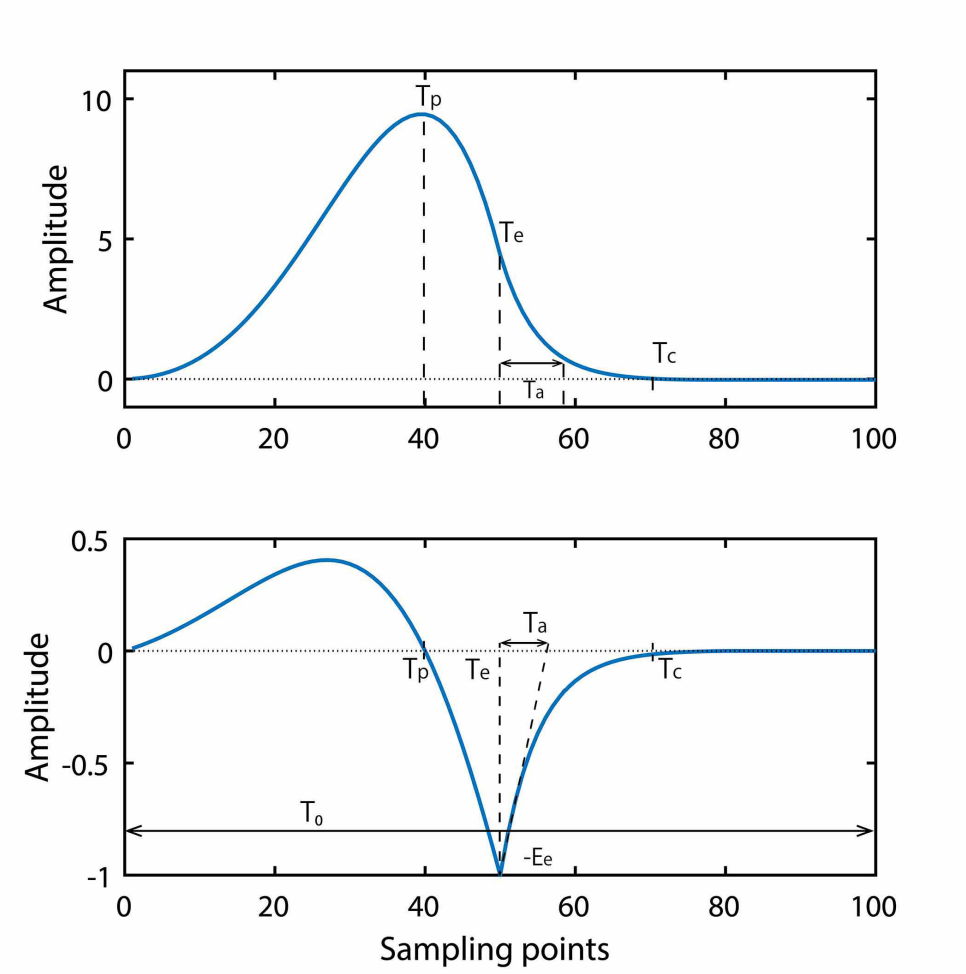

Figure 2-6: Glottal source waveform (up) and derivative glottal source waveform (bottom) of the LF model

of the LF glottal flow derivative for one fundamental period is given by:

$$u(n) = \begin{cases} E_1 e^{an} \sin(wn) & 0 \le n \le T_e \\ -E_2[e^{-b(n-T_e)} - e^{-b(T_0-T_e)}] & T_e \le n \le T_c \\ 0 & T_c \le n \le T_0 \end{cases} \tag{2.1}$$

where E_1, E_2, a, b and w are the parameters related to T_p, T_e, T_a ,E_e and T_0 [49].

- **FL model**

  Fujisaki and Ljungqvist proposed FL model to represent derivative glottal source waveform as the glottal excitation for vocal tract [53]. The equation of the FL

model consists of four parts with six shape parameters as shown in Eq. 2.2, in which six parameters are, W: open phase duration; S: pulse skew; D: the interval from glottal closure to maximum negative flow; A: slope at glottal opening; B: slope immediately before glottal closure and C: slop immediately after closure. The FL model waveform $u(n)$ for one period is given by:

$$u(n) = \begin{cases} A - \frac{2A+Ra}{R}t + \frac{A+Ra}{R^2}t^2, & 0 < t \leq R \\ a(t-R) + \frac{3B-2Fa}{F^2}(t-R)^2 - \frac{2B-Fa}{F^3}(t-R)^3, & R < t \leq W \\ C - \frac{2(C-\beta)}{D}(t-W) + \frac{C-\beta}{D^2}(t-W)^2, & W < t \leq E+D \\ \beta, & W+D < t \leq T \end{cases} \tag{2.2}$$

Where $a = \frac{4AR+6FB}{F^2-2R^2}$, $\beta = \frac{CD}{D-3(T-W)}$,and T is duration of one period glottal excitation.

- **RK model**

Rosenberg-Klatt (RK) model was proposed by Klatt for generating glottal source waveform [54], which three parameters are included for equation of the RK model. The waveform $u(n)$ for one period is given by:

$$u(n) = \begin{cases} 2an - 3bn^2, & 0 \leq n \leq OQ \times T \\ 0, & OQ \times T \leq n \leq T \end{cases} \tag{2.3}$$

$$a = \frac{27 \times AV}{4 \times (OQ^2 \times T)}, b = \frac{27 \times AV}{4 \times (OQ^3 \times T^2)}, \tag{2.4}$$

$$\tag{2.5}$$

Where T is duration of one period glottal excitation, open quotient (OQ) is the ratio of the open phase to the duration of whole period, AV is an amplitude, and a parameter for controlling the spectral tilt TL in dB down at 3k Hz.

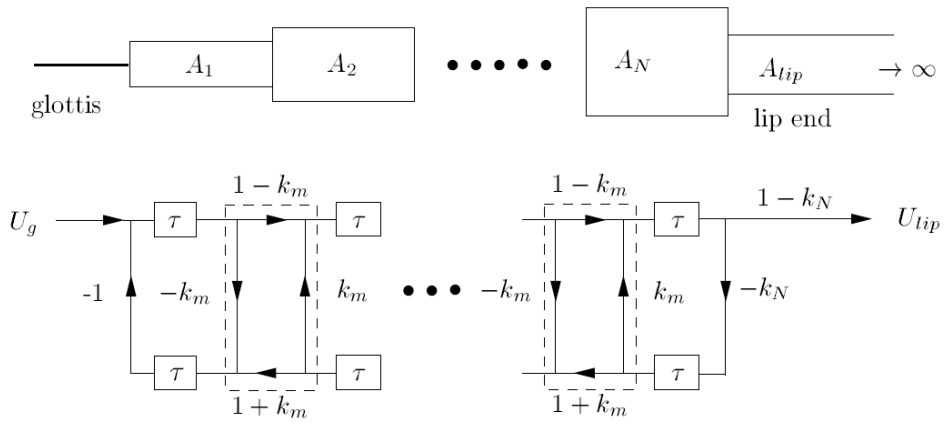

Figure 2-7: Acoustical straight tubes for vocal tract (up) and digital waveguide filter for vocal tract (bottom) [50]

2.3.3 Vocal tract filter model

The transmission of sound waveform in the vocal tract is often assumed as a straight tube, and a series of tubes are concatenated to represent the vocal tract that from glottis to lips by Sondhi [55] as shown in Fig. 2-7. A_i means cross-sectional areas of the acoustical tubes of vocal tract, and they have same length.

Source-filter model for voiced speech production, the vocal tract is generally assumed to be all pole filter, the N in Figure 2-7 is the number of acoustic tubes, in source-filter model, N is the orders of the all pole filter, and N can be calculated by following formula:

$$\begin{aligned} \frac{1}{2F_s} &= \frac{L}{c \times N}, \\ N &= \frac{2F_s L}{c} \end{aligned} \tag{2.6}$$

Where F_s is the sampling frequency, L is the length of vocal tract, and c is the speed of the sound in air. For example, the sound speed in air is often assumed to be 340m/sec, and the length of male's vocal tract is often assumed to be 17cm, so, N=$2 \times F_s \times \frac{0.17m}{340m/sec} = \frac{F_s}{1000}$. The well-known order selection rule for linear predictive coding (LPC) method is $(F_s/1000 + 2)$ [56].

There are a lot of methods to represent glottal source waveform transmition in the vocal tract. Among theses methods, digital waveguide filter is widely used to simulate characters of vocal tract [57].

The vocal tract is assumed to be lossless, because the factors of heat transfer and wall vibration, e.g., are neglected for the model. For the acoustic tubes, there are scattering occurring at the junction of two tubes, since the difference of two adjacent tubes. The scattering/reflection coefficient (k_m) are defined to the ratio of this difference:

$$k_m = \frac{A_m - A_{m-1}}{A_m + A_{m+1}} \tag{2.7}$$

Where k_m is the reflection coefficient at the junction between sections m and $m+1$.

Thus, the cross-sectional area of vocal tract tube at m section is:

$$A_m = \frac{1 + k_m}{1 - k_m} \times A_{m+1} \tag{2.8}$$

Because of each section is calculated in a relative manner, the final section in vocal tract (glottis section) is often assumed to be 1 ($A_{m+1} = 1$) [58].

LP analysis

For the well-known LP analysis, the vocal tract is modeled as a all-pole filter, the coefficient of this filter can be estimated by LP analysis with minimizing the prediction error in the least squares sense, and the current state can be predicted using the past state by these coefficient. The equation is as follow:

$$s(n) = -\sum_{i=1}^{p} a_i s(n-i) + e(n) \tag{2.9}$$

Where $s(n)$ is the speech signal at time n, a_i is the coefficient of the filter, p is the order, which is often set to: $p = F_s/1000 + 2$ (as mentioned in above), and $e(n)$ is the prediction error at time n.

The vocal tract shape (cross-section area of each section in vocal tract tube: A_m) could be calculated from the coefficient a_i. Reflection coefficients (k_m) could be calculated after obtaining the coefficient a_i of the vocal tract filter by the following equation

$$
\begin{aligned}
k(n) &= a_n(n), \\
a_{n-1}(m) &= \frac{a_n(m) - k(n)a_n(n-m)}{1 - k(n)^2}
\end{aligned}
\tag{2.10}
$$

Then, vocal tract shape could be calculated by Eqs. 2.8 and 2.9. Equations 2.10 and 2.8 are used to calculate the vocal tract shape in this study, a_i is estimated by the Auto-Regressive-eXogenous (ARX) model, see details in the next section.

Chapter 3

Estimation of glottal source waveform and vocal tract shape

The production process is often characterized by the glottal source and the acoustics of the vocal tract (the filter), which can be measured with instruments or estimated algorithmically. In recent years, some studies began to analyze emotions in speech using electromagnetic articulograph (EMA) [59–61], magnetic resonance imaging (MRI) [62–64], and high-speed videoendoscopy (HSV) [65]. However, these instrumental studies lacked flexibility for analyzing different degrees of emotional speech; moreover, it is also difficult and costly to collect sufficient recordings of emotional speech data. Therefore, many researchers attempted emotional speech analysis based on speech production models, e.g., the source-filter model [49].

There are now many methods for estimating glottal source waveforms and vocal tract shapes based on a source-filter model. A widely used method to estimate vocal tract filters is linear predication (LP) analysis, but the main problem with this method is that it is difficult to estimate vocal tract filters without glottal source effects from speech signals (source-tract interaction) [66]. To overcome this problem, Wong *et al.* estimated glottal source waveforms and vocal tract filters by LP analysis during the glottal closed phase, where there is no interaction [67]. This idea provides reliable estimations only in the long duration of glottal closure. However, it is difficult to find the glottal closed phase in real conditions, especially in the case of a very short glottal closed phase.

A simple and straightforward way to process speech signals to estimate glottal source

waveforms is inverse filtering, where glottal sources can be considered as residual signals [68, 69]. An improved method was proposed to deal with the residual signals by fitting a Lilijencrant-Fant (LF) model that is one of the widely used glottal source models [49, 70]. The advantage of this method is that a more accurate glottal source model is used, and the disadvantage of this method is source-tract interactions, as mentioned in the above paragraph.

Another method is to estimating glottal source waveforms and vocal tract shapes simultaneously based on an analysis-by-synthesis scheme. The main idea is that a glottal source model is employed as input glottal excitation to a vocal tract filter, and the auto-regressive eXogenous with the LF (ARX-LF) model is used, in which the glottal source signal is represented by the LF model glottal waveform derivative and the vocal tract transfer function is represented by the ARX filter [71]. The advantage of this method is that source-tract interaction was reduced because independent glottal sources and vocal tract models were used.

In recent years, The ARX-LF model has been widely used to estimate glottal source waveform and vocal tract shape from neutral voice[71, 72], and was recently used for analyzing singing[50]. Due to the difference in acoustic characteristics between neutral and emotional voice, an estimation approach approach for emotional voice is needed.

There are two difficult points: (1) multiple parameter fitting, and (2) Inaccurate GCI detection for the LF model, especially for emotional voice, GCI estimation is still challenge from emotional voice. In this study, for (1) multiple parameter fitting, initial values are provided from EGG signal. For (2) inaccurate GCI detection, GCI first estimated by SEDREAMS method, then corrected by EGG signal, finally GCI is shifted to obtain optimum parameter values for the ARX-LF model.

The accuracy of detected Glottal Closure Instant（GCI）and fitting multi-parameters are of great important for LF model [50, 72]. Although [70, 73] used a single shape parameter Rd [74] to fit multi-parameters of LF model. However, the glottal source of emotional speech is more complex, and single shape parameter Rd is difficult to describe complex and continuous varying degrees of emotional speech. For estimating GCI, it is still a challenge, especially in case of emotional voice.

Thus, the aim of this section is to propose schemes for estimating accurately of glottal

source waveform and vocal tract shape from emotional voice based on ARX-LF model, in which two problems (accurate estimation GCI and multi-parameters fitting) are revisited. In order to solve inaccurate GCI for the LF model, the estimated GCI is further shifted around estimated one. In order to solve multi-parameters fitting problem, there are two solutions, (1) providing initial value of the LF model parameters from EGG singal; (2) providing initial value of the LF model parameters from inverse filter method.

The remainder of this chapter introduces two kinds of proposed schemes for estimating glottal source waveform and vocal tract shape based on the ARX-LF model.

3.1 Proposed schemes based on source-filter model

Based on the source-filter theory of speech production, speech signals are modeled as output signals of a vocal tract filter with a glottal source excitation.

3.1.1 Auto-regressive eXogenous with Lilijencrant-Fant models

In this paper, ARX-LF model was used to estimate derivative glottal waves and vocal tract shapes from emotional speeches. Vocal tract filter is approximated by the ARX filter and derivative glottal wave is formulated by LF model [49]. Advantages of the LF model are: (1) LF model parameters have ability to describe different voice types [75], thus, it is possible to be used for emotional voice; (2) The LF model allows us to make use of large amount of findings in physiological, as well as perceptual correlates of model parameters [76]; (3) The LF model parameter can be measured from EGG signals, such as OQ [77].

3.1.2 Lilijencrant-Fant model

The LF model has six parameters to represent the derivative of the glottal flow. These are five parameters concerning time T_p, T_e, T_a,T_c, T_0 and one parameter concerning amplitude E_e. Where glottal opening instant (GOI) is set to 0, and T_0 is the end of the period, T_p is phase of maximum open of glottis, T_e is open phase of glottis, T_a is return phase, T_c is end of return phase and E_e is the amplitude at glottal close instant (GCI) point. The LF model in time domain is formulated as Equation 1.

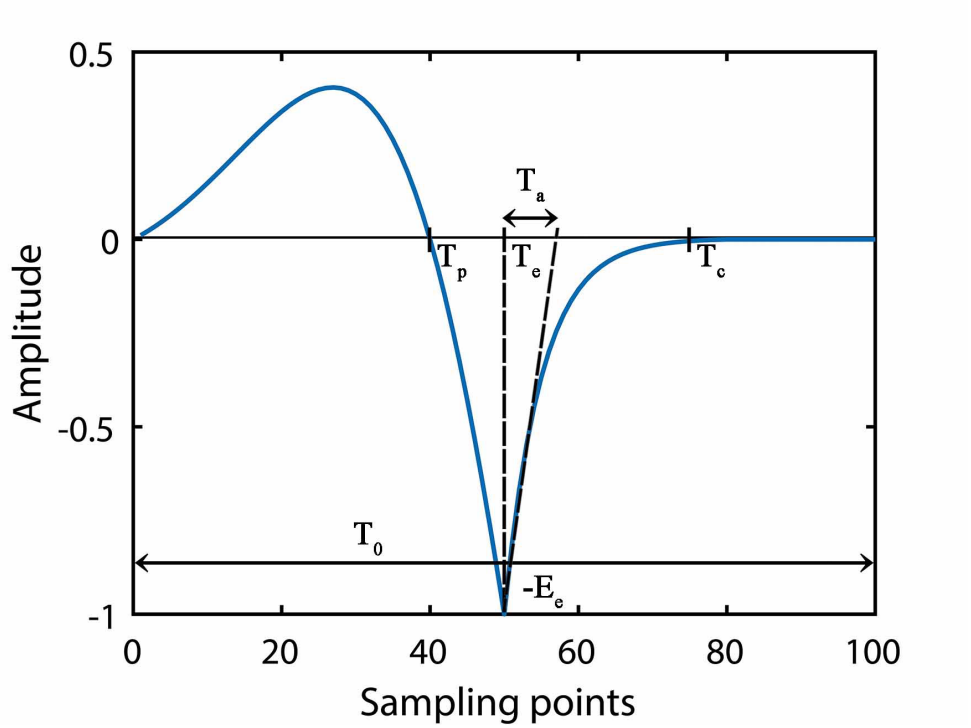

Figure 3-1: Lilijencrant-Fant model

$$u(t) = \begin{cases} E_1 e^{at} \sin(wt) & 0 \le t \le T_e \\ -E_2[e^{-b(t-T_e)} - e^{-b(T_0-T_e)}] & T_e \le t \le T_c \\ 0 & T_c \le t \le T_0 \end{cases} \tag{3.1}$$

The parameters E_1, E_2, a, b and w are implicitly related to T_p, T_e, T_a ,E_e and T_0 [49]. A typical LF waveform is depicted in Figure 3-1.

3.1.3 Auto-regressive eXogenous model

The ARX model simulates a vocal tract filter. Speech production process can be assumed to be a time-varying IIR system as follow:

$$s(n) + \sum_{i=1}^{p} a_i(n)s(n-i) = b_0 u(n) + e(n) \tag{3.2}$$

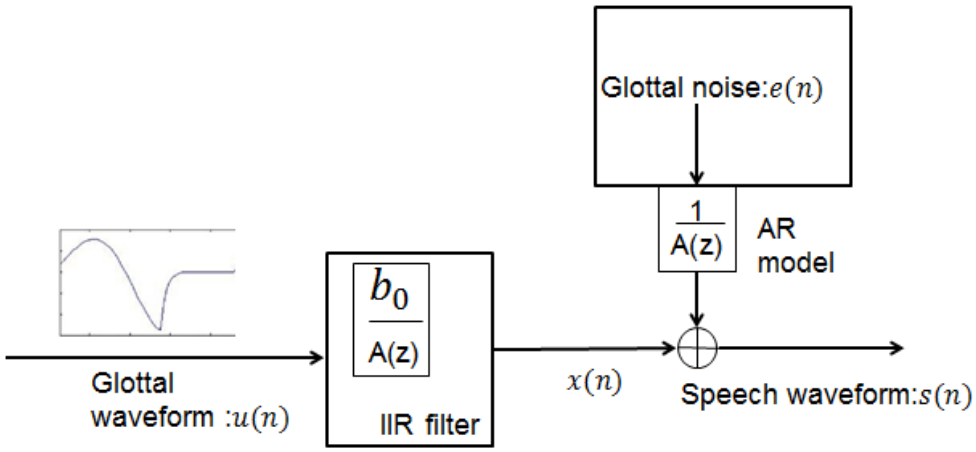

Figure 3-2: Auto-regressive eXogenous model

where $s(n)$ is the observed speech signal and $u(n)$ is the derivative of glottal waveform (LF waveform) at time n, a_i and b_0 are time-varying coefficients of the IIR filter, p is filter order, and $e(n)$ is the residual.

The output signal of the LF model acts as an input signal $u(n)$ to the vocal tract filter. Equation(2) is called ARX model, as illustrated in Figure 3-2. Output signal $x(n)$ plays a role in periodic components and $e(n)$ plays a role in non-periodic components in speech. In this model, the residual $e(n)$ in the speech production and its power in the voiced sound is obtained from the equation error. The vocal tract transfer function is defined as follow:

$$H(z) = \frac{B(z)}{A(z)} = \frac{b_0}{1 + a_1 z^{-1} + ... + a_p z^{-p}} \tag{3.3}$$

3.2 Proposed scheme with electroglottograph signal

Using ARX-LF model, the first step is to detect the position of the start point and end point of each glottal vibration period. As we known, among all events in one glottal vibration period, GCI located at the discontinuity in the derivative of glottal waveform, and the energy of excitation is the strongest in one glottal vibration period. Thus, compared with others, GCI is easily to be detected and frequently used for the estimation methods of glottal source waves. However, accurately estimate GCI is still a challenge, especially for emotional voice, and it is greatly affects the estimation results of ARX-LF

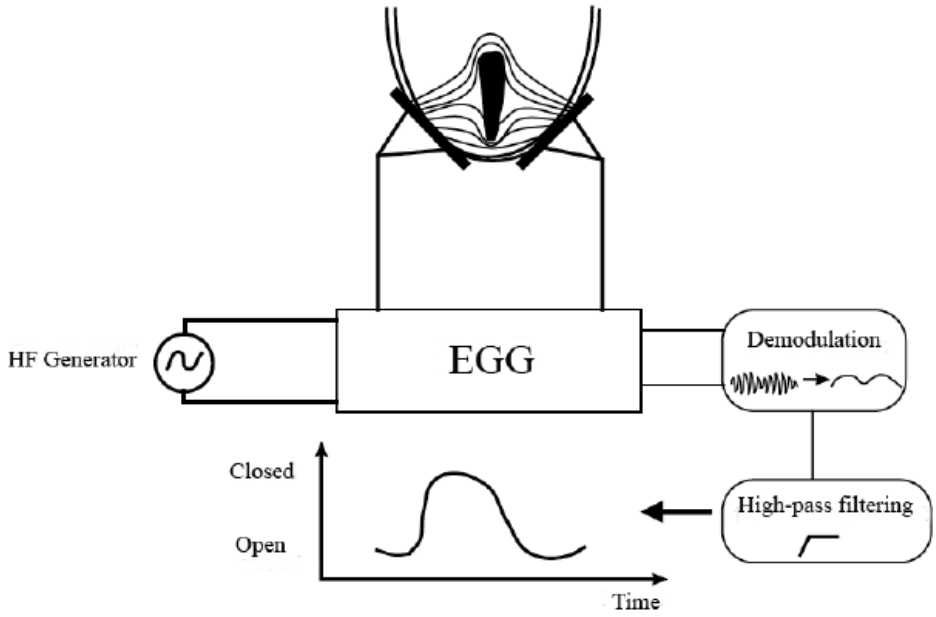

Figure 3-3: Theory of electroglottograph recorder [78]

model. This section describes solving the two problems, accuracy of GCI detection for LF model and fitting multi-parameters of LF model.

3.2.1 Estimation of open quotient from electroglottograph signal

In order to solve fitting multi-parameters of the LF model, reducing variable region of parameter values is considered, and a accuracy initial parameter value of LF model Open Quotient (OQ) is provided in advanced. OQ is always considered as T_e of LF parameters, and OQ can be calculated by GOI and GCI (OQ=$(GCI - GOI)/T_0$). Because most methods estimate GCI and GOI from dEGG signal as a standard values. Thus, the GCI and GOI are estimated from dEGG signal and OQ is calculated as a initial value for LF model in this paper. However, the accuracy of OQ directly calculated from GCI and GOI is difficult to satisfy the LF model, because estimation of GOI sometimes incorrect. In order to solve this problem, calculation of OQ is further processed using a smooth line (because OQ can not abruptly change largely between two continuous periods)for initial values of T_e for LF model as shown in Figure 3-4.

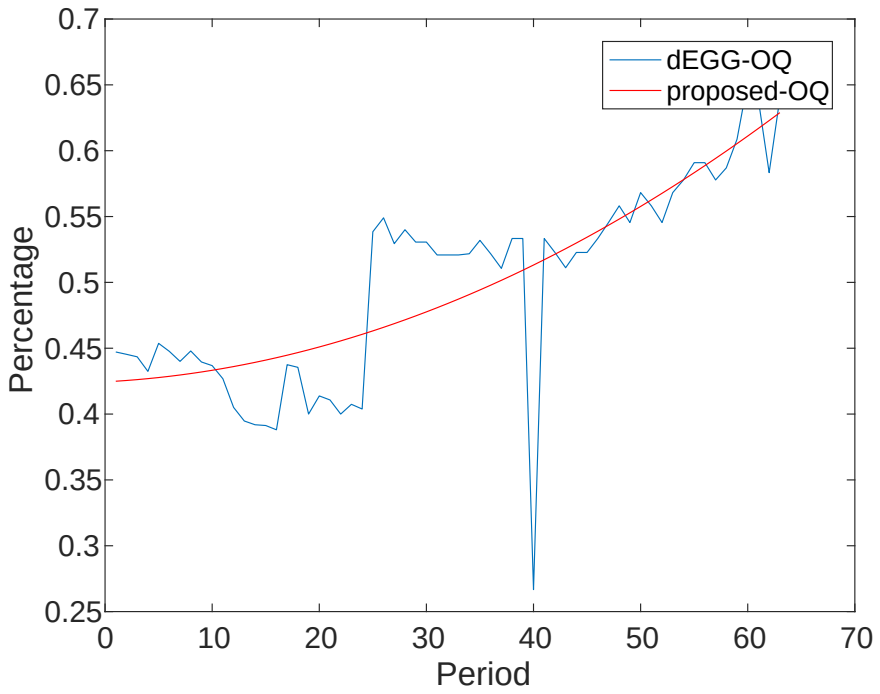

Figure 3-4: An example of smooth open quotient

Electroglottograph signal

EGG is a noninvasive measurement method for detecting of the vibration of vocal folds during voice production [78]. The EGG (see Figure 3-3) has two electrodes are placed on the surface of the neck. Thus, the EGG records voltage between two electrodes variate in the transverse electrical impedance of the larynx and nearby tissues. The impedance is higher when the vocal folds are open, and lower in vocal folds closure.

3.2.2 Glottal closure instant detection

Due to fitting the LF model parameters are based on speech signals, estimated GCI from dEGG signal is difficult directly applied to speech signals, this is because, there is a distance between EGG signal (at vocal folds) and speech signal (after lips). Therefore, the estimation of GCI from speech signals need to be considered. There are many different methods to estimate GCI from speech signals in recent years, such as the Dynamic Programming Projected Phase-Slope Algorithm (DYPSA) [79, 80], Yet Another GCI Algorithm (YAGA) [81], Zero-Frequency Filter (ZFF) [82], Speech Event Detection using the Residual Excitation and A Mean-based Signal (SEDREAMS) [83] and SE-VQ [84].

The performance of these were summarized by Drugman [85] and Kane [84], the SE-DREAMS and SE-VQ have comparable performance for different phonation types. Thus, the SEDREAMS method was selected to detect GCI_0 as a initial value for LF model in this paper.

This section provides two accurate initial values of parameters (OQ and GCI) for the ARX-LF model.

3.2.3 Scheme of analysis

The estimation procedure for a period of glottal source wave is plotted in Figure 3-5, in which two main processes are included. In the first process, optimal LF parameters and vocal tract coefficients can be obtained with fixed GCI. The initial values of LF parameters are used for synthesizing glottal wave $u(n)$, the glottal wave $u(n)$ is then exploited to re-synthesis speech wave $x(n)$ using the ARX model in the parameters of the vocal tract filter updated within each period in mean square error (MSE) sense for $e(n)$ with the helps of Least Mean Squares Error (LMSE) method. In each iteration of this optimization process, the parameters of LF model are changed randomly within a variation range, and glottal wave can be re-generated using these parameters.

For the initial values and the variation regions, since the GCI parameter is the most significant changes in glottis vibration within a period. We first estimate GCI_0 by SE-DREAMS method in proposed method. The distance between two continuous GCI_0 is regarded as one period T_0, and the estimated GCI_0 are initial for LF model. Open Quotient (OQ), which is equivalent to T_e of LF model, can be estimated from the differential electroglottograph (dEGG) signal. For other initial values of parameters, we set initial values vibration region: $0.65\times T_e \leq T_p \leq 0.85\times T_e$, the start point of one period is: GOI=GCI- ($T_e \times T_0$), $T_a \leq 0.1\times T_0$. For E_e, we use the maximal amplitude of the speech signal around GCI locations.

In the second process, we want to obtain more accurate LF parameters and vo-cal tract coefficients. The GCI parameters shift around the initial GCI, and the first process is updated again for the shifted GCI. For the given GCI, the iteration process in the minimization of MSE (MMSE) optimization is set to 2000. The optimal glottal source parameters and vocal tract coefficients are estimated by MMSE.

The LMSE method

ARX-LF model, Z transformation of speech waveform is $X(Z)$, AR filter is $B(Z)/A(Z)$, glottal source wave is $U(Z)$ and residual is $E(Z)$.

$$X(Z) = \frac{B(Z)U(Z) + E(Z)}{A(Z)} \tag{3.4}$$

Where,

$$\begin{aligned} A(Z)X(Z) &= B(Z)U(Z) + E(Z) \\ A(Z) &= 1 - \sum_{k=1}^{p} a_k(n) Z^{-k} \\ B(Z) &= b_0 \end{aligned} \tag{3.5}$$

Equation 3.5 in time domain is:

$$\begin{aligned} e = x_0 - X_a - Ub = x_0 - [X|U][\frac{a}{b}] = x_0 - Fh, \\ F = [X|U],\ h = [\frac{a}{b}], \\ e = \begin{bmatrix} e_n \\ e_{n-1} \\ \vdots \\ e_{n-N+1} \end{bmatrix},\ x_i = \begin{bmatrix} x_{n-i} \\ x_{n-i-1} \\ \vdots \\ x_{n-i-N+1} \end{bmatrix},\ u_j = \begin{bmatrix} u_{n-j} \\ u_{n-j-1} \\ \vdots \\ u_{n-j-N+1} \end{bmatrix} \\ a = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_p \end{bmatrix},\ U = u_0,\ b = b_0 \end{aligned} \tag{3.6}$$

u_j is LF model wave, because of sampling frequency was 12000 Hz, p was set to 14 in this

Table 3.1: The values of synthesis parameters

Source				Filter	Noise
f_0	OQ	T_p/T_e	T_a	Vowel types	SNR
$90 \sim 135$	0.3:0.05:0.9	0.65:0.05:0.85	0.03,0.08	2 vowels	30:10:50

paper. AR filter coefficients h were calculated by $e^{\mathrm{T}}e$ (LMSE):

$$\begin{aligned} e^{\mathrm{T}}e &= (x_0 - Fh)^{\mathrm{T}}(x_0 - Fh) \\ &= x_0^{\mathrm{T}}x_0 - x_0^{\mathrm{T}}Fh - h^{\mathrm{T}}F^{\mathrm{T}}x_0 + h^{\mathrm{T}}F^{\mathrm{T}}Fh \end{aligned} \tag{3.7}$$

$$\begin{aligned} -F^{\mathrm{T}}x_0 - F^{\mathrm{T}}x_0 + 2F^{\mathrm{T}}Fh &= 0 \\ F^{\mathrm{T}}Fh &= F^{\mathrm{T}}x_0 \end{aligned} \tag{3.8}$$

$$h = (F^{\mathrm{T}}F)^{-1}F^{\mathrm{T}}x_0 \tag{3.9}$$

3.2.4 Evaluation of proposed schemes using synthetic voices

This section describes the evaluation of estimation method of glottal source waves and vocal tract shapes. Especially when two accurate initial parameter values are provided. Frequently used method for evaluating estimation algorithm is to estimate synthetic vowels whose reference parameter values are known[72, 86, 87], and compare estimated parameter values with referenced parameter values. The synthesized voices are constructed by source-filter model, in which LF model as derivative glottal wave and vocal tract shapes that are /a/ and /i/ taken from [76]. In order to synthesize voices containing a wide range of pronunciation types, and simulate complex emotions as human as possible, two synthetic vowels with extensive changes of glottal source and vocal tract filter settings are used for evaluating the algorithm. Glottal source parameters are done in a similar method with [86, 87] in Table 3.1. Difference is that variable f_0 are more complex and it is estimated from real speech in this paper. Thus, the total numbers of synthesized LF waves are 14040 for /a/ and /i/.

The synthesized vowels (14040 periods) are estimated by proposed algorithm, note that although there is no EGG signal when using synthesized vowels, reference parameter

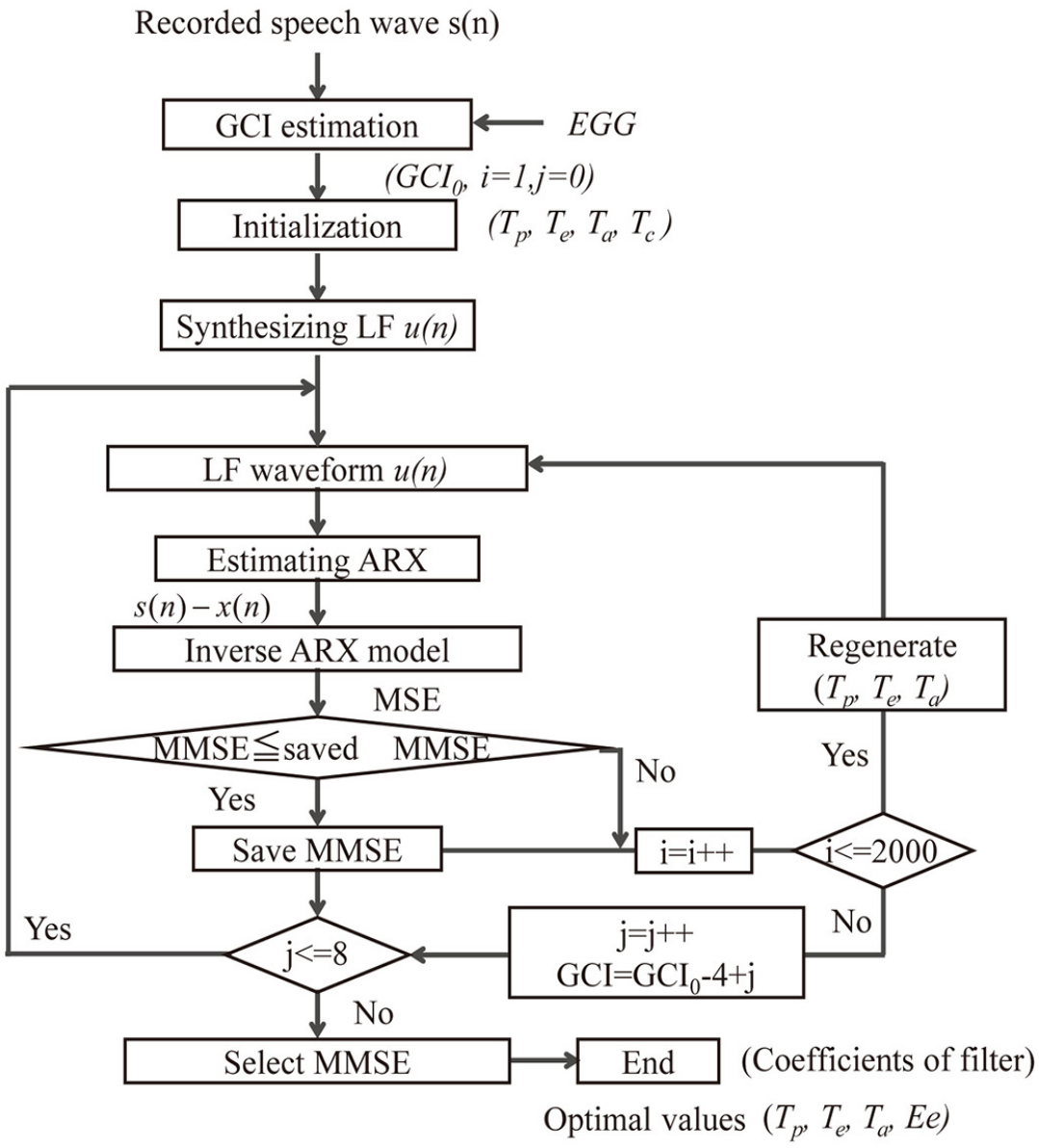

Figure 3-5: Estimation algorithm

T_e values were known in each period. Thus, we assume that reference parameter T_e values are as OQ (calculated from dEGG signal), and T_e of initial values vibration region for estimation algorithm is from that reference $T_e \times 0.95$ to reference $T_e \times 1.05$. For each parameter $\beta \in \{T_p, T_e, T_a, T_c, F_1, F_2\}$ and its estimated $\hat{\beta}$, the absolute percentage error γ is calculated as follows:

$$\gamma = \frac{|\hat{\beta} - \beta|}{\beta} \times 100\% \tag{3.10}$$

The average γ of estimated parameters were showed in Table 3.2,

Examples of the analysis results are shown in Figure 3-7, which demonstrates the

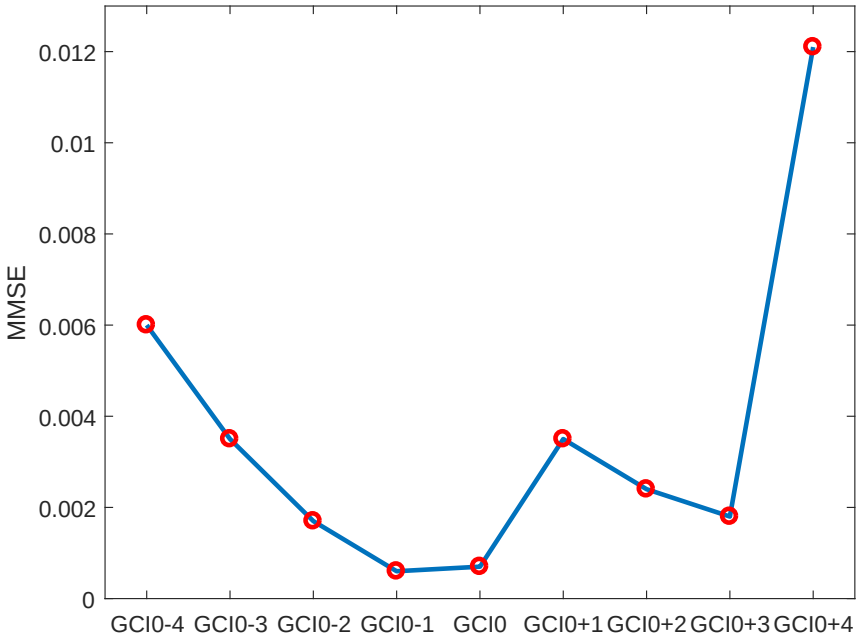

Figure 3-6: MMSE effected by inaccurate glottal closure instant detection

Table 3.2: Average γ (%)

	T_p	T_e	T_a	T_c	$F1$	$F2$
γ	5.75	3.41	34.6	5.65	2.05	0.59

effectiveness of the analysis approach for decomposing source and filter infor-mation. The glottal source wave and vocal tract shape are estimated from one synthesized vowel /a/. The solid lines in Figs. 4 (a), (b), and (c) plot the original glottal source wave, original vocal tract shape, and original speech wave, while, the dashed lines plot the estimated glottal source wave, estimated vocal tract shape, and estimated speech wave, respectively. Figs. 4 (d), (e), and (f) corre-sponded to (a), (b), and (c) in the frequency domain, respectively. For the vocal tract shape, it was calculated by Wakita's method [16]. A 44100-Hz sampling frequency with a 44th order of the vocal tract filter was applied to synthesize the voice in the synthesis step, while, a 12000-Hz sampling frequency with a 14th order of the vocal tract filter was utilized in the analysis step. According to equation 2.6, the length of original vocal tract shape (17 cm) is shorter than the length of the estimated vocal tract shape (21 cm) when the assumed sound speed is 340 m/s in the vocal tract (see Fig. 4(b)) when the assumed sound speed is 340 m/s in the vocal tract

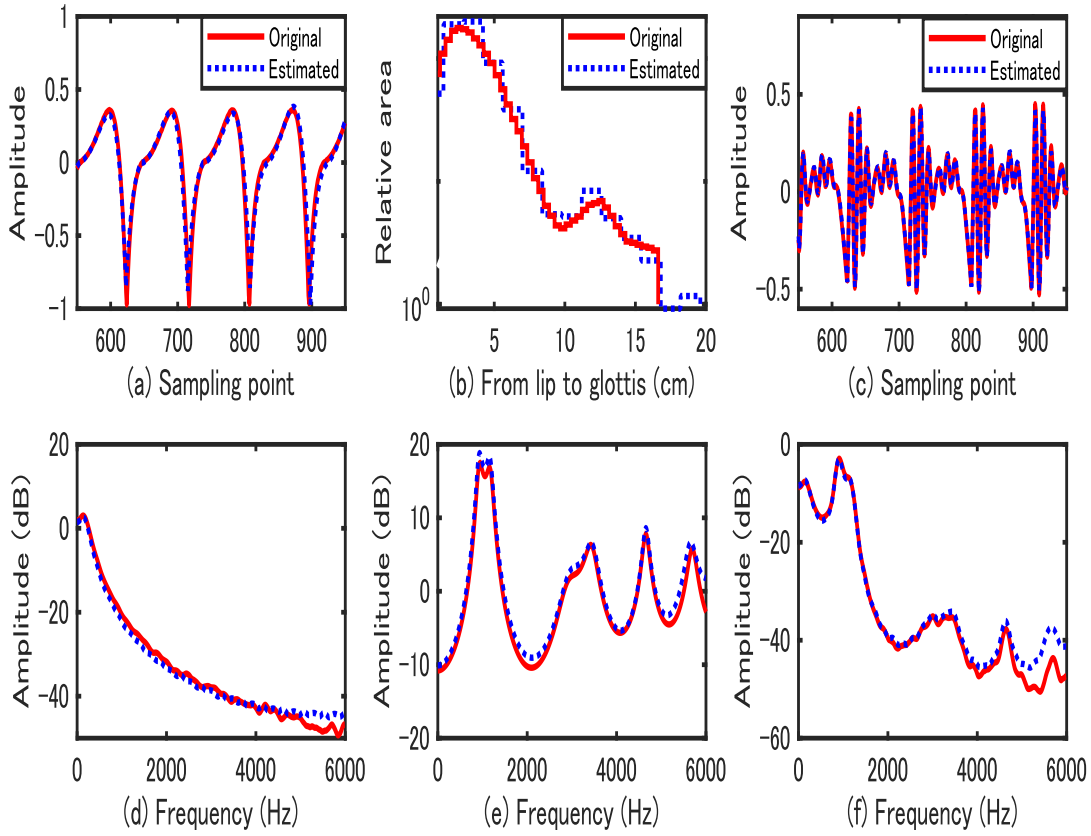

Figure 3-7: (a) Original glottal source wave (solid line) and estimated glottal source wave (dashed line), (b) original vocal tract shape (solid line) and estimated vocal tract shape (dashed line), (c) original speech wave (solid line) and estimated speech wave (dashed line); (d) (e) (f) corresponds with (a) (b) (c) in frequency domain, respectively.

(see Fig. 4(b)). These results suggest that the estimated glottal source wave, vocal tract shape and speech wave match the original data quite well in both the time and frequency domains.

Figure 3-6 shows how inaccurate GCI_0 detection affects the MMSE. GCIs are shifted 4 points for left and right centered at estimated GCI_0, MMSE are calculated at each GCI. We can see that accuracy of the estimated ARX-LF model is significantly affected by GCI locations and the best value of MMSE is around GCI_0, Thus, shifting GCI_0 can make estimation results more accurate.

Table 3.2 shows that most of the parameters of ARX-LF can be correctly estimated, and errors were less than 6 percentage except for T_a. Because T_a value is the smallest among those parameters, error of estimated T_a (34.6%) is the largest compared with others.

3.2.5 Estimation of glottal source waveform and vocal tract shapes from real emotional vowels

Data selection

Four actors (three males and a female) participated to utter the vowel sound /a/ with eight different emotional states: neutral, happy, sadness, afraid, disgusted, relaxed, surprised, and anger. The neutral sound /a/ was uttered one time as the reference. The other seven states were uttered in three different degrees (weak, normal, and strong), respectively. Therefore, there are 88 utterances ($1+7 \times 3=22$ for each speaker) in total. The speech signals and EGG signals were recorded together.

The actors' utterances were evaluated by listening experiments for the purpose of evaluation the speech emotional state form the point of view of the speech perception. By utilizing the dimensional approach, category and degree of emotions were described in the valence and arousal space. Since there are three different degrees for each emotions have been stored in the database, the dimensional approach was choose to evaluate emotions.

Ten natives participated in the listening test. For arousal and valence evaluations, a 7-point scale from -3 to 3 (-3: very negative to 3: very positive for valence and -3: very calm to 3: very excited for arousal) was used, and the mean value of the evaluation from all participants was calculated. The emotional speeches from four basic emotion categories (neutral, joy, sadness, and anger) with the strong degree were selected to further discuss the glottal source waves and vocal tract shapes. The mean values of the evaluation from selected speech data were described in the V-A space, as shown in Figure 3-8.

The parameters of the ARX-LF model corresponded to glottal source and vocal tract were extracted by utilizing the analysis-by-synthesis approach based on proposed schemes.

The parameters of the ARX-LF model corresponded to glottal source and vocal tract were extracted by utilizing the analysis-by-synthesis approach based on proposed schemes. The estimated results of glottal source waves and vocal tract shapes, which showed in Figures 3-9 and 3-10, were estimated from emotional speeches uttered by four different speakers. As shown in Figure 3-9, there are more high frequency components could be observed from the estimated spectra of glottal source waves. Same tendency has been observed in a previous study [88]. Figure 3-10 shows the front of vocal tract shape. Compared with sadness, anger shows the largest area function, while the area function

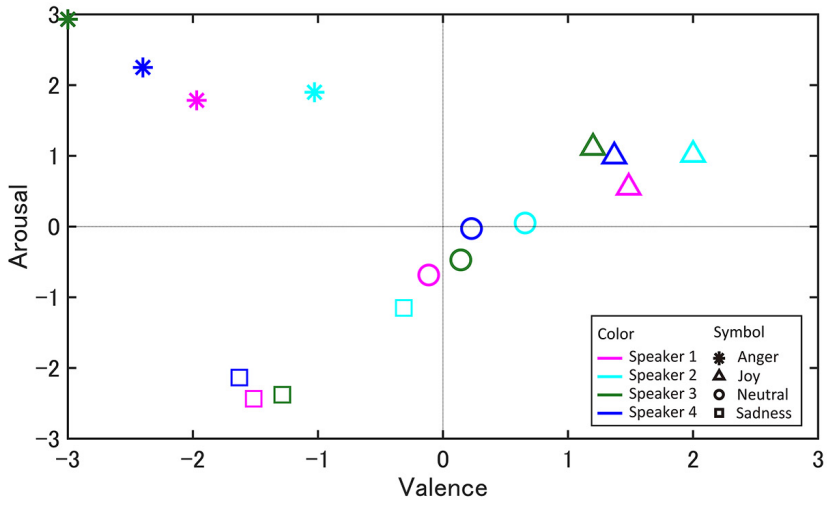

Figure 3-8: Selected voice data in V-A space

of neutral and joy was between sadness and anger. These results are consistent with the previous finds using the MRI and EMA data [61, 64].

3.3 Proposed scheme without electroglottograph signal

3.3.1 Introduction

There are now many methods for estimating glottal source waveforms and vocal tract shapes based on a source-filter model. A widely used method to estimate vocal tract filters is linear predication (LP) analysis, but the main problem with this method is that it is difficult to estimate vocal tract filters without glottal source effects from speech signals (source-tract interaction) [66]. To overcome this problem, Wong *et al.* estimated glottal source waveforms and vocal tract filters by LP analysis during the glottal closed phase, where there is no interaction [67]. This idea provides reliable estimations only in the long duration of glottal closure. However, it is difficult to find the glottal closed phase in real conditions, especially in the case of a very short glottal closed phase.

A simple and straightforward way to process speech signals to estimate glottal source waveforms is inverse filtering, where glottal sources can be considered residual signals [68, 69]. An improved method was proposed to deal with the residual signals by fitting a Lilijencrant-Fant (LF) model that is one of the widely used glottal source models [49, 70]. The advantage of this method is that a more accurate glottal source model is used, and

the disadvantage of this method is source-tract interactions, as mentioned in the above paragraph.

Another method is to estimating glottal source waveforms and vocal tract shapes simultaneously based on an analysis-by-synthesis scheme. The main idea is that a glottal source model is employed as input glottal excitation to a vocal tract filter, and the auto-regressive eXogenous with the LF (ARX-LF) model is used, in which the glottal source signal is represented by the LF model glottal waveform derivative and the vocal tract transfer function is represented by the ARX filter [71]. The advantage of this method is that there is no source-tract interaction because independent glottal sources and vocal tract models were used. However, it is difficult to optimize multiple parameters simultaneously.

To solve this problem, Li *et al.* proposed an iterative algorithm [89] to estimate accurate glottal source waveforms and vocal tract shapes, in which an electro-glotto-graph (EGG) signal was used to estimate initial values for the iteration. It is not convenient to always use EGG.

In this subsection, instead of EGG signal, we first obtained the initial values of the LF model parameters using an inverse filter [70]. Then, the accurate glottal source waveforms and vocal tract shapes were estimated simultaneously based on the ARX-LF model using the iterative algorithm [89] .

3.3.2 Estimation of glottal source waveform and vocal tract shape

The procedure for estimating glottal source waveforms and vocal tract shapes is shown in Fig. 3-11, and it includes two steps. In the first step, instead of accuracy, initial values are prepared for the next step. The main step is the second step, in which precise glottal source waveforms and vocal tract shapes are estimated simultaneously by the proposed scheme based on the ARX-LF model.

3.3.3 Initialization

The objective of this step is to determine the period for the LF model. In one period of the LF model waveform, the glottal closure instant (GCI) is a discontinuity location, as shown in Fig. 3-1, and it is easier to be detected than other locations in one period of the glottal

source waveform. Thus, GCI is detected first, and the distance between two continuous GCIs is regarded as one period T_0. The Speech Event Detection using the Residual Excitation And a Mean-based Signal (SEDREAMS) is used for detecting the GCI [85], because it provides more accurate results. The detected GCI by the SEDREAMS method is remarked as GCI_0.

John [87] evaluated three state-of-the-art the inverse filter methods: (1) closed-phase inverse filtering method (CPIF) [91], (2) iterative and adaptive inverse filtering (IAIF) [68], and (3) mixed-phase decomposition based on the complex-cepstrum (CCEPS) [69]. By testing on the synthetic speech, the IAIF showed the best performance among these three methods, the structure of IAIF is plot in Figure 3-12. Thus, the IAIF is used, and the LF model are used to obtain the initial values of the LF model for simultaneous estimation step. Dynamic programming (DyProg-LF) is employed to fit the LF model parameters, and estimated glottal source parameters (LF model) are remarked as T_p^0, T_e^0, T_a^0 and E_e^0. The detailed implementation of the DyProg-LF algorithm was described in [70, 92].

3.3.4 Implementation of simultaneous estimation

In this step, a simultaneous estimation algorithm is implemented to accurately estimate a glottal source waveform and vocal tract shape based on the ARX-LF model. There are two processes in this step. In the first process, the LF model parameters and vocal tract filter coefficients can be obtained with a fixed GCI. The initial values of the LF model parameters (T_p^0, T_e^0, T_a^0, E_e^0) are used for synthesizing glottal source waveform derivative $u(n)$, and $u(u)$ is then exploited to synthesize $x(n)$ using the ARX model, and the ARX model parameters can be estimated with mean square error (MSE) sense for $e(n)$ by the least square (LS) method. In each iteration of this optimization process, the LF model parameters are regenerated around the initial values of T_p^0, T_e^0, T_a^0 and E_e^0, and a glottal source waveform derivative can be re-generated using these parameters. Let a vector β:

$$\beta(n) = [-s(n-1) \cdots - s(n-p)u(n)]^{\mathrm{T}} \tag{3.11}$$

and the ARX model coefficients γ are represented as follows:

$$\gamma = [a_1 \cdots a_p b_0]^{\mathrm{T}} \tag{3.12}$$

Table 3.3: Average estimation errors (ε) for synthesized vowels of two methods

	Glottal source				**Vocal tract**	
	$T_p(\%)$	$T_e(\%)$	$T_a(\%)$	$E_e(\%)$	$F_1(\%)$	$F_2(\%)$
IAIF-Dyprog-LF	24.0	19.8	50.4	82.2	9.4	6.1
Proposed	11.4	9.5	60.0	23.2	2.3	1.1

γ can be calculated as follows:

$$\gamma = [\beta(n)\beta(n)^{\mathrm{T}}]^{-1}\beta(n)s(n) \tag{3.13}$$

In the second process, we can estimate more accurate LF model parameters and vocal tract coefficients. Since the performance of the ARX-LF model is affected by the accuracy of GCI which was reported by Lu [50], we suggested shifting the parameters of GCI around the value of GCI_0 further. Then, the first process works again for each shifted GCI. For the given GCI each time, the iteration processing in the minimization of mean square error (MMSE) optimization is set to 2000. The accurate glottal source parameters and vocal tract filter coefficients are finally estimated by MMSE.

The sampling frequency was 12000 Hz and p was set to 14 in this paper. The estimation length of the frame is three periods of glottal source waveforms, and the shift frame is one period of a glottal source waveform.

3.3.5 Evaluation of proposed schemes using synthetic voices

First, the synthetic vowels were used to test the proposed estimation method and the IAIF with Dyprog-LF method in [70]. The advantage of testing on synthetic vowels is that the accuracy of the proposed method can be investigated by comparing the estimated parameter values of glottal source waveforms and vocal tract shapes with referenced parameter values. Then, the glottal source waveforms and vocal tract shapes of real vowels are estimated to test the proposed estimation method and the IAIF with Dyprog-LF.

Synthesized vowels

The synthesized vowels are produced according to the source-filter model, in which a glottal source waveform derivative is generated by the LF model, and the vocal tract filters are taken from five Japanese vowels (/a/, /e/, /i/, /o/ and /u/) using Kawahara's method [76]. This is because of formant frequencies of vocal tract in these vowels vary widely, especially the first and second important formant frequencies (F_1 and F_2). A larger number of synthetic vowels with a wide range of LF model parameter values are used in this paper. The LF model parameters are varied: T_e: 0.3 to 0.9 with steps of 0.05; T_p/T_e: 0.65 to 0.8 with steps of 0.05 steps; T_a: 0.03, 0.08, within a suggested range in [86]. In order to synthesize more realistic vowels, the fundamental frequency (F_0) is obtained from a real vowel, 18 different F_0 ranged from 90 to 140 Hz are used for synthesizing vowels. The total number of synthesized vowels with 9360 ($4[T_p] \times 13[T_e] \times 2[T_a] \times 18[F_0] \times 5[filter]$) different conditions are used for testing the proposed method and the IAIF with Dyprog-LF method.

3.3.6 Results and discussion

Estimated values of the LF model parameters and F_1 and F_2 of the vocal tract were evaluated by the reference values Let the reference values as a vector $\theta \in \{T_p, T_e, T_a, E_e, F_1, F_2\}$ and the estimated values as vector $\hat{\theta}$. The error (ε) between estimation and reference values can be calculated as:

$$\varepsilon = \frac{|\hat{\theta} - \theta|}{\theta} \times 100\% \tag{3.14}$$

The average errors (ε) are listed in Table 3.3. Estimation errors of a glottal source are less than 13% except for those of T_a, since T_a was the smallest of all parameters as the denominator in Eq. 3.14 and the error was 58%. Estimation errors are less than 2% for the vocal tract. Figure 3-13 shows an example of estimated results, in which the glottal source waveform and vocal tract shape are estimated from a synthesized vowel /a/. As shown in Fig. 3-13, estimated glottal source waveforms and vocal tract shapes are very similar to the original ones in the time domain and frequency domain. The length of the vocal tract shapes are different between estimated and original one because the sampling frequency and the order of the vocal filters is different between the synthesis

Table 3.4: Average estimation errors (ε_{OQ}), and F_1 and F_2 are estimated by the ARX model and Praat from five males and five females

Male (M) and Female (F)	M1	M2	M3	M4	M5	F1	F2	F3	F4	F5
$\varepsilon_{OQ}(IAIF - Dyprog - LF)$ [%]	9.0	15.3	2.0	18.8	11.1	10.8	12.6	8.9	0.2	10.3
$\varepsilon_{OQ}(Proposed)$ [%]	12.6	13.0	1.5	0.2	7.1	4.9	1.5	2.7	9.0	10.8
$F_{1\ ARX}$ (Hz)	773	756	680	803	709	1031	1031	797	1125	1043
$F_{1\ Praat}$ (Hz)	734	740	675	788	686	1064	997	722	1098	1023
$F_{2\ ARX}$ (Hz)	1348	1189	1213	1313	1137	1611	1553	1025	1605	1459
$F_{2\ Praat}$ (Hz)	1334	1194	1213	1315	1129	1587	1506	1013	1587	1464

step (Kawahara's method: 44100-Hz sampling frequency with an order 44th order) and the analysis step (ARX-LF: 12000-Hz sampling frequency with 14th order).

Table3.3 shows that the estimation accuracy of the proposed method is higher than that of IAIF with Dyprog-LF.

3.3.7 Estimation of glottal source waves and vocal tract shapes from actual vowels

The voiced vowel (/a/) was pronounced by five male and five female Japanese speakers. The speech signals were recorded together with electroglottographic (EGG) signals. Thus, there are ten real voiced vowels used to test the proposed method and the IAIF with Dyprog-LF.

Results and discussion

There is no direct reference parameter available for the glottal sources and vocal tracts in real vowels. To evaluate glottal sources, as a reference value, we calculated the open quotient (OQ) to evaluate the accuracy of the proposed method. The recorded vowels were analyzed by the proposed method to estimate T_e, which is often considered as the

OQ_{LF}, and referenced OQ_{EGG} was calculated from a differentiated EGG (dEGG) signal by searching glottal opening instant (GOI) and GCI. Thus, the estimation errors can be calculated by Eq. 3.14. The estimation errors (ε) are listed in Table 3.4. Compared with the value of OQ obtained from the dEGG signal, the accuracy of the proposed method is higher than that of IAIF with Dyprog-LF.

Vocal tract parameters F_1 and F_2 were estimated by the proposed method and a widely used formant extractor (Praat), respectively. Results are shown in Table 3.4, where the values of F_1 and F_2 estimated by the proposed method are very similar to those extracted by Praat. Furthermore, for ten speakers, most of the values of F_1 estimated by proposed method are a little higher than those estimated by Prrat, and the values of F_2 of two the methods are mostly the same. Therefore, the proposed method can effectively estimate the vocal tract parameters.

Moreover, a continuous speech (/aiueo/) pronounced by a male speaker was challenged by the proposed method. It is impossible to discuss glottal source parameters since there was no EGG signal recorded together with a speech signal. The waveform and the spectrogram of the original speech signal and the resynthesized speech signals by the ARX-LF model are plotted in Fig. 3-14, which shows that the original speech signal is very similar to the resynthesized speech signal in the time and frequency domain. The spectrogram clearly shows that the formant frequencies are the same as the original one, which proves the high accuracy of the proposed method in estimating vocal tracts of continuous speech. And the signal-to-noise ratio (SNR) between original speech signal and residual (noise) is 23.1 dB. The spectrogram clearly shows that the formant frequencies are the same as the original one, which proves the high accuracy of the proposed method in estimating vocal tracts of continuous speech, and the mel-frequency cepstrum coefficients (order=14) distance between original speech and resynthesized speech signal is 0.93 dB. The synthesized speech can be perceived as well as original one by an informal perception test. Therefore, the proposed method is also suitable for continuous speech. The slight difference between the original speech and the resynthesized one may be caused by using only the ARX-LF model, in which $e(n)$ was not added in the synthesis process.

All of results demonstrate that the proposed method has higher estimation accuracy than that of IAIF with Dyprog-LF, and proposed method is applied for continuous speech.

3.3.8 Conclusion

In this section, we proposed a simultaneous estimation of glottal source waveforms and vocal tract shapes from speech signals based on the ARX-LF model. The estimation procedures contain two steps: obtaining the initial values of glottal source parameters, in which an inverse filter and the LF model are used, and using a simultaneous estimation procedure to obtain accurate glottal sources and vocal tract parameters with the ARX-LF model. We tested both the synthesized vowels and real vowels with the proposed method and IAIF with Dyprog-LF method. The results show that the proposed method has higher estimation accuracy than that of IAIF with Dyprog-LF. Moreover, the proposed method shows robustness for continuous speech. In future work, the proposed method will be used for voice conversion.

3.3.9 Analysis-by-synthesis system

Based on this proposed method, we developed a analysis-by-synthesis system that can estimate glottal source waveform and vocal tract shape directly from recorded speech signals. Japanese vowels /a/,/i/,/u/,/e/ and/o/ were analyzed by the system as shown in Figures 5-1, 3-16, 3-17, 3-18 and 3-19. The MATLAB GUI of analysis-by-synthesis system contains the recorded vowels waveform, estimated glottal source waveform, vocal tract shape and formant frequencies (F_1 and F_2).

One limitation of this system is that long time-consuming, 1 second voiced speech needs about 2 minutes runtime.

3.4 Summary

In this chapter, we proposed a method to estimate glottal source waveform and vocal tract shape basis on the ARX-LF mode by using an iterative algorithm, in which the initial values of the LF model were taken form EGG signals. Experimental results with synthetic and real speech signals showed the effectiveness of the proposed method. Moreover, instead of EGG signal, an inverse filter was first used to obtain the initial values of the LF model, the iterative algorithm was then applied to estimate glottal source waveform and vocal tract shape. Experimental results with synthetic and real speech signals showed the

effectiveness of the proposed method. An analysis-by-synthesis system was also developed for speech analysis.

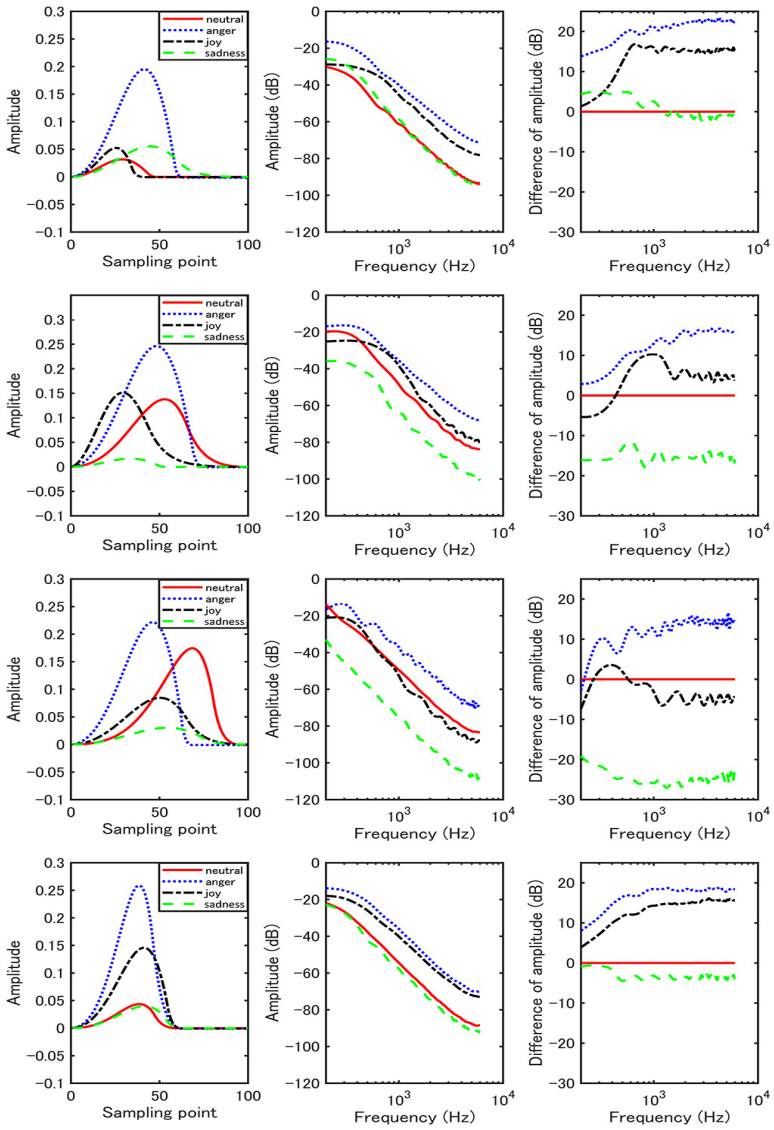

Figure 3-9: Results of four speakers (one speaker per row): (a) glottal source waves (first column); (b) spectra of glottal source wave (second column); (c) difference in spectra between neutral and other emotions (third column).

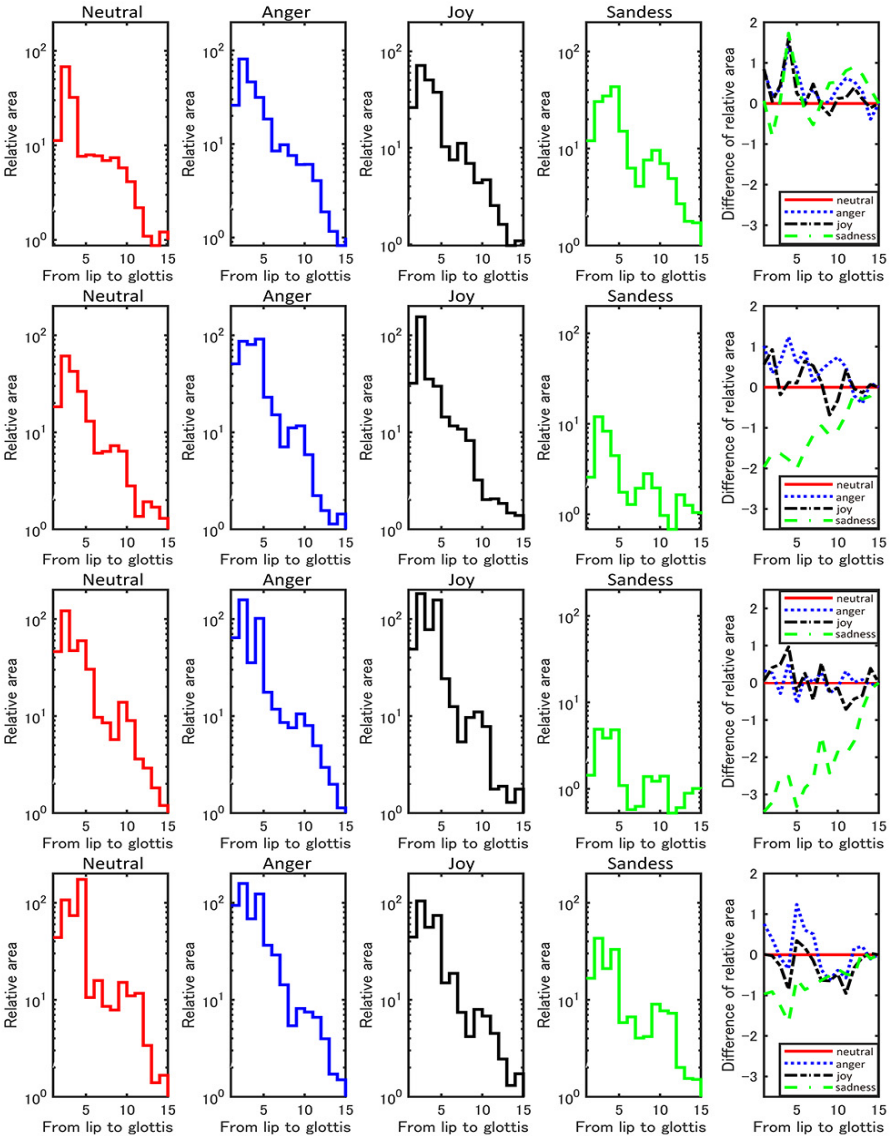

Figure 3-10: Results of four speakers (one speaker per row): vocal tract area functions and their differences between neutral and other emotion.

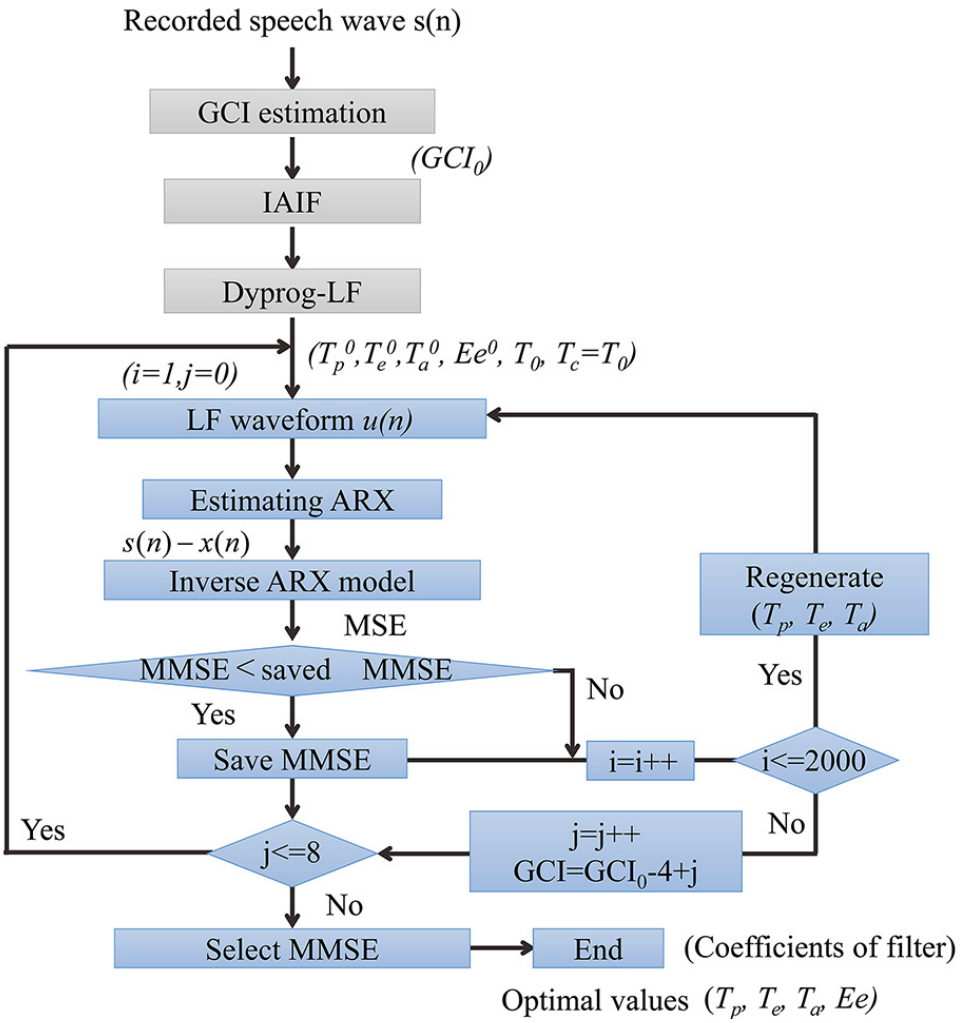

Figure 3-11: Estimation scheme of glottal source waveform and vocal tract shape

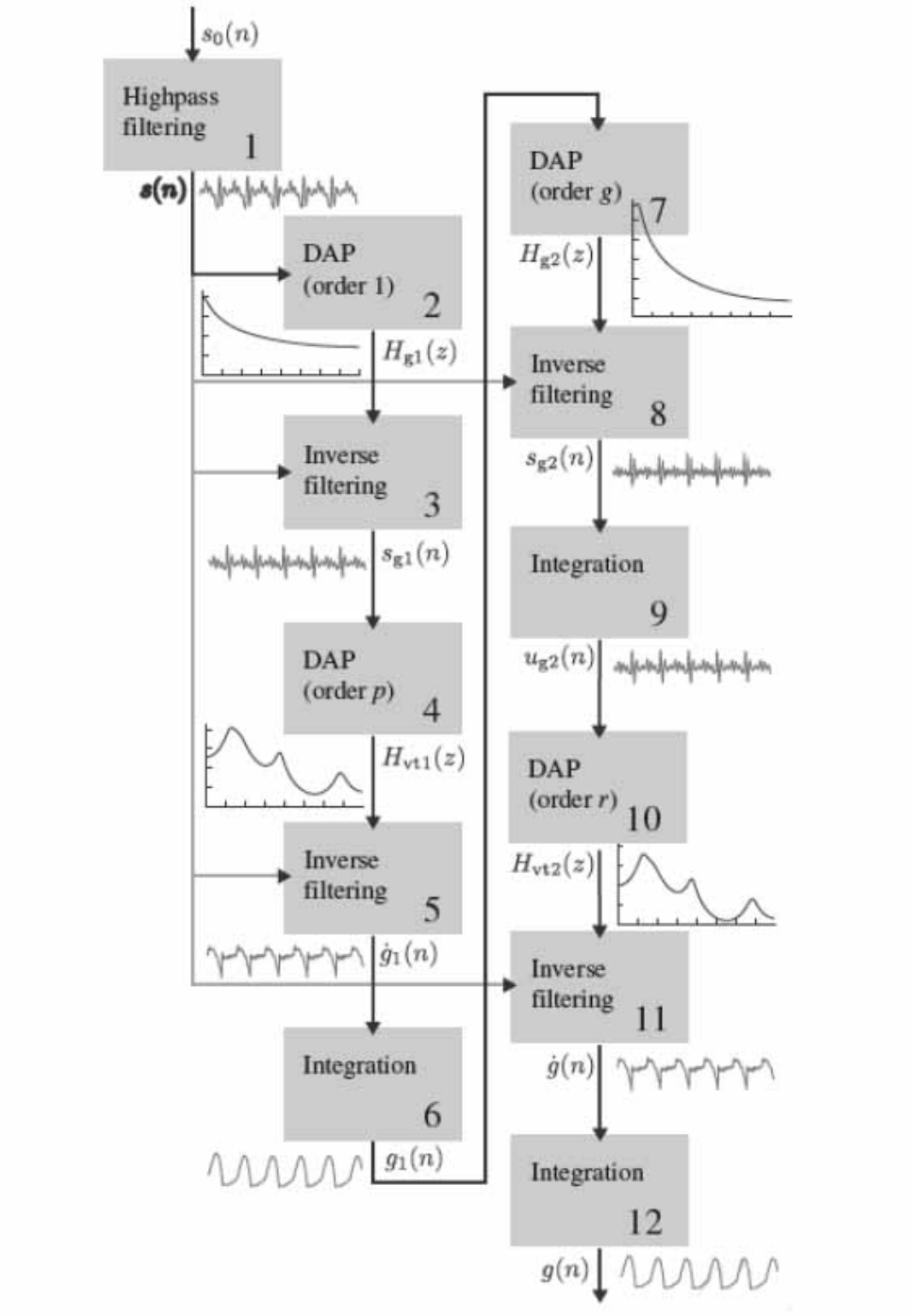

Figure 3-12: Structure of IAIF [90]

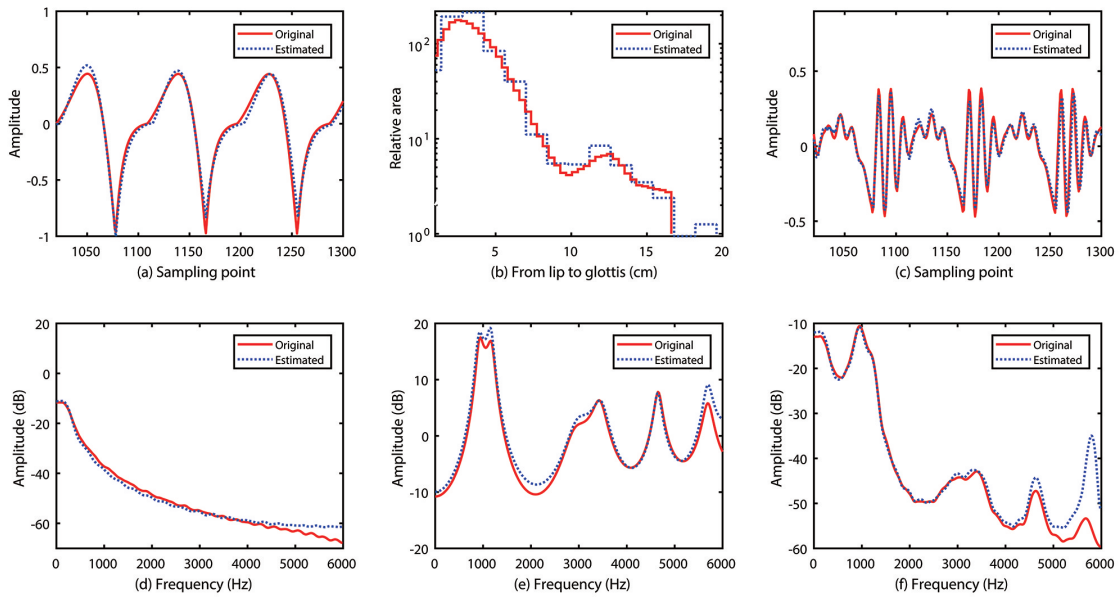

Figure 3-13: Original glottal source waveform and estimated glottal source waveform in time domain (a) and frequency domain (d), original vocal tract shape and estimated vocal tract shape (b) and its characteristic (e), original voice waveform and estimated voice waveform in time domain (c) and frequency domain (f).

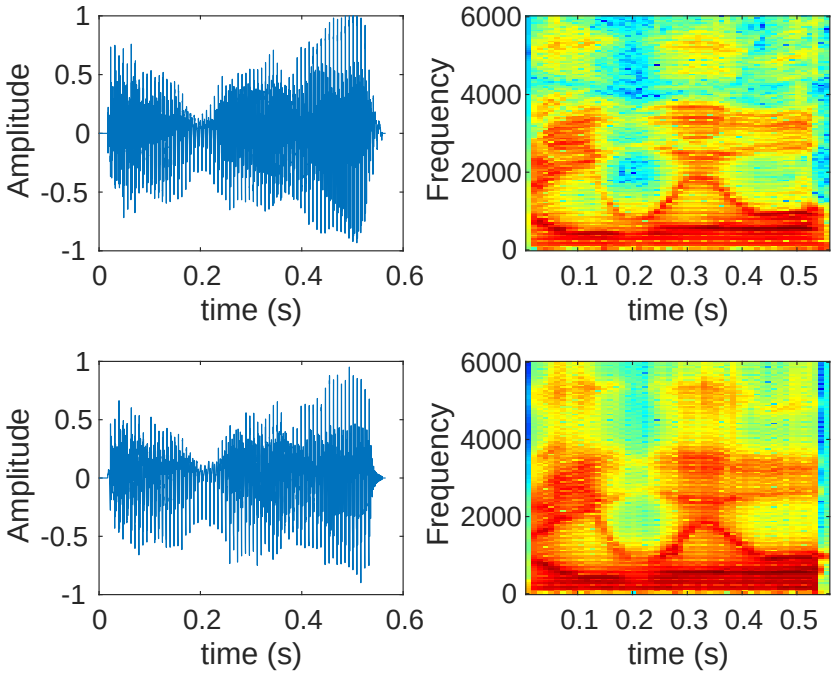

Figure 3-14: Original speech waveform and its spectrogram (top), re-synthesized speech waveform and its spectrogram (bottom)

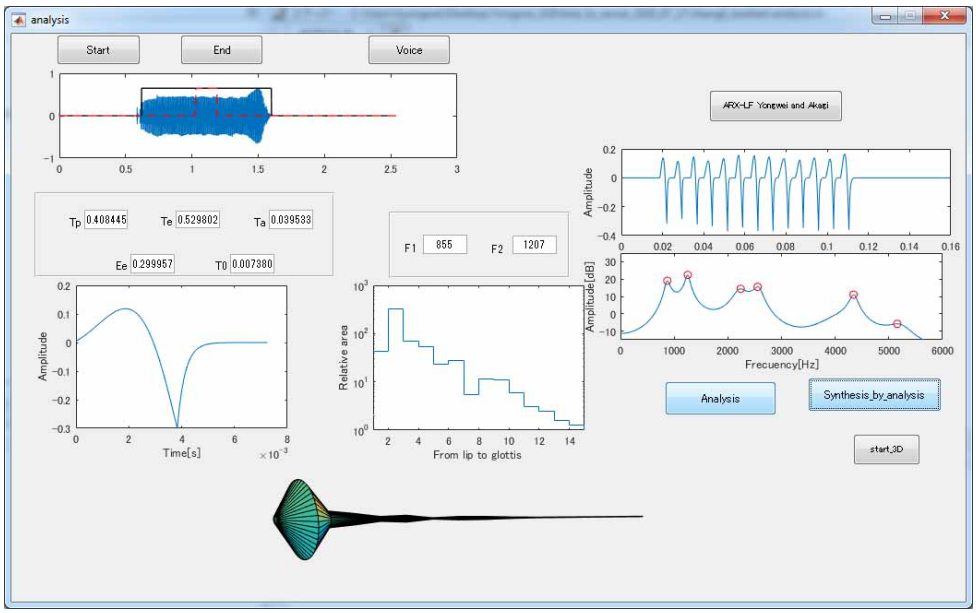

Figure 3-15: Result of vowel /a/

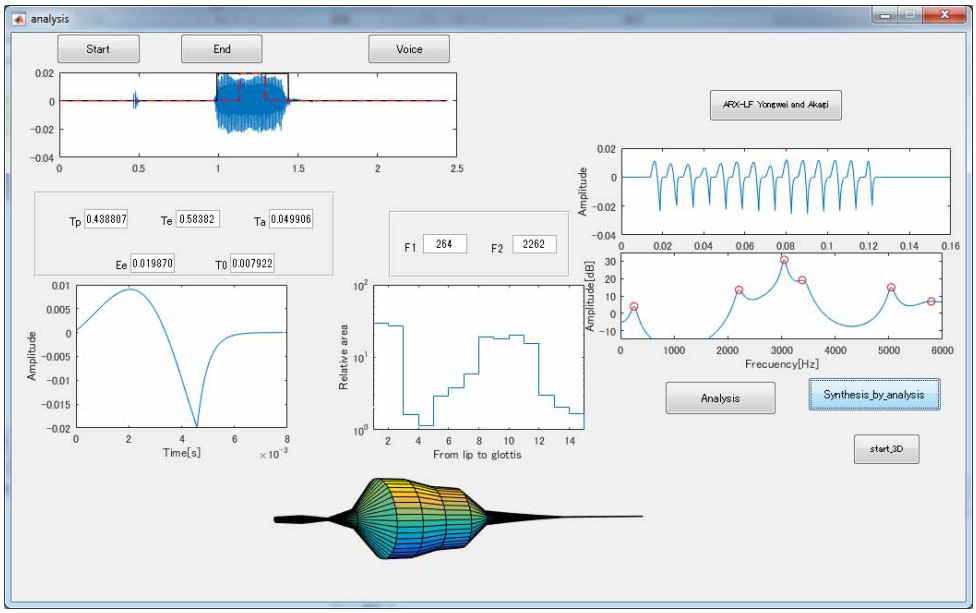

Figure 3-16: Result of vowel /i/

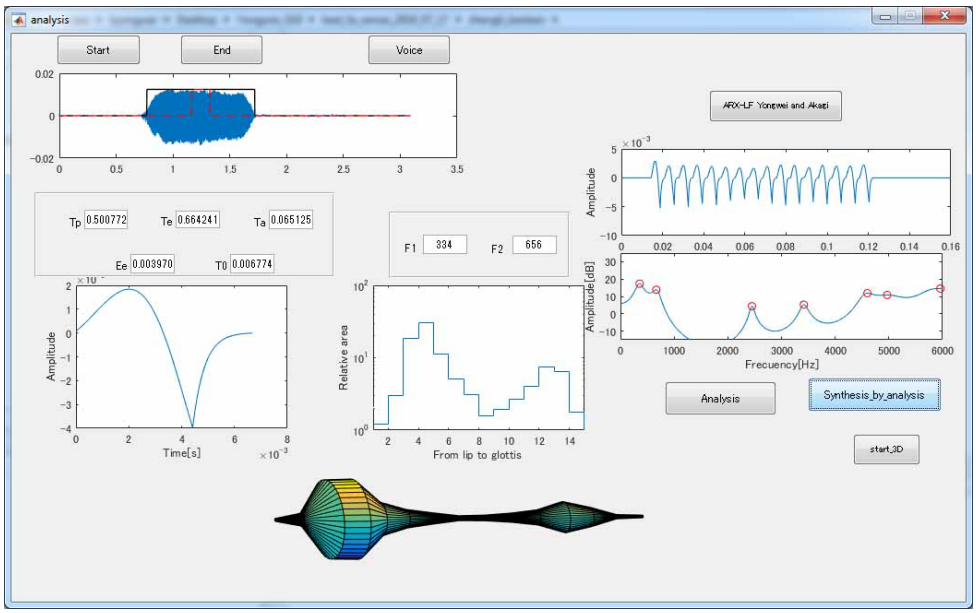

Figure 3-17: Result of vowel /u/

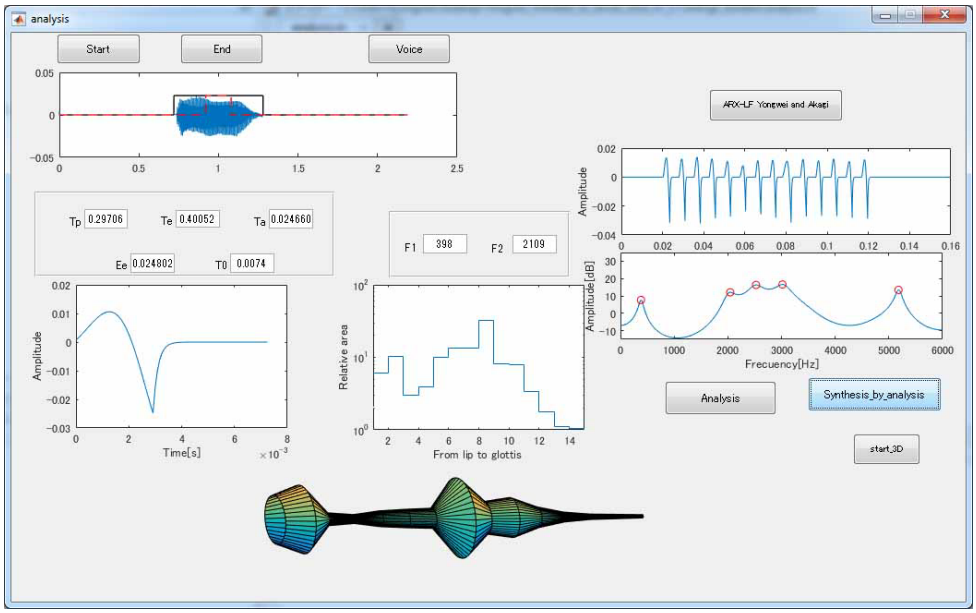

Figure 3-18: Result of vowel /e/

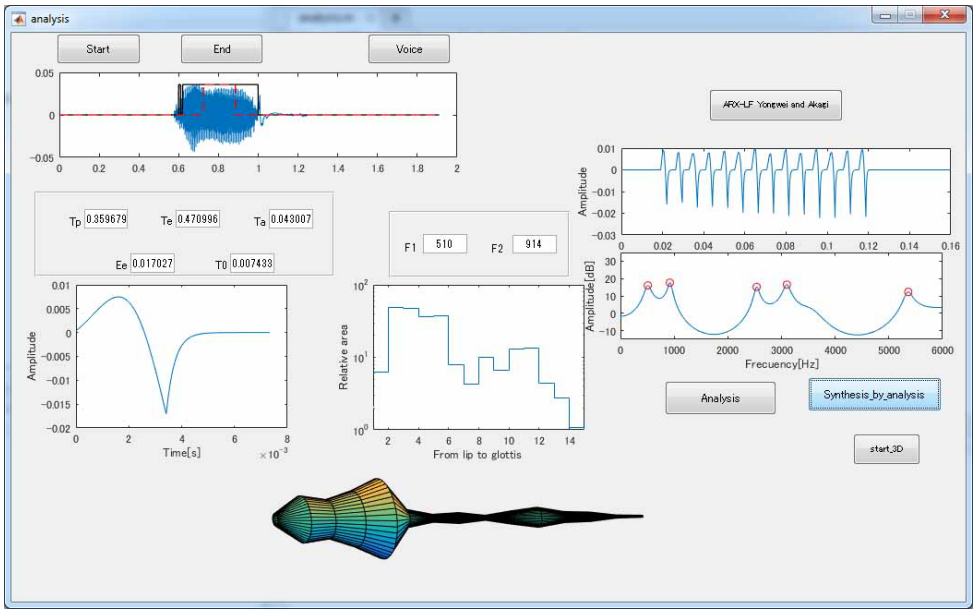

Figure 3-19: Result of vowel /o/

Chapter 4

Production related acoustic features

In order to discuss glottal source- and filter-related features, emotional speech uttered by four different speakers are analyzed by the proposed method with EGG signal as described in the previous section. The estimated results in Figures 4-1, 4-2, 4-3, 4-4, 3-9 and 3-10 are further discussed in this section.

4.1 Glottal source related features

From the Figures 4-1, 4-2, 4-3 and 4-4, the commonalities of glottal source waveform properties among speakers were found: (1) the T_0 of sadness and neutral were the largest, while that of anger and joy were the smallest. (2) the maximum negative peak of glottal source waveform derivative of anger and joy was often largest, while that of neutral and sadness were often the smallest.

Furthermore, since spectral tilts are frequently used to describe the characteristics of glottal source waves, they were adopted in discussing the commonalities of glottal source wave properties. The commonalities of glottal source waves spectra properties among speakers were summarized from the results shown in Figure 3-9. When compared with the spectral tilts of the glottal source waves of neutral, (1) those of anger and joy increased, and those of sadness decreased in the 200- to 700-Hz frequency range; (2) those of anger increased, but those of joy decreased, and those of sadness were the same as those of neutral in the 700- to 2000-Hz range; and (3) all spectral tilts had the same tendency over 2000 Hz.

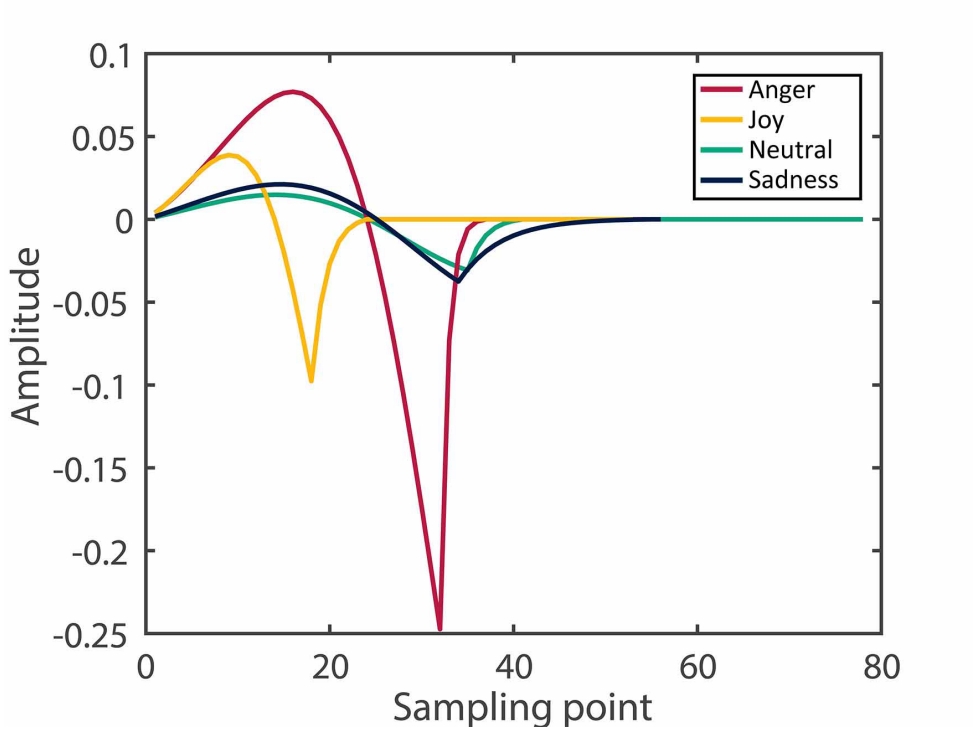

Figure 4-1: Estimated glottal source waveforms by proposed method (with electroglottograph signal) for speaker 1

Table 4.1: First formant frequency [Hz] estimated by the ARX model

	Anger	Joy	Neutral	Sadness
Speaker 1	861	902	762	650
Speaker 2	920	1020	885	961
Speaker 3	727	791	809	691
Speaker 4	793	861	703	709

4.2 Filter related features

For vocal tract shapes, the most notable characteristics are the vocal tract area functions. Thus, area functions normalized with the glottis were adopted in discussing the commonalities of vocal tract shape properties.

The commonalities of vocal tract shape properties among speakers are summarized in Figure 3-10 . The width of the front area function of anger was the largest, that of sadness was the smallest, and those of joy and neutral were in the middle. Moreover, the first formant frequency (F_1) values, which are listed in Table 4.1, were calculated among speakers and emotions from the ARX model. Table 4.1 shows that values of F_1 in joy and

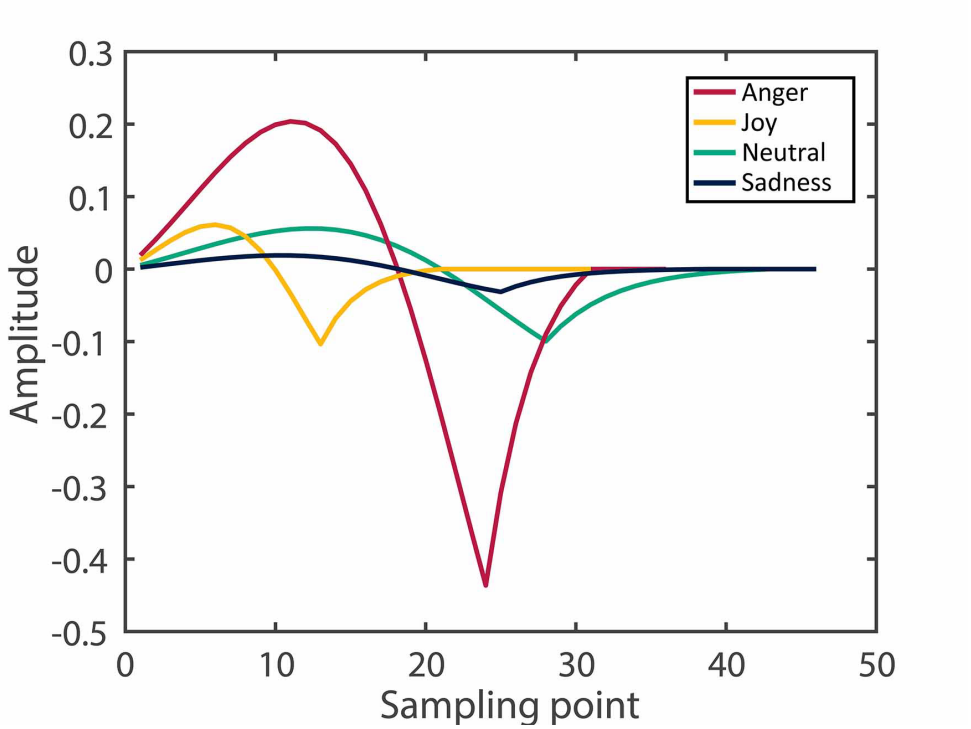

Figure 4-2: Estimated glottal source waveforms by proposed method (with electroglottograph signal) for speaker 2

anger speech are higher than those in neutral and sad speech. Table 4.1 and Fig. 6 show that a large mouth open area results in a higher F_1 value for anger and joy speech [61] reported a similar finding using EMA.

4.3 Discussion

The source- and filter- related features in emotional speech were discussed in the following: (1) For glottal source waves, the T_0 of sadness and neutral were the largest, while that of anger and joy were the smallest. It means that F_0 (T_0 related) are higher in joy and anger speech, while lower F_0 in sad and neutral speech. The maximum negative peak (E_e) of glottal source waveform derivative of anger and joy was often largest, while that of neutral and sadness were often the smallest. It means that intensity (E_e related) are higher in joy and anger speech, and lower intensity in sadness and neutral speech. Furthermore, spectral tilts of anger and joy increased, and those of sadness decreased in the 200- to 700-Hz frequency range; those of anger increased, but those of joy decreased, and those of sadness were the same as those of neutral in the 700- to 2000-Hz; all spectral tilts had the

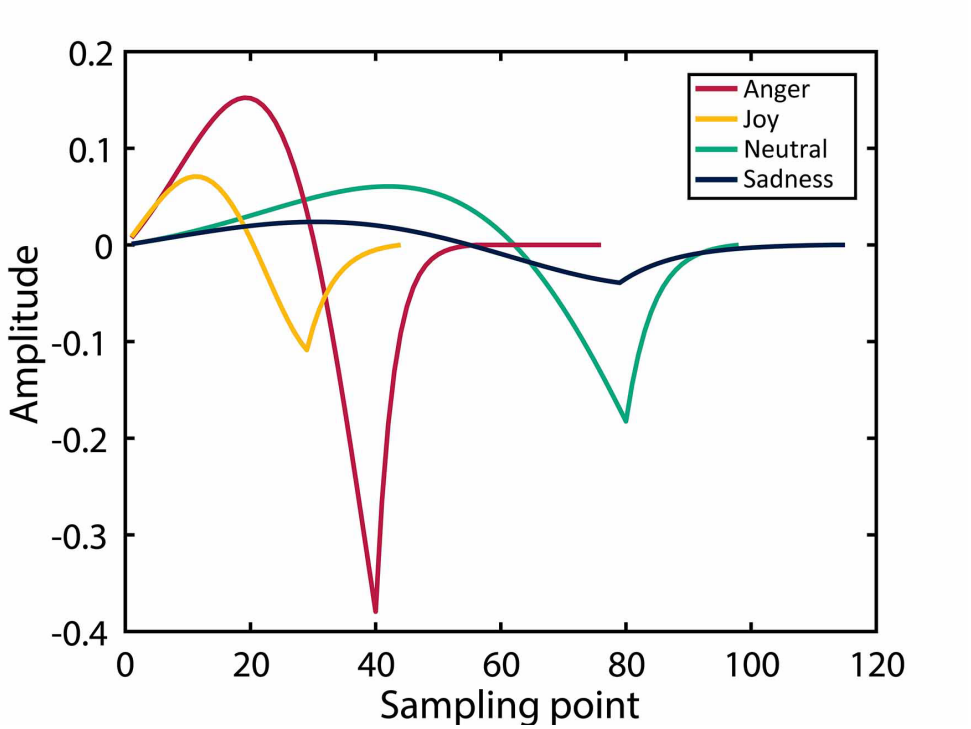

Figure 4-3: Estimated glottal source waveforms by proposed method (with electroglottograph signal) for speaker 3

same tendency over 2000 Hz. (2) For front area function of vocal tract shapes, the area function of anger was the largest, while that of sadness was the smallest, and those of joy and neutral were in the middle. These results are consistent with the previous findings using EMA and MRI. It results higher F_1 value for anger and joy speech.

The results are expected to be used for further discussion on emotional speech glottal source waves and vocal tract shapes among speakers and the applications to emotional speech processing from the point of view of speech production.

4.4 Summary

In this section, the source- and filter- related features in emotional speech were discussed. Source related features that F_0 ($T0$), intensity (E_e) and spectra tilt of glottal source waveform were summarized. Filter related features that F_1 (front area function of vocal tract shapes) was summarized.

For the next section, we discuss the contributions of glottal source waves and vocal tract shapes to the perception of emotions based on these results.

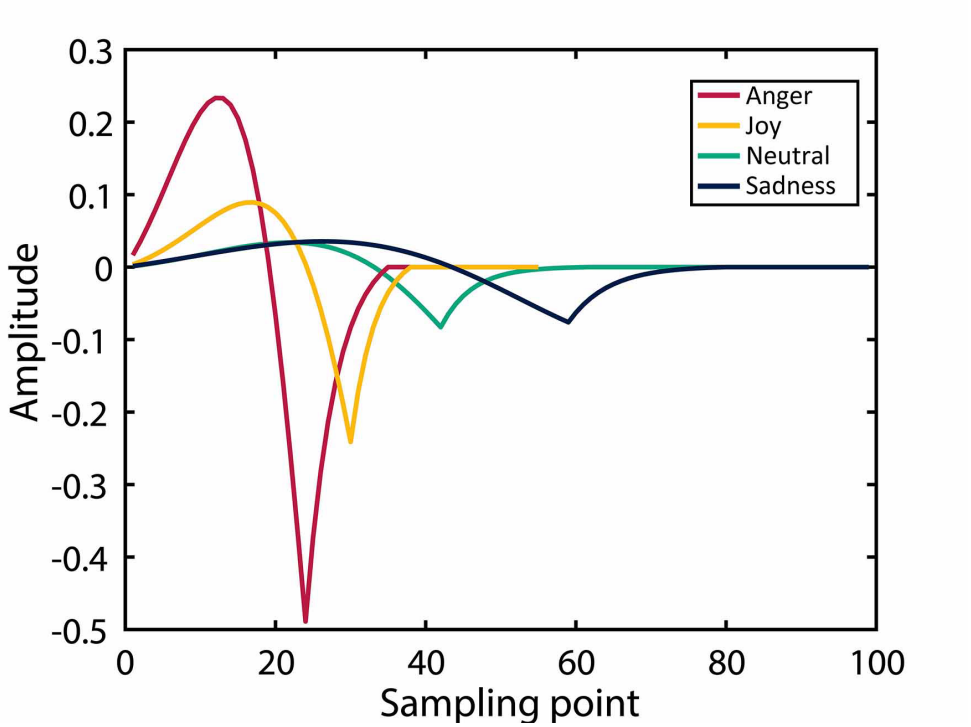

Figure 4-4: Estimated glottal source waveforms by proposed method (with electroglottograph signal) for speaker 4

Chapter 5

Effects of the glottal source and vocal tract cues on perception of emotional vowel

5.1 Pre-Experiment: Perception test for resynthesized vowels

5.1.1 Method

Stimuli

The Japanese voiced vowel (/a/) with four different emotional states (joy, neutral, anger and sadness) uttered by two professional male actors that were selected from previous section (Speaker 3 and Speaker 4 in previous section). In the following descriptions, actor's vowels are denoted original emotional vowels. The parameters of glottal source and vocal tract were estimated by proposed method (analysis-by-synthesis with EGG) based on the ARX-LF model, which was described in the Chapter 3.

The aim of this study is to discuss that effects of glottal source wave and vocal tract shape (periodic components) on perception of emotional speech in V-A spaces. Thus, the glottal noise (aperiodic components) is not considered in this study.

In order to achieve the aim, in first, whether estimated parameters of the ARX-LF model have ability to preserve emotional properties of original vowels on perception test

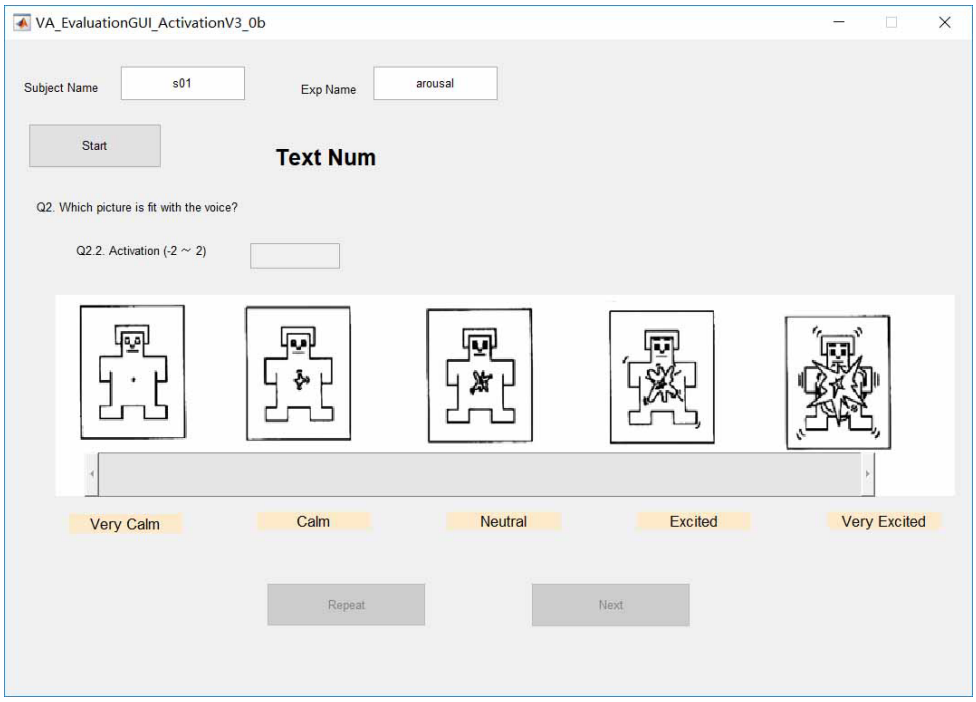

Figure 5-1: Experimental interface for evaluating arousal.

are needed to discuss.

In order to confirm that whether emotions of synthesized vowels (synthesized by estimated glottal source and vocal tract) are similar with emotions of original vowels on the emotional V-A spaces, and whether the naturalness of synthesized vowels similar with original vowels. Vowels are synthesized using analyzed glottal source and vocal tract shapes, there are 8 actual emotional vowels (neutral, joy, anger and sadness voices from the two actors), and 8 synthesized emotional vowels for the two speakers. Thus, there are total 16 emotional vowels in first perception test.

Note that in order to remove the interference of power among stimuli, the power of the all stimuli was normalized in root mean square (RMS) prior to the listeners.

Subjects

There were six males and four females, in total, ten normal-hearing people participated in the experiment. All listeners were Japanese and they were paid in this participation. The listeners were post-graduate students at the Japan Advanced Institute of Science and Technology, the age were between 23 an 30 years old.

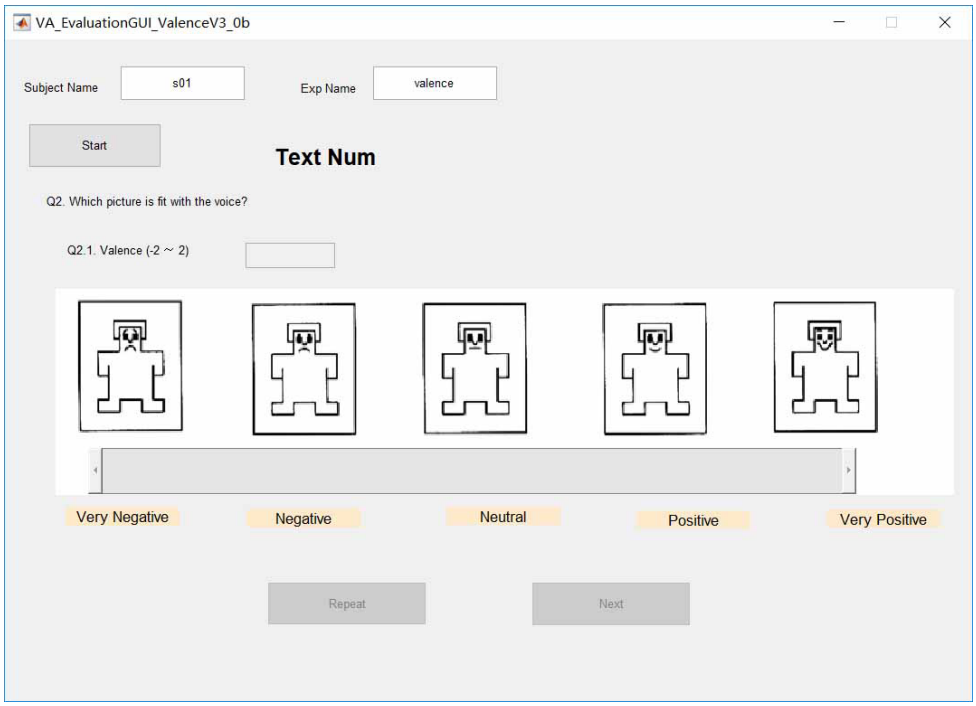

Figure 5-2: Experimental interface for evaluating valence.

Procedure

For each speaker, there was a total of sixteen resynthesized vowels. All stimuli were presented to each subject at a constant sound pressure level of 65 dB through HDA-200 headphones in a soundproof room. In the listening test, each subject was asked to listen a total of 16 tokens of resynthesized vowels and original vowels [2 speakers $\times$ 4 kinds glottal source parameters with 4 kinds of the vocal tract parameters=8 synthesized vowels] and 8 original emotional vowels [2 speakers with 4 emotional states]. The presentation orders of the stimuli were randomized across each subject. Each subject was asked to evaluate

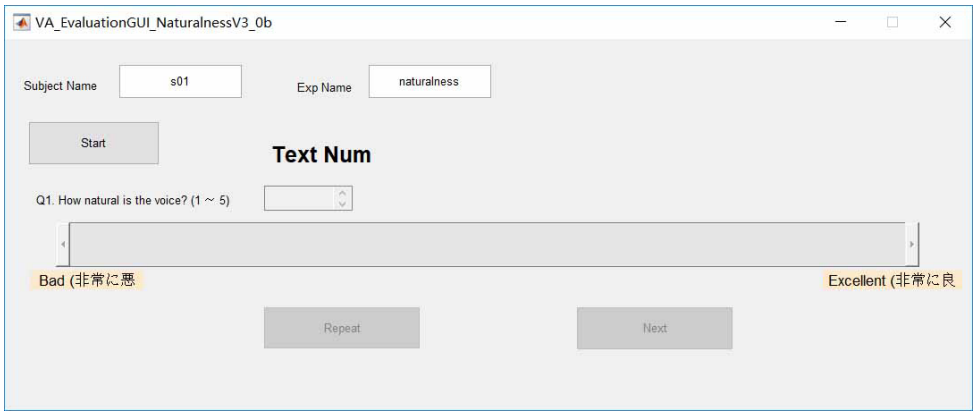

Figure 5-3: Experimental interface for evaluating naturalness.

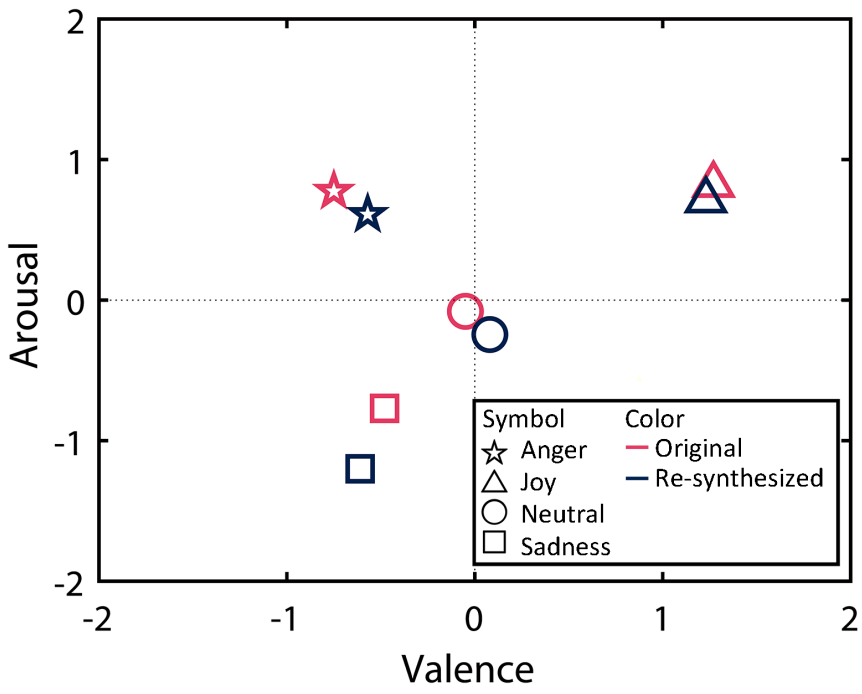

Figure 5-4: Positions of original voices (red color) and synthesized voices (blue color) on the V-A spaces for speaker 1.

a score for each emotional vowel based on his or her perceptual impression for emotional valence and arousal space, respectively. Before the listening test, the basic information for describing emotional dimensions and the meanings of valence and arousal were introduced to listeners. Regarding the evaluation of emotion in valence, please give a score of -2 for a very negative emotion and +2 for a very positive emotion. Regarding the evaluation of emotion in arousal, please give a score of -2 for a very calm emotion and +2 for a very excited emotion. The exact value provided by the listeners, which was dependent on his or her own perception/feeling [24]. The scores for the perceptual evaluation of emotional vowels in the valence and arousal (V-A) space, 40 points scale were ranged from -2 to +2 with a step of +0.1. The scores for the perceptual evaluation of the naturalness of the vowels ranged from +1 to +5 with a step of +1. The score +1 indicates very bad naturalness, with +5 very excellent naturalness.

Experimental MATLAB GUI for evaluating arousal, valence and naturalness are plotted in Figs 5-1, 5-2 and 5-3.

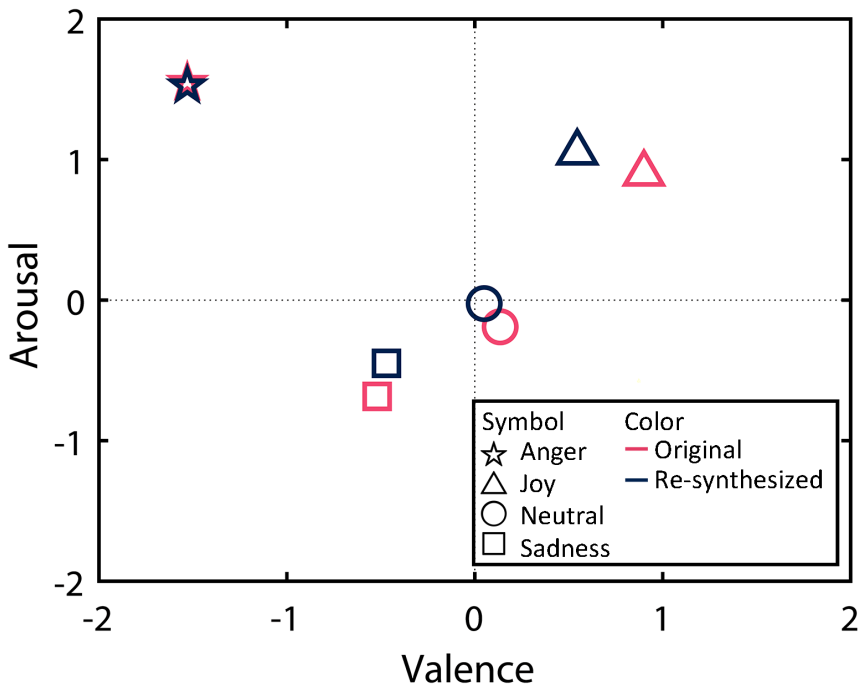

Figure 5-5: Positions of original voices (red color) and synthesized voices (black color) on the V-A spaces for speaker 2.

5.1.2 Experiment results and discussion

Figures 5-4 and 5-5 shows average scores of 10 subjects by evaluating positions of original voices and synthesized voices on V-A spaces. Using Analysis of variance (ANOVA), there is a significant difference if the significant level p <0.05 in this paper. Results of ANOVA indicated that there is no significant differences between original voices and synthesized voices on arousal [F (1, 159)=1.741, p= 0.2196] and Valence [F (1, 159)=0.123, p=0.739] spaces. Thus, emotions of synthesized vowels are similar with emotions of original vowels on the emotional V-A spaces.

Figure 5-6 shows average scores of 10 subjects by evaluating naturalness. ANOVA indicated that there is no significant differences between original voices and synthesized voices on naturalness [F(1, 159)=0.943, p=0.3569]. The scores of naturalness of vowels about 3, this is because that naturalness is affected by may factors, such as duration, interval and consonant. In our study, only vowel /a/ was used.

Synthesized voices by estimated glottal source waves and vocal tract shapes (period components) not only can represent impression of original voices on V-A spaces, but also naturalness. This results provide a guarantee for further discussion affects of glottal source

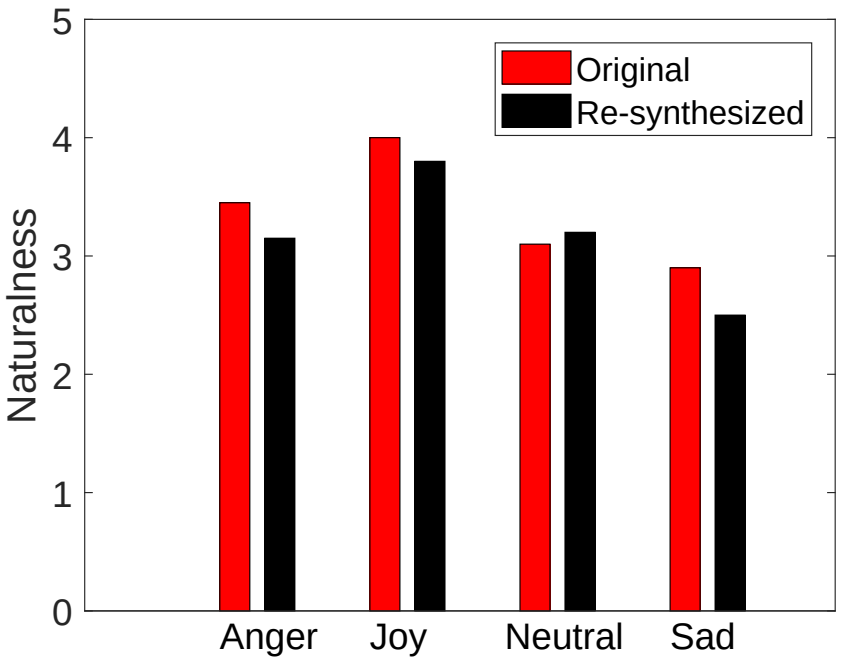

Figure 5-6: Naturalness of original voices (red color) and synthesized voices (black color) for speaker 1 and speaker 2

waves and vocal tract shapes on V-A spaces.

5.2 Experiment I: Effects of glottal source and vocal tract on perception of emotional vowels

5.2.1 Method

Stimuli

The parameters of glottal source and vocal tract were estimated by proposed method (analysis-by-synthesis with EGG) based on the ARX-LF model, which was described in the Chapter 3, same as previous section. Because of the estimation procedures of the proposed method was done one glottal period by one glottal period, the estimated glottal source and vocal tract (the ARX-LF model) parameters (except parameters of E_e and T_0 in the LF model) were first averaged among all periods for each emotional vowel, but parameters of T_0 and E_e in the LF model were kept as those of the original vowels. To examine the effects of glottal source (GS) and vocal tract (VT) cues on perception of

emotional vowels, the estimated glottal source parameters (averaged values) from one emotional state and the estimated vocal tract parameters (averaged values) from another emotional state were used to re-synthesize emotional vowels. Thus, there are total sixteen artificially-synthesized emotional vowels were synthesized for each speaker.

The ten normal-hearing Japanese who participated in the Pre-Experiment also participated in the experiment I, and were paid for their participation. In the listening test, each listener was asked to listen a total of 32 tokens of re-synthesized vowels [2 speakers $\times$ the glottal source parameters with 4 emotional states (neutral, anger, joy and sadness) $\times$ the vocal tract parameters with 4 emotional states (neutral, anger, joy and sadness)] and 8 original emotional vowels [2 speakers $\times$ 4 emotional states]. Each subject was asked to evaluate a score for each emotional vowel based on his or her perceptual impression for emotional valence and arousal space, respectively. The listening testing procedure was the same as that in the Pre-Experiment.

5.2.2 Results

The average perceptual scores of ten listeners, which evaluated for four original emotional vowels and sixteen re-synthesized emotional vowels of two speakers, are draw in Fig. 5-7. As seen in Fig. 5-7, because of the perceptual score of each emotional vowels synthesized with the averaged glottal source parameters (except parameters of E_e and T_0 in the LF model) and vocal tract parameters is close to that of original emotional vowel for two speakers, the averaged glottal source parameters and averaged vocal tract parameters still include much information in the valence and arousal space. The perceptual scores for emotional vowel synthesized with neutral glottal source parameters and neutral vocal tract parameters, which were located to almost (0,0) in the valence and arousal space, and the perceptual scores for the emotional vowels synthesized with the neutral glottal source parameters and arbitrary vocal tract parameters were also quite close to the (0,0) in the valence and arousal space. Moreover, it was showed that the perceptual scores for the emotional vowels synthesized with the joy glottal source parameters and arbitrary vocal tract parameters were positive in both valence and arousal. while the perceptual scores for the emotional vowels synthesized with the anger glottal source parameters and arbitrary vocal tract parameters were negative in valence, and positive in arousal. However, the

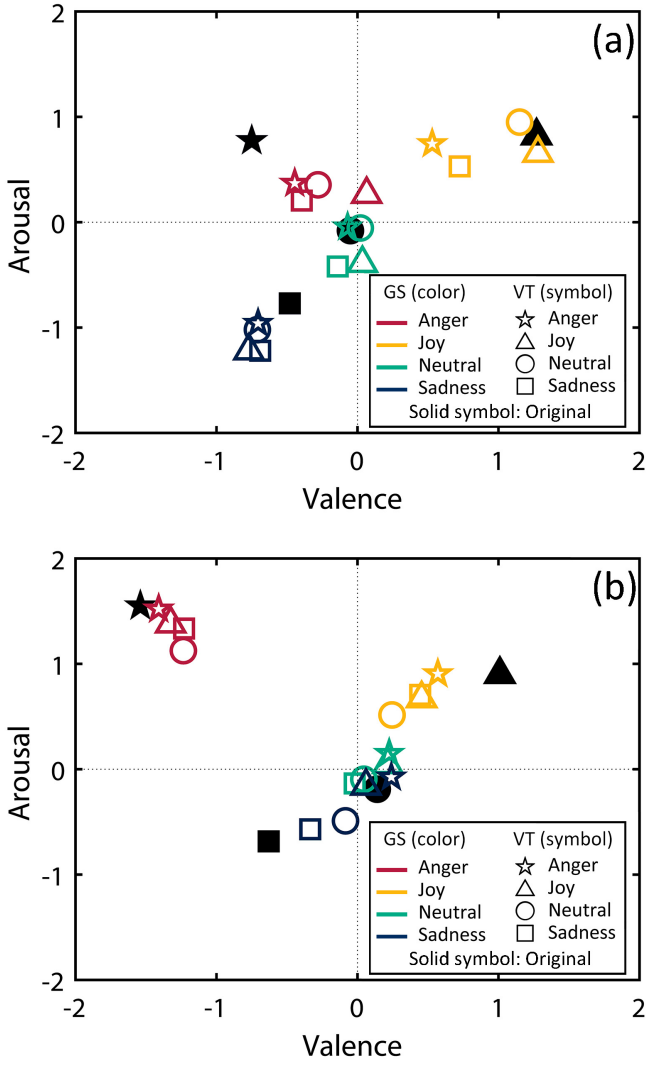

Figure 5-7: The perceptual results in the valence-arousal space of the emotional vowels synthesized with the arbitrary glottal source (GS) parameters and the vocal tract (VT) parameters for speaker 1 (a) and speaker 2 (b). The results of the original emotional vowels are also plotted in the V-A space with the solid symbols.

perceptual scores for emotional vowels synthesized with the sad glottal source parameters and arbitrary vocal tract parameters were negative in both valence and arousal, especially for the speaker 1. More important is that the perceptual scores differences (i.e., the distances in the valence and arousal space) for the emotional vowels synthesized with the given glottal source parameter and arbitrary vocal tract parameters were much smaller than those for the emotional vowels synthesized with the given vocal tract and arbitrary glottal source parameters. In other words, the vowels synthesized with the given glottal source parameters and arbitrary vocal tract parameters were centralized in the valence and arousal space, while the vowels synthesized with the given vocal tract parameters and

arbitrary glottal source parameters were more dispersedly distributed in the valence and arousal space.

To examine the effects of glottal source parameters (GS-Anger, GS-Joy, GS-neutral, GS-Sadness) and vocal tract parameters (VT-Anger, VT-Joy, VT-neutral, VT-Sadness), the perceptual scores (the perceptual scores in arousal and the perceptual scores in valence space) of the resynthesized emotional vowels were subjected to statistical analysis using the scores as the dependent variable, and the glottal source and vocal tract parameters as the two within subject factors. For the perceptual scores in *valence*, two-way analysis of variance (ANOVA) with repeated measures indicated the significant effects of glottal source [$F(3,\ 27)$=20.377, p=0.0000] and vocal tract [$F(3,\ 27)$=7.378, p=0.0009] for speaker 1, and glottal source [$F(3,\ 27)$=30.534, p=0.0000] for speaker 2. There was significant interaction between glottal source and vocal tract [$F(9,\ 81)$=3.298, p=0.0018] for speaker 1, and [$F(9,81)$=2.638, p=0.0099] for speaker 2. For the perceptual scores in *arousal*, two-way ANOVA with repeated measures indicated the significant effect of glottal source [$F(3,\ 27)$=33.729, p=0.0000] and vocal tract [$F(3,27)$=8.340, p=0.0004] for speaker 1, and glottal source [$F(3,\ 27)$=82.417, p=0.0000] and vocal tract [$F(3,\ 27)$=27.323, p=0.0000] for speaker 2. There was no significant interaction between glottal source and vocal tract for speaker 1, but significant interaction between glottal source and vocal tract [$F(9,\ 81)$=2.350, p=0.0207] for speaker 2.

5.2.3 Discussion

The perceptual scores for emotional vowels synthesized with the averaged glottal source parameters and averaged vocal tract parameters (rather than their dynamic varied values across all glottal periods) was close to (0, 0) in the valence and arousal space. Moreover, the perceptual scores of the synthesized emotional vowels with averaged glottal source and vocal tract parameters are close to that of original emotional vowels, which confirmed again the effectiveness of the ARX-LF model with the proposed analysis-by-synthesis approach [89]. The slight difference of perceptual scores between the synthesized emotional vowels (averaged glottal source parameters and vocal tract parameters) and the original ones may be caused by using the averaged parameters of the ARX-LF model. In comparison with the perceptual scores for emotional vowels synthesized with the given

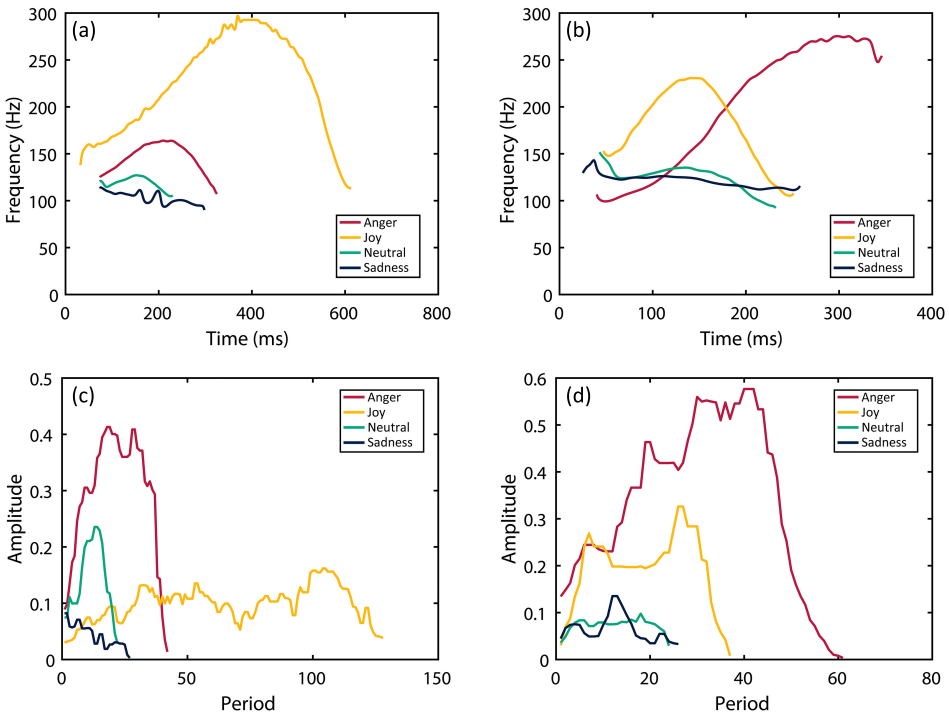

Figure 5-8: The F_0 shapes of the synthesized emotional vowels for speaker 1 (a) and speaker 2 (b); the amplitude at the glottal closure instant E_e of the synthesized emotional vowels for speaker 1 (c) and speaker 2 (d).

vocal tract and arbitrary glottal source parameters, the perceptual scores for emotional vowels synthesized with the given glottal source and arbitrary vocal tract parameters have the much smaller differences in the valence and arousal space. This indicates the dominant effect of the glottal source parameters (relative to the vocal tract parameters) in perception of emotions in vowels, and is consistent with results from many previous studies [93–95]. The perceptual scores in the valence and arousal space of the synthesized vowels with four typical emotions obtained in this experiment are similar to the results shown in [96] and [97]. Often the glottal source parameters are extracted from the glottal source waveform using the inverse filtering techniques [98, 99].

Although, it was clearly observed that the distribution of the perceptual scores on the valence and arousal space have similarities for two speakers, the relative distance from synthesized emotional vowels (sadness, joy and anger) to neutral vowels (center in the valence and arousal space) is also greatly. To discuss possible reasons for this difference, the F_0 contour and amplitude at GCI (E_e) of the emotional vowels were examined for two speakers, since F_0 and E_e (prosody features related parameter, pith and intensity)

are well-known to be important in emotional speech perception [38]. The F_0 contours and amplitudes (E_e) of the emotional vowels are plotted in Figs. 5-8 for two speakers.

As shown in Fig. 5-8 (a) and (b), the F_0 contour (e.g., the mean F_0, F_0 range and F_0 shape) of the emotional vowels differed greatly for the two speakers, especially for anger and joy vowels. Compared to the neutral vowel, the mean F_0 and F_0 range have the largest differences for joy in speaker 1, and anger vowel in speaker 2, which correspond to the largest distances of perceptual scores relative to the neutral vowel in the valence and arousal space. This indicates that the mean F_0 and F_0 range can substantially account for emotional vowel perception, similar with results in article [39]. For example, anger is characterized by a higher value of mean F_0, which might be partially caused by the heightened subglottal pressure during vowels in speech [94]. As shown in Fig. 5-8 (c) and (d), the E_e (amplitude at GCI) of emotional vowels were greatly different for two speakers. Comparing with the neutral vowel, the largest differences of E_e were for anger vowel in two speakers. This indicated the important contribution of the intensity-related E_e parameter for emotional vowel perception, this also observed by [100, 101]. Furthermore, Fig. 5-8 indicates large differences of duration in the emotional vowels in our experiment, which might also affect emotional speech perception as suggested by [39, 102].

5.3 Experiment II: Effects of glottal source and vocal tract to emotional vowel perception after neutralizing the fundamental frequency, intensity and duration cues

The results of Experiment I indicated that the glottal source cues play more important roles in emotional vowel perception than the vocal tract cues. In this study, six parameters were used to describe the glottal source waveform by the LF model. As Experiment I showed, F_0, $E_e$and duration are important cues for emotional vowel (especially arousal) perception, which was also reported in [100] and [39, 102]. A follow-up question is whether the other glottal source and vocal tract cues effects to emotional vowel perception. And how do these cues effect to emotional vowel perception in the valence and arousal space.

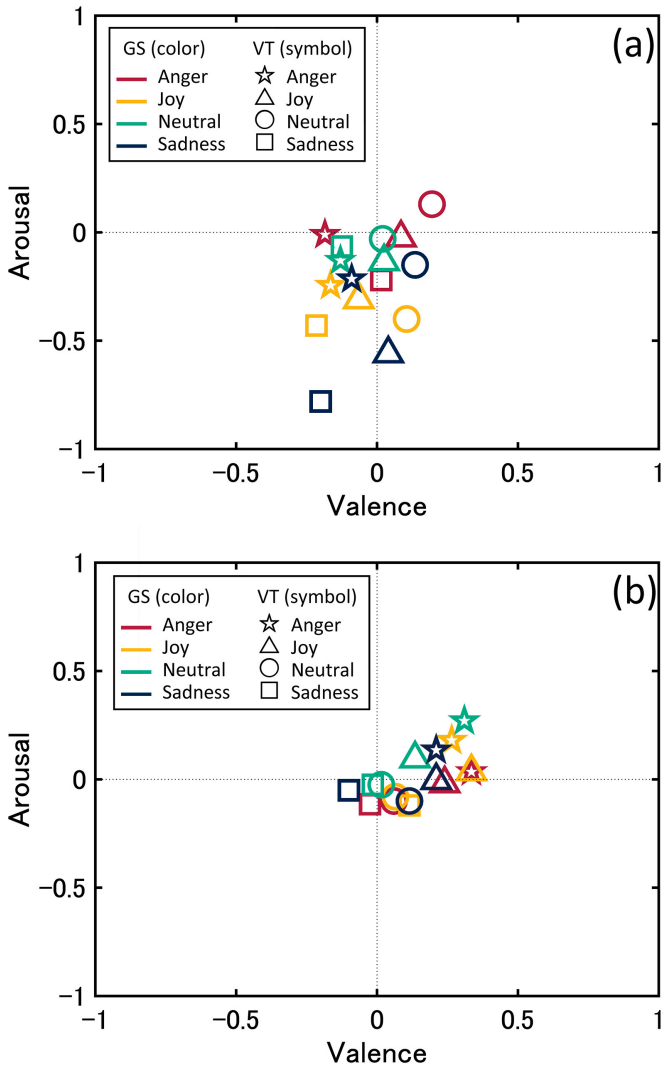

Figure 5-9: The perceptual scores in the valence and arousal space of the emotional vowels synthesized with the different, F_0 and E_e-neutralized glottal source and vocal tract parameters for speaker 1 (a) and speaker 2 (b).

This is the main contribution of this study, also is the task of Experiment II.

5.3.1 Method

The same voiced vowel (/a/) with four different emotional states (i.e., joy, anger, neutral and sadness) in the Pre-Experiment and Experiment I was also used in the Experiment II. The glottal source and vocal tract parameters in the ARX-LF model were averaged across all periods for re-synthesizing emotional vowels as in Experiment I. To examine the effects of the glottal source parameters (except F_0 and E_e) on emotional vowel perception in the valence and arousal space, the estimated T_0 (i.e., $1/F_0$) and E_e parameters of the

LF model of the emotional vowels (sadness, anger and joy) were replaced by those of the neutral vowels, while other parameters (T_p, T_e, T_a and T_c) of the LF model were kept for each emotional state (sadness, anger and joy vowels). In the synthesis process, the emotional vowels were re-synthesized using the glottal source parameters (referred to as F_0 and E_e-neutralized glottal source parameters) from one emotional state and the vocal tract parameters of another emotional state. In total, there were 32 tokens of emotional vowels that were re-synthesized for evaluating the perceptual scores in the valence and arousal space. Neutralized the T_0 and E_e parameters removed the effects of F_0 and intensity on the emotional vowel perception, which is equivalent to normalizing duration of the re-synthesized emotional vowels, and this helped to isolate the effects of the other glottal source cues to emotional vowel perception.

The ten normal-hearing Japanese who participated in Pre-Experiment and Experiment I also participated in the experiment II, and were paid for their participation. During the test, each subject listened to a total of 32 emotional vowel tokens [2 speakers $\times$ 4 sets of the glottal source parameters $\times$ 4 sets of the vocal tract parameters]. Each subject was asked to evaluate a score for each emotional vowel based on his or her perceptual impression for emotional valence and arousal space, respectively. The listening testing procedure was the same as that in the Pre-Experiment and Experiment I.

Table 5.1: The coefficients determination (R^2) of the linear regressions for the relationships between the perceptual scores in the valence and arousal space and F_1 or the spectral tilt of glottal source waveform.

	Speaker 1				Speaker 2			
	Anger	Joy	Neutral	Sadness	Anger	Joy	Neutral	Sadness
F_1-arousal	**0.74**	0.01	0.02	0.3	**0.99**	**0.97**	**0.97**	**0.79**
F_1-valence	0.46	**0.85**	**0.89**	**0.97**	**0.93**	**0.68**	**0.97**	0.54
Spectral tilt-arousal	**0.87**	**0.74**	**0.80**	**0.63**	**0.61**	**0.91**	**0.67**	0.03
Spectral tilt-valence	0.18	0.50	0.01	**0.80**	0.11	0.01	**0.63**	0.50

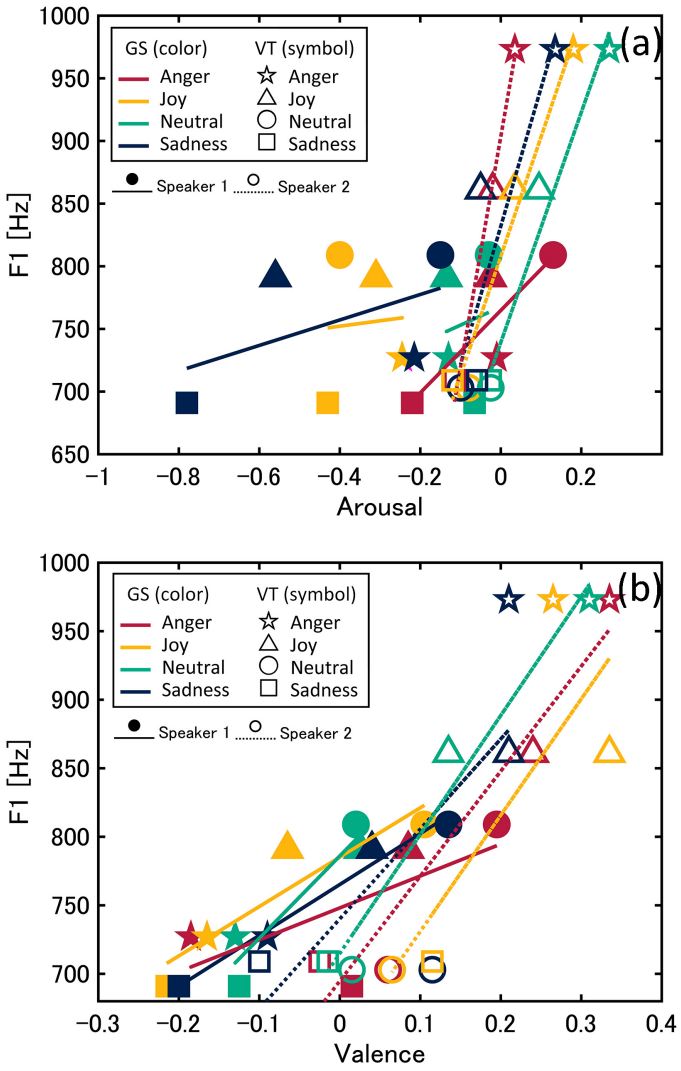

Figure 5-10: The relationship of the first formants (F_1) and the perceptual scores in arousal (a) and in valence (b), where the perceptual scores for four emotional vowels synthesized with the given glottal source and arbitrary vocal tract parameters were linearly regressed in the sense of minimum mean square error (MMSE).

5.3.2 Results

The average perceptual scores of ten listeners, which evaluated for four original emotional vowels and sixteen re-synthesized emotional vowels with the F_0 and E_e-neutralized glottal source parameters and vocal tract parameters of two speakers, were draw in Fig. 5-9.

As seen in Fig. 5-9, the perceptual scores for the emotional vowels synthesized with the F_0 and E_e-neutralized glottal source parameters (anger, joy,and sadness) moved towards the center (0,0) of the synthesized neutral vowel with neutral glottal source and neutral vocal tract parameters in the valence and arousal space. The perceptual scores of most of

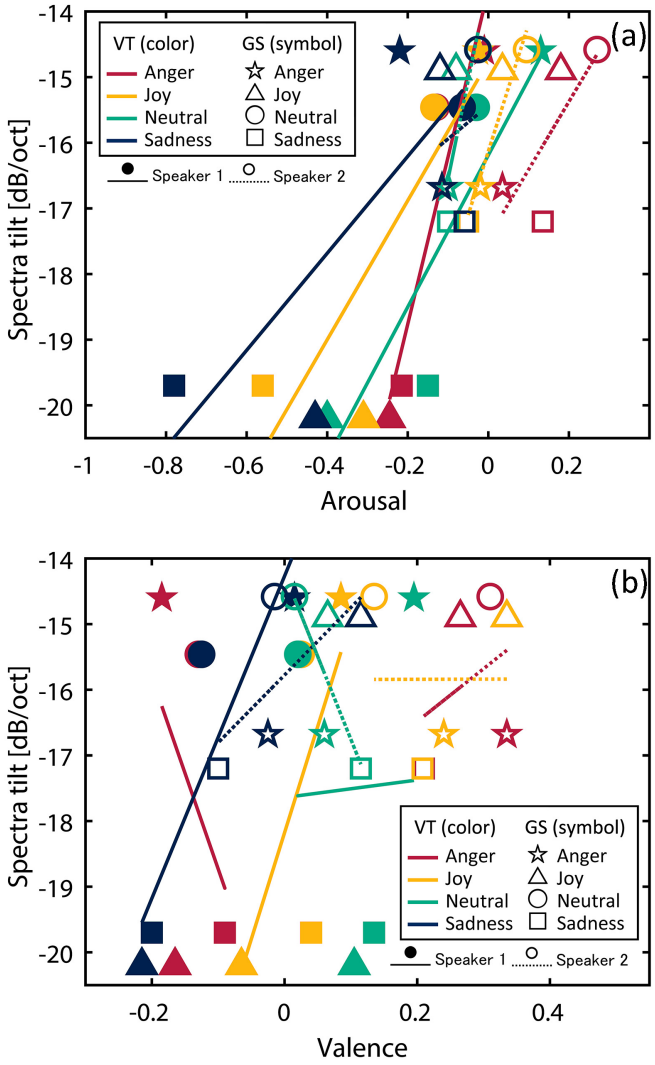

Figure 5-11: The relationship of the spectral tilt of glottal source waveform and the perceptual scores in arousal (a) and in valence (b), , where the perceptual scores for four emotional vowels synthesized with the given glottal source and arbitrary vocal tract parameters were linearly regressed in the sense of minimum mean square error (MMSE).

the synthesized emotional vowels with the F_0 and E_e-neutralized glottal source parameters were negative in the valence and arousal space for speaker 1. Although the perceptual scores in the valence of the emotional vowels synthesized with different sets of glottal source and vocal tract parameters were centralized, those in arousal space were relatively scattered. Moreover, for the emotional vowels synthesized with the given vocal tract parameters, the perceptual scores in arousal were often the lowest for vowels with the sadness glottal source parameters, and were often the highest for vowels with the anger glottal source parameters. For speaker 2, it was showed that the perceptual scores of the synthesized emotional vowels were relatively scattered in valence space. For the emotional

vowels synthesized with the given glottal source parameters, the perceptual scores were the lowest for vowels with sadness vocal tract parameters in valence space, followed by those for the emotional vowels synthesized with the neutral and joy vocal tract parameters, and the scores for the emotional vowels synthesized with the anger vocal tract parameters were the highest in valence.

To examine the effects of glottal source parameter (GS-Anger, GS-Joy, GS-Neutral, GS-Sadness) and vocal tract parameter (VT-Anger, VT-Joy, VT-Neutral, VT-Sadness), the perceptual scores (the perceptual scores in arousal and valence space) of emotional vowels synthesized with the F_0 and E_e-neutralized parameters in the LF model were subjected to statistical analysis using the scores as the dependent variable, and the glottal source and vocal tract parameters as the two within subject factors. For the perceptual scores in *valence*, two-way ANOVA with repeated measures indicated the significant effects of vocal tract $[F(3, 27) = 6.267, p = 0.0023]$ for speaker 1, and glottal source $[F(3, 27) = 3.800, p = 0.0215]$ and vocal tract $[F(3, 27) = 6.005, p = 0.0028]$ for speaker 2. There were no significant interactions between glottal source and vocal tract for both speakers. For the perceptual scores in *arousal*, two-way ANOVA with repeated measures indicated the significant effects of glottal source $[F(3, 27) = 24.988, p = 0.0000]$ and vocal tract $[F(3, 27) = 8.132, p = 0.0005]$ for speaker 1, and vocal tract $[F(3, 27) = 15.138, p = 0.0000]$ for speaker 2. There was a significant interaction between glottal source and vocal tract $[F(9, 81) = 4.208, p = 0.0002]$ for speaker 1, and there was no significant interaction between glottal source and vocal tract $[F(9, 81) = 0.486, p = 0.8799]$ for speaker 2.

5.3.3 Discussion

As shown in Figs. 5-7 and 5-9, in comparison, the perceptual scores for emotional vowels synthesized with the F_0 and E_e-neutralized glottal source parameters shows movements towards those of the synthesized neutral vowels, which further demonstrates the important effects of the glottal source parameters (F_0, E_e and duration) on emotional vowels perception, similar to results in articles [39] and [100].

After removing the effects of the F_0, E_e and duration cues, the vocal tract information significantly effects to perception scores in the valence and arousal space for both speakers, which was consistent with the findings in article [103]. In contrast, the F_0 and E_e-

neutralized glottal source cues performed differently for emotional vowel perception in valence and arousal space for the two speakers. Moreover, the F_0 and E_e-neutralized glottal source cues to the valence of the perceived emotion for speaker 2, whereas the F_0 and E_e-neutralized glottal source cues to the arousal of the perceived emotion for speaker 1. The different effect of the F_0-modified glottal source cues was also reported in [103].

To further analyze the effects of the F_0 and E_e-neutralized glottal source and vocal tract cues on the emotional vowel perception in the valence and arousal space, the spectral tilt of glottal source waveform and the first formant frequency (F_1) of vocal tract were adopted for describing characterize the F_0 and E_e-neutralized glottal source and vocal tract cues, respectively. In this study, the spectral tilts of glottal source waveform were calculated by STRAIGHT [104] between 180-Hz and 700-Hz, and the averaged first formant frequencies (F_1) of vocal tract were calculated from the ARX model parameters.

The relationships of F_1 and the averaged perceptual scores of the re-synthesized emotional vowels in valence and arousal space are plotted in Fig. 5-10, in which the four perceptual scores corresponding to the given glottal source parameters and arbitrary vocal tract parameters are linearly regressed in the sense of minimum mean square error (MMSE). As seen in Fig. 5-10, it is clearly showed that there was a great scattering of F_1 of the synthesized emotional vowels with arbitrary glottal source and arbitrary vocal tract, as well as the relative differences in F_1 of the emotional vowels compared to those of the neutral vowels for the two speakers. The differences in F_1 of different emotional vowels were also reported in many studies [45, 105]. After linearly regressing the perceptual scores in the valence and arousal space of the emotional vowels synthesized with the given glottal source parameters, it was found that the varying range in F_1 of the re-synthesized emotional vowels for speaker 2 was much bigger than that for speaker 1. This was attributed to the individualized characteristics of F_1 in emotional vowels [61]. To further discuss the reliability of the linear regressions in Fig. 5-10, their coefficients of determination (R^2) were calculated, and were listed in the Table 5.1.

As seen in Fig. 5-10 and Table 5.1, it was showed that F_1 varied proportionally to the perceptual scores in valence for both speakers, while in arousal for only speaker 2. This indicated that emotions with higher averaged values of F_1 were usually positive perceptual scores in valence space, which is in line with the results in [40, 103]. Furthermore, it was

showed that the slopes of the regression lines of the perceptual scores for speaker 2 were much higher than those for speaker 1, which indicates the significant differences in the degree of F_1 variation against the perceptual scores in valence and arousal for different speakers. Reasons of this difference might be that different degree of the vocal tract were used by different speakers.

The relationships of the spectral tilt of the glottal source waveforms and the perceptual scores of the re-synthesized emotional vowels in arousal and valence space were plotted in Fig. 5-11, in which the four perceptual scores corresponding to the given vocal tract parameters and arbitrary glottal source parameters are linearly regressed in the MMSE sense. Fig. 5-11 indicates that the varying range of the spectral tilts of the glottal source waveforms for speaker 2 were largely bigger than those for speaker 1. The difference in the range of the spectral tilt of the glottal source waveforms could be from the different styles (e.g., individuality and degree of emotions) in the emotional vowel productions for different speakers. That individuality affects emotional speech was also found in article [106]. It is further showed that the spectral tilt of the glottal source waveforms varied proportionally to the perceptual scores in arousal space for both speakers, as seen in Fig. 5-11 (a). Noted also that the slopes of the regression lines for speaker 2 are similar to those for speaker 1, which demonstrated that for a given amount of spectral tilt change of the glottal source waveform, the change of perceptual scores of emotional vowels in arousal is roughly similar. Fig. 5-11 (b) demonstrates that no clear consistent patterns of spectral tilts of glottal source waveform for perceptual scores in valence, as is demonstrated also by the low coefficients values of determination in Table 5.1. This result is similar to the previous findings in article [103] which suggested no significant relation of the glottal source waveform with emotional perception in valence space.

Figs. 5-10 and 5-11 clearly indicate large differences in the range of F_1 and spectral tilt of the glottal source waveform for the both speakers. Specifically, the ranges of F_1 for speaker 2 were much bigger than those for speaker 1, while the ranges of spectral tilt of glottal source waveform for speaker 1 were much larger than those for speaker 2. These differences suggested that the speaker factor might have a significant effect on emotional vowel perception in valence and arousal space.

5.4 Conclusion

In this chapter, three experiments were performed to investigate the effects of the glottal source and vocal tract cues on emotional vowel perception in terms of valence and arousal space. Using the ARX-LF model with the proposed analysis-by-synthesis approach, the glottal source and vocal tract parameters were first estimated from the emotional Japanese vowels, followed by the controllable modifications of glottal source and vocal tract parameters, and then exploited for resynthesizing emotional vowels. The resynthsized emotional vowels were presented to native Japanese listeners with normal hearing for evaluating perceptual scores on valence and arousal space. From these results, the following findings were obtained:

1. The ARX-LF model-based analysis-by-synthesis approach thus helped to analyze the perceptual effects of the glottal source and vocal tract cues on emotional vowels perception in valence and arousal space. The effect of the glottal source cues was dominant on perception of emotional vowels relative to the vocal tract cues, which was in line with previous results in articles [93–95]. Furthermore, perception of emotions in vowels were found to be highly speaker-dependent, which was attributed in part to the large variation in emotional vowel production, such as inter-speaker differences in F_0.

2. The vocal tract cues effected to emotion perception in valence and arousal space, after neutralizing the effects of the F_0, E_e and duration cues. F_1 varied proportionally to the perceptual scores in valence and arousal, and the scattering of F_1 of the emotional vowels was different among different speakers, which was attributed to the individualized vocal tract characteristics during production of emotional vowels. The positive proportionality of the spectral tilt of the glottal source waveform to arousal perception was observed, with no clear relation to perception in valence space. Furthermore, the range of spectral tilt of the glottal source waveform was different for each speaker, most likely due to different styles of emotional speech production among speakers.

3. Comparison of results of Experiments I and II showed that the GS cues play an important role in perception of emotions in vowels, and that the emotions in the

synthesized vowels vary greatly across speakers, regardless of the neutralization of the F_0, E_e and duration parameters.

One limitation of this study is the small number of speakers (only two speakers). Though many differences in glottal source waveforms and vocal tract cues were found for the two speakers, some common findings were also found; for example, F_0 has an important role in perception of emotions in vowels, which was consistent with the previous study reported by [45].

The use of voiced vowel /a/ in the present study minimized the effects of other factors (e.g., prosodic and linguistic content) in order to examine the relative effects of the glottal source and vocal tract cues to emotion perception in vowels. It is believed that this study is useful to provide insight into emotion perception in speech. However, perception of emotions in speech is based on the combined effects of prosodic (both glottal and vocal tract information), and linguistic factors. All these factors should be taken into account in future work, allowing the acoustic parameters (e.g., F_0, duration and energy) to be modulated in a natural fashion and to be more easily generalizable to real-life utterances.

Also, this study focused on acoustic and perceptual characteristics of emotional vowels in Japanese. Since culture and language play important roles in emotional expression, future work is necessary to examine different language scenarios with respect to production and perception of emotional speech.

5.5 Contribution of parameters of the Lilijencrant-Fant model to emotional voice perception

The glottal source waveform and vocal tract related acoustic features were discussed in previous section, and results showed that glottal source plays a decisive role on perception of emotional vowel in V-A space. In this section, we focus on discussion of contributions of the LF model parameters on emotional voice perception for emotional voice conversion/synthesis.

The LF model parameters are divided into two categories: prosody-related parameters (E_e and $T0$), and the spectra-related parameters (T_p, T_e and T_a) of shape. Since the parameter T_c of the LF model is often set to T_0, it is not discussed in this section. It is

well known that controlling the fewer parameters is better for voice conversion/synthesis. Fant described a new parameter R_d [74, 107], it can be calculated from the T_0-normalized parameters (R_k, R_g and R_a). The relationships are following:

$$R_d = (\frac{1}{0.11}) \times (0.5 + 1.2 \times R_k) \times (\frac{R_k}{4 \times R_g} + R_a) \tag{5.1}$$

where, the parameters R_k, R_g and R_a are defined as following:

$$R_k = \frac{(T_e - T_p)}{T_p} \tag{5.2}$$

$$R_g = \frac{T_0}{2 \times T_p} \tag{5.3}$$

$$R_a = \frac{T_a}{T_0} \tag{5.4}$$

5.5.1 Contribution of prosody-related parameters of the Lilijencrant-Fant model to emotional voice perception

As seen in Figs. 5-7 and 5-9, in comparison, the perceptual scores in the valence and arousal space of the synthesized emotional vowels shows movements towards those of the synthesized neutral vowels, which demonstrates parameter $T_0(1/F_0)$ and E_e (intensity) in glottal source cues , also duration are greatly effects to perception in the valence and arousal space.

5.5.2 Contribution of spectra-related parameter of the Lilijencrant-Fant model to emotional voice perception

As seen in 5-9, parameter T_p, T_e and T_a also contribute to the valence and arousal space after remove the effects of the $F_0(1/T_0)$ and intensity (E_e) and duration cues.

The relationship between R_d and valence, between R_d and arousal are plotted in Fig 5-12. Fig 5-12 clearly showed that R_d is negatively correlated with arousal, high arousal level has low R_d value. But, no clearly relationship between valence and R_d.

The results can guide us to convert emotional voice using the ARX-LF model.

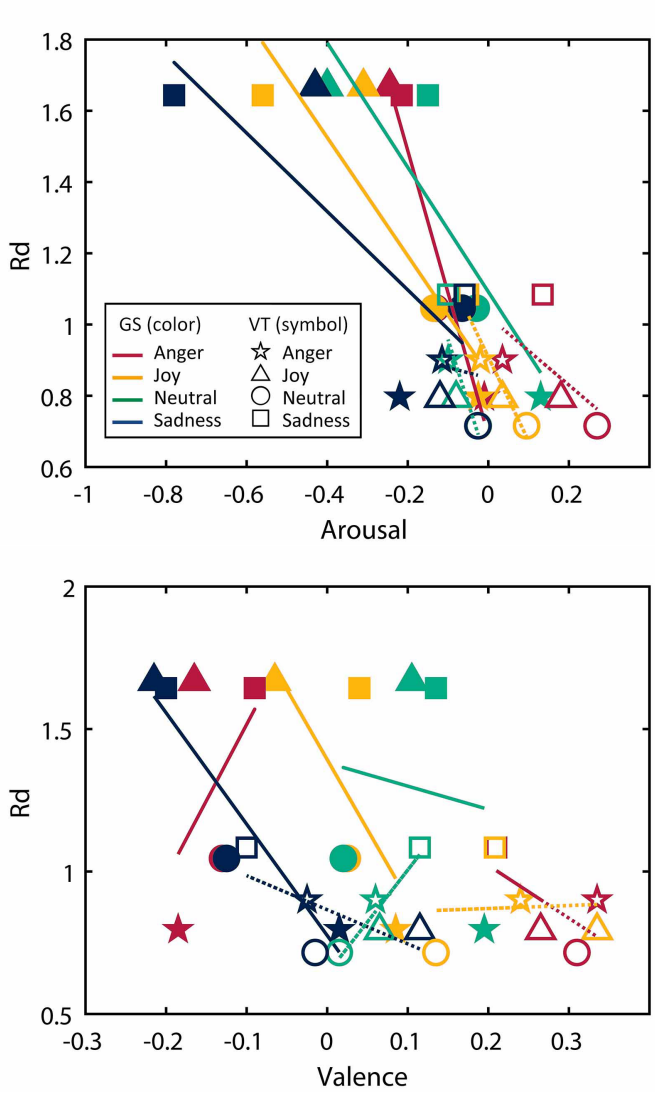

Figure 5-12: Relationship of parameter R_d to emotional voice perception.

Chapter 6

Conclusion

In this chapter, first, we summarized all of work of this study. Contributions are then discussed. Finally, future work is introduced.

6.1 Summary

The purpose of this research is to investigate the effects of glottal source and vocal tract cues on perception of emotional vowels from the point of view of speech production. The investigation starts with the speech production (source and filter) and links to the emotional speech perception (valence and arousal) via acoustic features. Thus, the investigation has three steps:

1. Estimating glottal source waveform and vocal tract shape from emotional vowels.
2. Discussing related acoustic features of glottal source waveform and vocal tract filter.
3. Discussing the effects of these features on perceptions of emotional vowel in V-A space.

For the first step:

Because existing methods are difficult to estimate glottal source waveform and vocal tract shape accurately, we first developed an analysis-by-synthesis method based on the ARX-LF model for estimating glottal source waveform and vocal tract shape of emotional vowel simultaneously. Considering that it is difficult to optimize multiple parameters of the LF mdoel, the EGG signal was employed to provide initial values for LF model.

Moreover, since GCI affects the accuracy of the LF model, the GCI are further shifted around estimated GCI. The proposed method was tested on synthetic and real voiced speech signals. The glottal source waveform and vocal tract shape were then estimated from Japanese vowel /a/ with four emotional states (neutral, joy, sadness and anger) by the proposed method. The spectral tilt and character of vocal tract shape were consistent with previous results.

Furthermore, we developed an analysis-by-synthesis system that has ability to estimate glottal source wave and vocal tract shape form speech signals for neutral voices. The experimental results outperform traditional widely used inverse filter method both on synthetic and real vowels.

For the second step: the following findings were obtained:

Source related features:

(1) The T_0 ($1/F_0$) of sadness and neutral were the largest, while that of anger and joy were the smallest.

(2) the maximum negative peak (intensity related parameter E_e) of glottal source waveform derivative of anger and joy was often the largest, while that of neutral and sadness were often the smallest.

(3) When compared with the spectral tilts of the glottal source waveform of neutral, those of anger and joy increased, and those of sadness decreased in the 200- to 700-Hz frequency range.

Filter related features:

The width of the front area function of anger was the largest, that of sadness was the smallest, and those of joy and neutral were in the middle. The large mouth open area results in higher F1 value stand for anger and joy.

For the third step: the following findings were obtained:

(1) The effect of the glottal source information was dominant on perception of emotional vowels relative to the vocal tract information. The prosody features of glottal source-related (pitch/F_0, intensity/E_e, and duration) play an important role in perception of emotions in vowels. Moreover, emotions in vowels were found to be speaker-dependent.

(2) After removing the effects of prosody features (F_0, intensity, and duration), glottal source and vocal tract cues still contribute to perception of emotional vowels in V-A

space. The positive proportionality of the spectral tilt of the glottal source waveform to arousal space was observed, with no clear relation to valence space. F_1 (vocal tract) varied proportionally to the perceptual scores in valence and arousal.

6.2 Contributions

The main contribution is that the results of this research can help us to understand production and perception of emotions in vowels. Although previous studies of emotional speech analysis were based on the acoustical features, such as F_0, intensity, and duration. However, there are many source and filter related features contributing to the perception of emotions. This study gives a systematic investigation of emotional vowels that from production to perception.

The second contribution of this research is that the results can enlighten the fields of emotional speech recognition, synthesis, and conversion. It provides an important basis information for the fields of emotional engineering.

The third contribution of this research is that we developed a analysis-by-synthesis method that can simultaneously estimate glottal source waveform and vocal tract shape based on ARX-LF model. It is expected to be used in many speech signal processing fields, such as speech analysis and speech synthesis.

6.3 Future work

This study focused on the investigation of Japanese vowels. As we know, language also play important roles in emotional expression. Thus, we will examine the different languages scenarios with respect to production and perception of emotional speech, such as Chinese and English.

It is well known that emotional speech contains not only vowels but also consonants. Present study investigated the effects of glottal source waveform and vocal tract shape cues on perception of emotional vowels. Thus, consonants will be analyzed in the future.

Also, in the future, we will focus on the emotional speech conversion. The conversion rules of glottal source parameters and vocal tract parameters will be constructed, and

emotional speech conversion system will be developed based on our proposed analysis-by-synthesis tool.

Bibliography

[1] S. Tomkins, *Affect imagery consciousness: Volume I: The positive affects*. Springer publishing company, 1962.

[2] C. E. Izard, *Human emotions*. Springer Science & Business Media, 2013.

[3] C. E. Izard, *Patterns of emotions: A new analysis of anxiety and depression*. Academic Press, 2013.

[4] M. Wenger, "Emotion as visceral action: An extension of lange's theory," 1950.

[5] E. Gellhorn, "Motion and emotion: The role of proprioception in the physiology and pathologyof the emotions.," *Psychological Review*, vol. 71, no. 6, p. 457, 1964.

[6] H. F. Dunbar, *Emotions and bodily changes: a survey of literature on psychosomatic interrelationships*. Columbia Univ. Press, 1938.

[7] H. Fujisaki, "Information, prosody, and modeling-with emphasis on tonal features of speech," in *Speech Prosody 2004, International Conference*, 2004.

[8] P. N. Juslin and K. R. Scherer, "Vocal expression of affect," *The new handbook of methods in nonverbal behavior research*, pp. 65–135, 2005.

[9] E. R. Skinner, "A calibrated recording and analysis of the pitch, force and quality of vocal tones expressing happiness and sadness; and a determination of the pitch and force of the subjective concepts of ordinary, soft, and loud tones," *Communications Monographs*, vol. 2, no. 1, pp. 81–137, 1935.

[10] E. Kramer, "Judgment of personal characteristics and emotions from nonverbal properties of speech.," *Psychological Bulletin*, vol. 60, no. 4, p. 408, 1963.

[11] M. Alpert, R. L. Kurtzberg, and A. J. Friedhoff, "Transient voice changes associated with emotional stimuli," *Archives of General Psychiatry*, vol. 8, no. 4, pp. 362–365, 1963.

[12] W. A. Hargreaves, J. Starkweather, and K. Blacker, "Voice quality in depression.," *Journal of Abnormal Psychology*, vol. 70, no. 3, p. 218, 1965.

[13] C. Caffi and R. W. Janney, "Toward a pragmatics of emotive communication," *Journal of pragmatics*, vol. 22, no. 3-4, pp. 325–373, 1994.

[14] C. Lisetti, "Affective computing," in *Springer-Verlag*, London Limited, 1998, pp. 71–73.

[15] K. R. Scherer, "Vocal communication of emotion: A review of research paradigms," *Speech communication*, vol. 40, no. 1-2, pp. 227–256, 2003.

[16] J. A. Russell, "A circumplex model of affect.," *Journal of personality and social psychology*, vol. 39, no. 6, p. 1161, 1980.

[17] D. A. Sauter, F. Eisner, A. J. Calder, and S. K. Scott, "Perceptual cues in nonverbal vocal expressions of emotion," *The quarterly journal of experimental psychology*, vol. 63, no. 11, pp. 2251–2272, 2010.

[18] K. R. Scherer, "Emotion as a multicomponent process: A model and some cross-cultural data," *Review of Personality & Social Psychology*, vol. 5, pp. 37–63, 1984.

[19] M. Grimm, K. Kroschel, E. Mower, and S. Narayanan, "Primitives-based evaluation and estimation of emotions in speech," *Speech Communication*, vol. 49, no. 10-11, pp. 787–800, 2007.

[20] H. Wu, D. Jiang, Y. Zhao, and H. Sahli, "Dimensional emotion driven facial expression synthesis based on the multi-stream DBN model," in *Signal & Information Processing Association Annual Summit and Conference (APSIPA ASC), Asia-Pacific*, IEEE, 2012, pp. 1–6.

[21] E. Brunswik, "Historical and thematic relations of psychology to other sciences," *The Scientific Monthly*, vol. 83, no. 3, pp. 151–161, 1956.

[22] K. R. Scherer, "Personality inference from voice quality: The loud voice of extroversion," *European Journal of Social Psychology*, vol. 8, no. 4, pp. 467–487, 1978.

[23] C.-F. Huang and M. Akagi, "A three-layered model for expressive speech perception," *Speech Communication*, vol. 50, no. 10, pp. 810–828, 2008.

[24] R. Elbarougy and M. Akagi, "Improving speech emotion dimensions estimation using a three-layer model of human perception," *Acoustical science and technology*, vol. 35, no. 2, pp. 86–98, 2014.

[25] Y. Xue, Y. Hamada, and M. Akagi, "Voice conversion for emotional speech: Rule-based synthesis with degree of emotion controllable in dimensional space," *Speech Communication*, vol. 102, pp. 54–67, 2018.

[26] C. E. Williams and K. N. Stevens, "Emotions and speech: Some acoustical correlates," *The Journal of the Acoustical Society of America*, vol. 52, no. 4B, pp. 1238–1250, 1972.

[27] I. R. Murray and J. L. Arnott, "Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion," *The Journal of the Acoustical Society of America*, vol. 93, no. 2, pp. 1097–1108, 1993.

[28] D. Erickson, "Expressive speech: Production, perception and application to speech synthesis," *Acoustical Science and Technology*, vol. 26, no. 4, pp. 317–325, 2005.

[29] A. Paeschke, "Global trend of fundamental frequency in emotional speech," in *Speech Prosody*, 2004, pp. 671–674.

[30] K. R. Scherer, R. Banse, H. G. Wallbott, and T. Goldbeck, "Vocal cues in emotion encoding and decoding," *Motivation and emotion*, vol. 15, no. 2, pp. 123–148, 1991.

[31] Y. Hashizawa, S. Takeda, M. D. Hamzah, and G. Ohyama, "On the differences in prosodic features of emotional expressions in Japanese speech according to the degree of the emotion," in *Speech Prosody*, 2004, pp. 655–658.

[32] T. Ehrette, N. Chateau, C. d'Alessandro, and V. Maffiolo, "Prosodic parameters of perceived emotions in vocal server voices," in *Speech Prosody*, 2002, pp. 259–262.

[33] K. Maekawa, "Phonetic and phonological characteristics of paralinguistic information in spoken Japanese," in *Fifth International Conference on Spoken Language Processing*, vol. 2, 1998, pp. 635–638.

[34] L. Leinonen, T. Hiltunen, I. Linnankoski, and M.-L. Laakso, "Expression of emotional–motivational connotations with a one-word utterance," *The Journal of the Acoustical society of America*, vol. 102, no. 3, pp. 1853–1863, 1997.

[35] I. Yanushevskaya, C. Gobl, and A. N. Chasaide, "Mapping voice to affect: Japanese listeners," in *Proceedings of Speech Prosody*, 2006, pp. 1–4.

[36] S. J. L. Mozziconacci, *Speech variability and emotion: Production and perception.* Technische Universiteit Eindhoven Eindhoven, 1998.

[37] P. N. Juslin and P. Laukka, "Impact of intended emotion intensity on cue utilization and decoding accuracy in vocal expression of emotion.," *Emotion*, vol. 1, no. 4, p. 381, 2001.

[38] T. Bänziger and K. R. Scherer, "The role of intonation in emotional expressions," *Speech communication*, vol. 46, no. 3-4, pp. 252–267, 2005.

[39] N. Audibert, V. Aubergé, and A. Rilliard, "The prosodic dimensions of emotion in speech: The relative weights of parameters," in *Interspeech'2005-Eurospeech-9th European Conference on Speech Communication and Technology*, 2005, pp. 525–528.

[40] H. Mori and H. Kasuya, "Voice source and vocal tract variations as cues to emotional states perceived from expressive conversational speech," in *INTERSPEECH*, Antwerp, Belgium, 2007, pp. 27–31.

[41] B. Schuller, G. Rigoll, and M. Lang, "Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture," in *Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP'04). IEEE International Conference on*, IEEE, vol. 1, 2004, pp. I–577.

[42] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan, *et al.*, "The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing," *IEEE Transactions on Affective Computing*, vol. 7, no. 2, pp. 190–202, 2016.

[43] D. G. Childers and C. Ahn, "Modeling the glottal volume-velocity waveform for three voice types," *The Journal of the Acoustical Society of America*, vol. 97, no. 1, pp. 505–519, 1995.

[44] K. R. Scherer, "Vocal affect expression: A review and a model for future research.," *Psychological bulletin*, vol. 99, no. 2, p. 143, 1986.

[45] D. Erickson, K. Yoshida, C. Menezes, A. Fujino, T. Mochida, and Y. Shibuya, "Exploratory study of some acoustic and articulatory characteristics of sad speech," *Phonetica*, vol. 63, no. 1, pp. 1–25, 2006.

[46] C. M. Lee, S. Yildirim, M. Bulut, A. Kazemzadeh, C. Busso, Z. Deng, S. Lee, and S. Narayanan, "Emotion recognition based on phoneme classes," in *ICSLP*, 2004, pp. 2193–2196.

[47] M. Airas and P. Alku, "Emotions in vowel segments of continuous speech: Analysis of the glottal flow using the normalised amplitude quotient," *Phonetica*, vol. 63, no. 1, pp. 26–46, 2006.

[48] F. Ringeval and M. Chetouani, "A vowel based approach for acted emotion recognition," in *Ninth Annual Conference of the International Speech Communication Association*, 2008.

[49] G. Fant, J. Liljencrants, and Q.-G. Lin, "A four-parameter model of glottal flow," *STL-QPSR*, vol. 4, no. 1985, pp. 1–13, 1985.

[50] H.-L. Lu, "Toward a high-quality singing synthesizer with vocal texture control," PhD thesis, Stanford University, 2002.

[51] T. Raitio, *Voice source modelling techniques for statistical parametric speech synthesis*. Aalto University, 2015.

[52] G. Fant, "Acoustic theory of speech production," 1970.

[53] H. Fujisaki and M. Ljungqvist, "Estimation of voice source and vocal tract parameters based on ARMA analysis and a model for the glottal source waveform," in *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'87.*, IEEE, vol. 12, 1987, pp. 637–640.

[54] D. H. Klatt and L. C. Klatt, "Analysis, synthesis, and perception of voice quality variations among female and male talkers," *the Journal of the Acoustical Society of America*, vol. 87, no. 2, pp. 820–857, 1990.

[55] M. M. Sondhi, "Resonances of a bent vocal tract," *The Journal of the Acoustical Society of America*, vol. 79, no. 4, pp. 1113–1116, 1986.

[56] J. D. Markel and A. J. Gray, *Linear prediction of speech*. Springer-Verlag, 1976.

[57] P. R. Cook, *Identification of control parameters in an articulatory vocal tract model, with applications to the synthesis of singing*. Stanford University, 1991.

[58] H. Wakita, "Direct estimation of the vocal tract shape by inverse filtering of acoustic speech waveforms," *IEEE Transactions on Audio and Electroacoustics*, vol. 21, no. 5, pp. 417–427, 1973.

[59] D. Erickson, "Acoustic and articulator analysis of sad Japanese speech," *Proc. Fall Meet. Phonet. Soc. Jpn., 2004*, 2004.

[60] A. Li, Q. Fang, F. Hu, L. Zheng, H. Wang, and J. Dang, "Acoustic and articulatory analysis on mandarin Chinese vowels in emotional speech," in *Chinese Spoken Language Processing (ISCSLP), 2010 7th International Symposium on*, Tainan,Taiwan: IEEE, 2010, pp. 38–43.

[61] D. Erickson, C. Zhu, S. Kawahara, and A. Suemitsu, "Articulation, acoustics and perception of mandarin Chinese emotional speech," *Open Linguistics*, vol. 2, no. 1, 2016.

[62] S. Lee, E. Bresch, J. Adams, A. Kazemzadeh, and S. Narayanan, "A study of emotional speech articulation using a fast magnetic resonance imaging technique," in *INTERSPEECH*, Pittsburgh,USA, 2006, pp. 1792–1795.

[63] J. Kim, A. Toutios, Y.-C. Kim, Y. Zhu, S. Lee, and S. Narayanan, "USC-EMO-MRI corpus: An emotional speech production database recorded by real-time magnetic resonance imaging," in *International Seminar on Speech Production (ISSP)*, Cologne, Germany, 2014, pp. 226–229.

[64] T. Kitamura, "Similarity of effects of emotions on the speech organ configuration with and without speaking," in *INTERSPEECH*, 2010, pp. 909–912.

[65] G. Degottex, E. Bianco, and X. Rodet, "Usual to particular phonatory situations studied with high-speed videoendoscopy," in *International Conference on Voice Physiology and Biomechanics*, Tampere, Finland, 2008, pp. 19–26.

[66] L. R. Rabiner and R. W. Schafer, *Digital processing of speech signals*. Prentice-hall Englewood Cliffs, NJ, 1978, vol. 100.

[67] D. Wong, J. Markel, and A. Gray, "Least squares glottal inverse filtering from the acoustic speech waveform," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 27, no. 4, pp. 350–355, 1979.

[68] P. Alku, "Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering," *Speech Communication*, vol. 11, no. 2-3, pp. 109–118, 1992.

[69] T. Drugman, B. Bozkurt, and T. Dutoit, "Complex cepstrum-based decomposition of speech for glottal source estimation," *in Proc. Interspeech Conf.*, 2009.

[70] J. Kane and C. Gobl, "Automating manual user strategies for precise voice source analysis," *Speech Communication*, vol. 55, no. 3, pp. 397–414, 2013.

[71] D. Vincent, O. Rosec, and T. Chonavel, "Estimation of LF glottal source parameters based on an ARX model," in *INTERSPEECH*, Lisbon, Portugal, 2005, pp. 333–336.

[72] Q. Fu and P. Murphy, "Robust glottal source estimation based on joint source-filter model optimization," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 2, pp. 492–501, 2006.

[73] G. Degottex, A. Roebel, and X. Rodet, "Phase minimization for glottal model estimation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 5, pp. 1080–1090, 2011.

[74] G. Fant, "The LF-model revisited. transformations and frequency domain analysis," *Speech Trans. Lab. Q. Rep., Royal Inst. of Tech. Stockholm*, vol. 2, no. 3, p. 40, 1995.

[75] D. G. Childers and C. Ahn, "Modeling the glottal volume-velocity waveform for three voice types," *The Journal of the Acoustical Society of America*, vol. 97, no. 1, pp. 505–519, 1995.

[76] H. Kawahara, K.-I. Sakakibara, H. Banno, M. Morise, T. Toda, and T. Irino, "Aliasing-free implementation of discrete-time glottal source models and their applications to speech synthesis and f0 extractor evaluation," in *Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2015 Asia-Pacific*, IEEE, 2015, pp. 520–529.

[77] D. G. Childers and C. Lee, "Vocal quality factors: Analysis, synthesis, and perception," *the Journal of the Acoustical Society of America*, vol. 90, no. 5, pp. 2394–2410, 1991.

[78] T. Dubuisson, "Glottal source estimation and automatic detection of dysphonic speakers," PhD thesis, University of Mons, 2012.

[79] A. Kounoudes, P. A. Naylor, and M. Brookes, "The DYPSA algorithm for estimation of glottal closure instants in voiced speech," in *ICASSP*, 2002, pp. 349–352.

[80] P. A. Naylor, A. Kounoudes, J. Gudnason, and M. Brookes, "Estimation of glottal closure instants in voiced speech using the DYPSA algorithm," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 1, pp. 34–43, 2007.

[81] M. R. Thomas, J. Gudnason, and P. A. Naylor, "Estimation of glottal closing and opening instants in voiced speech using the YAGA algorithm," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 1, pp. 82–91, 2012.

[82] K. S. R. Murty and B. Yegnanarayana, "Epoch extraction from speech signals," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 8, pp. 1602–1613, 2008.

[83] T. Drugman and T. Dutoit, "Glottal closure and opening instant detection from speech signals," in *INTERSPEECH*, 2009, pp. 2891–2894.

[84] J. Kane and C. Gobl, "Evaluation of glottal closure instant detection in a range of voice qualities," *Speech Communication*, vol. 55, no. 2, pp. 295–314, 2013.

[85] T. Drugman, M. Thomas, J. Gudnason, P. Naylor, and T. Dutoit, "Detection of glottal closure instants from speech signals: A quantitative review," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 3, pp. 994–1006, 2012.

[86] T. Drugman, B. Bozkurt, and T. Dutoit, "A comparative study of glottal source estimation techniques," *Computer Speech & Language*, vol. 26, no. 1, pp. 20–34, 2012.

[87] J. Kane and C. Gobl, "Evaluation of automatic glottal source analysis," in *International Conference on Nonlinear Speech Processing*, Springer, 2013, pp. 1–8.

[88] M. Schröder, R. Cowie, E. Douglas-Cowie, M. Westerdijk, and S. Gielen, "Acoustic correlates of emotion dimensions in view of speech synthesis," in *Seventh European Conference on Speech Communication and Technology*, 2001, pp. 87–90.

[89] Y. Li, K.-I. Sakakibara, D. Morikawa, and M. Akagi, "Commonalities of glottal sources and vocal tract shapes among speakers in emotional speech," in *International Seminar on Speech Production (ISSP)*, Springer, 2017, pp. 24–34.

[90] M. Airas, "TKK Aparat: An environment for voice inverse filtering and parameterization," *Logopedics Phoniatrics Vocology*, vol. 33, no. 1, pp. 49–64, 2008.

[91] B. Yegnanarayana and R. N. Veldhuis, "Extraction of vocal-tract system characteristics from speech signals," *IEEE transactions on Speech and Audio Processing*, vol. 6, no. 4, pp. 313–327, 1998.

[92] J. Kane, I. Yanushevskaya, A. Ní Chasaide, and C. Gobl, "Exploiting time and frequency domain measures for precise voice source parameterisation," in *Speech Prosody*, 2012, pp. 143–146.

[93] R. Sun, E. Moore, and J. F. Torres, "Investigating glottal parameters for differentiating emotional categories with similar prosodics," in *ICASSP*, Taibei, Taiwan: IEEE, 2009, pp. 4509–4512.

[94] J. Sundberg, S. Patel, E. Bjorkner, and K. R. Scherer, "Interdependencies among voice source parameters in emotional speech," *IEEE Transactions on Affective Computing*, vol. 2, no. 3, pp. 162–174, 2011.

[95] T. Waaramaa, A.-M. Laukkanen, M. Airas, and P. Alku, "Perception of emotional valences and activity levels from vowel segments of continuous speech," *Journal of voice*, vol. 24, no. 1, pp. 30–38, 2010.

[96] D. C. Rubin and J. M. Talarico, "A comparison of dimensional models of emotion: Evidence from emotions, prototypical events, autobiographical memories, and words," *Memory*, vol. 17, no. 8, pp. 802–808, 2009.

[97] P. Gangamohan, S. R. Kadiri, and B. Yegnanarayana, "Analysis of emotional speech-A review," in *Toward Robotic Socially Believable Behaving Systems-Volume I*, Springer, 2016, pp. 205–238.

[98] P. Alku, "Glottal inverse filtering analysis of human voice production-A review of estimation and parameterization methods of the glottal excitation and their applications," *Sadhana*, vol. 36, no. 5, pp. 623–650, 2011.

[99] M. Rothenberg, "A new inverse-filtering technique for deriving the glottal air flow waveform during voicing," *The Journal of the Acoustical Society of America*, vol. 53, no. 6, pp. 1632–1645, 1973.

[100] V. Auberge and M. Cathiard, "Can we hear the prosody of smile?" *Speech Communication*, vol. 40, no. 1, pp. 87–97, 2003.

[101] J. Tao, Y. Li, and S. Pan, "A multiple perception model on emotional speech," in *Affective Computing and Intelligent Interaction and Workshops, 2009. ACII 2009. 3rd International Conference on*, IEEE, 2009, pp. 1–6.

[102] N. Audibert, D. Vincent, V. Auberge, and O. Rosec, "Expressive speech synthesis: Evaluation of a voice quality centered coder on the different acoustic dimensions," in *Proc. Speech Prosody*, vol. 2006, Dresden, Germany, 2006, pp. 525–528.

[103] A.-M. Laukkanen, E. Vilkman, P. Alku, and H. Oksanen, "On the perception of emotions in speech: The role of voice quality," *Logopedics Phoniatrics Vocology*, vol. 22, no. 4, pp. 157–168, 1997.

[104] H. Kawahara, I. Masuda-Katsuse, and A. De Cheveigne, "Restructuring speech representations using a pitch-adaptive time–frequency smoothing and an instantaneous-frequency-based f0 extraction: Possible role of a repetitive structure in sounds," *Speech communication*, vol. 27, no. 3, pp. 187–207, 1999.

[105] M. Goudbeek, J. P. Goldman, and K. R. Scherer, "Emotion dimensions and formant position," in *INTERSPEECH*, 2009, pp. 1575–1578.

[106] M. Bulut and S. Narayanan, "On the robustness of overall F0-only modifications to the perception of emotions in speech," *The Journal of the Acoustical Society of America*, vol. 123, no. 6, pp. 4547–4558, 2008.

[107] G. Fant, "The voice source in connected speech," *Speech communication*, vol. 22, no. 2-3, pp. 125–139, 1997.

Publications

Journal

[1] **Yongwei Li**, Junfeng Li and Masato Akagi, "Contributions of the glottal source and vocal tract cues to emotional vowel perception in the valence-arousal space," The Journal of the Acoustical Society of America, Vol. 144, no. 2, pp. 908-916, 2018.

Book Chapter

[2] **Yongwei Li**, Ken-Ichi Sakakibara, Daisuke Morikawa and Masato Akagi, "Commonalities of glottal sources and vocal tract shapes among speakers in emotional speech," Studies on Speech Production, Lecture Notes in Computer Science, Springer, Vol. 10733, pp. 24-34, 2018.

International Conference

[3] **Yongwei Li**, Ken-Ichi Sakakibara and Masato Akagi, "Estimation of glottal source waveforms and vocal tract shapes from speech signals based on ARX-LF model," The 11th International Symposium on Chinese Spoken Language Processing (ISCSLP), November, 2018.

[4] **Yongwei Li**, Ken-Ichi Sakakibara, Daisuke Morikawa and Masato Akagi, "Commonalities of glottal sources and vocal tract shapes among speakers in emotional

speech," The 11th international seminar on speech production (ISSP), pp.79-81, October, 2017.

[5] **Yongwei Li**, Daisuke Morikawa and Masato Akagi, "A method to estimate glottal source waves and vocal tract shapes for widely pronounced types using ARX-LF model," The Journal of the Acoustical Society of America, 140 (4), pp. 2963-2963, 2016.

[6] **Yongwei Li** and Msato Akagi, "Glottal source analysis of emotional speech," RISP International workshop on nonlinear circuits, communications and signal processing (NCSP ' 14), pp.513-516, March, 2014.

Domestic Conference

[7] **Yongwei Li**, Ken-Ichi Sakakibara and Masato Akagi, "Simultaneous estimation of glottal source waveform and vocal tract shape from speech signal based on ARX-LF model," ASJ Fall Meeting, September, 2018.

[8] **Yongwei Li**, Ken-Ichi Sakakibara and Masato Akagi, "Relationships between features of glottal sources and vocal tract shapes and perceived positions on valence and activation in emotional speech," ASJ Fall Meeting, September, 2017.

[9] **Yongwei Li**, Ken-Ichi Sakakibara, Daisuke Morikawa and Masato Akagi, "Commonalities and difference of glottal sources and vocal tract shapes among speakers in emotional speech," ASJ Spring Meeting, 3-Q-28, pp. 1475-1478, March, 2017.

[10] **Yongwei Li**, Daisuke Morikawa and Masato Akagi, "Estimation of glottal source waves and vocal tract shapes of emotional speech using ARX-LF model," ASJ Fall Meeting, 2-Q-36, pp. 215-218, September, 2016.

[11] **Yongwei Li**, Yasuhiro Hamada and Masato Akagi, "Analysis of glottal source waves for emotional speech using ARX-LF model," IPSJ SIG technical report, Vol. 2015-MUS-107 No.69, pp. 1-4, May, 2015.

[12] **Yongwei Li**, Yasuhiro Hamada and Masato Akagi, "Analysis of glottal source waves for emotional speech using ARX-LF model„" ASJ Spring Meeting, 2-Q-44, pp. 407-410, March, 2015.

[13] Yasuhiro Hamada, Elbarougy Reda, **Yongwei Li** and Masato Akagi, "Study on method to control acoustic features related to the position on the Valence-Activation space," ASJ Spring Meeting, 2-Q-45, pp 411-414, March, 2015.