HOKKAIDO UNIVERSITY

DOCTORAL THESIS

Analysis on Acoustical and Perceptual Characteristics of Whispered Speech and Whisper-to-Normal Speech Conversion

---

Author:

Hideaki KONNO

Supervisor:

Dr. Mineichi KUDO

A thesis submitted in fulfillment of the requirements

for the degree of Doctor of Philosophy

in the

Division of Computer Science

Graduate School of Information Science and Technology

December 2016

HOKKAIDO UNIVERSITY

Abstract

Graduate School of Information Science and Technology

Doctor of Philosophy

Analysis on Acoustical and Perceptual Characteristics of Whispered Speech and Whisper-to-Normal Speech Conversion

by Hideaki KONNO

Conversion of whispered speech into normal speech has been recently studied in order to enable people, who have injured their vocal cords, to communicate using their own voice without help of medical equipment such as an electric artificial larynx. This thesis tries to make clear the characteristics of Japanese whispered speech through analyses of comparing it with normally phonated one, and proposes a method of improving prosody of converted speech using the knowledge obtained by the analyses.

There are phonemic quality and prosody as properties of speech. Phonemic quality is a sort of timbre by which phonemes are distinguished, whereas prosody indicates accent, intonation, rhythm etc., which is observed in a longer period than that of phonemes. Phonemic quality conveys most of linguistic information of speech, while prosody plays an important role for carrying non-linguistic information such as emotion in addition to carrying linguistic information.

The vocal cords vibration classifies speech as voiced or unvoiced one, and is concerned with phonemic quality of consonants. Prosody of normal

speech is affected by the pitch which is mainly determined by the fundamental frequency (F_0), that is, the number of vibration of vocal cords. Therefore, the vibration of the vocal cords relates to both of phonemic quality and prosody of normal speech.

Whispering is a mode of speech used in daily situations such that people cannot speak loudly. Whispered speech is generated by a sound source of turbulence air flow passing through the vocal cords which is not vibrating. Therefore, whispered speech is always unvoiced and has no F_0. Nevertheless, whispered speech can communicate phonemic quality and prosody like normal speech does. This study approached this secret by the method of mathematical analysis.

The most important acoustical characteristics related to the phonemic quality are the spectral envelopes, or the vocal tract characteristics, represented by the formant frequencies of vowels. Whispered vowels have formant structure as voiced vowels. However, in many languages, it has been reported that the formant frequencies of whispered vowels are different from those of voiced vowels even though both modes of vowels are the same phoneme. Although the reason of difference is explained in the side of speech generation, it is not clear how the difference effects on the phonemic quality of whispered vowels in the side of speech perception. On the other hand, pitch is one of the attributes of auditory sensation that causes the impression being from low to high, and conveys speech prosody. Being different from normal speech, whispered speech has no F_0, hence it may be appropriate that another feature substitutes for F_0. However, the acoustic features related with pitch of whispered speech have not clarified enough, and the method of representing the pitch quantitatively has not established yet. Therefore, in this study, the author has carried out the acoustic analyses and perceptual experiments and then investigated the method of converting whispered speech into normal speech.

It is necessary to generate F_0 contour for conversion of whispered to normal speech. The conventional study gave weight to the compensation of the formants or spectral envelope, and generation of the F_0 contour has not studied enough. Although the method of generating F_0 contour for the natural conversion has been developed recently, no investigation has been done on the viewpoint of reproducing prosody of whisper. Thus, this study aimed to keep the original prosody of whisper in the converted speech by estimating pitch of whisper quantitatively and generating F_0 contour based on that. The outline of this thesis is as follows.

Chapter 1 describes background and objective of this study.

In Chapter 2, after introducing the conventional study related to phonemic quality of whisper, acoustic analyses and auditory experiments of Japanese whispered speech are described. As for whispered vowels, the low frequency formants of whispered vowels tend to shift to higher frequency regions compared with those of normally phonated vowels. By the auditory experiments using synthetic speech, the change of phonemic quality occurred by the difference of glottal sources even if the formant frequencies are unchanged. This could be caused by the change of pitch. Moreover, the author showed the identification rate of whisper decreased by 20% to normal speech by phoneme identification experiments using Japanese 110 monosyllables, and showed a possibility to reflect the balance of energy over low and high frequency region upon perception and generation of voiced or unvoiced consonants differed from normally phonated consonants.

Chapter 3 analyzes whispered vowels of various pitch following description of the past studies of pitch of whisper, and confirms that formant frequencies and distribution of spectral energy of whispered vowels changes according to pitch. In addition, this chapter explains on construction of the multiple regression model to predict pitch values of whispered vowels by applying the principal analysis and the multiple regression analysis to results of mel filter bank analysis.

Chapter 4 introduces research on whisper to normal speech that has been studied so far, and proposes the conversion method that uses pitch values estimated from whispered speech described in Chapter 3 as F_0. Auditory experiments demonstrated that the correctly perceived rate of Japanese word accent was increased from 55.5% to 72.0% compared with that when a constant F_0 was used.

Chapter 5 and 6 describe discussions, future works, and conclusions.

The contributions of this study can be summarized as follows: 1) described the difference of formant frequencies between whispered and voiced Japanese vowels considering its effect to phonemic quality of whispered vowels, 2) developed the recording procedure that is able to obtain a pitch value of each whispered vowel, and analyzed quantitatively the relation between pitch values and spectral shape of the whispered vowels, 3) developed the method of making equations to predict pitch contour from whispered speech using the results of the analyses, and proposed the converting method from whispered to normal speech preserving prosody of the whisper.

Acknowledgments

I wish to express my appreciation to a lot of people who have supported and helped me.

I am deeply grateful to Dr. Mineichi Kudo, the supervisor of my doctoral thesis. Without his support, advice, and encouragement throughout the days of my doctoral course, I could not have completed this thesis.

I would like to thank the sub-supervisors Dr. Hideyuki Imai and Dr. Masanori Sugimoto for appropriate and insightful comments and advice.

I would like to express grateful thanks to Dr. Masaru Shimbo, Emeritus Professor of Hokkaido University, who taught me the basis of research at his laboratory when I was a senior and a master student of Hokkaido University. He had also been advising me for many years after I finished my master course and began to work.

I would like to thank Jun Toyama who led me to the field of speech research when I was a senior student and coached me for my master's degree.

I would appreciate to Dr. Kazumi Murata, Emeritus Professor of Hokkaido University, who guided me when I worked at Hokkaido Institute of Technology that is currently named Hokkaido University of Science.

I am grateful to the staff of Hakodate Campus, Hokkaido University of Education where I currently work, especially to Dr. Hideo Kanemitsu and Dr. Nobuyuki Takahashi.

Thanks to my parents, Eiji and Motoko Konno, I could study at an undergraduate and a master course of Hokkaido University. I am very grateful to their financial support and warm eyes.

I wish to express my special thanks to my daughters, Kanae and Haruna Konno, and lastly my wife Fumie Konno for their understanding to my study, patience and moral support.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

Whispering is a mode of speech used in daily situations such that people need to keep a conversation among a few people, or not to disturb others in public places such as a library, a movie theater, a meeting room, etc. Whispered speech is generated by a sound source of turbulence air flow passing through the vocal cords which is not vibrating. Therefore, whispered speech is always unvoiced and has no fundamental frequency (F_0). There are some other differences of their respective acoustic characteristics between whispered and normally phonated speech. Nevertheless, whispered speech can communicate the same linguistic information as ordinary speech. In this respect, whispering is an interesting object for studying speech recognition and perception.

In addition, whispering has become more important these days due to the widespread use of smart phones. Besides, some people, whose larynx has been affected by an accident or disease, may only converse by whispering. Thus, there is a need for an effective method to convert whispered speech into normal speech to make the whispered speech more intelligible over the phone or to improve communication for impaired persons.

1.2 Background and objective

The nature of whispered speech has been investigated from some points of view. For example, Schwartz [51] reported that the long-term power spectra of whispered speech show a reduction in the intensity, and flattening of the spectrum compared with those of normally phonated speech as well.

Recently, various techniques for converting whispered speech to normal speech have been developed with the aim of improving intelligibility and naturalness [16], [43], [48], [52], [62], and conversion from other kinds of speech, e.g., body-conducted speech, to whispered speech has also been studied [14], [61].

Two things are necessary for conversion of whispered speech to normal speech: modification of the vocal tract characteristics and generation of a F_0 contour. The vocal tract characteristics have to be modified because they are different in whispered speech and normally phonated speech even if the same phonemes are being uttered. In whispered speech, perceived pitch may be generated by changing the vocal tract characteristics. Therefore, we need to recover the vocal tract characteristics in a synthetic system. Another ingredient for normal speech synthesis is the F_0 contour, which conveys intonation and pitch accent in normally phonated speech. The methodology of generating the F_0 contour is classified into five groups: 1) the use of a fixed F_0 for any word [52], 2) use of a contour model of F_0 for individual words [16], 3) estimation of the F_0 contour on the basis of the power of the speech [48], 4) application of the F_0 contour estimated from the normal speech which has the same phrase as the whispered speech [47], [61], [62], and 5) estimation of the F_0 contour from the spectral shape including formants [43], [44]. Note that the first two groups give the same fixed F_0 contour regardless of the uttered words or

only a slight variation of the F_0 contour. The third group does not correct pitch contour unless appropriate feedback control of speech gain is made. The fourth group needs pairs of normal and whispered speech of the same sentence for designing the system. This is sometimes too costly and impractical, e.g., in cases in which the normal voice has been lost due to disease. Therefore, in this study, we used the fifth method in which a dynamic pitch contour is generated depending on the uttered whispered speech [34].

In this study, we first considered the relationship between phonemic quality and acoustic characteristics of whispered speech. Next, the recording procedure was developed for obtaining whispered vowels having their perceived pitch values. Then, the relationship between the pitch value and short-term spectral features of five Japanese whispered vowels was quantitatively analyzed. Based on the results, we derived a regression formula to estimate the pitch value from the spectrum of whispered vowels. Using this estimation technique, we proposed the method of whispered to normal speech conversion preserving prosody of the original whispered speech.

1.3 Organization of thesis

Chapter 1 has described motivation, background, objective of the current thesis. Chapter 2 describes analyses on phonemic quality of whispered vowels, and identification experiments of whispered speech. Chapter 3 describes relation between perceived pitch of whispered vowels and their spectral properties. The model to predict pitch of whispered vowel is also shown. Chapter 4 describes a method of convert whisper to normal phonated speech using estimated pitch of whisper. Finally, Chapter 5 and Chapter 6 describe discussions, future works, and conclusions.

Chapter 2

Phonemic Quality and Intelligibility

2.1 Background

It is a well-known fact that two or three low order formant frequencies, denoted as F_1, F_2, and F_3, represent phonemic quality of normally phonated vowels. Whispered vowels have formants as well, and their formant frequencies have been researched so far.

Smith [54] investigated the first to third formants of four whispered and voiced vowels in American English using narrow band spectrum analyzer. He reported that the formants of whispered vowels to be higher in frequency than those of the same vowels produced with voicing. Kallail and Emanuel [24] and Kallail and Emanuel [23] analyzed formants of five vowels spoken by 20 female and 15 male subjects, respectively. They reported the same trend shown by Smith [54], and it was strongly evident for F_1. The formant comparison between whispered and voiced vowels was carried out for many languages, and the same trend was reported, e.g., for Japanese by Konno, Toyama, Shimbo, *et al.* [36], Matsuda, Mori, and Kasuya [40], Ito, Takeda, and Itakura [18], for Swedish by Eklund and Traunmüller [4], for Serbian by Jovičić [20], for Mandarin by Li and

Xu [39], and for British English by Sharifzadeh, McLoughlin, and Russell [53].

On the other hand, whispered speech is produced without vibration of vocal folds, and thus it has no fundamental frequency (F_0). Consequently, every speech including voiced consonant is produced in an unvoiced way. Because of that reason, intelligibility of whispered speech is thought to be degraded rather than normally phonated speech.

There have been reported some experimental results on intelligibility of whispered speech. Kallail and Emanuel [22] presented the result of identification experiments using five whispered and normally phonated vowels, in which normal vowel samples were identified correctly more often than whispered vowel samples. Tartter [59] described the correct identification rate of whispered 10 vowels is 82% that is 10% lower than that of normal vowels. As for consonants, she reported 64% for 18 whispered consonants [57]. In the experiments using whispered speech vocoded from normal speech, consonants are generally a little less reliable in the whispered speech than in voiced speech [17]. Although there are some other reports on intelligibility of whisper and acoustic characteristics of whispered consonants [18], [21], [47], the mechanism causing such degradation has not been made clear.

2.2 Characteristics of Japanese whispered vowels

This section summarizes the acoustical and perceptual characteristics of Japanese whispered vowels described in Konno, Toyama, Shimbo, *et al.* [36].

2.2.1 Acoustic analyses

In order to make clear the difference between Japanese whispered and normally phonated vowels, acoustic analyses were carried out.

Three male speakers uttered sustained five Japanese vowels in normal phonation and in whisper. Each vowel was uttered five times by each speaker, and duration of 0.5 ms was used for analyses. The linear predictive coding (LPC) analysis with the adaptive inverse filtering (AIF) [49] was adopted for precise analyses of formants. Fig. 2.1 shows spectra of normally phonated and whispered /a/ of the same speaker obtained by the fast Fourier transform (FFT), LPC, and LPC with AIF. Flattened spectra by LPC with AIF mean successful extraction of vocal tract characteristics of vowels.

F_1 and F_2 of each speech sample are plotted in Fig. 2.2, and normalized plot using F_3 are shown in Fig. 2.3. In Fig. 2.2, the low frequency formants of whispered vowels tend to shift to higher frequency regions compared with those of normally phonated vowels. In Fig. 2.3, the distributions of whispered vowels are not the same as phonated ones although the normalization decreases the differences.

2.2.2 Vowel identification experiments

In order to investigate the reason of the formant frequency differences between whispered and normally phonated vowels, phoneme identification experiments using synthetic whispered and voiced vowels were conducted.

The stimuli were made by Klatt cascade synthesizer [26] with various formants and two kinds of glottal sources. A voiced source having F_0 of 120 Hz was used for voiced vowels, and an unvoiced source for whispered vowels was made from Gaussian noise by low pass filtering. The spectral

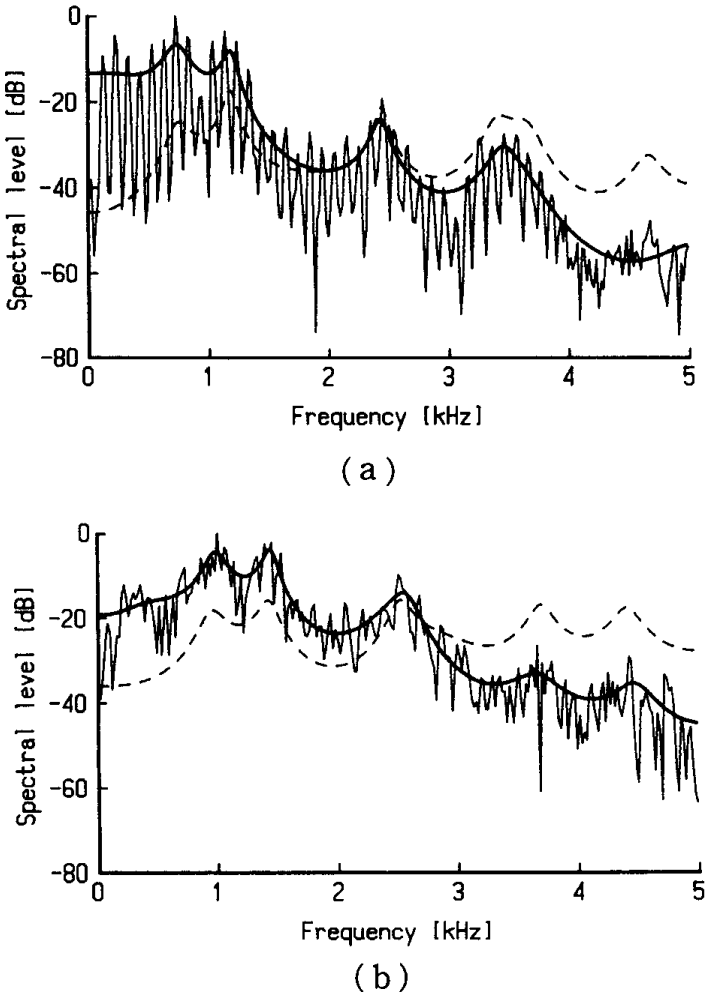


FIGURE 2.1: Spectra of (a) phonated and (b) whispered vowel /a/ uttered by speaker A: 512 points FFT (thin lines), LPC (thick lines) and LPC with the adaptive inverse filtering (dashed lines) (reprinted from [36]).

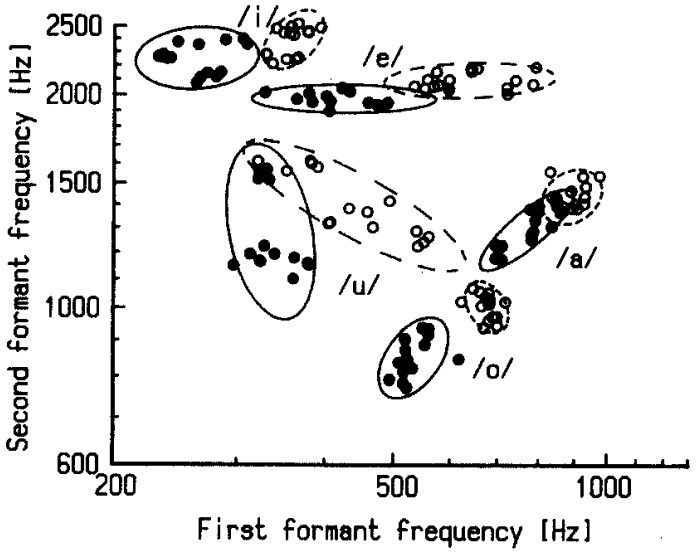


FIGURE 2.2: The distribution of phonated (solid lines) and whispered (dashed lines) vowels in the $\log F_1$-$\log F_2$ plane (reprinted from [36]).

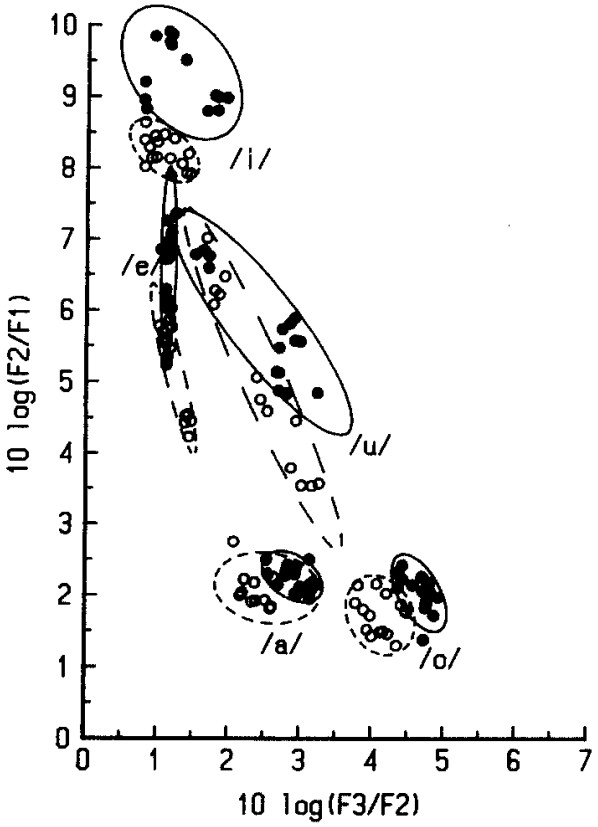


FIGURE 2.3: The distribution of phonated (solid lines) and whispered (dashed lines) vowels in the $\log(F_3/F_2)$-$\log(F_2/F_1)$ plane (reprinted from [36]).

tilts of voiced and unvoiced sources were -12 dB/oct and -6 dB/oct, respectively.

The configuration of formant frequencies are shown in Fig. 2.4. The formant bandwidths of stimuli were approximated by the following equation [56]

$$B_n = 50(1 + \frac{F_n^2}{6 \times 10^6}), n = 1, 2.$$

The number of formants for the synthesis was six, and the third to sixth formant were set to the default values of the synthesizer.

The resultant identification rates of stimuli are in Fig. 2.4, and shifts of phoneme boundaries calculated from the identification rates are shown in Fig. 2.5. These figures show that phonemic quality of vowels changes by the difference of glottal sources even if the formant frequencies are unchanged. In other words, phonemic quality of synthetic vowels depends not only on formants but also on their glottal sources. Moreover, similarity between formant shifts of real speech in Fig. 2.2 and phoneme boundary shifts of synthetic vowels in Fig. 2.5 suggests that the formant shift of whispered vowels against normally phonated ones are concerned with maintaining the phonemic quality of whispered vowels.

Hirahara and Kato [13] reported the formant shifts accompanied with increasing F_0s of voiced vowels. The shifts are similar to the shifts of phoneme boundary in Fig. 2.5. Since F_0 is the primary cue for the pitch perception, pitch of whispered and voiced vowels might be one of the cues for phonemic quality of vowels. On the other hand, Fujisaki and Kawashima [6] suggested that the spectral tilts of noise-excited vowels have influence on their phonemic quality. Therefore, the phonemic quality of whispered and voiced vowels may be related to the difference of their spectral tilts.

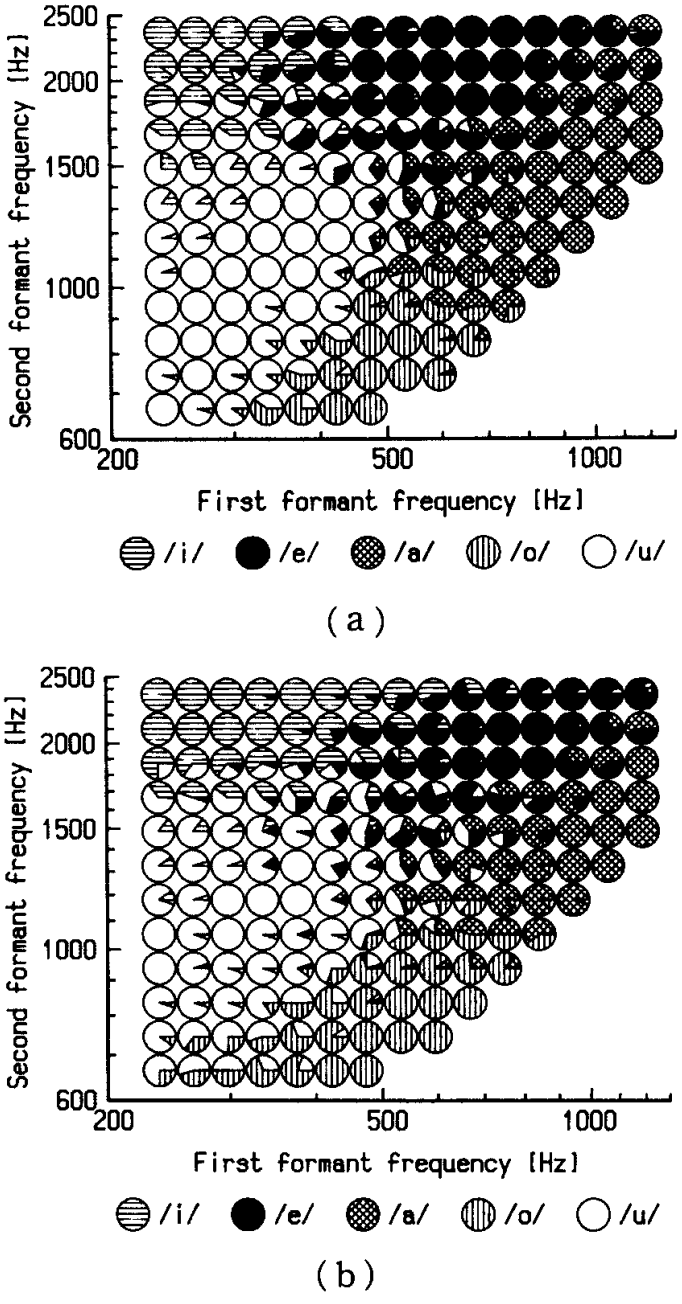


FIGURE 2.4: Configuration of stimuli in the $\log F_1$-$\log F_2$ plane and their phoneme identification rates represented by circular graphs: (a) phonated (voice source excited) vowels; (b) whispered (voiceless source excited) vowels (reprinted from [36]).

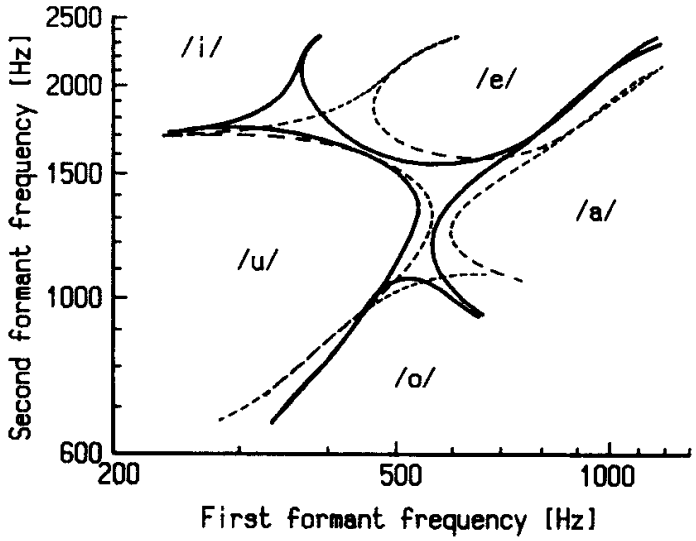


FIGURE 2.5: Perceptual vowel boundaries of phonated (voice source excited) and whispered (voiceless source excited) stimuli in the $\log F_1$-$\log F_2$ plane, shown by solid and dashed lines, respectively (reprinted from [36]).

2.3 Intelligibility of Japanese whispered syllables

This section describes the experiments to compare the intelligibility of Japanese whispered and normally phonated syllables, and to investigate perceptual cues for phonemic quality of whispered speech.

2.3.1 Experimental condition

An auditory listening experiment was conducted. A female speaker uttered 110 monosyllables in two modes of normal and whispered. The form of monosyllables are V (vowel: /i/, /e/, /a/, /o/ or /u/), CV (consonant-vowel) or a uvular nasal [N]. They covers all Japanese morae which can be uttered in isolation. The speech was recorded in a soundproof booth using a condenser microphone (Sony C-48) and a digital recorder (Fostex FR-2) with 48 kHz sampling rate and 16 bit quantization. One more

speech was vocoded from the corresponding normal speech by replacing the original vocal source with a Gaussian noise using a source-filter vocoder, TANDEM-STRAIGHT [25]. The sampling rate of the vocoded speech is also 48 kHz.

Ten listeners (five male and five female students) were presented those three modes of monosyllables by a D/A converter (Luxman DA-100) with a 48 kHz sampling rate through headphones (Sennheiser HDA 200) at around 65 dB SPL. A subject (listener) was asked to identify and to input the syllables by Japanese kana characters from a computer keyboard after presentation of stimuli (syllables). Mistyping was correctable until the enter key was pushed.

Three modes are completely separated and each set of randomly ordered 110 syllables with twice presentation is examined at a round.

2.3.2 Results

Table 2.1 shows the classification of 110 syllables used to aggregate the results. For simplicity, we use only one of four places of articulation such as labial, alveolar, velar, and glottis. According to Koizumi [27], /z/-sound was categorized as a voiced affricate, not a fricative.

First, the 110 syllables were classified in 9 groups by the manner of articulation and voiced/unvoiced. Three modes of normal, vocoded, whispered are compared in the correct identification rate in Fig. 2.6, where "vowel" represents syllables consist of V (vowel only), and the other groups mean CV (consonant-vowel) syllables except that [N] is included in "nasal." Typical errors occured in voiced/unvoiced plosive and affricate groups. The vowel parts in the syllables were perfectly identified even in whispered and vocoded speech, except that /u/ and /nu/ was misidentified

TABLE 2.1: Classification of syllables by manner of articulation, place of articulation, and voiced/unvoiced.

Manner of articulation	Place of articulation	Voiced/ Unvoiced	Syllable
Vowel	—	Voiced	i e a o u
Semivowel	Labial	Voiced	wa
	Velar	Voiced	ya yo yu
Nasal	Labial	Voiced	mi me ma mo mu mya myo myu
	Alveolar	Voiced	ne na no nu
	Velar	Voiced	ni nya nyo nyu N
Liquid	Alveolar	Voiced	ri re ra ro ru rya ryo ryu
Plosive	Labial	Voiced	bi be ba bo bu bya byo byu
	Labial	Unvoiced	pi pe pa po pu pya pyo pyu
	Alveolar	Voiced	de di da do
	Alveolar	Unvoiced	te ti ta to
	Velar	Voiced	gi ge ga go gu gya gyo gyu
	Velar	Unvoiced	ki ke ka ko ku kya kyo kyu
Affricate	Alveolar	Voiced	ze za zo
	Alveolar	Unvoiced	zu tsu ji je ja jo ju chi che cha cho chu
Fricative	Labial	Unvoiced	fu fi fe fa fo
	Alveolar	Unvoiced	se sa so su shi she sha sho shu
	Velar	Unvoiced	hi hya hyo hyu
	Glottis	Unvoiced	he ha ho

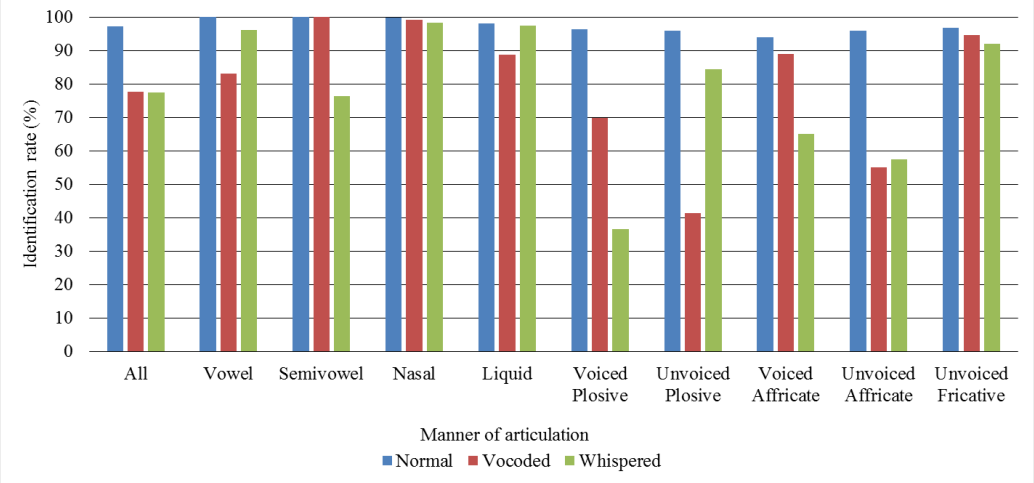


FIGURE 2.6: Intelligibility of syllables grouped by manner of articulation [30].

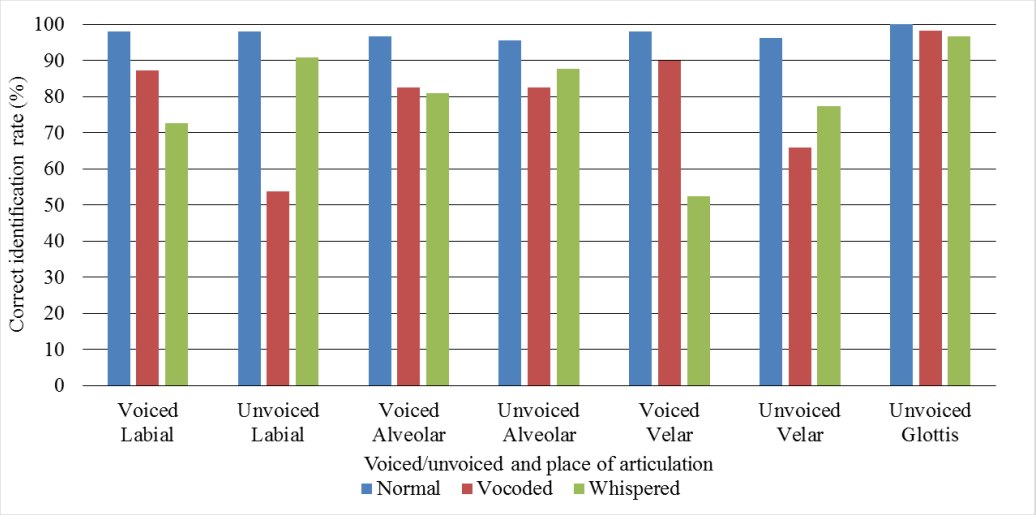


FIGURE 2.7: Intelligibility of syllables excluding isolated vowels grouped by place of articulation.

once as [N] respectively. Thus, misidentification is due to mislabeling, insertion or deletion of consonants.

Next, the syllables excluding isolated vowels were classified in 7 groups by the place of articulation and voiced/unvoiced. The correct identification rate are shown in Fig. 2.7. Comparing results of whispered syllables within each place of articulation, the correct identification rate of the voiced consonants are less than that of the unvoiced consonants.

Such errors were examined more in detail by confusion matrices shown in Tables 2.2, 2.3, and 2.4. Those tables show that which group of syllables (in row) is identified into which group (in column) in normal speech (Table 2.2), in vocoded speech (Table 2.3) and in whispered speech (Table 2.4). A typical difference of misidentification between whispered speech and vocoded speech is seen in plosive and affricate consonants. In vocoded speech (Table 2.3), unvoiced plosives and affricates are misidentified into voiced those, e.g., /p/ (unvoiced labial plosive) to /b/ (voiced labial plosive), /k/ (unvoiced velar plosive) to /g/ (voiced velar plosive), [tʃ] (unvoiced alveolar affricate) to [dʒ] (voiced alveolar affricate). While, in whispered speech, voiced plosives are misidentified into unvoiced those, e.g.,

/b/ (voiced labial plosive) to /p/ (unvoiced labial plosive), /g/ (voiced velar plosive) to /k/ (unvoiced velar plosive).

TABLE 2.2: Confusion matrix of normally phonated syllables. (VO: vowel, SV: semivowel, NA: nasal, LI: liquid, PL: plosive, AF: affricate, FR: fricative, LA: labial, AL: alveolar, VE: velar, GL: glottis, V: voiced, U: unvoiced).

			Responses																			
			VO	SV		NA			LI	PL						AF		FR			
			—	LA	VE	LA	AL	VE	AL	LA	LA	AL	AL	VE	VE	AL	AL	LA	AL	VE	GL
			V	V	V	V	V	V	V	V	U	V	U	V	U	V	U	U	U	U	U
Stimuli	VO	—	V	100																		
	SV	LA	V		20																	
		VE	V			60																
	NA	LA	V				180															
		AL	V					80														
		VE	V						80													
	LI	AL	V							157	3											
	PL	LA	V							7	153											
		LA	U							2		158										
		AL	V							3			77									
		AL	U									4		76								
		VE	V												155	5						
		VE	U												9	151						
	AF	AL	V														169	11				
		AL	U														5	115				
	FR	LA	U																97			3
		AL	U																	179	1	
		VE	U																		80	
		GL	U																			60

TABLE 2.3: Confusion matrix of vocoded syllables (VO: vowel, SV: semivowel, NA: nasal, LI: liquid, PL: plosive, AF: affricate, FR: fricative, LA: labial, AL: alveolar, VE: velar, GL: glottis, V: voiced, U: unvoiced).

			Responses																			
			VO	SV		NA			LI	PL						AF		FR			
			—	LA	VE	LA	AL	VE	AL	LA	LA	AL	AL	VE	VE	AL	AL	LA	AL	VE	GL
			V	V	V	V	V	V	V	V	U	V	U	V	U	V	U	U	U	U	U
Stimuli	VO	—	V	83								5							4		1	7
	SV	LA	V		20																	
		VE	V			60																
	NA	LA	V				159		1													
		AL	V					79	1													
		VE	V				1		99													
	LI	AL	V	4			1	1		143	8	1									2	
	PL	LA	V	3						6	118	22						7		3	1
		LA	U	7		1				28	**63**	**47**	1	3	2			5		2	1
		AL	V							13		2	32	33							
		AL	U							5	6	10	**21**	**34**	4						
		VE	V	1		7				4	1				129	7		1			1
		VE	U	5		2					2	3		1	**54**	**84**				2	7
	AF	AL	V														171	9				
		AL	U														**54**	**66**				
	FR	LA	U									2							93			5
		AL	U																	180		
		VE	U			6															74	
		GL	U																1			59

TABLE 2.4: Confusion matrix of whispered syllables (VO: vowel, SV: semivowel, NA: nasal, LI: liquid, PL: plosive, AF: affricate, FR: fricative, LA: labial, AL: alveolar, VE: velar, GL: glottis, V: voiced, U: unvoiced).

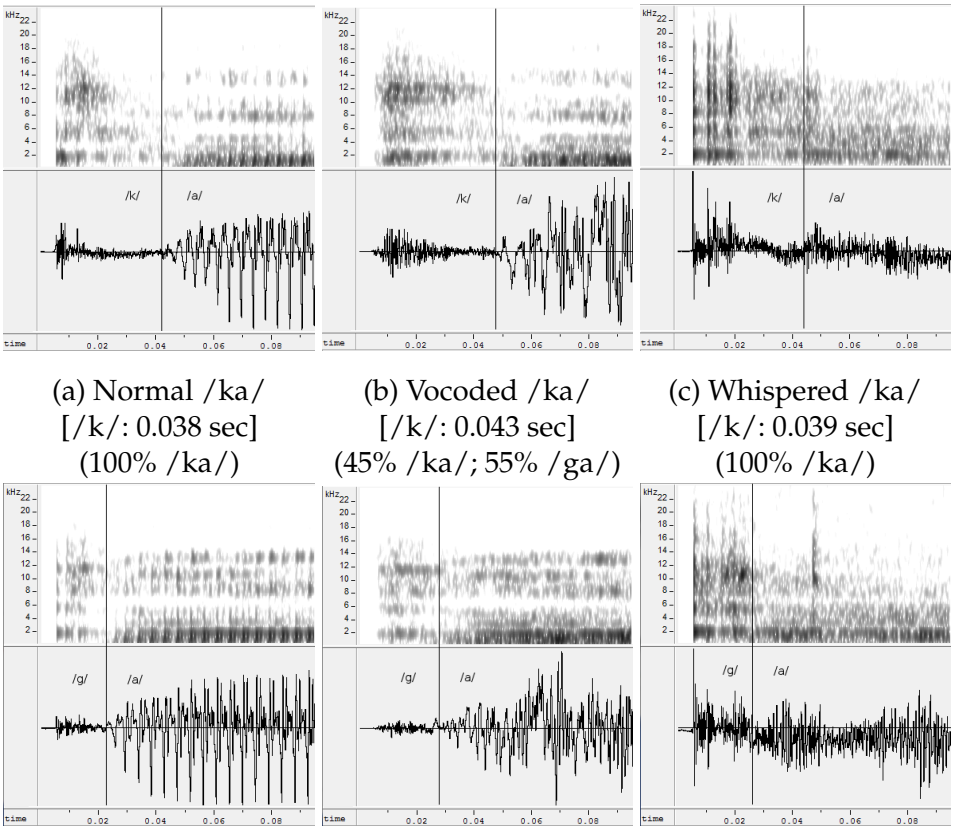
			Responses																			
			VO	SV		NA			LI	PL						AF		FR			
			—	LA	VE	LA	AL	VE	AL	LA	LA	AL	AL	VE	VE	AL	AL	LA	AL	VE	GL
			V	V	V	V	V	V	V	V	U	V	U	V	U	V	U	U	U	U	U
Stimuli	VO	—	V	96			2		1		1											
	SV	LA	V		20																	
		VE	V			41			1	10											8	
	NA	LA	V				155	4	1													
		AL	V					80														
		VE	V			1			99													
	LI	AL	V	1						156				3								
	PL	LA	V							6	**72**	**81**		1								
		LA	U							1	20	139										
		AL	V							12	4	9	**14**	**41**								
		AL	U							15		1	5	59								
		VE	V							10		1		1	**60**	**87**	1					
		VE	U							4				1	16	139						
	AF	AL	V													1	118	59		2		
		AL	U														51	69				
	FR	LA	U				1												97			2
		AL	U																	180		
		VE	U			21															59	
		GL	U	2																		58

2.3.3 Discussion

As described in the previous subsection, the contrastive error direction was found on identification of voiced or unvoiced consonant between whispered and vocoded syllables. In order to find the reason, we have examined spectrograms and waveforms of a voiced plosive /g/ in syllable /ga/ and an unvoiced plosive /k/ in /ka/. These are shown in Fig. 2.8 with identification rates. In these figures, voice-onset-time (VOT) of normal speech or plosive duration of vocoded/whispered is also shown. Note that the frequency range is 0 to 24 kHz and the unit of time is second. In Fig. 2.8, the following things are observed in plosive parts: 1) the high-frequency part is more enhanced in whispered speech compared with vocoded and normal speech, 2) the beginning of /k/ in whispered speech has many spliced bursts in the waveform, and consequently, in the spectrogram. These observation derives the following hypothesis. If the voice source is a noise and the high-frequency is enhanced/weakened, then the consonant sounds as unvoiced/voiced. Indeed, this explains why /k/ in vocoded speech is wrongly listened as /g/ (Fig. 2.8 (b)), /g/ in whispered speech as /k/ (Fig. 2.8 (f)), and those are correctly identified (Fig. 2.8 (c), (e)). In Fig. 2.9, we show the spectrograms and waveforms of whispered /ku/ and /gu/, since /gu/ has the best identification rate (80%) in the voiced plosives. The hypothesis described above also explains this figure of whisper. If our interpretation is correct, we could improve the intelligibility of whispered speech by controlling the balance of energy over frequency.

2.4 Summary

In Section 2.2, the formant frequency and phonemic quality of isolated whispered vowels were compared with those of phonated vowels by means



(a) Normal /ka/ [/k/: 0.038 sec] (100% /ka/)

(b) Vocoded /ka/ [/k/: 0.043 sec] (45% /ka/; 55% /ga/)

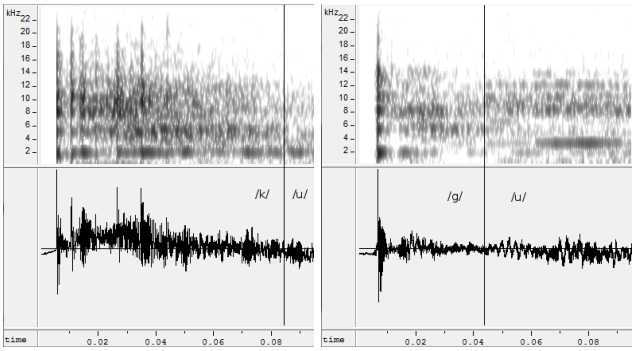
(c) Whispered /ka/ [/k/: 0.039 sec] (100% /ka/)

(d) Normal /ga/ [/g/: 0.018 sec] (100% /ga/)

(e) Vocoded /ga/ [/g/: 0.023 sec] (95% /ga/; 5% /ka/)

(f) Whispered /ga/ [/g/: 0.020 sec] (40% /ga/; 60% /ka/)

FIGURE 2.8: Spectrograms and waveforms of three modes of /ka/ and /ga/. Voice-onset-time for normal speech or duration from burst of plosive to beginning of vowel for vocoded/whispered speech is within brackets. Identification rates are shown in parentheses.



(a) Whispered /ku/ [/k/: 0.080 sec] (85% /ku/; 15% /gu/)

(b) Whispered /gu/ [/g/: 0.038 sec] (80% /gu/; 20% /ku/)

FIGURE 2.9: Spectrograms and waveforms of whispered /ku/ and /gu/. Duration from burst of plosive to beginning of vowel for whispered speech is within brackets. Identification rates are shown in parentheses.

of speech analysis and phoneme identification tests using synthetic vowels. The low frequency formants of whispered vowels tend to shift to higher frequency regions compared with those of phonated vowels. The perceptual vowel boundaries of whispered and phonated vowels in the $\log F_1$–$\log F_2$ plane were also different. The relation between the shift of formant frequencies and the phonemic quality of whispered vowels is investigated, and the possibility of perceived pitch being one of the cues for vowel perception was discussed.

In Section 2.3, the author conducted a quantitative evaluation on degradation of the intelligibility of Japanese whispered speech, and obtained 80% of intelligibility of normally uttered ones for 110 whispered monosyllables. The same degree of degradation was also observed in vocoded speech that was excited by noise. The difference between whispered and vocoded speech was seen in individual groups of the manner of articulation: voiced plosives/affricates were often misidentified into unvoiced those in whispered speech, while the reverse misidentification, from unvoiced to voiced, often occurs in vocoded speech. This difference was

explained by the difference of the spectral configuration, that is, a larger balance is put on high-frequency to enhance unvoiced nature and a large balance is put on low-frequency to enhance voiced nature. In other words, the balance of energy over frequency may be one of the voicing cues for whispered consonants. This study also shows a possibility to increase the intelligibility of whispered speech by reducing the high-frequency energy of the voiced consonant parts.

Chapter 3

Pitch of Whispered Vowels

In this chapter, we investigate the spectral properties of five whispered Japanese vowels uttered in isolation and having different pitch. Pitch of whispered vowels was measured in terms of the manner in which the speakers were listening to pure tones while uttering and adjusting its frequency so that the pitch matched the utterance, or speakers changed the pitch of utterances to match the given pure tones [32]. Acoustic analyses were carried out on formant frequencies, a spectral tilt, and a peak frequency of wide-ranging spectral shape using second-order LPC method named a global peak. Moreover, we derived a multiple regression function to convert the outputs of a mel-scaled filter bank of whispered speech into the perceived pitch value [29], [34].

3.1 Background

3.1.1 Pitch

Pitch[1], as one of the attributes of sensation, conveys speech prosody such as intonation and accent of pitch-accent languages such as Japanese. The pitch of normally phonated speech corresponds to its fundamental frequency (F_0), which is the number of vocal-fold vibrations per second.

[1]ANSI [1] defines pitch as "that attribute of auditory sensation in terms of which sounds may be ordered on a scale extending from low to high."

However, whispered speech has no F_0 because the sound source of whispering is turbulent airflow [38] and the vocal folds do not vibrate. Nevertheless, whispered speech can communicate the prosody [8], [9], [28], [58] and convey the sensation of pitch. Moore [46] stated that "assigning a pitch value to a sound is generally understood to mean specifying the frequency of a pure tone that has the same subjective pitch as the sound."

3.1.2 Acoustic analysis and perceptual experiments on pitch of whispered speech

Whispered vowels having different pitch have been analyzed so far. According to Meyer-Eppler [45], in five whispered German vowels, as the pitch level increases, a specific range of the spectrum rises upward in one group of vowels, and only the power in a high-frequency region increases in another group. In five whispered Japanese vowels, Hirahara [12] showed that the degree of increase in formant frequencies varies over vowels when the pitch increases. A significant change of the first formant frequency (F_1) and the second formant frequency (F_2) on Japanese vowel /a/ was reported by Higashikawa, Nakai, Sakakura, *et al.* [10], and, similarly, a large change of F_2 was reported for Mandarin by Chen and Zhao [2].

A speech perceptional approach has also been taken to reveal the relationship between the pitch of whispered vowels and formant frequencies. By the paired comparison, Harbold [7] demonstrated the relative pitch rankings of twelve voiced vowels spoken in monotone and twelve whispered vowels, and suggested that tonality may not be the sole determinant of listeners' pitch judgments. The pitch ranking order of whispered vowels was almost identical to that of voiced vowels. Thomas [60] reported

that perceived pitch of whispered vowels were very closely to F_2, and McGlone and Manning [42] stated the perception of pitch in both whispered and voiced vowels was strongly influenced by acoustic information in the area of F_2 and/or above. Konno, Toyama, Shimbo, *et al.* [37] showed that both formant frequencies in a low-frequency region and spectral tilt affect the perception of pitch in whispered vowels. Higashikawa and Minifie [11] showed that the pitch of the whispered vowel /a/ can be increased by increasing F_1 and F_2. Although these studies suggested that formant frequencies are concerned with perceived pitch of whisper in common, the quantitative relation between pitch and formants is unknown.

3.2 Spectral properties of whispered vowels stepwise increasing pitch

3.2.1 Recordings and speech samples

In this experiment, a male and a female subjects aged around 22 years uttered five Japanese whispered vowels of /i/, /e/, /a/, /o/ and /u/ in different levels of pitch. Their speech was recorded by a condenser microphone (SONY C-48) in a soundproof booth. The other pieces of equipment are a microphone amplifier (AVARON DESIGN M5), a digital audio tape deck (SONY DTC-2000ES) and a digital memory recorder (FOSTEX FR-2). Sampling rate is 48 kHz and and the quantization is 16 bits.

At the beginning, the speaker was asked to utter the lowest whispered vowels as he/she could and to search the frequency by tuning a dial of a pure tone generator and by listening to the generated tone through a headphone. Then the frequency of the guidance tone is raised in a half-tone step, e.g., C, C#, D, ..., and they uttered so as to match the guidance tone. In this way, they spoke 5 to 9 whispered vowels with different level

of pitch in the range of 90 Hz to 160 Hz by the male speaker and of 400 Hz to 600 Hz by the female speaker.

As a post-processing, we cut out a part of 500 ms from the recorded signals such that the resultant part includes a target vowel only and then applied a high-pass filter with cut-off frequency of 50 Hz. The filter was applied in order to suppress a noise caused by a strong breath. The energy of cut signals was normalized to a constant before analysis. To preserve the original characteristics of the spectrum, no pre-emphasis was applied.

3.2.2 Pitch comparison test

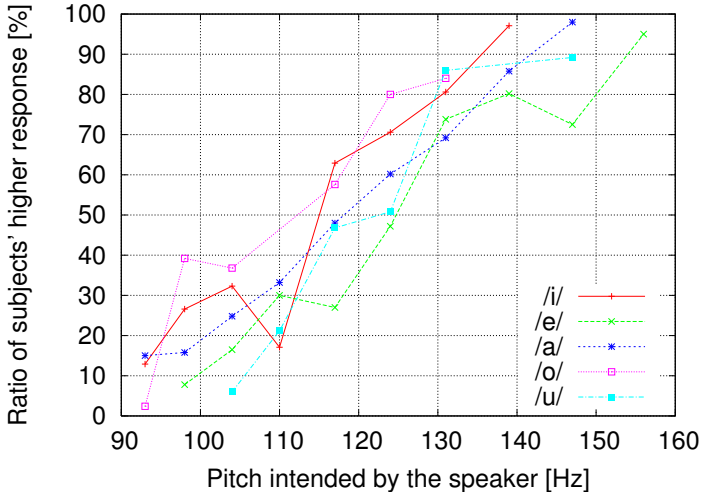
Before the acoustic analyses, we made sure the pitch of speech samples by paired comparison tests. The stimulus pairs were made from all combinations of speech samples in each vowel of each speaker, and presented to the subjects through headphones (SENNHEISER HDA200) in random order. The subjects judged the second sample was lower or higher than the first one. The number of subjects were five for the male samples, and three for the female samples.

The results are shown in Fig. 3.1 [31], [33]. In this figure, for each speech sample, the ratio of "higher" response against all judgments are shown. As the pitch intended by the speaker becomes higher, the sample tended to be judged as "higher". Therefore, the pitch values of the speech samples were appropriate.

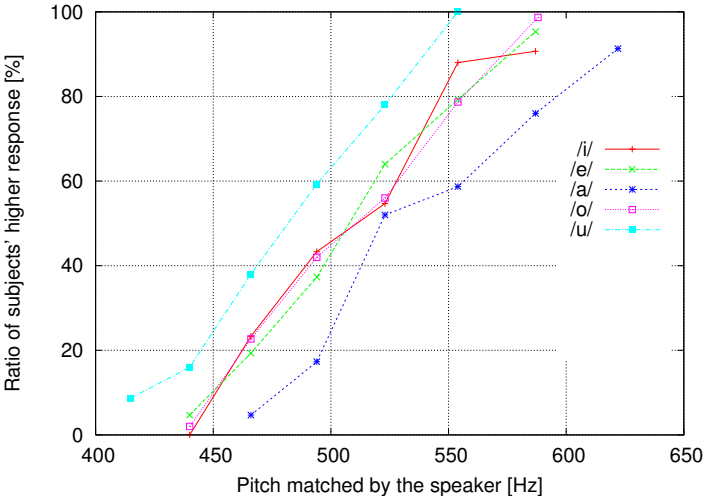
3.2.3 Acoustic analysis

Difference in spectrum

The speech samples were transformed to spectra in the frequency domain by FFT at 48 kHz sampling rate with shifting windows of length 42.7 ms (2048 points). The window is the Hamming window and the shift



(a) the male speaker



(b) the female speaker

FIGURE 3.1: Results of pitch comparison test.

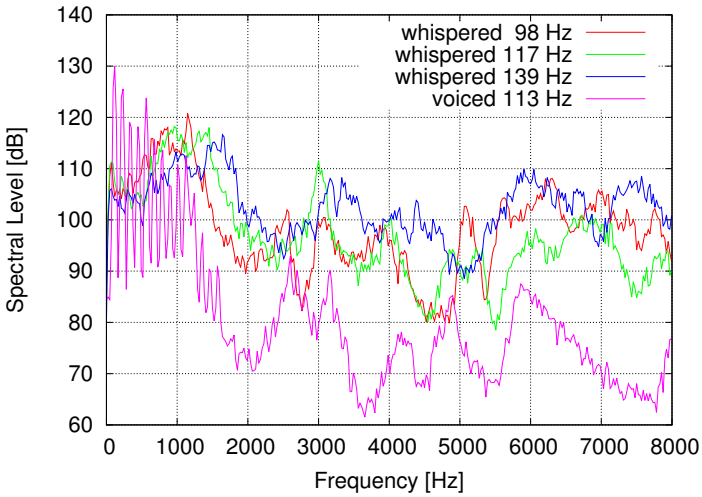


FIGURE 3.2: FFT spectra of voiced and whispered vowel /a/ uttered by the male speaker [29].

length is a half of the window length. The spectra collected by moving windows over a sample were averaged with respect to the same speaker, same vowel, and same pitch. An example of the averaged spectrum of whispered /a/ uttered by a male speaker is shown in Fig. 3.2 with that of normally phonated (voiced) /a/ by the same speaker as a reference. We see a clear difference between whispered and voiced vowels. There exists a clear harmonic structure less than 1 kHz in the voiced vowel but no such structure is seen in the whispered vowel. On the contrary, the energy of high-frequency part of whispered vowel is larger than that of voiced vowel. In addition, we can observe some characteristics in the whispered vowel according to the increase of level of pitch: 1) formants (spectral peaks) with frequencies less than 2 kHz shifted upward and 2) the tilt of the spectrum becomes more flat. The shift of formants is the observations already known in previous studies [12], [45]. We investigate these observations quantitatively in the following sub-sections.

Formant frequencies

To make clear the amount of shift of formants, we applied the linear predictive coding (LPC) analysis to the speech samples. The samples were down-sampled to 12 kHz and LPC analysis with the order of 14 through 17 was applied. This time, the frame length is 256 points and the frame-shift length is 128 points. The five formants were calculated from the roots of LPC equation, taking into consideration the pole bandwidth and the continuity between them. The results [31], [33] are shown in Fig. 3.3 and 3.4. The first two to three formants less than 5 kHz tend to move upward as the pitch intended by the speaker increases.

Spectral tilt and global peak frequency

Two more quantitative spectral properties, i.e., spectral tilt and global peak frequency, were used to present the spectral shape.

In order to obtain the spectral tilt, speech spectrum was approximated by the frequency characteristics of the second order auto-regression model with the constraint that two poles are at 0 Hz, that is, are on the real axis of the z-plane.

The model equation is

$$\frac{1}{A(z)} = \frac{1}{1 + az^{-1} + \frac{1}{4}a^2z^{-2}}$$

where

$$-2 < a < 0, a \in R.$$

The spectral tilt is represented by differential of the approximated spectrum at 1 kHz as

$$\frac{d}{df}\left|\frac{1}{A(e^{j2\pi f})}\right|\Bigg|_{f=1000}$$

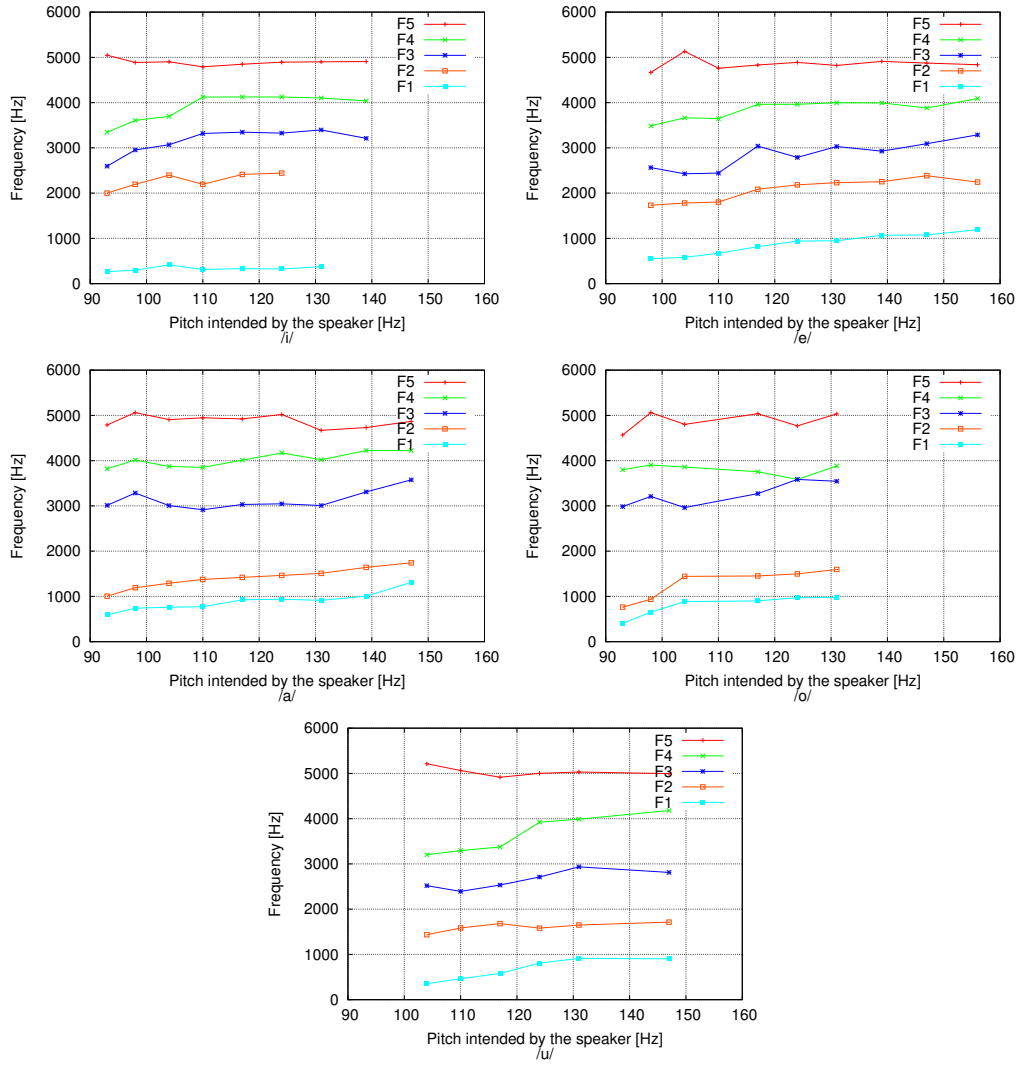


FIGURE 3.3: The first to fifth formant frequencies (F_1–F_5) of the five Japanese vowels, i.e., /i, e, a, o, u/, of the male speaker. Formant frequencies are not plotted if the formants disappear.

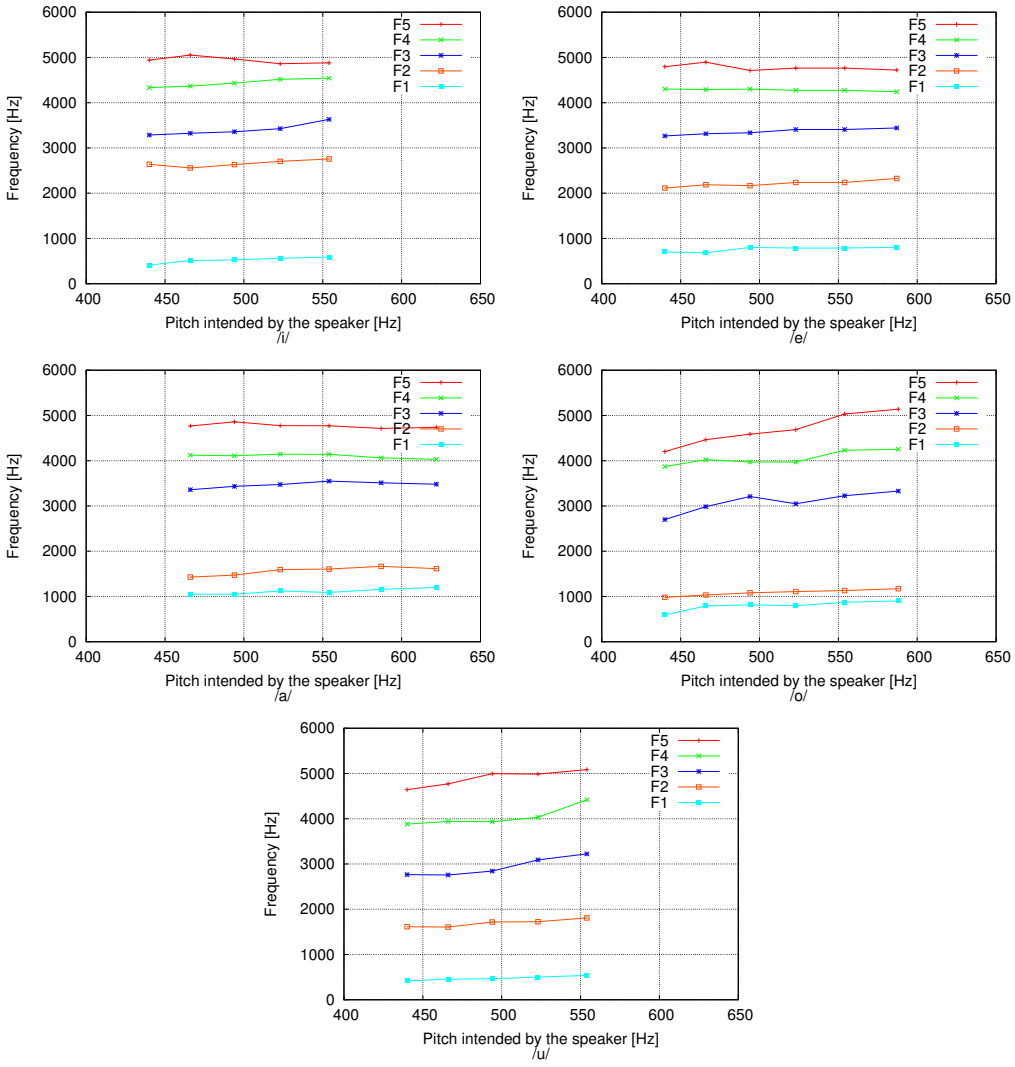


FIGURE 3.4: The first to fifth formant frequencies (F_1–F_5) of the five Japanese vowels, i.e., /i, e, a, o, u/, of the female speaker.

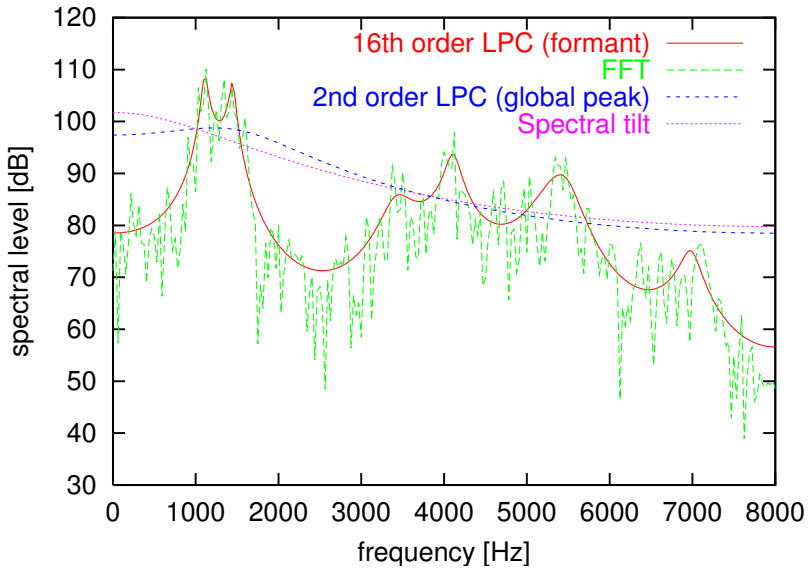


FIGURE 3.5: An example of spectral envelopes of whispered /a/ related to the analysis of spectral tilt and global peak frequency.

where f is a variable of frequency. a is determined from the speech signal in each speech frame so that it minimizes

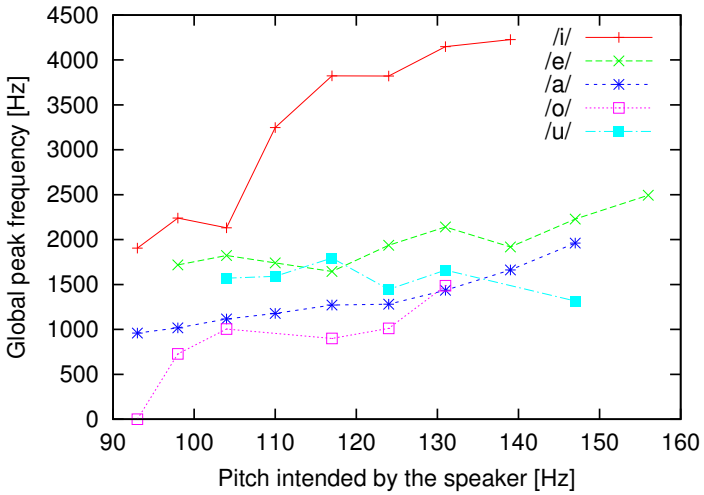
$$\sum_{i=0}^{N-1} e(i)^2 = \sum_{i=0}^{N-1} (x(i) + ax(i-1) + \frac{1}{4}a^2 x(i-2))^2$$

where $x(i)$ is the observed speech signal, and N is the number of samples per frame. On the other hand, the global peak frequency was calculated as the pole frequency using the ordinary 2nd order LPC analysis [35].

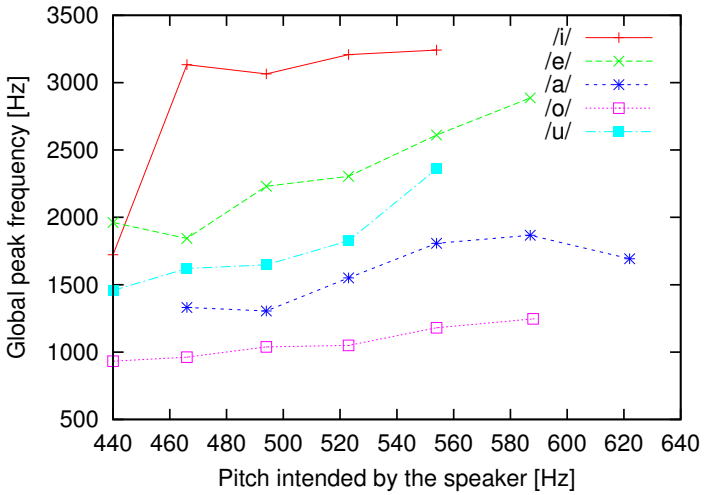
Figure 3.5 shows the examples of various spectra of whispered /a/ related to the spectral tilt and global peak frequency.

The conditions of analysis such as sampling frequency, window, frame length, and frame period are as same as those of formant analysis shown in the subsection 3.2.3.

Results are shown in Figs.3.6 and 3.7. Figure 3.6 shows the global peak frequencies of vowels whispered by the male and female speaker. For both speakers, increment of the pitch intended by the speaker raises the global peak frequencies. The values of spectral tilt shown in Figure 3.7 also tend

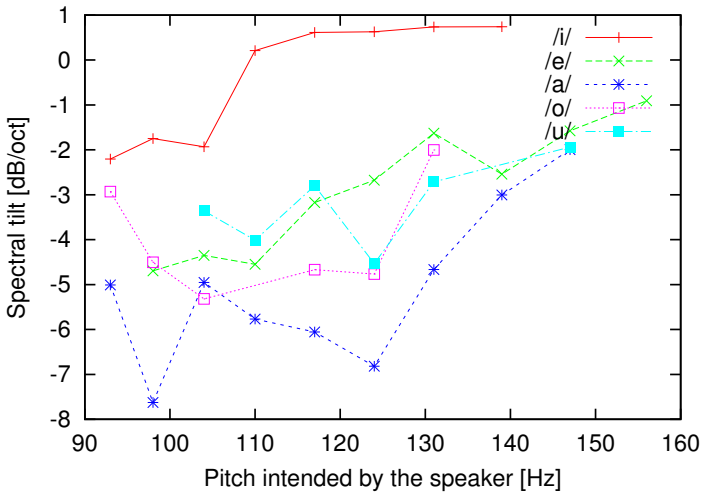


(a) the male speaker

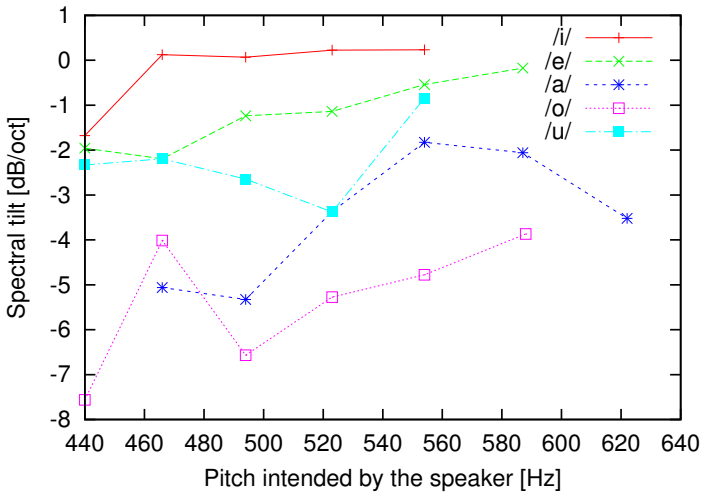


(b) the female speaker

FIGURE 3.6: Global peak frequencies of the whispered vowels.



(a) the male speaker



(b) the female speaker

FIGURE 3.7: Spectral tilt of the whispered vowels.

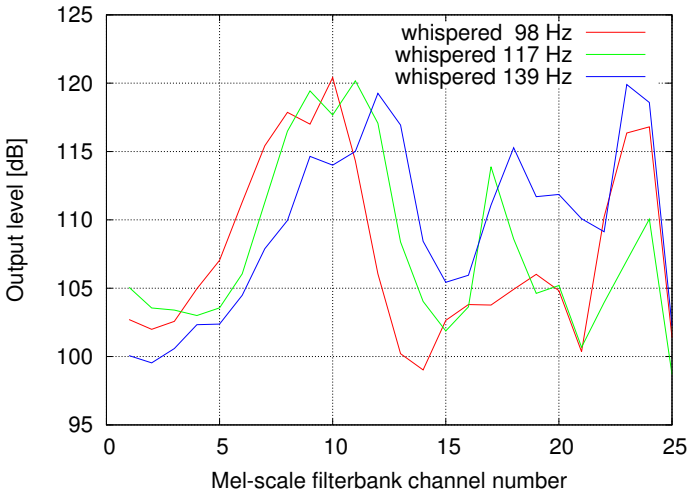


FIGURE 3.8: Output of filter bank of /a/ uttered by a male speaker [29].

to increase, that is, flatten the spectrum as the pitch rises. However the tendency is less obvious than the global peak.

Mel scale filter bank output

To analyze the difference in spectra in a global manner, we examined the outputs of a filter bank equally arranged along the mel scale [29]. The samples were down-sampled to 16 kHz. The number of channels of the filter bank was 25. The software package of HTK was used for the analysis. We took an average for the output of each filter bank channel. An example of the results is shown in Fig. 3.8. From this figure, we see the increase of intensity of region #15–#20 (2.3 kHz–4.2 kHz) as the pitch level increases. To evaluate this tendency analytically we applied the principal component analysis (PCA) to the averaged outputs of the mel-scale filter bank. PCA was carried out on all the outputs of all vowels and all values of pitch in each speaker.

A few principal components were sufficient in the sense of squared error to explain the data expressed in vectors of the outputs of the filter bank channels. In /a/ of the male speaker, for example, the first five of

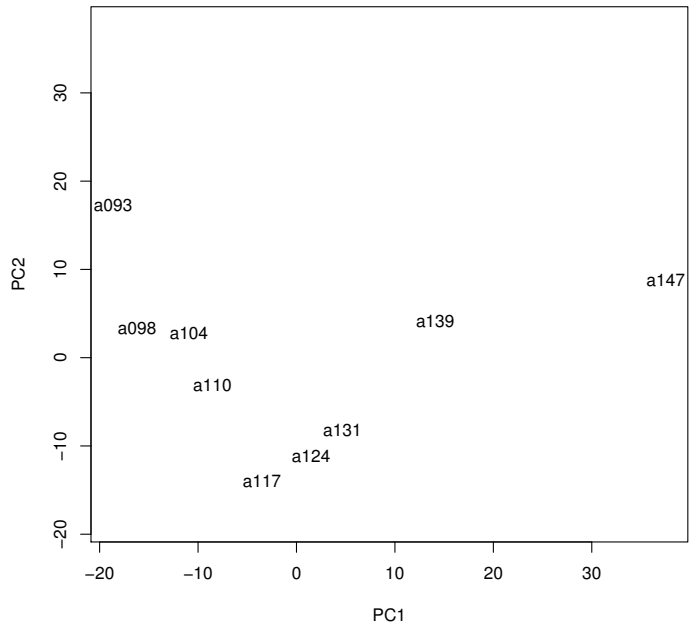


FIGURE 3.9: Configuration of /a/ uttered by a male speaker in the plane of the first two principal components. A label along each point shows the vowel and its pitch frequency [29].

them contributed for reconstruction by 64%, 85%, 91%, 95%, 97%, respectively, when they are combined in turn. As an example, all /a/'s uttered by the male speaker are plotted in the plane spanned by the first two principal components in Fig. 3.9, where nine samples of /a/ with different nine levels of pitch are displayed, where the number attached to a point indicates the intended pitch frequency. It can be observed that the first component corresponds to the height of pitch faithfully. Indeed, an almost linear relationship can be seen between the pitch frequency and the value of first principal component in Fig. 3.10. The corresponding first eigenvector is shown in Fig. 3.11. This eigenvector shows that the components lower than #11 (1.3 kHz) work negatively to the increase of pitch frequency, while those components between #11 and #22 work positively. This eigenvector suggests that if we want to increase the pitch level, decrease the energy in low frequencies under 1 kHz and increase the formant

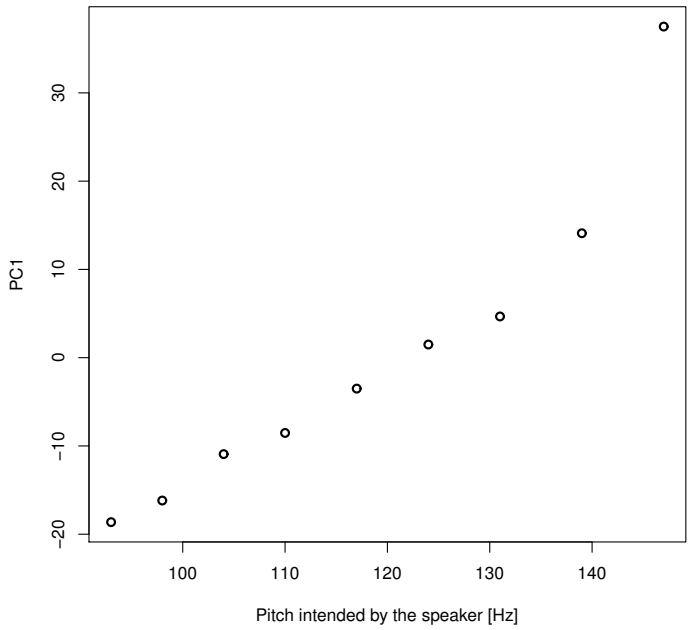


FIGURE 3.10: The first principal component values of each sample of /a/ uttered by a male speaker [29].

frequencies between 1 kHz to 5 kHz upward. Indeed, through the eigenvectors, the decrease of energy below 550 Hz is observed in common to all whispered vowels when the speaker intended to increase the level of pitch.

To examine how well the pitch level can be predicted from a small number of principal components, we carried out a multivariate linear regression of the pitch frequency by the three principal components. An example of the results is shown in Fig. 3.12. The average p-value over all vowels was 4.9%. This means by three major factors we can explain the reasons why the pitch was increased or decreased in whispered speech. The most influential factor is a combination of energy in a low frequency part and formant shift in a middle frequency part as described in the case of the first eigenvector.

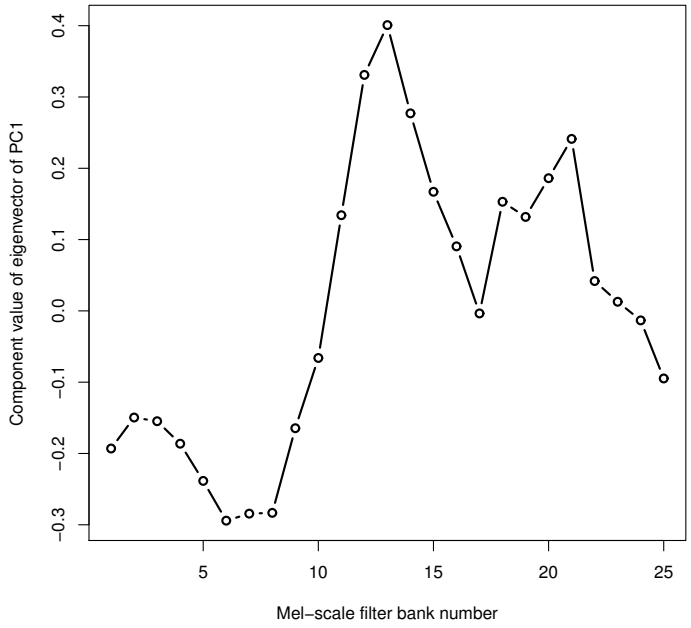


FIGURE 3.11: Eigenvector of the first principal component for /a/ uttered by a male speaker [29].

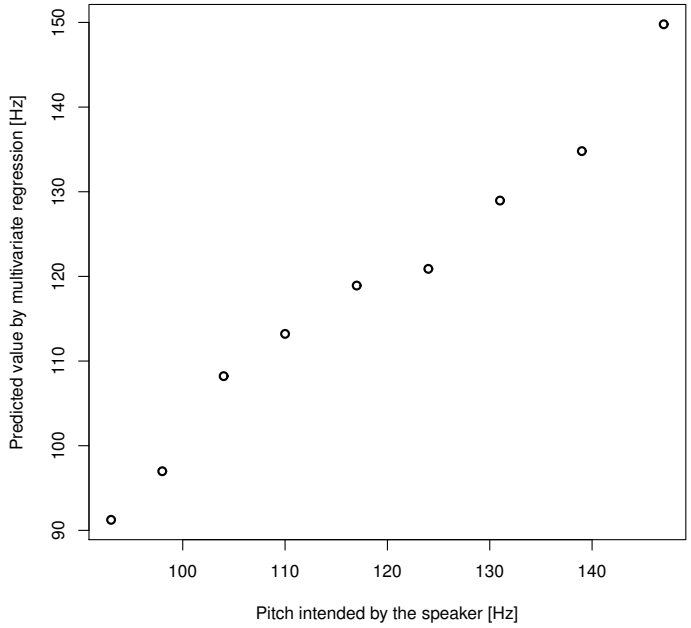


FIGURE 3.12: Results of a multivariate linear regression [29].

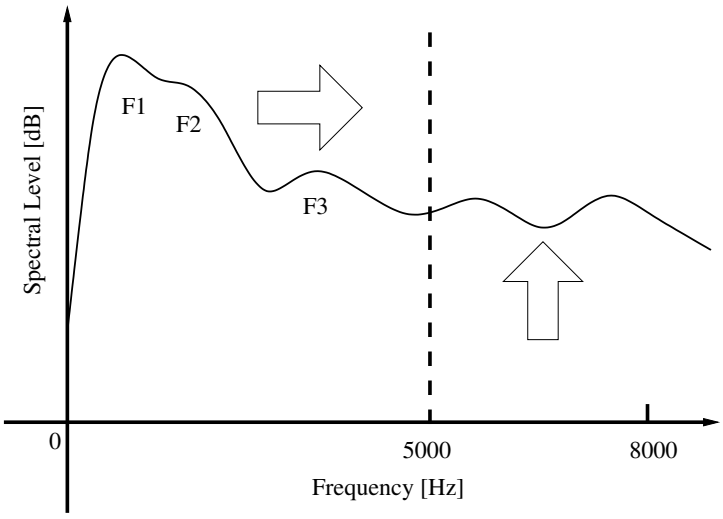


FIGURE 3.13: Illustrative explanation of the change of spectral shape by the increment of whispered pitch [29].

3.2.4 Discussion

There are two cues known to explain the perception of pitch. They are the place cue and the temporal cue [50]. The place cue corresponds to the characteristic frequency of auditory nerves and is applicable for all sounds in the audible frequency range. The temporal cue is obtained by temporal intervals of impulses in neural activities and is said to be effective for the frequency range less than 5 kHz where the phase-locking can occur. Our results might be related to this phase-locking. One possible explanation on the basis of phase-locking is as follows. Even in a whispered waveform, there exists a periodicity corresponding to formant frequencies, although the periodicity is not clear. If we assume that both of the place and temporal cues could influence on perceiving pitch of whispers, the region of frequency less than 5 kHz should be impact more than the region of frequency more than 5 kHz on the perception of pitch in whispered speech. Hence, increasing the pitch of whispered speech may cause upward shift of the formants less than 5 kHz firstly and/or enhancement of the power of high frequency range more than 5 kHz.

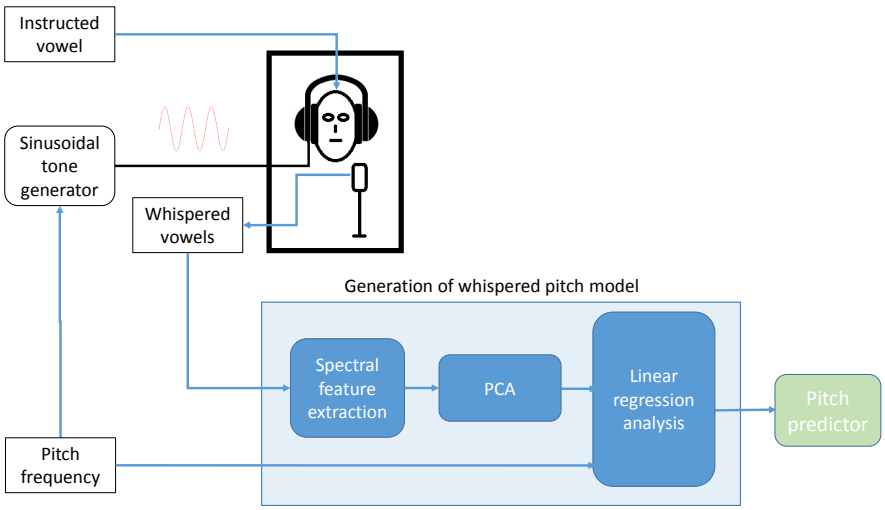


FIGURE 3.14: Generating a pitch predictor [34].

3.3 Construction of pitch prediction model

This section describes the method in which we constructed an pitch predictor from spectral information of whispered vowels [34].

3.3.1 Generation of vowels with a perceived pitch

It is generally difficult to measure the pitch in uttered whispered vowels because there is no fundamental frequency F_0. Therefore, we developed a novel method for connecting the uttered whispered vowel to the perceived pitch as follows.

We first generate a pure tone of a known frequency as a guide tone. A subject is asked to listen to the guide tone and then utter a specified vowel in such a way that the guide tone and the uttered vowel have the same perceived pitch. That is, the subject fulfills the roles of both a listener and a speaker simultaneously, as shown in Fig. 3.14

We used all Japanese vowels, /i/, /e/, /a/, /o/ and /u/, spoken in whisper by three male subjects (Speakers M1–M3) aged 22–46 years and

TABLE 3.1: Speakers' attributes. #Data shows the number of collected speech samples over five vowels for the speaker.

Speaker ID	Gender(M/F)	Age	Nationality	#Data
M1	M	22	Japanese	38
M2	M	23	Japanese	42
M3	M	46	Japanese	33
F1	F	22	Japanese	28
F2	F	22	Japanese	42
F3	F	21	Japanese	32

three female subjects (Speakers F1–F3) aged 21–22 years (Table 3.1). They whispered each of the five vowels separately. In a soundproof booth, we recorded their whispers using a microphone (Sony C-48), a microphone amplifier (Avaron Design M5), and a digital recorder (Sony DTC-2000 ES or Fostex FR-2). The sampling rate was 48 kHz and the quantization level was 16 bits.

Before the recordings, we determined the lowest frequency of the guide tone. First, the speaker was instructed to utter the whispered vowel as low as possible; the speaker then tuned the guide tone to that pitch by the following method (method of adjustment): the speaker turns the frequency dial of the pure tone generator to match the pitch of the pure tone, which is presented through headphones (Sennheiser HDA200), with the whispered vowel being uttered by the speaker. The initial position of the frequency dial was about 350 Hz for Speaker F1 and about 100 Hz for the other speakers. Secondly, the nearest higher tone in the music scale (standard tuning; A4 = 440 Hz) was chosen as the first guide tone for that vowel of the speaker.

The speaker uttered one vowel at a time in such a way that the pitch was as close as possible to the presented guide tone coming from the headphones into their ears (Fig. 3.14). We repeated this process as long as the speaker could produce the vowel, increasing the tone by a half step of the musical scale each time.

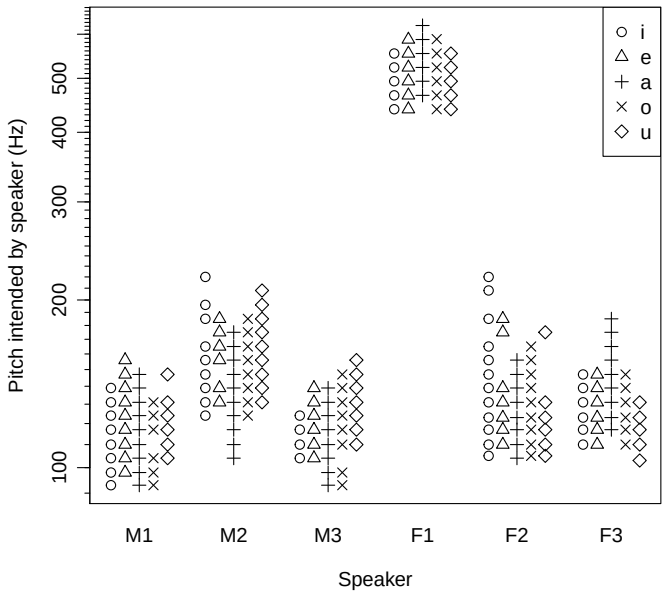


FIGURE 3.15: Intended pitch by the speakers of all speech samples.

Next, we cut a duration of 0.5 seconds from the recorded vowel data[2] and removed the low frequency noise by a high-pass filter whose cutoff frequency is 50 Hz. Furthermore, we normalized the data so as to have the same sound pressure level. We did not apply any high frequency emphasis. Hereafter, we defined the frequency of the guide pure tone the "pitch intended by the speaker."

Figure 3.15 shows the intended pitches of all speech samples which were used in the following analyses. We excluded some samples when they violated the monotonic order of pitch. The fact that Speaker F1's sample pitch is higher than that of the others suggests that the initial frequency of the pure tone in the pitch matching procedure affects the pitch decision from the faint pitch of whisper.

[2]Only the duration of a sample /i/ of 124 Hz uttered by Speaker M3 is 0.4 seconds.

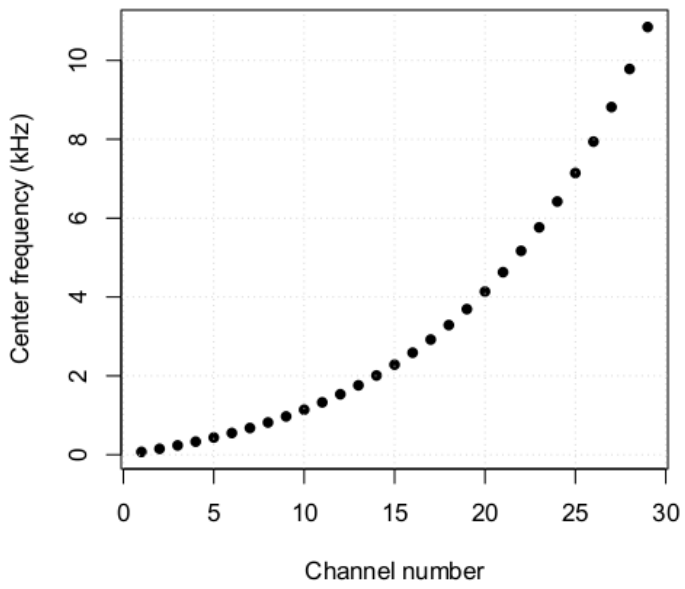


FIGURE 3.16: Center frequencies of mel filter bank channels [34].

3.3.2 Mel filter bank analysis

We analyzed the speech samples by means of the output energies of a 29-channel mel filter bank. The central frequencies of the channels are shown in Fig 3.16. Before the analysis, we down-sampled the data from 48 kHz to 24 kHz. The frame length is 25 ms and the frame period is 10 ms. In each frame, the outputs of the filter bank are measured in decibels. We averaged the outputs over frames in a vowel of a pitch intended by the speakers.

The results are shown in Figs. 3.17, 3.18, and 3.19. In these figures, only the highest and lowest pitch samples are shown. From Fig. 3.17, we can see that an increase of the pitch intended by the speaker causes the following tendencies:

T1: In the low frequency range under 500 Hz (around channel #5), the amount of energy decreases,

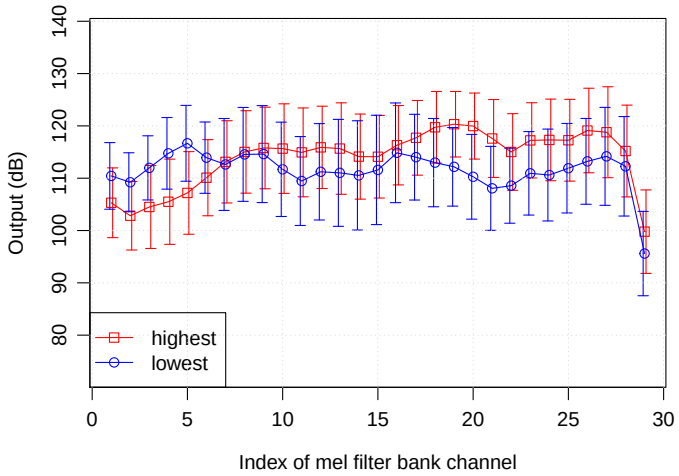


FIGURE 3.17: Average and standard deviation of mel filter bank outputs over five whispered Japanese vowels uttered by six speakers. Only the lowest and highest pitch samples in each speaker's vowels are used.

T2: In the middle and higher ranges over 3 kHz (around channel #17), the energy of the spectrum increases, and

T3: T1 and T2 result in a flatter spectral tilt.

These tendencies are also seen in Figure 3.18. In addition, for every vowel, in the frequency range under 5 kHz (around channel #22), the spectrum shifts toward the higher frequency region. For each speaker in Figure 3.19, the tendencies T1 – T3 are also observed. In these figures, we have shown the data for two pitch extremes, but the same trend can be observed for analyses on a finer scale (Fig. 3.20).

These findings are consistent with the observation of previous studies (e.g., [45]) in the sense that the contour of a spectrum affects the perceived pitch as a whole; therefore, our study contributes towards the quantification of this tendency.

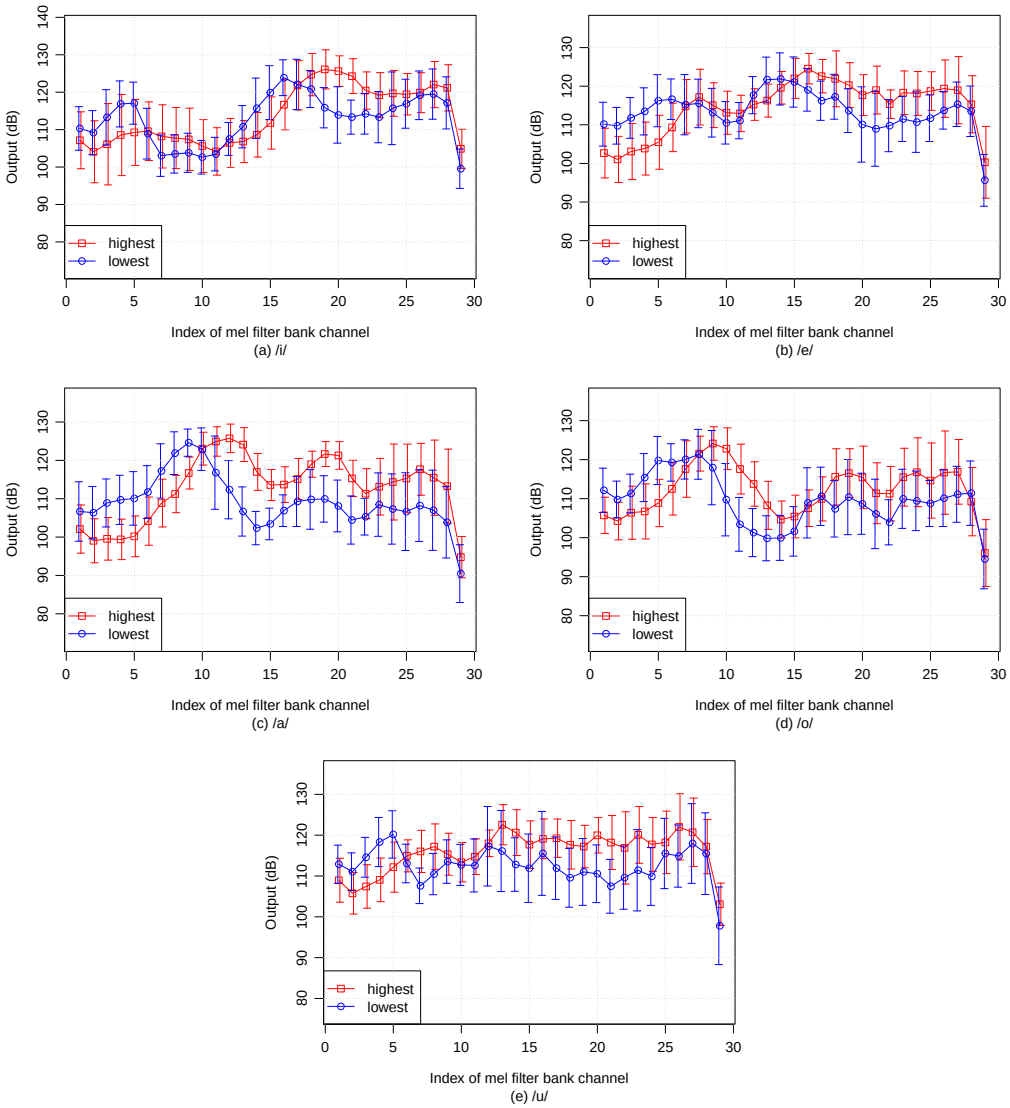


FIGURE 3.18: Outputs of mel filter bank for five whispered Japanese vowels. Only the lowest and highest pitch samples are used. The average plots with plus and minus standard deviation bars are drawn from 279 speech frames for /i/ and 288 frames for /e/, /a/, /o/ and /u/ for the six speakers.

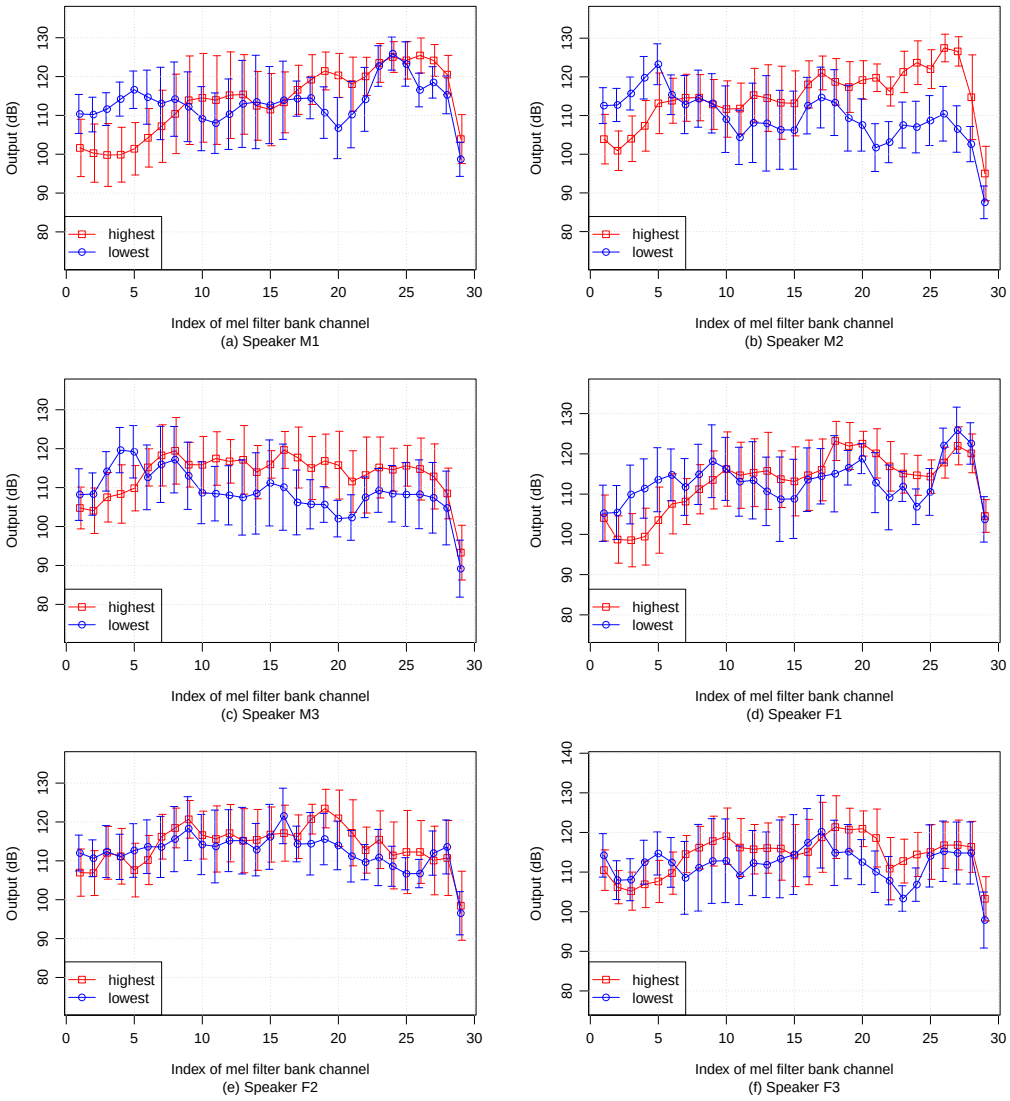


FIGURE 3.19: Outputs of mel filter bank for five whispered Japanese vowels. Only the lowest and highest pitch samples are used. The average plots with plus and minus standard deviation bars are drawn using speech frames of all vowel samples for every speaker.

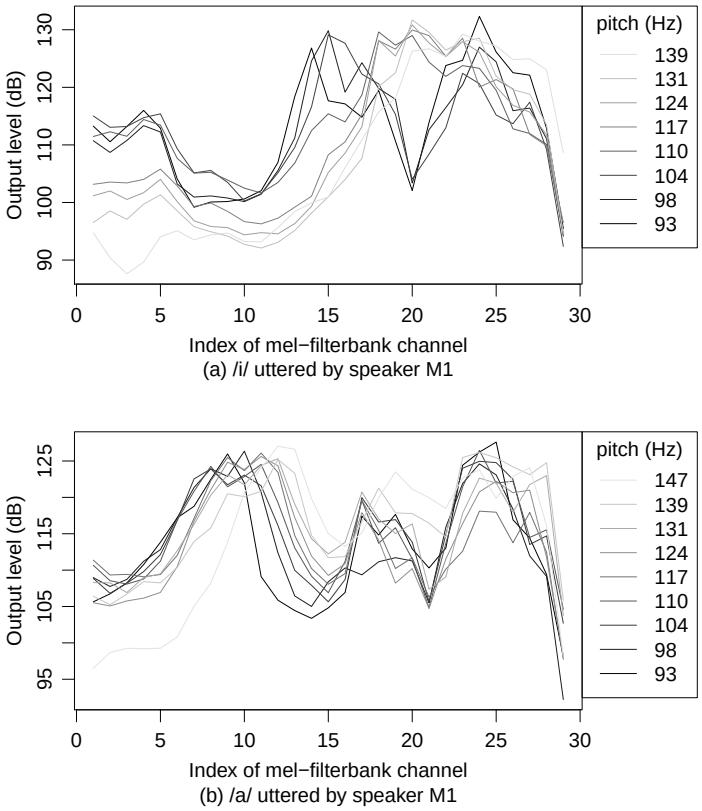


FIGURE 3.20: Outputs of mel filter bank of whispered Japanese /i/ and /a/ uttered by male speaker M1 in response to detailed pitch change [34].

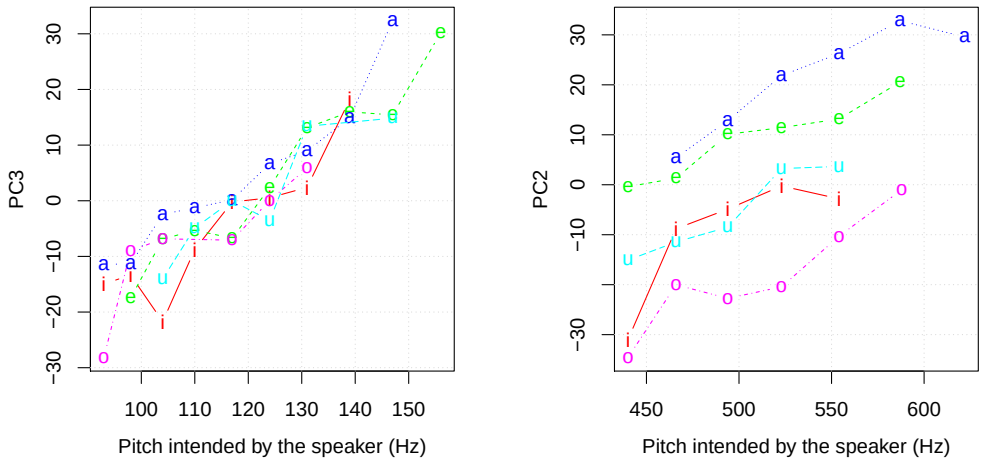
3.3.3 Pitch prediction of whispered vowel

Multiple regression

One of the goals of this thesis is to estimate pitch from the observed spectral information and to generate the fundamental frequency F_0 for whisper-to-normal speech conversion. However, since whispered pitch is faint and subjective, it is difficult to determine the corresponding guide tone. Therefore, some subjects will choose different guide tones to others. Indeed, the intended pitches by Speaker F1 were higher than those by the other subjects. Consequently, we need to construct a specific pitch predictor for each individual.

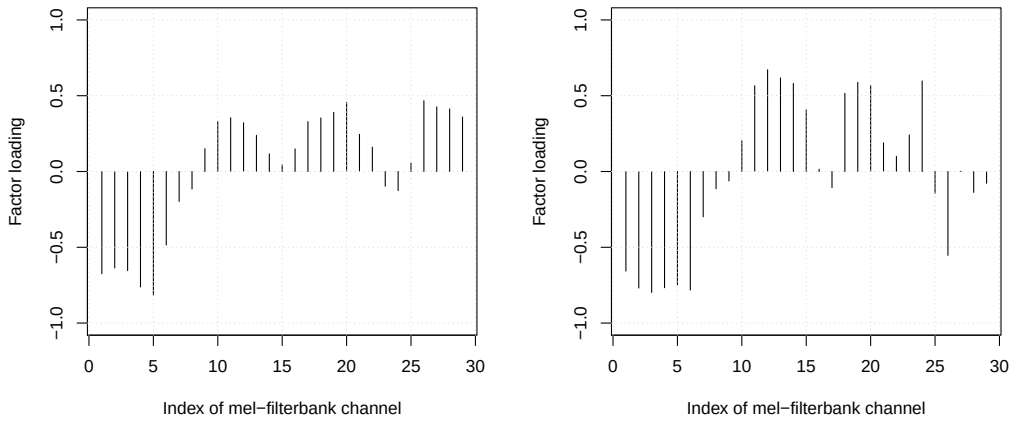
We designed a two-step pitch predictor that utilizes principal component analysis (PCA) followed by multiple regression analysis (MRA). PCA is carried out to reduce the dimensionality and to calculate uncorrelated variables, and MRA is applied to estimate the value of F_0. Applying PCA, 29 output values of the filter bank uniformly arranged over the mel scale were reduced to the smallest number of principal components so as to keep the contribution rate of 95%. We then estimated the coefficients of a linear regression by regarding these transformed variables as the input variables and the logarithm of the pitch intended by the speaker as the output variable. Next, through t-testing, we kept only significant input variables for which the p-value was smaller than 0.1. Finally, we applied MRA again with those contributing input variables to obtain the F_0 predictor.

The results for the six subjects are shown in Figs. 3.21–3.23. In the PCA step, the number of chosen principal components whose cumulative contribution rate is over 95% was 8 or 9, e.g., 9 for M1 and 8 for F1. In general, the first principal component is the most contributing for approximation



(a) Speaker M1 (b) Speaker F1

FIGURE 3.21: Principal component (PC3 for M1 and PC2 for F1) which contributes most to predict the pitch intended by the speaker. [34]



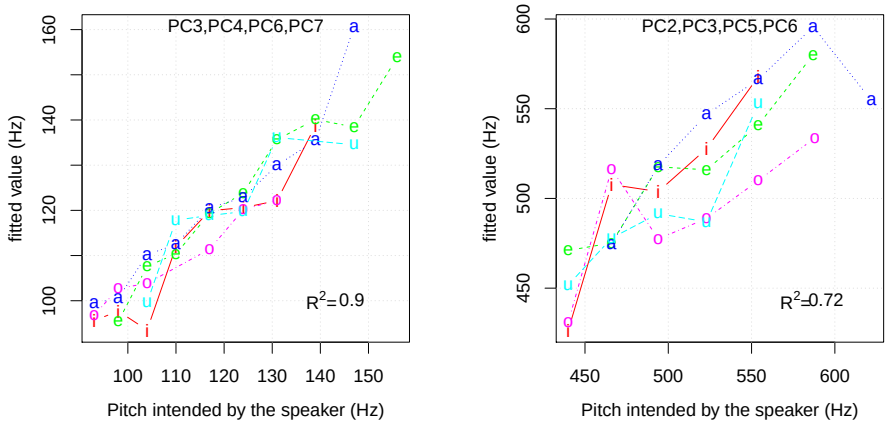
(a) Speaker M1 (b) Speaker F1

FIGURE 3.22: Factor loading of the most pitch-contributing principal component corresponding to Fig. 3.21 [34].

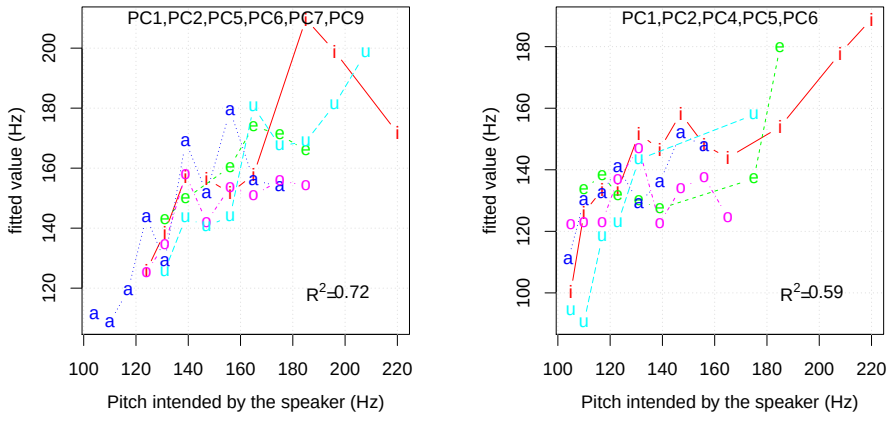
of the given data, but it is not always the most contributing one for the regression. Therefore, we carried out multiple linear regressions with these principal components as inputs and the logarithm values of the pitches intended by the speaker as outputs, and ranked them by their p-values in t-testing (the smaller, the more contributing). For example, the third principal component was found to contribute most in the regression for M1 and the second principal component was most contributing for F1 as shown in Fig. 3.21. From Fig. 3.21, we can see that the principal components are proportional to the pitch intended by the speaker to some extent. However, we also note that the value varies largely for the vowels uttered by Speaker F1. From Fig. 3.22, we can see how such a leading principal component reflects the original features (filter bank outputs). In Fig. 3.22, the *factor loading values* show the correlation between each input variable and the value of the leading principal component. For both M1 and F1, the factor loading values take on large negative values in the channels 1–5 (i.e., below 500 Hz), and large positive values in channel #10 and above (starting at about 1 kHz). This does not violate the T1–T3 tendency.

Next, we selected only the principal components that contributed to regression with a p-value smaller than 0.1. The number of components ranged from four to six depending on the subject. We then carried out MRA again with the chosen principal components. The performance of the pitch predictors is shown in Fig. 3.23. The quality of fit is evaluated by the value of the "coefficient of determination" R^2 [3]. The average value of R^2 was 0.75 (0.90 for M1, 0.72 for M2, 0.74 for M3, 0.72 for F1, 0.59 for F2, and 0.83 for F3, respectively). In Fig. 3.23, we can see that the predicted values are close to the intended values and their ordering approximately follows the pitch intended by speaker.

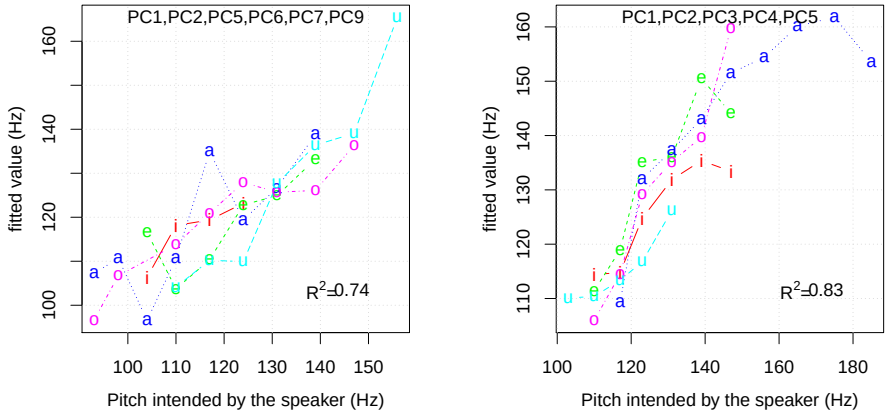
[3] R^2 falls in the 0–1 range, and is one for a perfect fit and zero for a failed fit.



(a) Speaker M1 (b) Speaker F1



(c) Speaker M2 (d) Speaker F2



(e) Speaker M3 (f) Speaker F3

FIGURE 3.23: Regression fit for all six speakers in five vowels. The coefficient of determination R^2 and the selected principal components are indicated inside each graph.

3.4 Summary

The results of the acoustic analysis of vowels step-wise increasing pitch show a tendency of upward shift of F_1, F_2, global peaks, and flattened spectral tilts on overall vowels with increasing pitch. The intended pitches by the speakers were not correspondent to a specific formant frequency. We applied multivariate analysis such as the principal component analysis to the data in order to make clear which part of frequency contributes much to the change of pitch. Two or three formants of less than 5 kHz are shifted upward and the energy is increased in high frequency region over 5 kHz. Moreover, we investigated the quantitative relationship between spectral information and pitch in more detail. From the results, we derived a multiple regression function to convert the outputs of a mel-scaled filter bank of whispered speech into the perceived pitch value.

Chapter 4

Whisper to Normal Speech Conversion Preserving Pitch

This chapter describes a method for converting whispered speech to corresponding normal speech, and subjective evaluation experiments of the method using Japanese pitch accent words [34].

4.1 Background

4.1.1 Conventional conversion systems

The typical systems converting whispered speech to corresponding normal speech are shown in Fig. 4.1 and Table 4.1.

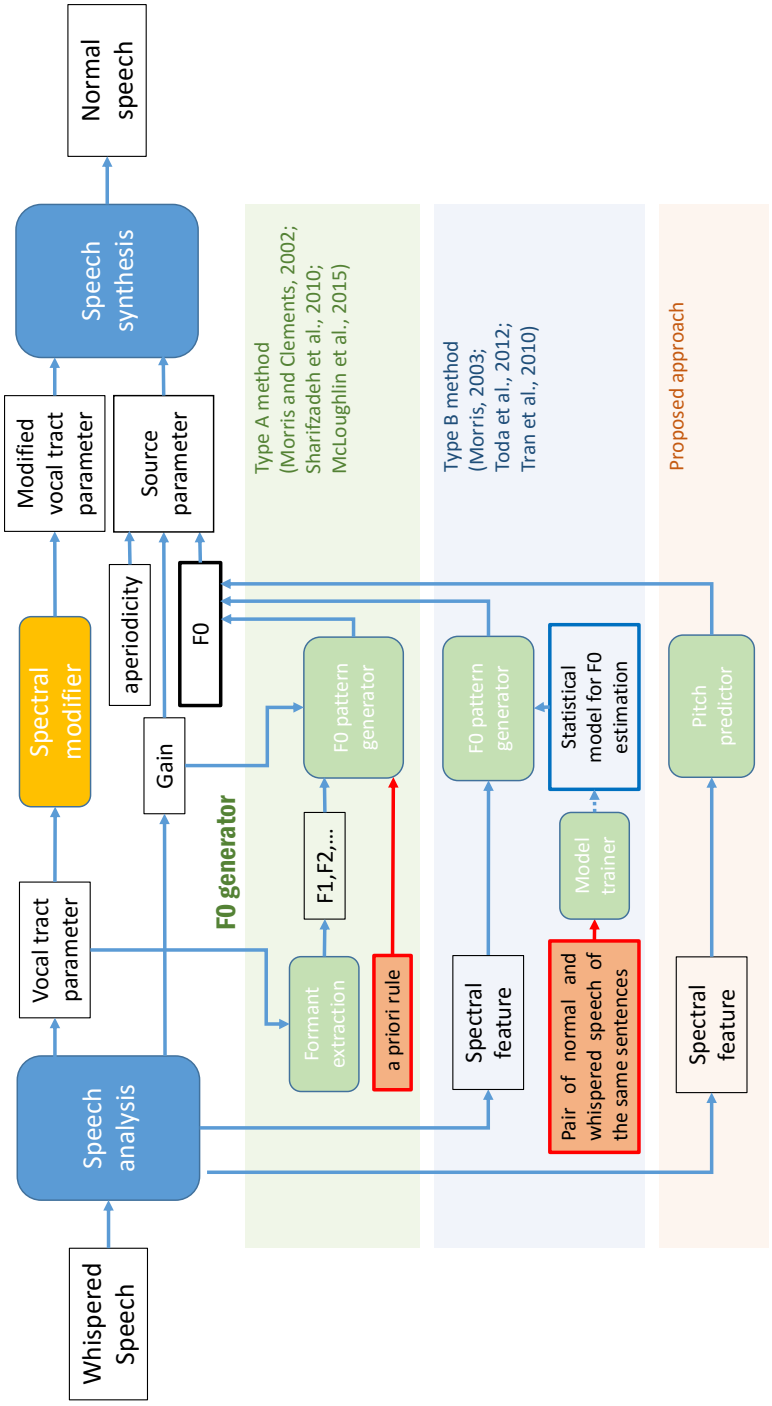


FIGURE 4.1: Block diagram of whispered speech to normal speech converters. (Reprinted from [34])

Conversion of whispered speech to normal speech requires 1) modification of vocal tract information and 2) generation of the fundamental frequency F_0 (Fig. 4.1). Modification of vocal tract information is typically carried out by shifting formant frequencies and altering formant bandwidths or by spectrum estimation using a Gaussian mixture model.

There are mainly two types of methods for F_0 generation. In one approach (Type A in Fig. 4.1), the formants or the gain obtained from the whispered speech are used for generating F_0 (e.g., [43], [48], [52]). This approach typically needs additional information such as the rule for F_0 generation as a function of formants or gain. In general, it is difficult to construct such a rule appropriately. In the other approach (Type B in Fig. 4.1), a model of F_0 is constructed in advance in a feature space, such as a Gaussian mixture model in the cepstral coefficient space [61], [62] or a jump Markov linear system with speech gain and linear prediction cepstral coefficients (LPCC) [47] , and F_0 is generated by referring to the extracted spectral information of the whispered speech. In this case, we need the pairs of normal speech and whispered speech of the same sentences with time mapping of phonemes to train the model. We summarize the methodology in Table 4.1.

In the approach proposed in this study, we do not need any *a priori* rules as in the Type A approach or any extra data other than whispered speech (unlike in Type B approaches). By our approach, whispered speech's original intonation, accent and tone are expected to remain preserved in the converted normal speech.

4.1.2 Pitch and accent

Pitch-accent languages (e.g., Japanese) and tone languages (e.g., Chinese) have lexically distinct words that differ only in pitch [41]. Using such

TABLE 4.1: Methods for conversion of whispered speech to normal speech. In the analysis and synthesis (A&S) method, the abbreviations used are MELP for mixed excitation linear prediction, CELP for Code-Excited Linear Prediction, and MLSA for mel log spectrum approximation. The column 'Extra info.' indicates which additional information other than the whispered speech is necessary.

Type	Method of F_0 generation (author(s), year)	A&S method	Extra info.
A	Real-time pitch control by gain [48]	MELP	Rule for F_0 generation
A	Constant F_0 [52]	CELP	—
A	Harmonic derivative of formant frequencies [43]	LPC and sine wave speech-based reconstruction	Rule for F_0 generation
B	F_0 prediction from gain and LPCC using JMLS [47]	MELP	Normal speech
B	F_0 prediction from spectral parameter using GMM [61]	STRAIGHT and MLSA	Normal speech
P	Pitch prediction from spectrum (Proposed)	LPC	—

TABLE 4.2: Conditions of LPC analysis and synthesis for whisper to normal speech conversion

Sampling rate	16 kHz (down-sampled from 48 kHz)
Window	Hamming Window of 400 points (25 ms)
Frame period	80 points (5 ms)
LPC order	20
Voice source	Impulse sequence invoked by generated F_0

words in four pitch-accent or tone languages, Jensen [19] reported the correct recognition rates of the whispered word meaning. The rates were 53% to 73% in Norwegian, 100% in Swedish, 71% to 85% in Slovenian, and 73% to 88% in Chinese (Mandarin). As for Japanese whispered words, Sugito, Higasikawa, Sakakura, *et al.* [55] reported an approximately 90% accuracy in perceptually recognizing correct accents. In Mandarin, 72% of whispered words were judged with a correct tone [28]. Thus, whispered speech is considered to carry pitch-accent or tonal cues.

4.2 Whisper to normal speech conversion using pitch of whisper

We constructed a system for conversion of whispered speech to normal speech as shown in the last block in Fig. 4.1. Speech was synthesized using an LPC vocoder, which utilized the spectral information and the pitch estimated at a frame level. Other details are shown in Table 4.2.

Unlike previous techniques (upper two blocks in Fig. 4.1), this system generates F_0 in an online fashion according to the spectral information obtained by the mel filter bank without relying on *a priori* rules and availability of parallel normal speech. Using the relation between spectral information and pitch of whisper, it is anticipated that pitch accents could be recovered correctly in the converted normal speech. To confirm this, we conducted an accent recovery experiment.

TABLE 4.3: Speech material for subjective evaluation: eight Japanese 2-morae words

Japanese Word (2 morae)	Meaning	
	w. first accent (e.g., /áme/)	w. second accent (e.g., /amé/)
/ame/	rain	candy
/asa/	morning	hemp
/fuji/	(mount) Fuji	wisteria
/haru/	(season) spring	(**verb**) tense
/iki/	breath	stylish
/kami/	God	paper
/tsuyu/	dew	rainy season
/yoi/	good	drunkenness

4.3 Accent recovery experiment

We examined to what degree the accents are kept by this conversion in an auditory experiment. As described in the literature, e.g., McCawley [41], the accent in Japanese words is a pitch accent (low/high), not a stress accent (weak/strong) as seen in English words. Therefore, the correct perception of Japanese accents depends on how well we distinguish F_0 changes in a word. We used eight 2-morae Japanese words (Table 4.3). The meaning of each word differs depending on the accent position such as /áme/ meaning "rain" and /amé/ meaning "candy."

We compared three kinds of stimuli generated from a whispered word: 1) the original whispered word, 2) the synthesized word with a constant value of F_0, and 3) the synthesized word with varying F_0 values estimated by the proposed method. Here, the constant value in 2) is the median value of F_0's estimated by the proposed method for the word. Sixteen whispered words (eight words with two different accents) uttered by one male speaker (M3) and one female speaker (F2) are converted to the other two kinds of synthesized words.

Eight subjects (4 males and 4 females) aged 16–22 years, different from M3 and F2, were asked to answer if the accent is put on the first mora or

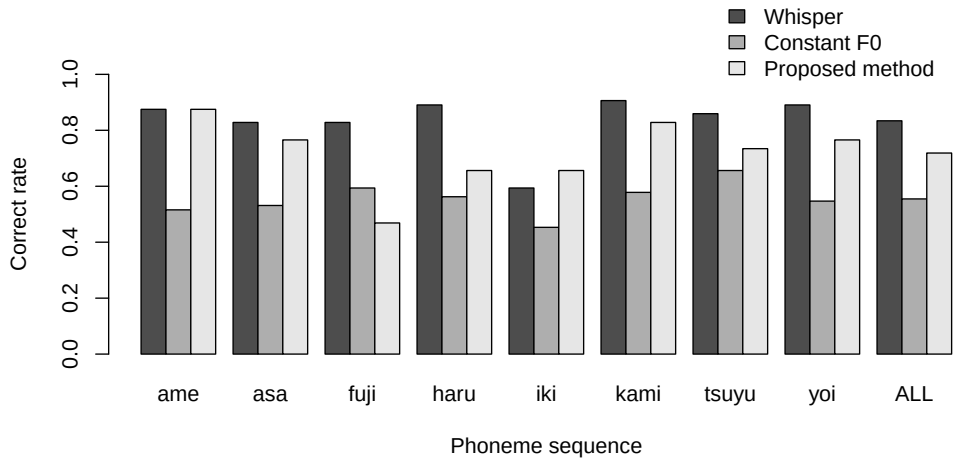


FIGURE 4.2: Accent recognition rates over eight words in three of "whispered speech," "synthesized speech with a constant F_0," and "synthesized speech by the estimated F_0" (proposed method) [34].

the second mora while listening to the stimuli presented through headphones (Sennheiser HD650). In practice, they chose one of two words displayed on the PC screen, deciding between two kanji words representing two different meanings (e.g., "rain" or "candy" in kanji).

We recorded the number of correct answers, i.e., the number of times that a subject chose the correct word written in kanji in two trials, in each of which sixteen words were presented once in a random order.

The results are shown in Fig. 4.2. For reference purposes, the accents of 83% of the whispered words were correctly judged. For the synthesized words, 72% of the words with estimated F_0 were correctly identified and 55% of the words with a constant F_0 were correctly identified. The word /fuji/ was the only word for which the correctly identified synthesized words with a constant F_0 was higher than the percentage of correctly identified synthesized words with estimated F_0. This could be because F_0 did not change significantly in the proposed method and another acoustic characteristic, such as power, was dominant for accent identification. An

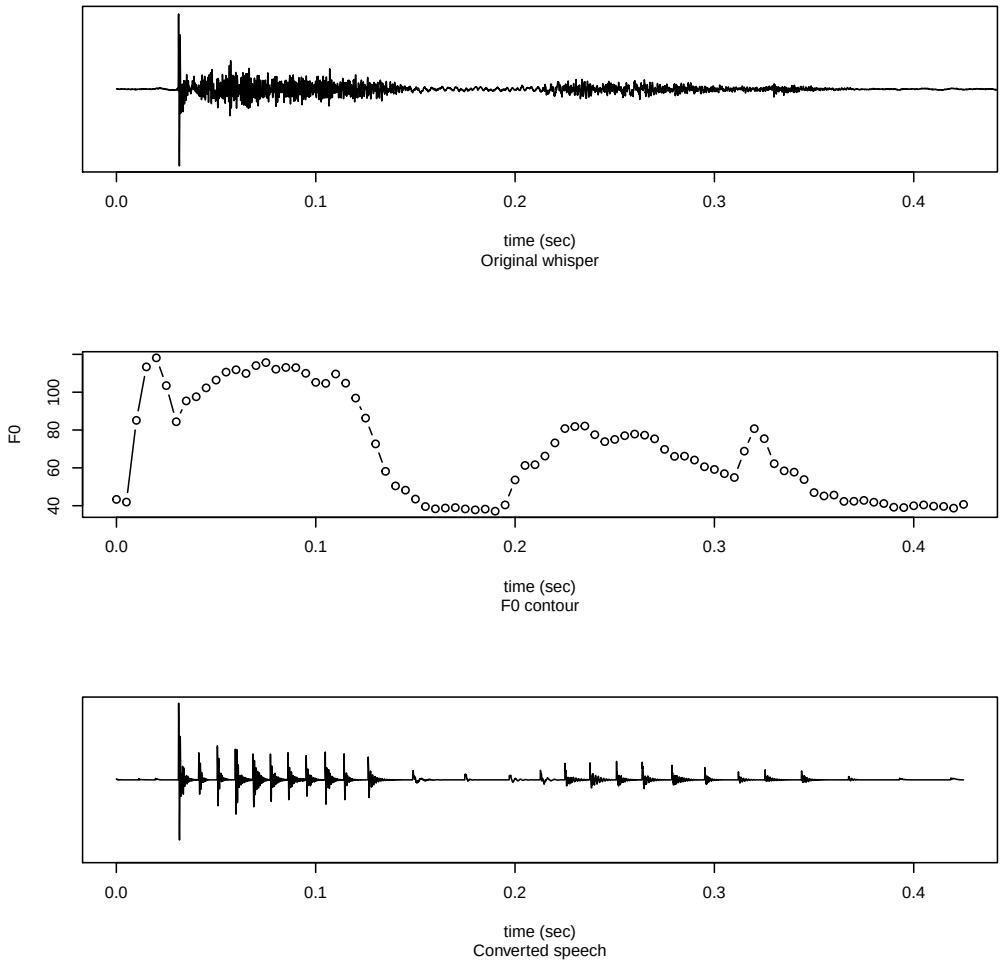


FIGURE 4.3: Waveforms and generated F_0 contour from a whispered word (/áme/ uttered by Speaker F2) [34].

example of synthesized words is shown in Fig. 4.3 with the contour of estimated F_0 and the waveform. In this example of /áme/, the first accent is confirmed by high values of F_0 in the first 0.1 seconds. By comparing the waveforms between the original and converted speech, we notice that the latter is more jaggy and has distinct pitch according to F_0. Therefore, the accent position was clearly kept, in other words the accent was "recovered," but the sound quality was still not satisfactory due to the poor ability of the synthesizer. Although the generated F_0's for silent segments are quite low, this does not cause a large problem since there is no speech present in these parts. Since the frequencies of the F_0 contour are relatively low compared with the real speech of female speakers, it causes the jaggy waveform and the unnatural nature of the converted speech. However, our method recovers the word accent better than the constant F_0 without any adjustment of the F_0 contour.

4.4 Summary

We conducted experiments in which speakers uttered five whispered Japanese vowels in accordance with the pitch of a guide pure tone. From the results, we derived a multiple regression function to convert the outputs of a mel-scaled filter bank of whispered speech into the perceived pitch value. Using this estimated pitch value as F_0, we constructed a system for conversion of whispered speech to normal speech. Auditory experiments demonstrated that the correctly perceived rate of Japanese word accent was increased from 55.5% to 72.0% compared with that when a constant F_0 was used.

Chapter 5

Discussion and Future Work

The goal of this study was to find out the acoustic correlates of phonemic quality and pitch of whispered speech, and to estimate the pitch values of whisper. whispered words were able to convert to normally phonated words using the estimated pitch values. By the subjective evaluation, it was shown that 72% of them had the same pitch accent of the original whispered words.

Contributions of this thesis are as follows:

1. Described the difference of formant frequencies between whispered and voiced Japanese vowels considering its effect to phonemic quality of whispered vowels.

2. Developed the recording procedure which can obtain a pitch value of each whispered vowel, and analyzed quantitatively the relation between pitch values and spectral shape of the whispered vowels.

3. Developed the method of making equations to predict pitch contour from whispered speech using the results of the analyses, and proposed the converting method from whispered to normal speech preserving prosody of the whisper.

Remaining works related to this thesis are

- experiments on whispered speech intelligibility with increasing the number of speakers to reveal the perceptional cues of whispered voiced consonants,
- increasing accuracy of pitch prediction by finding other features and by improving the algorithm and the conditions,
- construction of speaker-independent pitch prediction model, and
- improvement of intelligibility and naturalness of converted speech.

Chapter 6

Conclusion

This study was carried out to reveal the acoustical and perceptual characteristics of whispered speech, and to improve prosody of normal speech converted from whisper.

First, acoustic analyses and auditory experiments related to phonemic quality of Japanese whispered speech was carried out. As for whispered vowels, the low frequency formants of whispered vowels tend to shift to higher frequency regions compared with those of normally phonated vowels. By the auditory experiments using synthetic speech, the change of phonemic quality occurred by the difference of glottal sources even if the formant frequencies are same. This may be caused by the change of pitch. Moreover, the author showed the identification rate of whisper decreased by 20% to normal speech by phoneme identification experiments using Japanese 110 monosyllables, and showed a possibility to reflect the balance of energy over low and high frequency region upon perception and generation of voiced or unvoiced consonants differed from normally phonated consonants.

Next, we confirmed that formant frequencies and distribution of spectral energy of whispered vowels changes according to the pitch and succeeded in constructing the multiple regression model to estimate a pitch value of whispered vowels by applying the principal component analysis and the multiple regression analysis to the results of mel filter bank

analysis. Moreover, whispered-to-normal speech conversion method was proposed that uses the pitch values estimated from whispered speech as the F_0 contour. Auditory experiments demonstrated that the correctly perceived rate of Japanese word accent was increased from 55.5% to 72.0% compared with that of a constant F_0 contour was used.

These results will be applicable to smart phone applications for gaining intelligibility of whispered speech[15]. Although silent speech interface (SSI) has also been able to use for such purposes[3], SSI needs extra apparatus other than ordinary microphones. Moreover, if prosody of whispering is obtained, it will be used for speech recognition[64], emotion recognition[5], [63], speech understanding and dialog system using whispered speech for interfaces of humanoid robots etc. As a pathological usage, it might be used to improve the pitch of alaryngeal speech.

Nowadays, associating with rapid progress of ICT technology, e.g., vocaloids, unpredictable usage of speech could happen. The study of whispering should contribute such an unexpectable matters.

Bibliography

[1] ANSI, *American national standard acoustical terminology*, ANSI S1.1-1994. American National Standards Institute, 1994.

[2] X. Chen and H. Zhao, "Relationship between fundamental and formant frequency in whispered Mandarin", in *Audio, Language and Image Processing, 2008. ICALIP 2008. International Conference on*, 2008, pp. 303–306. DOI: 10.1109/ICALIP.2008.4590067.

[3] B. Denby, T. Schultz, K. Honda, T. Hueber, J. M. Gilbert, and J. S. Brumberg, "Silent speech interfaces", *Speech Communication*, vol. 52, no. 4, pp. 270–287, 2010. DOI: 10.1016/j.specom.2009.08.002.

[4] I. Eklund and H. Traunmüller, "Comparative study of male and female whispered and phonated versions of the long vowels of Swedish", *Phonetica*, vol. 54, no. 1, 1997. DOI: 10.1159/000262207.

[5] M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases", *Pattern Recognition*, vol. 44, no. 3, pp. 572–587, 2011. DOI: 10.1016/j.patcog.2010.09.020.

[6] H. Fujisaki and T. Kawashima, "The roles of pitch and higher formants in the perception of vowels", *IEEE Transactions on Audio and Electroacoustics*, vol. 16, no. 1, pp. 73–77, 1968. DOI: 10.1109/TAU.1968.1161952.

[7] G. J. Harbold, "Pitch ratings of voiced and whispered vowels", *The Journal of the Acoustical Society of America*, vol. 30, no. 7, pp. 600–601, 1958. DOI: 10.1121/1.1909704.

[8] W. F. L. Heeren and V. J. van Heuven, "The interaction of lexical and phrasal prosody in whispered speech", *The Journal of the Acoustical Society of America*, vol. 136, no. 6, pp. 3272–3289, 2014. DOI: 10.1121/1.4901705.

[9] W. F. L. Heeren and C. Lorenzi, "Perception of prosody in normal and whispered French", *The Journal of the Acoustical Society of America*, vol. 135, no. 4, pp. 2026–2040, 2014. DOI: 10.1121/1.4868359.

[10] M. Higashikawa, K. Nakai, A. Sakakura, and H. Takahashi, "Perceived pitch of whispered vowels-relationship with formant frequencies: a preliminary study", *Journal of Voice*, vol. 10, no. 2, pp. 155 –158, 1996. DOI: 10.1016/S0892-1997(96)80042-7.

[11] M. Higashikawa and F. D. Minifie, "Acoustical-perceptual correlates of "whisper pitch" in synthetically generated vowels", *Journal of Speech, Language, and Hearing Research*, vol. 42, no. 3, pp. 583–591, 1999. DOI: 10.1044/jslhr.4203.583.

[12] T. Hirahara, "Acoustic analysis of whispered vowels in different notes", ATR, ATR Tech. Report. TR-A-0120, 1991, (in Japanese), pp. 1–28.

[13] T. Hirahara and H. Kato, "The effect of F0 on vowel identification", in *Speech Perception, Production and Linguistic Structure*, Y. Tohkura, E. Vatikiotis-Bateson, and Y. Sagisaka, Eds., Tokyo: Ohmsha, 1992, pp. 89–112.

[14] T. Hirahara, M. Otani, S. Shimizu, T. Toda, K. Nakamura, Y. Nakajima, and K. Shikano, "Silent-speech enhancement using body-conducted vocal-tract resonance signals", *Speech Communication*, vol. 52, no. 4, pp. 301 –313, 2010. DOI: 10.1016/j.specom.2009.12.001.

[15] B. den Hond, *Whisper tech turns secrets into normal speech*, 2016. DOI: 10.1016/S0262-4079(16)31556-1.

[16] C. Huang, X. Y. Tao, L. Tao, J. Zhou, and H. B. Wang, "Reconstruction of whisper in Chinese by modified MELP", in *Computer Science Education (ICCSE), 2012 7th International Conference on*, 2012, pp. 349–353. DOI: 10.1109/ICCSE.2012.6295089.

[17] T. Irino, Y. Aoki, H. Kawahara, and R. D. Patterson, "Comparison of performance with voiced and whispered speech in word recognition and mean-formant-frequency discrimination", *Speech Communication*, vol. 54, no. 9, pp. 998 –1013, 2012. DOI: 10.1016/j.specom.2012.04.002.

[18] T. Ito, K. Takeda, and F. Itakura, "Analysis and recognition of whispered speech", *Speech Communication*, vol. 45, no. 2, pp. 139 –152, 2005. DOI: 10.1016/j.specom.2003.10.005.

[19] M. K. Jensen, "Recognition of word tones in whispered speech", *Word*, vol. 14, no. 2-3, pp. 187–196, 1958. DOI: 10.1080/00437956.1958.11659663.

[20] S. T. Jovičić, "Formant feature differences between whispered and voiced sustained vowels", *Acta Acustica united with Acustica*, vol. 84, no. 4, pp. 739–743, 1998.

[21] S. T. Jovičić and Z. Šarić, "Acoustic analysis of consonants in whispered speech", *Journal of Voice*, vol. 22, no. 3, pp. 263 –274, 2008. DOI: 10.1016/j.jvoice.2006.08.012.

[22] K. J. Kallail and F. W. Emanuel, "The identifiability of isolated whispered and phonated vowel samples", *Journal of phonetics*, vol. 13, no. 1, pp. 11–17, 1985.

[23] K. J. Kallail and F. W. Emanuel, "An acoustic comparison of isolated whispered and phonated vowel samples produced by adult male subject", *Journal of Phonetics*, vol. 12, no. 2, pp. 175–186, 1984.

[24] ——, "Formant-frequency differences between isolated whispered and phonated vowel samples produced by adult female subjects", *Journal of Speech, Language, and Hearing Research*, vol. 27, no. 2, pp. 245–251, 1984. DOI: 10.1044/jshr.2702.251.

[25] H. Kawahara, M. Morise, T. Takahashi, R. Nisimura, T. Irino, and H. Banno, "Tandem-STRAIGHT: a temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, F0, and aperiodicity estimation", in *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, 2008, pp. 3933–3936. DOI: 10.1109/ICASSP.2008.4518514.

[26] D. H. Klatt, "Software for a cascade/parallel formant synthesizer", *The Journal of the Acoustical Society of America*, vol. 67, no. 3, pp. 971–995, 1980. DOI: 10.1121/1.383940.

[27] T. Koizumi, *An Introduction to Phonetics*. Tokyo: Daigaku-shorin, 1996, (in Japanese).

[28] Y.-Y. Kong and F.-G. Zeng, "Temporal and spectral cues in Mandarin tone recognition", *The Journal of the Acoustical Society of America*, vol. 120, no. 5, pp. 2830–2840, 2006. DOI: 10.1121/1.2346009.

[29] H. Konno, H. Kanemitsu, N. Takahashi, and M. Kudo, "Acoustic characteristics related to the perceptual pitch in whispered vowels", in *Automatic Speech Recognition and Understanding (ASRU), 2013 IEEE Workshop on*, 2013, pp. 245–249. DOI: 10.1109/ASRU.2013.6707737.

[30] H. Konno, R. Sato, H. Imai, and M. Kudo, "Deterioration of intelligibility in whispered japanese speech", in *Asia-Pacific Signal and Information Processing Association, 2014 Annual Summit and Conference (APSIPA)*, 2014, pp. 1–4. DOI: 10.1109/APSIPA.2014.7041723.

[31] H. Konno, H. Kanemitsu, and M. Kudo, "Characteristics of whispered vowels uttered to be matching to pure tones' pitch", in *Proceedings of 2012 Autumn Meeting of the Acoustical Society of Japan*, (in Japanese), 2012, pp. 427–428.

[32] H. Konno, H. Kanemitsu, J. Toyama, and M. Shimbo, "Spectral properties of Japanese whispered vowels referred to pitch", *The Journal of the Acoustical Society of America*, vol. 120, no. 5, pp. 3378–3378, 2006. DOI: 10.1121/1.4781611.

[33] H. Konno, K. Kosugi, and M. Ito, "Characteristics of whispered vowels uttered in various pitch", in *Proceedings of 2011 Spring Meeting of the Acoustical Society of Japan*, (in Japanese), 2011, pp. 582–583.

[34] H. Konno, M. Kudo, H. Imai, and M. Sugimoto, "Whisper to normal speech conversion using pitch estimated from spectrum", *Speech Communication*, vol. 83, pp. 10 –20, 2016. DOI: 10.1016/j.specom.2016.07.001.

[35] H. Konno, N. Kurahashi, H. Kanemitsu, and M. Shimbo, "Spectral features of whispered vowels uttered under different conditions", in *Proceedings of 2002 Autumn Meeting of the Acoustical Society of Japan*, (in Japanese), vol. I, 2002, pp. 375–376.

[36] H. Konno, J. Toyama, M. Shimbo, and K. Murata, "A study on the formant frequency and phonemic quality of Japanese whispered vowels", *The Journal of the Acoustical Society of Japan*, vol. 50, no. 8, pp. 623–630, 1994, (in Japanese). [Online]. Available: http://ci.nii.ac.jp/naid/110003110709/en/.

[37] ——, "The effect of formant frequency and spectral tilt of unvoiced vowels on their perceived pitch and phonemic quality.", *IEICE technical report. Speech*, vol. 95, no. 565, pp. 39–45, 1996, (in Japanese).

[Online]. Available: http://ci.nii.ac.jp/naid/110003296570/en/.

[38] J. Laver, *Principles of phonetics*. Cambridge University Press, 1994.

[39] X.-L. Li and B.-L. Xu, "Formant comparison between whispered and voiced vowels in mandarin", *Acta Acustica united with Acustica*, vol. 91, no. 6, pp. 1079–1085, 2005.

[40] M. Matsuda, H. Mori, and H. Kasuya, "Formant structure of whispered vowels", *The Journal of the Acoustical Society of Japan*, vol. 56, no. 7, pp. 477–487, 2000, (in Japanese). [Online]. Available: http://ci.nii.ac.jp/naid/110003110932/en/.

[41] J. McCawley, *The phonological component of a grammar of Japanese*, ser. Monographs on linguistic analysis. Mouton, 1968.

[42] R. E. McGlone and W. H. Manning, "Role of the second formant in pitch perception of whispered and voiced vowels", *Folia Phoniatrica et Logopaedica*, vol. 31, no. 1, pp. 9–14, 1979. DOI: 10.1159/000264144.

[43] I. V. McLoughlin, H. R. Sharifzadeh, S. L. Tan, J. Li, and Y. Song, "Reconstruction of phonated speech from whispers using formant-derived plausible pitch modulation", *ACM Transactions on Accessible Computing*, vol. 6, no. 4, 12:1–12:21, May 2015. DOI: 10.1145/2737724.

[44] I. V. McLoughlin, J. Li, and Y. Song, "Reconstruction of continuous voiced speech from whispers.", in *Proceedings of INTERSPEECH 2013*, ISCA, 2013, pp. 1022–1026.

[45] W. Meyer-Eppler, "Realization of prosodic features in whispered speech", *The Journal of the Acoustical Society of America*, vol. 29, no. 1, pp. 104–106, 1957. DOI: 10.1121/1.1908631.

[46] B. C. J. Moore, *An introduction to the psychology of hearing*, Sixth. Brill, 2013.

[47] R. W. Morris, "Enhancement and recognition of whispered speech", PhD thesis, Georgia Institute of Technology, Atlanta, GA, USA, 2003.

[48] R. W. Morris and M. A. Clements, "Reconstruction of speech from whispers", *Medical Engineering & Physics*, vol. 24, no. 7-8, pp. 515–520, 2002. DOI: 10.1016/S1350-4533(02)00060-7.

[49] T. Nakajima, T. Suzuki, H. Ohmura, S. Ishizaki, and K. Tanaka, "Estimation of vocal tract area function by adaptive deconvolution and adaptive speech analysis system", *The Journal of the Acoustical Society of Japan*, vol. 34, no. 3, pp. 157–166, 1978, (in Japanese). [Online]. Available: http://ci.nii.ac.jp/naid/110003109146/en/.

[50] J. Pickles, *An Introduction to the Physiology of Hearing*. Academic Press, 1988.

[51] M. F. Schwartz, "Power spectral density measurements of oral and whispered speech", *Journal of Speech, Language, and Hearing Research*, vol. 13, no. 2, pp. 445–446, 1970. DOI: 10.1044/jshr.1302.445.

[52] H. R. Sharifzadeh, I. V. McLoughlin, and F. Ahmadi, "Reconstruction of normal sounding speech for laryngectomy patients through a modified CELP codec", *Biomedical Engineering, IEEE Transactions on*, vol. 57, no. 10, pp. 2448–2458, 2010. DOI: 10.1109/TBME.2010.2053369.

[53] H. R. Sharifzadeh, I. V. McLoughlin, and M. J. Russell, "A comprehensive vowel space for whispered speech", *Journal of Voice*, vol. 26, no. 2, e49–56, 2012. DOI: 10.1016/j.jvoice.2010.12.002.

[54] W. F. Smith, "A formant study of whispered vowels", PhD thesis, The University of Oklahoma, 1973. [Online]. Available: http://hdl.handle.net/11244/3689.

[55] M. Sugito, M. Higasikawa, A. Sakakura, and H. Takahashi, "Perceptual, acoustical, and physiological study of Japanese word accent in whispered speech", *IEICE technical report. Speech*, vol. SP91-1, no. 1, pp. 1–8, 1991, (in Japanese).

[56] C. C. Tappert, J. Mártony, and G. Fant, "Spectrum envelopes for synthetic vowels", *STL-QPSR*, vol. 4, no. 5, pp. 2–6, 1963. [Online]. Available: http://www.speech.kth.se/prod/publications/files/qpsr/1963/1963_4_3.

[57] V. C. Tartter, "What's in a whisper?", *The Journal of the Acoustical Society of America*, vol. 86, no. 5, pp. 1678–1683, 1989. DOI: 10.1121/1.398598.

[58] V. C. Tartter and D. Braun, "Hearing smiles and frowns in normal and whisper registers", *The Journal of the Acoustical Society of America*, vol. 96, no. 4, pp. 2101–2107, 1994. DOI: 10.1121/1.410151.

[59] V. C. Tartter, "Identifiability of vowels and speakers from whispered syllables", *Perception & Psychophysics*, vol. 49, no. 4, pp. 365–372, 1991. DOI: 10.3758/BF03205994.

[60] I. B. Thomas, "Perceived pitch of whispered vowels", *The Journal of the Acoustical Society of America*, vol. 46, no. 2B, pp. 468–470, 1969. DOI: 10.1121/1.1911712.

[61] T. Toda, M. Nakagiri, and K. Shikano, "Statistical voice conversion techniques for body-conducted unvoiced speech enhancement", *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, no. 9, pp. 2505–2517, 2012. DOI: 10.1109/TASL.2012.2205241.

[62] V.-A. Tran, G. Bailly, H. Lœvenbruck, and T. Toda, "Improvement to a NAM-captured whisper-to-speech system", *Speech Communication*, vol. 52, no. 4, pp. 314 –326, 2010. DOI: 10.1016/j.specom.2009.11.005.

[63] D. Ververidis and C. Kotropoulos, "Emotional speech recognition: Resources, features, and methods", *Speech Communication*, vol. 48, no. 9, pp. 1162–1181, 2006. DOI: 10.1016/j.specom.2006.04.003.

[64] K. Vicsi and G. Szaszák, "Using prosody to improve automatic speech recognition", *Speech Communication*, vol. 52, no. 5, pp. 413–426, 2010. DOI: 10.1016/j.specom.2010.01.003.