Contribution of temporal modulation cues on the perception of speaker individuality and vocal emotion

Zhi Zhu

Japan Advanced Institute of Science and Technology

Doctoral Dissertation

Contribution of temporal modulation cues on the perception of speaker individuality and vocal emotion

Zhi Zhu

Supervisor: Professor Masashi UNOKI

School of Information Science
Japan Advanced Institute of Science and Technology

March 2018

Abstract

The ability of speech communication should be the biggest difference between human and any other animal. Human speech carries not only the language message (linguistic information) but also nonlinguistic information such as speaker individuality and vocal emotion. The speaker individuality information can be defined as the information that is used by human to distinguish such a specific speaker from any other person. The vocal emotion information can be defined as the information that is used by human to identify the emotion state of speaker from speech. Both speaker individuality and vocal emotion play an important role in the speech communication of our daily life. Understanding the mechanism of how human can perceive nonlinguistic information from speech should be very important for the clarification of the mechanism of speech perception. However, the perceptual process of speaker individuality and vocal emotion is still not fully clarified at present.

Previous studies about the perception of nonlinguistic information were always based on the source-filter theory from the viewpoint of speech production. The basic reason is that nonlinguistic information can be thought to be derived from human vocal organs. The contributions of typical acoustic features conveyed in speech, such as F0, spectral envelope, intensity, and speech rates, were investigated. However, it was found that such typical acoustic features have difficulty to account for the human response from cochlear-implant (CI) listeners. A probable reason is that, for CI listeners, the temporal modulation cues provided by the temporal envelope are used as primary cues, however, the typical acoustic features can not represent the features of the temporal envelope well. The temporal modulation cues provided by the temporal envelope are also considered to be important for perceiving nonlinguistic information.

Why the temporal modulation cues provided by temporal envelope of speech should be important and needed to be clarified. At first, from the viewpoint of auditory, temporal envelope plays an important role in human auditory system. The signal processing in peripheral auditory system can be roughly modeled as band-pass filtering (auditory filterbank) and envelope extracting (inner-hair cell model). The sound signal is first divided into several narrow band signals by auditory filterbank. Then the temporal envelope of each band is extracted as the mechanism of inner hair cells. Furthermore, it is suggested that human auditory system carries out a kind of modulation frequency analysis on the temporal envelope that can be modeled as a modulation filterbank. The auditory system should analysis the modulation frequency components at the early stage close to the periphery. Therefore, the temporal modulation cues provided by temporal envelope may contribute the perception of nonlinguistic information.

For speech perception, the temporal envelope has also been proved to be an important cue in the perception of linguistic information. Studies using noise-vocoded speech (NVS) demonstrated that human can perceive linguistic information with using the temporal envelope as a primary cue. NVS can be generated by dividing speech signal into several narrow bands and replacing the carriers in each narrow band with band-limited noise. The spectral cues provided will be poorer and poorer with less number of channels. It is shown that NVS with only four bands is sufficient to achieve good vowel, consonant, and sentence recognition. Furthermore, previous studies also showed that the low modulation frequency components of temporal envelope should contribute to the perception of linguistic information. If the temporal modulation cues are so that important to speech perception, they should also contribute to the perception of nonlinguistic information.

The clarification of the contribution of temporal modulation cues to the perception of speaker individuality and vocal emotion should be important to the development of CI devices. CI devices provide the

temporal envelope information as a primary cue, however, the temporal fine structure information is not effectively encoded. As the poor spectral cue, CI listeners have problem with identifying the speaker or the emotion from only speech. It is necessary to clarify the contribution of temporal modulation cues on the perception of nonlinguistic information to optimize the CI device and improve the performance of speaker and vocal-emotion recognition of CI listeners and also for the clarification of the perceptual process of nonlinguistic information. Furthermore, the clarification of the contribution of temporal modulation cues to the perception of speaker individuality and vocal emotion will also deepen our understanding of the temporal modulation information and speech perception. To clarify the contribution of temporal modulation cues, an analysis method purely based on the modulation frequency analysis mechanism of the auditory system is necessary.

The ultimate research goal of the present study is to clarify the contribution of temporal modulation cues to the perception of speaker individuality and vocal emotion. To reach that goal, at first, the role of temporal envelope and modulation frequency information in speaker and vocal emotion recognition was investigated to confirm whether temporal modulation cues actually contribute to the perception of speaker individuality and vocal emotion. Speaker and vocal-emotion recognition experiments using NVS were carried out to investigate the effects of different temporal and spectral resolutions of NVS on the perception of speaker individuality and vocal-emotion. The spectral and temporal modulation cues will be reduced when the spectral and temporal resolution decrease.

For spectral cue, the speaker distinction performance was not sensitive to the spectral resolution, at least in the limited set of stimuli in the present study. For vocal-emotion recognition, the spectral resolution was important for the recognition of only neutral, joy, and cold anger NVS, but not sadness or hot anger NVS. For temporal modulation cues, the results showed that the recognition rates were significantly decreased with lower upper limit of modulation frequency for both speaker and vocal emotion. In the other word, it was more difficult to recognize the speaker or vocal emotion from NVS if the temporal modulation cues provided were reduced. Therefore, it was confirmed that the temporal modulation cues contribute to the perception of speaker individuality and vocal emotion. Compared to the perception of linguistic information, the temporal modulation cues provided by higher modulation frequency bands are suggested to be important for the perception of speaker individuality and vocal emotion.

At the next step, the relationship between the modulation spectral features and the perceptual data obtained from speaker and vocal-emotion recognition experiments was analyzed to clarify the exactly contribution of temporal modulation cues on the perception of speaker individuality and vocal-emotion. The modulation spectral features were extracted from the modulation spectrogram of speech data. The modulation spectrogram was calculated by the process of auditory filterbank, temporal envelope extraction and modulation filterbank. The correlation between the discriminability index of modulation spectral features and the perceptual data was calculated to demonstrate the relationship between modulation spectral features and the perception of speaker individuality and vocal-emotion.

For speaker individuality, there were positive correlations between the modulation spectral features and the perceptual data of speaker distinction experiment. For vocal emotion, similar results were also obtained, however, the correlations were roughly higher than that of speaker distinction experiments. The results showed that the modulation spectral features were useful to account for the perceptual data of speaker and vocal-emotion recognition experiments using NVS. Therefore, modulation spectral features were suggested to be important cues contribute to the perception of speaker individuality and vocal emotion.

Finally, applications of the temporal modulation information in simulating CI listeners' response and vocal-emotion conversion of NVS were discussed. At first, the feasibility of using NVS to simulate CI listeners' response in vocal emotion recognition was investigated by carried out vocal-emotion recognition experiments using both NVS and original emotional speech with NH and CI listeners. The results

showed that the vocal-emotion recognition paradigm using NVS can be used to investigate vocal emotion recognition by CI listeners. Furthermore, it was also suggested that the modulation spectral features can also be used to account the performance of CI listeners in the vocal-emotion recognition.

Effect of the modification of modulation spectrogram on the vocal-emotion recognition was then investigated. A method based on a linear prediction (LP) scheme was proposed to modify the modulation spectrogram and its features of neutral speech to match that of emotional speech. The temporal envelopes were modulation-filtered by using IIR filters to modify the modulation spectrum from neutral to emotional speech. The IIR filters were derived from the relation of modulation characteristics of neutral and vocal emotions on a LP scheme. On the acoustic frequency domain, the average amplitude of the temporal envelope was corrected using the ratio of the average amplitude between neutral and emotional speech. Finally, a vocal-emotion recognition experiment using NVS generated by the converted temporal envelope was carried out. The results showed that the modulation spectrogram of neutral speech can be successfully converted to that of emotional speech by the proposed method. The results of the evaluation experiment confirmed the feasibility of vocal emotion conversion on the modulation spectrogram for NVS.

In conclusion, the fact that the temporal modulation cues contribute to the perception of speaker individuality and vocal emotion was confirmed by the speaker and vocal-emotion recognition experiments using NVS. Furthermore, the investigation of modulation spectral features demonstrated that there were high correlations between modulation spectral features and the perceptual data obtained from speaker and vocal-emotion recognition experiments. Therefore, the modulation spectral features could be important cues contribute to the speaker and vocal-emotion recognition with NVS. These results further proved that the temporal modulation cues play an important role in the perception speaker individuality and vocal-emotion.

Keywords: Speech perception, speaker individuality, vocal emotion, temporal cue, noise-vocoded speech, modulation spectral feature

Acknowledgments

First and foremost, I would like to express my deepest appreciation to my supervisor, Professor UNOKI Masashi, for his tremendous guidance and support. I feel very lucky to have him to be the supervisor of my master and PhD study. Without his invaluable guidance, motivation, and great care, my research work would never be successful. Studying with him over these years was a wonderful experience that I would never forget.

I would like to acknowledge my vice supervisor, Professor AKAGI Masato, for his invaluable suggestions and comments in my research. The discussions with Professor AKAGI make my research progress continuously. My sincerest gratitude goes to Professor DANG Jianwu, for his may insightful discussions and invaluable guidance for this thesis and my research like.

I would also like to express my sincere thanks to Doctor FURUKAWA Shigeto, for his invaluable guidance and many insightful discussions that directly contributed to this thesis. I would like to acknowledge Professor ARAKI Yukiko, for her great suggestions an comments for the thesis. I would like to acknowledge Professor KITAMURA Tatsuya, for his valuable data of perceptual speaker similarity.

Most importantly, I am forever indebted to my mother and father. I am profoundly indebted to my wife, CHENG Yu, for her unconditional support and encouragement. I dedicate this thesis to my parents and my wife, for their unfailing love, support and patience.

Table of Contents

List of Figures

List of Tables

Acronym and Abbreviation

CI	Cochlear Implant
NH	Normal Hearing
NVS	Noise–Vocoded Speech
ERB_N	Equivalent Rectangular Bandwidth
MSF	Modulation Spectral Feature
MSCR	Modulation Spectral Centroid
MSSP	Modulation Spectral Spread
MSSK	Modulation Spectral Skewness
MSKT	Modulation Spectral Kurtosis
MSFT	Modulation Spectral Flatness
MSTL	Modulation Spectral Tilt
LP	Linear Prediction
CC	Correlation Coefficient

Chapter 1

General introduction

1.1 Introduction

The biggest difference between human and any other animal is that human can communicate through speech. One famous test to judge artificial intelligence's (AI) ability to exhibit intelligent behavior equivalent to human is so called Turing test. Turing thought that a perfect AI should be indistinguishable from a human in natural language conversation with text. Furthermore, I think the final goal of AI should be the ability of speech communication, because speech carries much more information than the language message. For this goal, we must understand how human can perceive the various information contained in speech at first.

From the viewpoint of information generation, Fujisaki divided the information contained in speech signal with 3 different categories [2]. Those are linguistic information, paralinguistic information and non-linguistic information.

- **Linguistic information**: the symbolic information that is represented by a set of discrete symbols and rules for their combination.
- **Paralinguistic information**: the information that is not inferable from the written counterpart but is deliberately added by the speaker to modify or supplement the linguistic information.
- **Nonlinguistic information**: the other information such like factors as the age, gender, idiosyncrasy, physical and emotional states of the speaker, etc.

As Fujisaki's definition, the linguistic information is the language message that the speaker wants to convey. The paralinguistic information contains the intentions, attitudes, and speaking styles of the speaker which should be under the control of the speaker and attached to the language message. The two largest categories in nonlinguistic information should be **speaker individuality** and **vocal emotion**. The speaker individuality information can be defined as the information that is used by listener to distinguish such a specific speaker from other people. The vocal emotion information is defined as the information that is used by human to identify the emotion state of speaker from speech. Speaker individuality and vocal emotion play important roles in the speech communication of our daily life. However, the perception process of speaker individuality and vocal emotion is still not fully clarified at present.

Previous studies about the perception of nonlinguistic information were based on the source-filter theory from the viewpoint of speech production. Obviously, the basic reason is that nonlinguistic information is thought to be derived from human vocal organs. For speaker individuality, the F0 contour, spectral envelope, and the formants of speech have been proved to contribute speaker recognition [3–6]. For vocal emotion, previous works also focused on the acoustic features conveyed in speech, such as F0, spectral envelope, intensity, and speech rate [7–9]. For both speaker individuality and vocal emotion, the time-averaged acoustic features were investigated sufficiently. However, the temporal modulation cues provided by the dynamic components of speech are also considered to be important for perceiving nonlinguistic information.

From the viewpoint of auditory perception, the temporal modulation cues provided by the temporal envelope is very important. The signal processing in peripheral auditory system can be roughly modeled as band-pass filtering (auditory filterbank) and envelope extracting (inner-hair cell model) [10, 11]. The sound signal is first divided into several narrow band signals by auditory filterbank. Then the temporal envelope of each band is extracted as the mechanism of inner hair cells. Furthermore, it is suggested that human auditory system carries out a kind of modulation frequency analysis on the temporal envelope that can be modeled as a modulation filterbank [12]. In this study, the temporal modulation cues are defined as the cues provided by the modulation frequency components of temporal envelope.

The auditory system should analysis the modulation frequency components at the early stage close to the periphery. Therefore, the temporal modulation cues provided by temporal envelope may contribute the perception of nonlinguistic information. To clarify the perceptual process of nonlinguistic information, it is necessary to clarify the contribution of temporal cues. The importance to clarify the contribution of temporal modulation cues on the perception of nonlinguistic information will be described in the next section.

1.2 Motivation

1.2.1 The contribution of temporal cue to speech perception

The importance of temporal modulation cues in the perception of linguistic information has been studied by many researchers. The temporal envelope of speech has been proved to be an important cue for speech perception from the studies using noise-vocoded speech (NVS) [13–16]. NVS is generated by replacing the temporal fine structure of the sub-band of speech with band-limited noise, so the spectral cue is reduced dramatically and the temporal modulation cues are preserved [17]. Shannon *et al.* showed that NVS with only four bands is sufficient to achieve good vowel, consonant, and sentence recognition [13]. Therefore, human can successfully perceive linguistic information using the temporal envelope of speech as a primary cue.

Drullman *et al.* investigated the important modulation frequency bands for speech perception by low- and high-pass filtering the temporal envelope of speech [18, 19]. They showed that the modulation frequency bands from 4 to 16 Hz contained important cues related to linguistic information. Xu *et al.* attempted to elucidate the importance of temporal envelope for phoneme recognition using NVS [20]. The spectral resolution was manipulated by varying the number of channels of NVS and the temporal resolution was manipulated by varying the lowpass cutoff frequencies used to extract the temporal envelope. The results showed that vowel recognition plateaued at the 4 Hz upper limit of modulation frequency. Tachibana *et al.* used a similar experimental paradigm and demonstrated that the modulation frequency components below 8 Hz is important for sentence recognition [14].

Rosen developed a framework for describing the acoustic structure of speech based on temporal aspects [21]. From Rosen's viewpoint, speech signal can be comprised of three main temporal features: envelope, periodicity, and fine-structure. The envelope cues contain the modulation frequency band between about 2 and 50 Hz. This low modulation frequency band should include the information about variations of intensity, duration, attach, decay, and segmental cues of speech. From the previous studies described above, it has been proved that such temporal modulation cues are very important for the perception of linguistic information. Therefore, temporal modulation cues also have potential to be important cues in the perception of nonlinguistic information. To understand the perceptual process in human auditory system, the contribution of temporal modulation cues in the perception of nonlinguistic information must be clarified.

1.2.2 Perception of nonlinguistic information by cochlear implant listeners

As CI listeners using the temporal envelope as a primary cue, it is very important to clarify the contributions of temporal modulation cues on the perception of nonlinguistic information. CI system mimic the signal processing of the auditory peripheral system with four main steps: bandpass filterbank, envelope extraction, amplitude compression, and impulse signal generation [22]. As the number of channels of the bandpass filterbank in CI system is so limited, CI device can only provide poor spectral cue. CI devices provide the temporal envelope as a primary cue, and the temporal fine structure information is not effectively encoded. NVS was usually used as a CI simulation with normal-hearing (NH) listeners to predict the response of CI listeners. As the poor spectral cue, CI listeners have problem with identifying the speaker or the emotion of speaker from only speech.

Vongphoe and Zeng evaluated whether the temporal envelope is sufficient to support both speech recognition and speaker recognition with NVS [23]. Their results showed a disassociation between speech and speaker recognition when primarily temporal cues are used: participants recognized vowels accurately but speakers poorly. Gonzalez and Oliver investigated speaker recognition as a function of the number of channels in both noise and sinewave-vocoded speech as CI simulations [24]. The performance of speaker recognition was shown to be poorer with fewer number of channels. However, Krull *et al.*

showed that training results in improved recognition rates of speaker in CI simulations [25]. Moreover, child CI listeners succeeded in differentiating their mothers' utterances from those of other people [26]. CI listeners' differentiation of speakers was facilitated by long-term familiarity, which suggested that temporal modulation cues have the potential to effectively support CI listeners to distinguish speakers.

It has also been known that CI listeners' performances of vocal-emotion recognition are poorer than NH listeners, as the poor spectral cues provided by CI device [23, 24, 27, 28]. Luo *et al.* showed that vocal-emotion recognition of NH listeners using NVS was significantly improved as the cut-off frequency of modulation low-pass filter was increased from 50 to 500 Hz [28]. The modulation frequency bands between 50 and 500 Hz mainly included the periodic information related to F0 [21]. However, the contribution of the temporal modulation cues defined as the modulation frequency band below 50 Hz is still unknown. By comparing the performances of vocal-emotion recognition by CI listeners and HN listeners using NVS, Chatterjee *et al.* [27] found that the mean performance of CI listeners was similar to that of NH listeners with 8-band NVS. They then analyzed the F0, intensity, and duration of stimuli. However, it was found that the acoustic analyses could not account for all of the perceptual data of the vocal-emotion recognition experiment with NVS.

As the CI device provide the temporal envelope as a primary cue, it is necessary to clarify the contribution of temporal cues in the perception of nonlinguistic information to optimize the CI device and improve the performance of speaker and vocal-emotion recognition of CI listeners. It seems like it is difficult to account for the perceptual data of experiments using NVS with traditional acoustic features such like F0, duration, and intensity. Acoustic analysis based purely on the temporal features may be useful, because the temporal envelope is used as a primary cue in the perception of NVS.

1.3 Research goal

Previous studies related to human auditory system showed that the temporal modulation information plays an important role. The auditory system should carry out an modulation frequency analysis at the early stage close to the periphery. The temporal modulation cues were suggested to contribute the perception of various information from sound wave.

From the previous studies about speech perception, it has been proved that temporal modulation cues are important for the perception of linguistic information. Therefore, temporal modulation cues also have potential to be important cues in the perception of nonlinguistic information. As the CI device provide the temporal envelope as a primary cue, discussing the contribution of temporal modulation cues in the perception of nonlinguistic information is also important for optimizing the CI device and improving the performance of speaker and vocal-emotion recognition of CI listeners.

The ultimate research goal of the present study is to clarify the contribution of temporal modulation cue to the perception of speaker individuality and vocal-emotion. To reach that goal, at first, the role of temporal envelope and modulation frequency components in the perception of speaker individuality and vocal-emotion is investigated to confirm that whether the temporal modulation cues contribute speaker and vocal-emotion recognition.. Then, to clarify the exact contribution of temporal modulation cues in the perception of speaker individuality and vocal-emotion, the modulation spectral features of speech are analyzed to account for the perceptual data obtained from the speaker and vocal-emotion recognition experiments using NVS.

In addition, applications of the temporal modulation information in simulating CI listeners' response and vocal-emotion conversion of NVS are discussed. In this work, the contribution of temporal modulation cues on the perception of speaker individuality and vocal-emotion are discussed together to indicate the difference in the perception of linguistic information and nonlinguistic information and to clarify the common roles of temporal modulation cue in the perception of nonlinguistic information.

1.4 Outline of thesis

The rest of this dissertation consists of five chapters and is organized as follows. Figure 1.1 shows the organization of this dissertation.

Chapter 2 introduces the background knowledge and previous studies about nonlinguistic information and the perception of temporal modulation components At first, the previous studies about the perception of speaker individuality and vocal emotion based on the acoustical features of speech are reviewed. Then the previous studies about the contribution of temporal modulation cues in human auditory system and the perception

of linguistic information are reviewed to expound the importance of temporal modulation in both psychoacoustic and speech perception.

Chapter 3 purpose to clarify the role of temporal modulation cue in speaker and vocal-emotion recognition. Speaker and vocal-emotion recognition experiments are carried out to confirm whether the temporal modulation cues provided by the temporal envelope of speech contribute to the perception of speaker individuality and vocal emotion. In the experiments, speaker distinction and vocal emotion recognition are conducted by NH listeners under different upper limit of modulation frequency and the number of channels of NVS stimuli. The spectral and temporal modulation cues will be further reduced when the number of channels and upper limit of modulation frequency decrease, respectively. The experimental paradigm used in this experiment can also clarify the important modulation frequency bands for speaker and vocal-emotion recognition.

Chapter 4 describe the relationship between modulation spectral features and the perceptual data obtained in the speaker and vocal-emotion recognition experiments to clarify the exact contribution of temporal modulation cues on the perception of speaker individuality and vocal-emotion. An auditory-based method is used to calculate the modulation spectrogram of speech and the modulation spectral features are extracted from the modulation spectrogram. The correlation between the modulation spectral features and the perceptual data is then analyzed to discuss whether the modulation spectral features will contribute to speaker or vocal emotion recognition. In order to investigate the relationship between modulation spectral features and the perceptual data of speaker and vocal-emotion experiments, an discriminability index d' is used. The d' of each modulation spectral feature present the physical distance of the distributions of modulation spectral feature with different speaker or vocal-emotion and the d' of the perceptual data present the psychological distance of different speaker or vocal-emotion. The correlation between the d' values of modulation spectral features and the perceptual data is calculated to demonstrate the relationship between modulation spectral features and the perception of speaker individuality and vocal-emotion.

Chapter 5 discusses the applications of the temporal modulation information in simulating CI listeners' response and vocal-emotion conversion of NVS. The feasibility of using NVS to simulate CI listener's response in vocal emotion recognition is discussed by

carried out vocal-emotion recognition experiments using both NVS and original emotional speech with NH and CI listeners. Effect of the modification of modulation spectrogram on the vocal-emotion recognition is then investigated. A method based on a linear prediction (LP) scheme is proposed to modify the modulation spectrogram and its features of neutral speech to match that of emotional speech.

Chapter 6 summarizes this study and emphasizes its contributions to this research field as well as other research fields. Furthermore, future works about deepening the understanding of the perceptual process of speech, development of CI device and other research field are introduced.

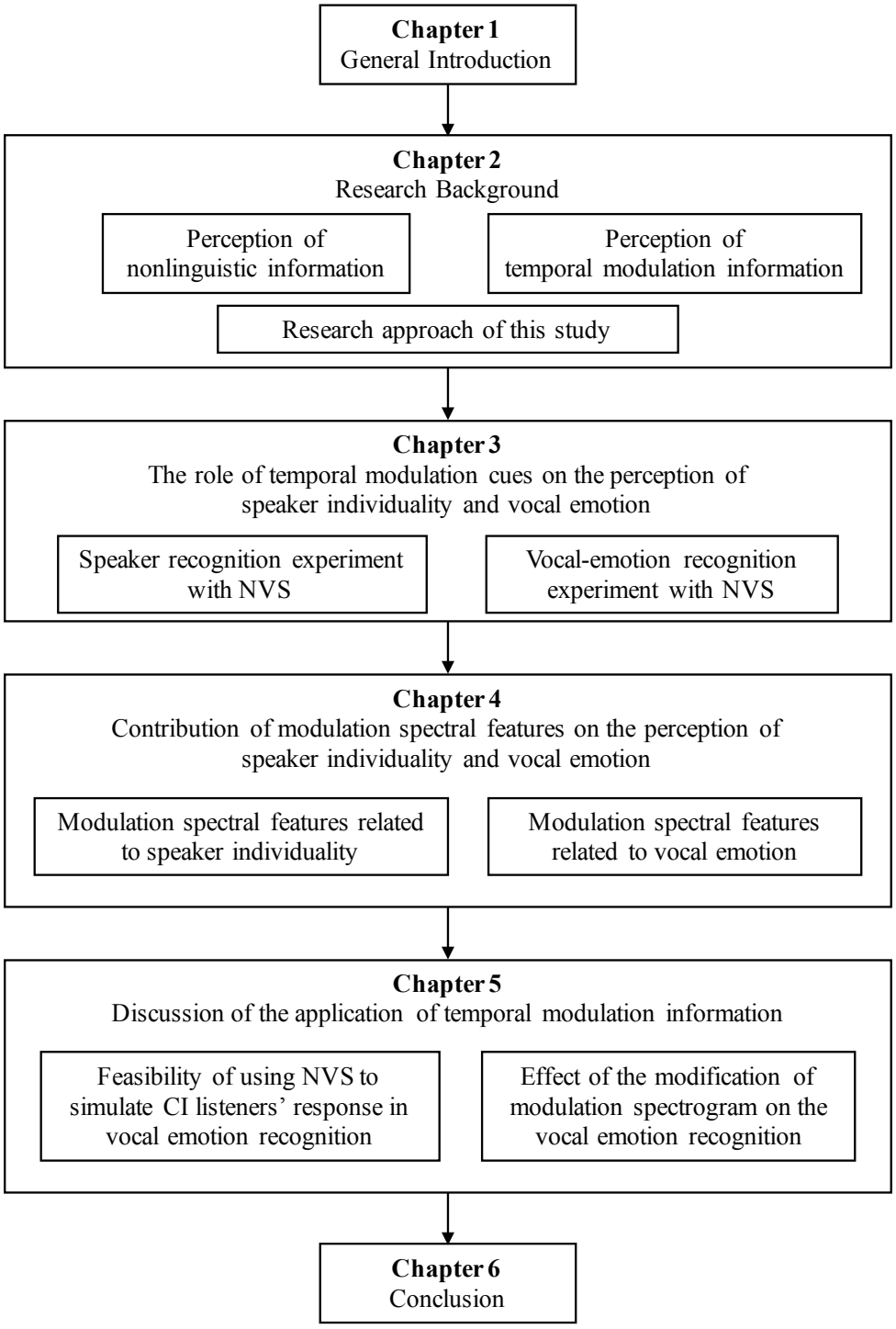


Figure 1.1: Organization of this dissertation.

Chapter 2

Research background

2.1 Introduction

In this chapter, the background knowledge and previous studies about nonlinguistic information and the perception of temporal envelope are introduced. At first, the previous studies of speaker individuality and vocal emotion and the approach they used are reviewed. Then the previous studies about the importance of temporal envelope in human auditory system and the perception of linguistic information are reviewed to expound the contribution of temporal modulation cues in both psychoacoustic and speech perception. Finally, the background knowledge and previous studies are summarized and the research approach used in this study is explained.

2.2 The perception of nonlinguistic information

2.2.1 Speaker individuality

Speaker individuality can be divided into two categories that are inherent and acquired features. The inherent features are inborn characteristics derived from the individual difference of vocal organs (vocal fold and vocal tract). The acquired features are postnatal characteristics derived from the individual difference of the speaking style of speaker. In this study, the speaker individuality is defined as the physical features that is used by human to distinguish such a specific speaker from any other person. Therefore, only the physical features that contribute to the perception of speaker individuality is discussed

and both inherent and acquired features are included in based on this definition.

Previous studies about the perception of speaker individuality almost based on the source-filter theory [29] from the concept of speech production. The source-filter theory consider the vocal fold as a sound source and the vocal tract as a kind of linear filter. Based on this theory, previous studies investigated the features related to vocal fold source (F0, etc.) and vocal tract filter (spectral envelope, formant, etc.) related to speaker individuality.

Numerous acoustic features related to the perception of speaker individuality were investigated so far. Ito *et al.* reported that the acoustic parameters affecting the perception of speaker individuality are important in the order of spectral envelope, F0, and the dynamic features of speaking style (tempo, etc.) [30], and the features about spectral envelope are especially important. Hashimoto *et al.* analysis the contributions of acoustic features (F0, spectrum, and duration) affecting speaker identification quantitatively with hearing experiments [31]. They found that the spectral envelope and F0 have remarkable contributions, and it is also reported that the degree of contribution depends on the difference of acoustic feature between different speakers. Kasuya *et al.* then investigated the contribution of static and dynamic features of vocal tract to speaker identification based on the ARX model [32]. It was reported that the contribution of static features is larger than that of dynamic features. Related to this result, Kitamura *et al.* reported that the spectral trajectory patterns do not affect speaker identification remarkably [33].

Some studies focus attention on one particular acoustic feature of vocal tract or vocal ford related to speaker identification. Kuwabara *et al.* investigated the role of formant frequencies and bandwidths in the perception of speaker individuality [34]. They found that the frequency shift of the formants below F3 affected the perception of speaker individuality and the F3 is the most important feature. Kitamura *et al.* focused on the effect of spectral envelope especially [5, 35, 36]. As a result, the spectral envelope components above 1740 Hz carried more speaker individuality information and the components on lower frequency band seem mostly relate to the linguistic information. Moreover, it was suggested that in such high frequency band, the peaks are more important than the dips to speaker identification. From the knowledge of physiology, Kitamura *et al.* then found that such speaker individuality information appeared in the high frequency band

is derived from the shape of hypopharyngeal cavity that does not have much movement during speaking [6]. Aimino *et al.* showed that the nasal sound is effective for speaker identification based on the individuality differences in the physiological characteristics of articulatory organs such as nasal cavity and nasopharynx [37]. Akagi and Ienaga found that the contour of F0 contributes to the perception of speaker individuality [3].

2.2.2 Vocal emotion

Based on the expression of vocal-emotion, emotional speech can be divided into two categories: spontaneous emotional speech and acting emotional speech. The spontaneous emotional speech based on the emotional state of speaker which can not be controlled be speaker. The acting emotional speech based on the purpose of expressing the emotion with speakers' controlling. The mechanism of speech production of spontaneous and acting emotional speech may be different [38, 39]. In this study, only acting emotional speech data are used for convenience.

For the perception of vocal emotion, previous works focused on the acoustic features conveyed in speech, such as F0, spectral envelope, intensity, and speech rate and using such acoustic features to modeling the perception of vocal-emotion [7, 40]. Scherer *et al.* presented emotion speech stimuli (14 kinds of emotion) to NH listeners to label the emotion of each stimulus [8]. At the same time, they also extracted 29 different acoustical features (F0, intensity, speaking rate, duration, time averaged spectrum, etc.) of each emotion speech signal. An emotion classification model was then constructed using multiple regression analysis analyzed the contribution of each acoustical features. The emotion recognition rates of human response and regression model were 48% and 40% and the trend of confusion of human response and regression model was very similar. Therefore, it is suggested that humans are also using such kinds of acoustical features to recognize the vocal-emotion of speech.

Recently, researchers focus on the structure of model. Huang and Akagi proposed a three-layered model with semantic primitives as a middle layer between vocal emotion and acoustic features [9]. The bottom layer was acoustic features and fuzzy inference system was used to built the model. The three-layered model was found to be useful in vocal-emotion recognition system [41] and vocal-emotion conversion system [42, 43]. However,

the acoustic features used in these studies were still based on source-filter model and the waveform of speech.

In summary, previous studies usually use the acoustical features based on source-filter model to investigate the perception of vocal emotion and speaker individuality. However, in the study of the perception of nonlinguistic information by CI listeners, it was found that the traditional acoustical features did not work well to account for the perceptual data [27]. As the CI device provided poor spectral cues, the traditional acoustical features can be perceived by CI listeners. On the other hand, the CI device provided temporal modulation cues as primary cues and the temporal modulation cues also play an important role in the human auditory system. For clarify the perception process of nonlinguistic information, a method based purely on the auditory perception is necessary.

2.3 The perception of temporal envelope

2.3.1 Human auditory system and modulation filterbank

This work focus on the contribution of temporal modulation cue to the perception of speaker individuality and vocal-emotion. The temporal modulation cue plays an important role in human peripheral auditory system [44]. The peripheral auditory system is composed of the outer, middle, and inner ear. The outer ear collects the sound wave and the middle convert the sound wave from air vibration to liquid vibration in cochlea effectively.

The cochlea works as an frequency analyzer and transforms the sound wave into neural signal. A membrane called the basilar membrane runs along the length of the cochlea. Sound waves produces traveling waves along the basilar membrane. The basilar membrane works as a filter bank (auditory filterbank), splitting the complex sound wave into several frequencies. Gammatone filterbank is widely used as a model of auditory filterbank in the auditory system [45]. Furthermore, a gammachirp filterbank was proposed to simulate the auditory filterbank [46]. The Equivalent rectangular bandwidth (ERB_N) was always used to measure the bandwidth of auditory filterbank [47] Glasberg and Moore then proposed a frequency scale called the ERB_N-number scale based on ERB_N). ERB_N-number is conceptually similar to the Bark scale proposed by Zwicker and Terhardt [48] and the

mel scale of pitch. The ERB_N-number scale and Bark scale are all based on peripheral auditory system. The Bark scale is based on the critical bandwidth, however, the ERB_N-number scale is based on the auditory filter shape measure by nothched-noise which should can simulate the frequency analysis function of auditory system better.

The movement of the basilar membrane causes a displacement of the inner hair cells and the inner hair cells transform such movement into neutral signal. As the mechanism of inner hair cells, such function can be modeled as a process of envelope extraction and amplitude compression. Therefore, the signal process in peripheral auditory system can be computationally modeled as a bandpass filterbank, envelope extraction and amplitude compression [10, 11]. The CI device basically use such signal process to simulate human peripheral auditory system [22,49]. Dau *et al.* proposed a computational model of human auditory signal processing and perception using modulation filterbank after the process of envelope extraction [50,51]. The results showed that such model works better than the previous models without modulation filterbank. Recently the modulation filterbank was widely used in the speech intelligibility predictor and auditory system modeling [52–54] There both physiological [55] and psychology [12] evidence suggested the existence of modulation filterbank in auditory system. Auditory system may also have a modulation frequency analyzer to analysis the modulation frequency components of the envelope of speech. Therefore, from the view point of auditory perception, the temporal modulation cues should play an important role in the perception of various information from speech.

2.3.2 Contribution of temporal modulation cues on the perception of linguistic information

For speech perception, previous studies have proved that the temporal modulation cues are important for the perception of linguistic information. Shannon *et al.* showed that NH listeners can recognize the linguistic information of NVS using the temporal modulation cues as primary cues with poor spectral cues [13]. NVS can be generated by dividing speech signal into several narrow bands and replacing the carriers in each narrow band with band-limited noise. The spectral cues provided will be poorer and poorer with less number of channels. Therefore, NVS was also usually used as a CI simulation in the studies of the speech perception by CI listeners to simulate the poor spectral cues provided

by CI device [25, 27, 56–58]. It is shown that NVS with only four bands is sufficient to achieve good vowel, consonant, and sentence recognition [15,59]. Many other studies using NVS showed that importance of temporal modulation cues in speech perception [60–64] Therefore, human can perceive the linguistic information with the temporal envelope as a primary cue.

For the contributions of modulation frequency components, Rosen firstly developed a framework for describing the acoustic structure of speech based on temporal aspects [21]. The envelope cues contain the modulation frequency band between about 2 and 50 Hz including the information about variations of intensity, duration, attach, decay, and segmental cues of speech. Drullman *et al.* investigated the important modulation frequency bands for speech perception by low- and high-pass filtering on the temporal envelope [18, 19]. They measured the speech-reception threshold for sentences in noise with reduce the high or low modulation frequency components of speech. The results showed that the modulation frequency bands from 4 to 16 Hz contained important cues related to linguistic information. Studies using NVS with similar experimental method also showed such modulation frequency bands are important for speech perception [14,20].

2.4 The research approach of this study

In this chapter, the previous studies of the perception of speaker individuality and vocal emotion were reviewed at first. The previous studies of nonlinguistic information almost focused on the acoustical features of speech based on the concept speech production such like F0, spectral envelope, etc. However, in the study of the perception of nonlinguistic information by CI listeners, it was found that the traditional acoustical features did not work well to account for the perceptual data [27]. The traditional acoustical features based on speech production have difficult to present the temporal modulation cues of speech.

Why should we clarify the contribution of temporal modulation cues on the perception of speaker individuality and vocal emotion? One important reason is that the temporal envelope plays an important role in auditory system. The auditory system should analysis the modulation frequency components at the early stage close to the periphery. Furthermore, some essential studies showed the importance of temporal modulation cues in the

perception of linguistic information. Therefore, the temporal modulation cues provided by temporal envelope should also contribute the perception of nonlinguistic information. For clarify the perceptual process of nonlinguistic information, a method based purely on the auditory perception is necessary.

To clarify the contribution of temporal modulation cues, it must be confirmed that the temporal modulation information actually contribute to the perception of speaker individuality and vocal emotion. Shannon *et al.* showed that the temporal envelope contribute to the perception of linguistic information [13]. Drullman *et al.* investigated the important modulation frequency bands for speech perception by low- and high-pass filtering on the temporal envelope [18, 19]. These important studies demonstrated us the temporal modulation cues provided by the temporal envelope of speech contribute to the perception of linguistic information. In this study, the approaches of using NVS and low-pass filtering the temporal envelope are combined to confirm the contribution of temporal modulation cues on the perception of speaker individuality and vocal emotion. Speaker and vocal-emotion recognition experiments using NVS are carried out with low-pass filtering the temporal envelope and varying the number of channels.

Furthermore, it is also necessary to clarify the exact features of temporal modulation information related to the perception of speaker individuality and vocal-emotion. Modulation frequency analysis has been shown to be useful for many research fields such as auditory physiology, psychoacoustics, speech perception, and signal analysis and synthesis [65–68]. The modulation spectral features were also widely used in speech technologies about nonlinguistic information such as speaker recognition [69–74] and vocal emotion recognition [75–77]. The fact that temporal modulation information is important to speech perception and can be used in speaker or vocal-emotion recognition system showed that the temporal modulation cues ought to play an important role in the perception nonlinguistic information. The modulation spectral features must be more useful to account for the perceptual data of speaker or vocal-emotion recognition experiments than the traditional acoustic features based on speech production. Therefore, in this study, the relationship between modulation spectral features and the perceptual data obtained in the speaker and vocal-emotion recognition experiments is analyzed to clarify the exact contribution of temporal modulation cues on the perception of nonlinguistic information.

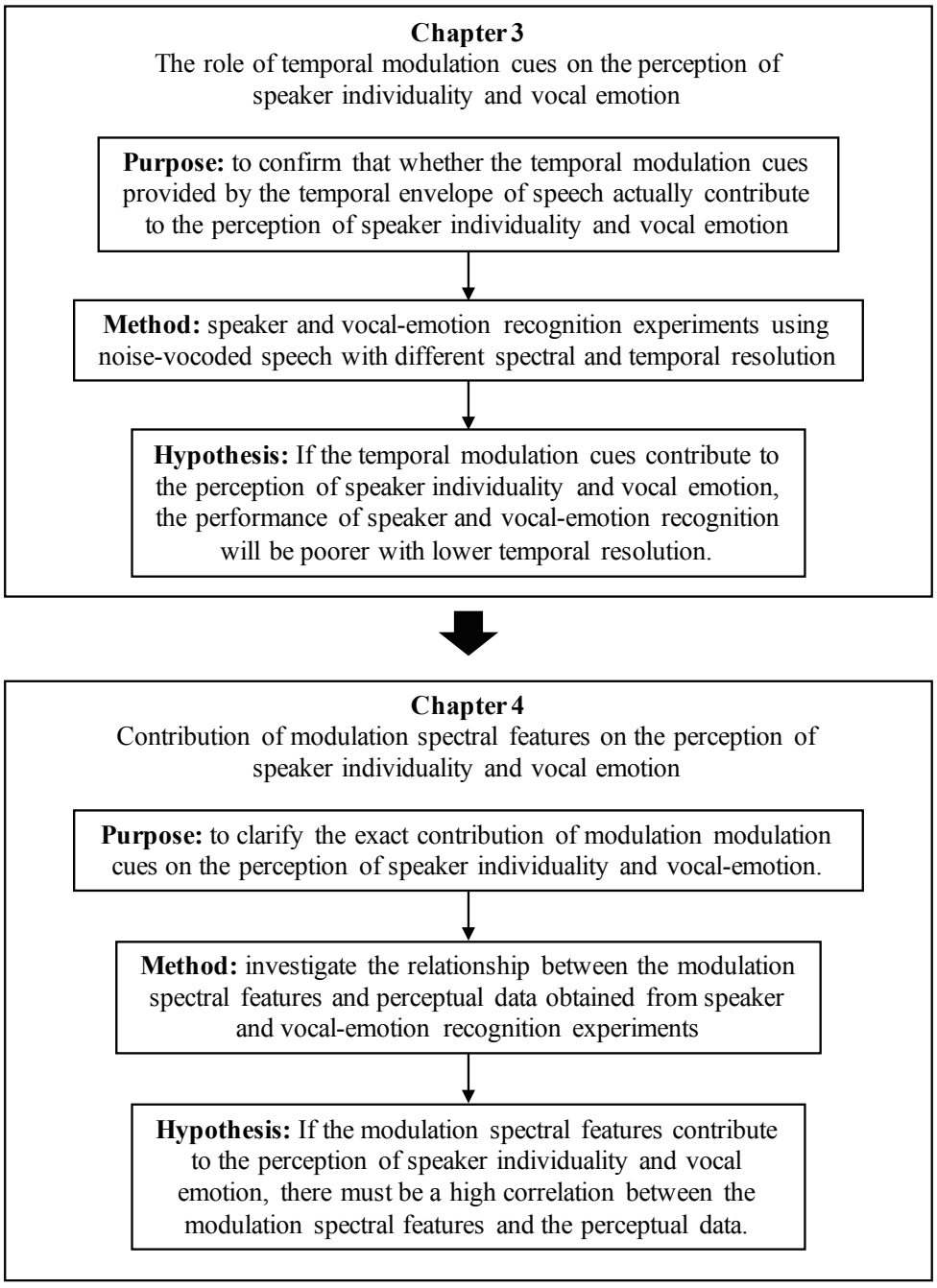


Figure 2.1: The research approach of this study.

Chapter 3

The role of temporal modulation cues on the perception of speaker individuality and vocal emotion

3.1 Introduction

This chapter aims to clarify the role of temporal modulation cues in speaker and vocal-emotion recognition using NVS and confirm that whether the temporal modulation information actually contribute to the perception of speaker individuality and vocal emotion. Furthermore, the effects of different spectral resolutions are also investigated. In the experiment, speaker distinction and vocal emotion recognition are conducted by NH listeners under different upper limit of modulation frequency (0, 0.5, 1, 2, 4, 8, 16, 32, and 64 Hz) of NVS. The temporal modulation cues provided by NVS will be poorer with lower upper limit of modulation frequency. In addition, the role of temporal modulation cues in the different spectral resolutions condition are also investigated by varying the number of channels (4, 8, and 16). The spectral and temporal modulation cues are reduced further when the number of channels and upper limit of modulation frequency decrease, respectively. If the temporal modulation cues contribute to the perception of nonlinguistic information, the performance of speaker or vocal-emotion recognition will be poorer with lower temporal resolution of NVS. Therefore, this experimental paradigm can also clarify the important modulation frequency bands for speaker and vocal-emotion recognition.

3.2 Signal Processing: Noise-Vocoded Speech

Figure 3.1 illustrates the schematic diagram of the signal processing to generate NVS. First, to reduce the effect of the average intensity, the active speech levels of all speech signals were normalized to −26 dBov by using the P.56 speech voltmeter [78]. The speech signal $s(n)$ was then divided into several frequency bands by using a band-pass filterbank, as following,

$$s(k,n) = s(n) * h_{BPF}(k,n) \tag{3.1}$$

where n is time, $h_BPF(k,n)$ is the transform function of the band-pass filter in channel k, and $s(k,n)$ is the sub-band signal in channel k. At the same, the Gaussian white noise with the same length of the speech signal was also divided into several frequency bands using the same band-pass filterbank, as following,

$$WN(k,n) = WN(n) * h_{BPF}(k,n) \tag{3.2}$$

where $WN(n)$ is Gaussian white noise and $WN(k,n)$ is the band-limited noise in channel k.

The bandwidth and boundary frequencies of the band-pass filters (6th-order Butterworth infinite impulse response (IIR) filter) were defined using ERB_N (Equivalent Rectangular Bandwidth) and ERB_N-number scale [44]. The ERB_N-number scale is comparable to a scale of distance along the basilar membrane, so the frequency resolution of the auditory system can be faithfully replicated by dividing frequency bands in accordance with the ERB_N-number. The relationship between ERB_N-number and acoustic frequency is defined as follows:

$$\mathrm{ERB_N - number} = 21.4\log_{10}\left(\frac{4.37f}{1000} + 1\right) \tag{3.3}$$

where f is acoustic frequency in Hz. The boundary frequencies of the band-pass filters were defined from 3 to 35 ERB_N-number with bandwidth as 2, 4, or 8 ERB_N. Therefore, the numbers of channels of the band-pass filterbank were 16, 8, or 4. The number of channels determines the frequency resolution of NVS: higher frequency resolution will be obtained with more channels. Table 3.1 shows the boundary frequencies of the 4-, 8-, and 16-band band-pass filters in Hz and ERB_N-number. Figure 3.2, 3.3, and 3.4 shows the frequency response of the ERB_N-number based band-pass filterbank.

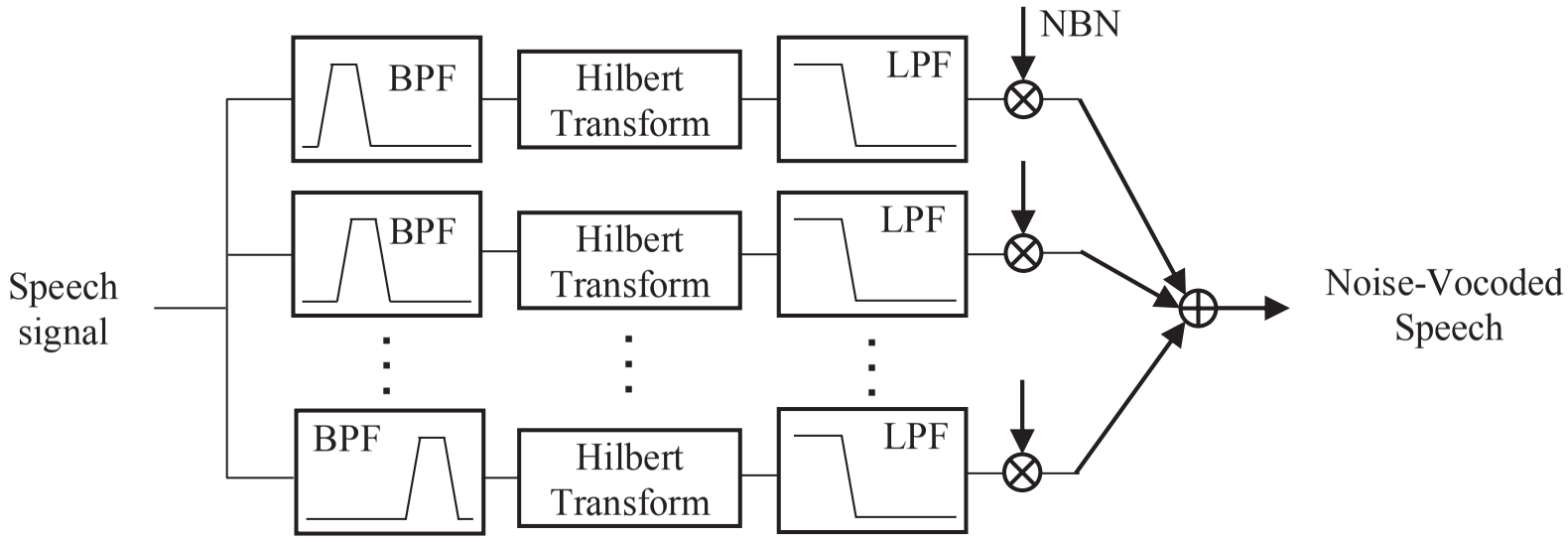


Figure 3.1: Schematic diagram of noise-vocoder method used to generate stimuli (BPF: band-pass filter; LPF: low-pass filter; and NBN: narrow-band noise).

Table 3.1: The boundary frequencies of the band-pass filters in Hz and ERB_N-number.

4-band	8-band	16-band	ERB_N-number	Hz
1	1	1	3	87.18
		2	5	163.1
		3	7	257.2
	2	4	9	373.8
		5	11	518.5
	3	6	13	698.0
2		7	15	920.5
	4	8	17	1197
		9	19	1539
	5	10	21	1963
3		11	23	2489
	6	12	25	3142
		13	27	3951
	7	14	29	4955
4		15	31	6200
	8	16	33	7743
			35	9657

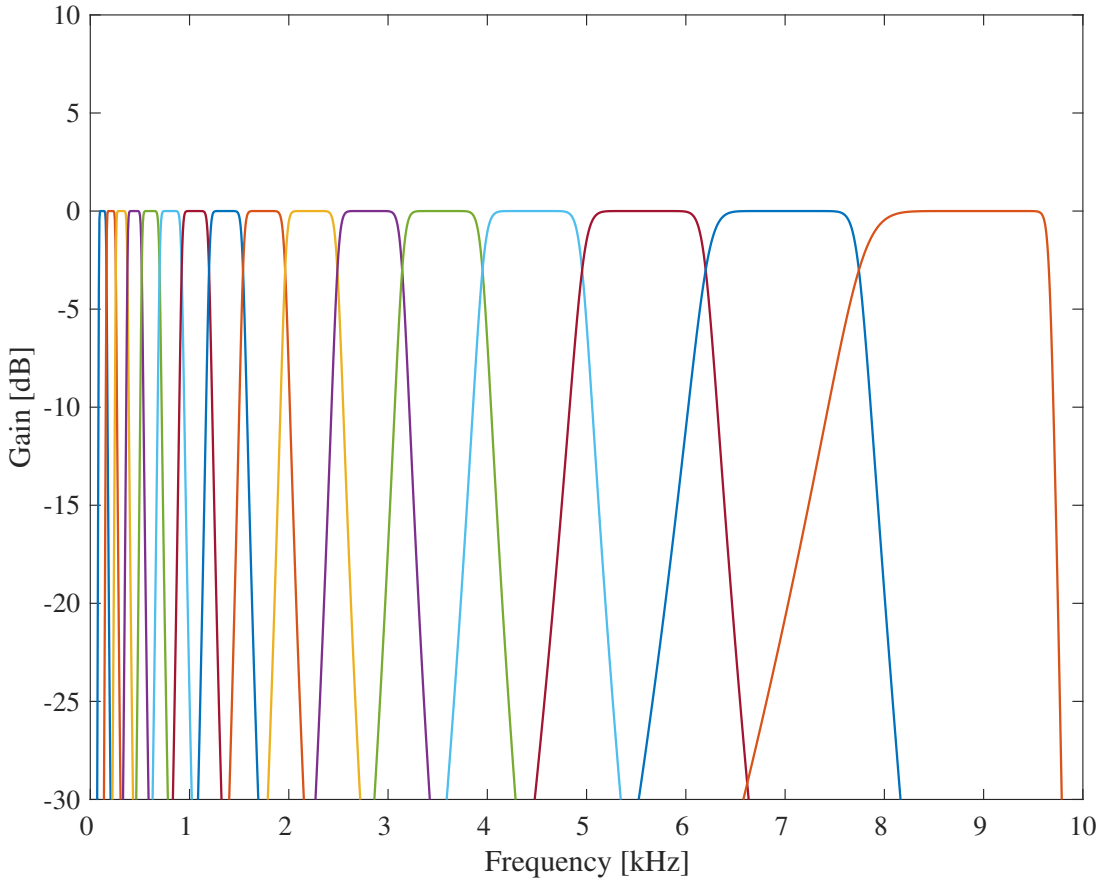


Figure 3.2: Frequency response of the ERB_N-number based 16-band band-pass filterbank.

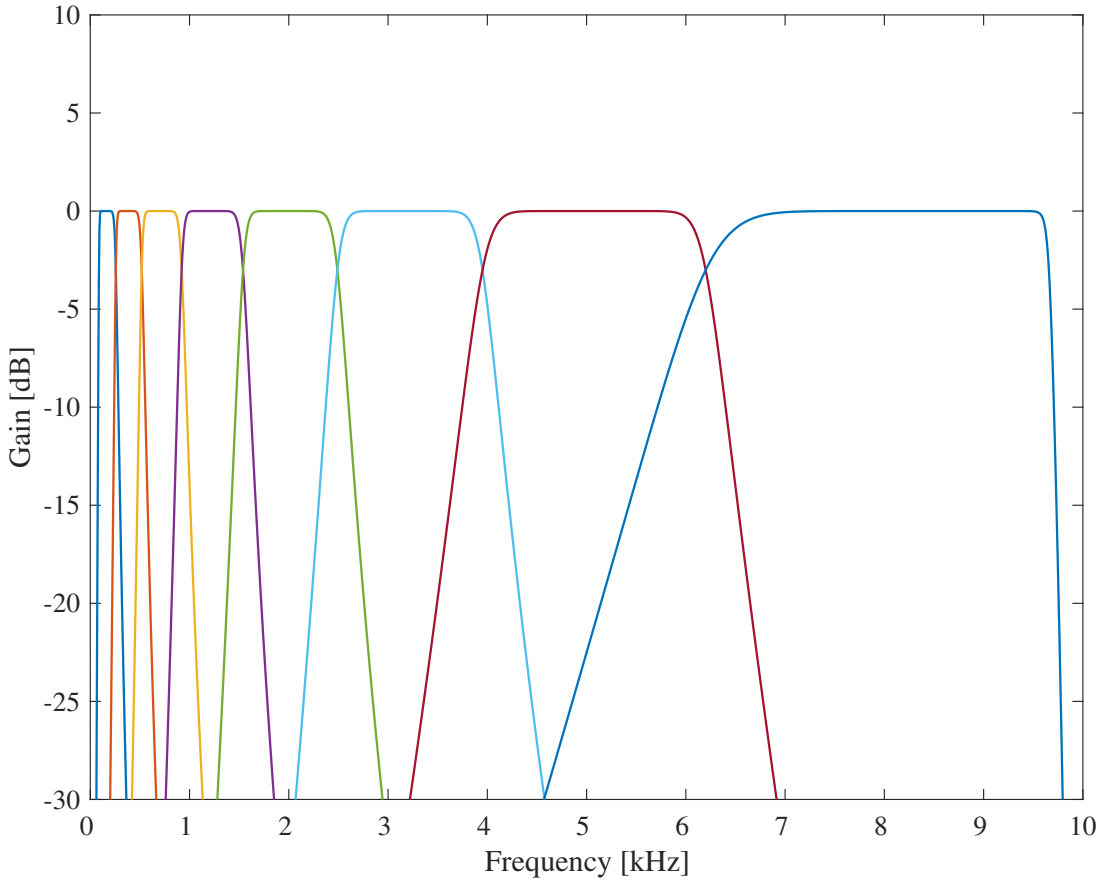


Figure 3.3: Frequency response of the ERB_N-number based 8-band band-pass filterbank.

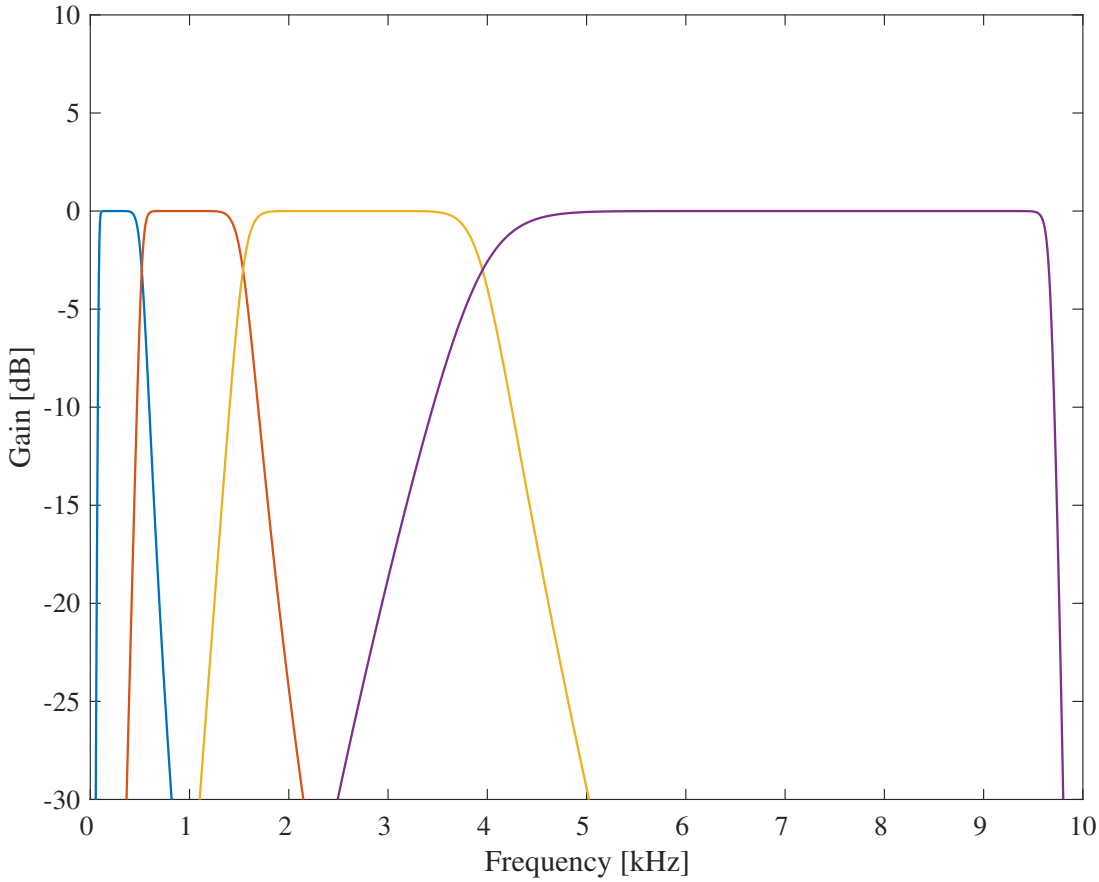


Figure 3.4: Frequency response of the ERB_N-number based 4-band band-pass filterbank.

Then, the temporal envelope $e(k,n)$ of the output signal from each band-pass filter was extracted using the Hilbert transformation and performing a low-pass filter (2nd-order Butterworth IIR filter).

$$e(k,n) = |s(k,n) + j\mathcal{H}[s(k,n)]| * h_{LPF}(n), \tag{3.4}$$

where $\mathcal{H}$ denotes the Hilbert transform and $h_L(n)$ is the impulse response of the low-pass filter. The cut-off frequency of the low-pass filter determined the upper limit of modulation frequency. The upper limit of modulation frequency relates to the temporal resolution that higher temporal resolution will be obtained with higher upper limit of modulation frequency. To investigate the role of temporal cues in the perception of nonlinguistic information, the cut-off frequencies of the low-pass filter were 0.5, 1, 2, 4, 8, 16, 32, and 64 Hz. Moreover, there was an additional "0 Hz" condition where only the direct current component of the Hilbert envelope was extracted.

Finally, the temporal envelope in each channel served to amplitude modulation with the band-limited noise that was generated by band-pass filtering white noise at the same boundary frequency. All amplitude-modulated band-limited noises were summed to generate the NVS stimulus, as following,

$$NVS(n) = \sum_{i=1}^{K} e(k,n) \times WN(k,n), \tag{3.5}$$

where $NVS(n)$ is the NVS signal and the K is the number of channels which could be 4, 8, or 16.

Figure 3.5 shows the spectrogram of a Japanese speech data. Figure 3.6 shows the spectrogram a 16-band NVS with the upper limit of modulation frequency is 64 Hz generated from the speech data in figure 3.5. Figure 3.7 shows the spectrogram a 4-band NVS with the upper limit of modulation frequency is 4 Hz generated from the same speech data. For NVS, there is no such harmonic structure and the power in such one band is almost flat. With the decreasing of the number of channels, the spectral solution is also decreased and the spectral cue is poorer. Also with the decreasing of the upper limit of modulation frequency, the NVS is smoothed further and the temporal modulation cue is poorer.

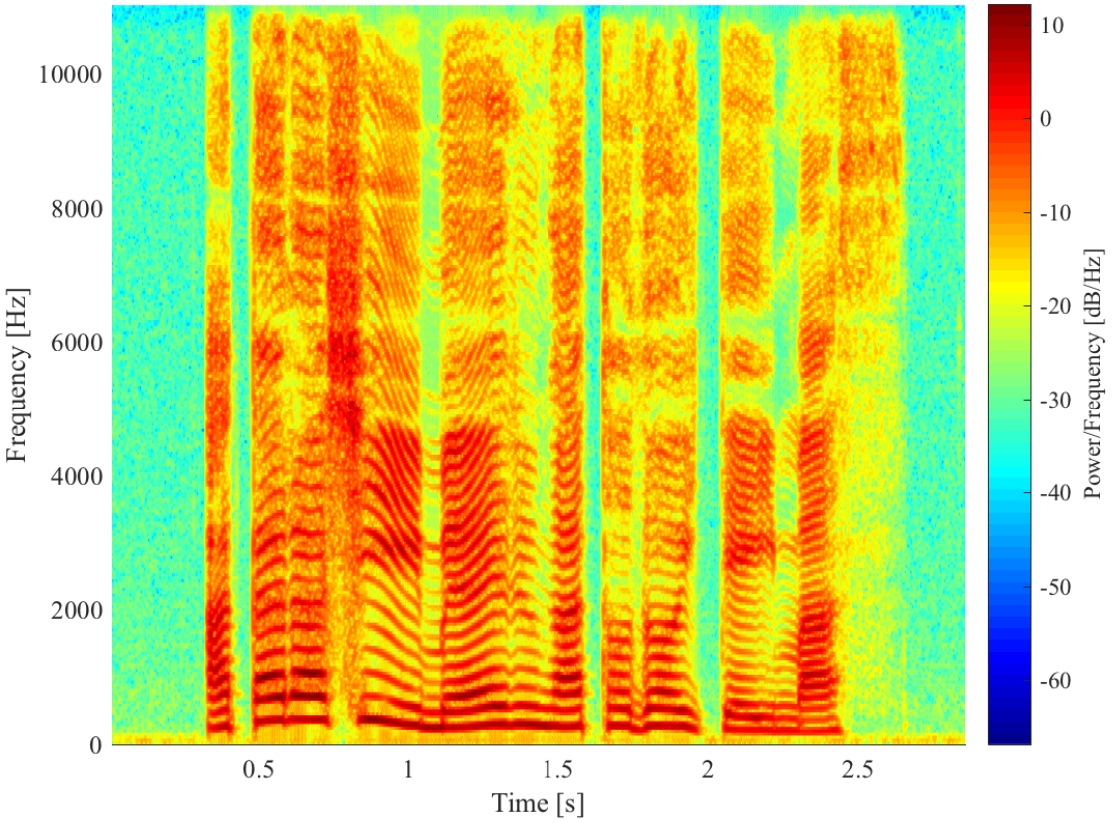


Figure 3.5: Spectrogram of original speech.

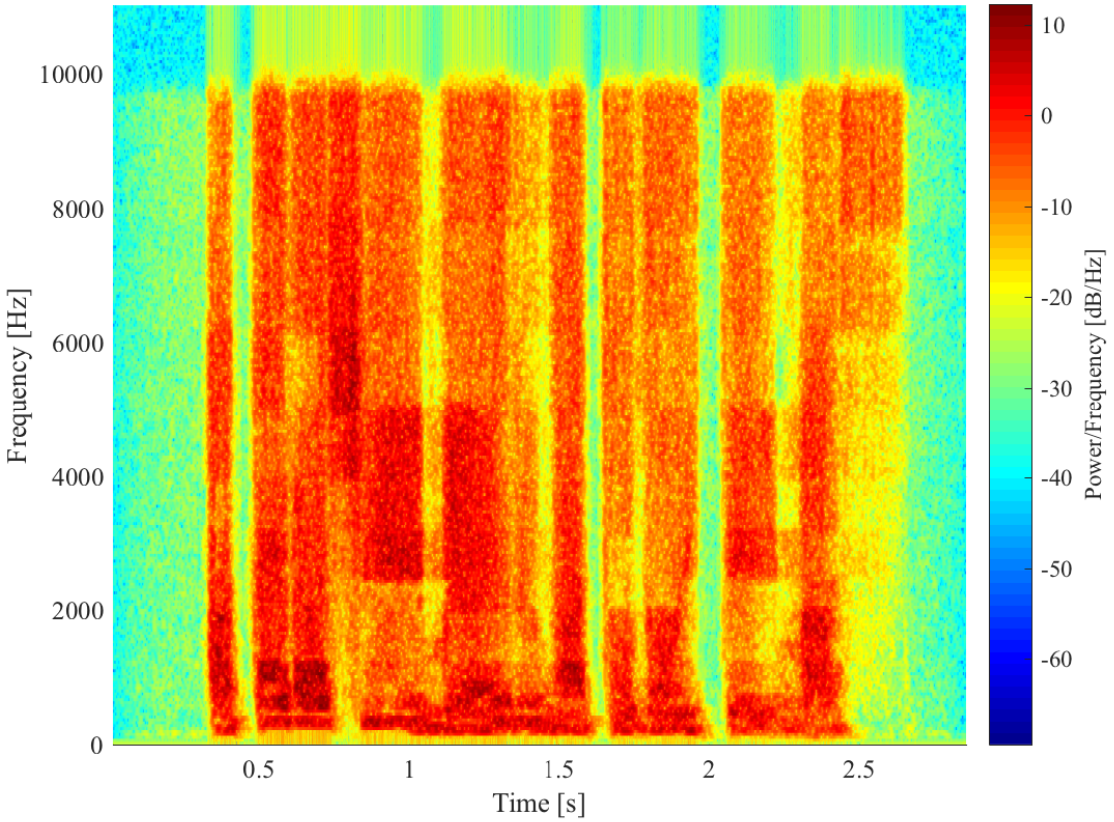


Figure 3.6: Spectrogram of the 16-band NVS and the upper limit of modulation frequency is 64 Hz.

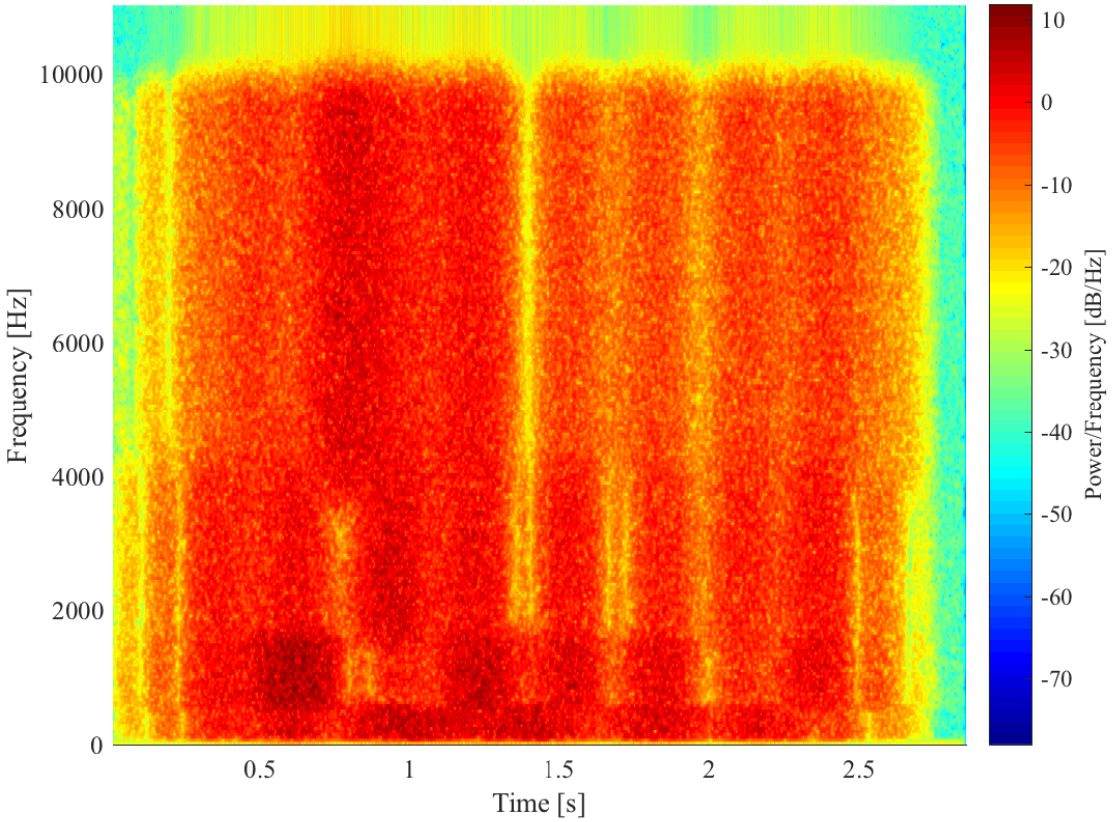


Figure 3.7: Spectrogram of the 4-band NVS and the upper limit of modulation frequency is 4 Hz.

3.3 speaker distinction experiment using noise-vocoded speech with different temporal resolution

3.3.1 Speech Data

The speech data were selected from ATR Japanese speech database set C. All the speech data were recorded at a 20 kHz sampling frequency. Each sentence was uttered for about four to five seconds.

In this study, the XAB method was used in the speaker distinction experiment. In the XAB method, one trial consists of three different speech signals (X, A, and B). The speakers of A and B are different, and the speaker of X is the same speaker of either A or B. Participants are asked to select which speaker, A or B, is more similar to the speaker of X. It is assumed that the similarity of the speaker pair A and B will affect the results of speaker distinction rates dramatically. Two highly similar speakers may be difficult to distinguish between even when the spectral and temporal cues are reserved. On the other hand, the two highly dissimilar speakers may be easy to distinguish between even when the spectral and temporal cues are reduced. This kind of bias is not undesirable.

Kitamura *et al.* measured the perceptual similarity of 20 female and 20 male Japanese speakers in ATR speech database set C [1]. NH listeners listened to the same two same sentences spoken by two speakers and were asked to rate the similarity of these two speakers from 1 to 5. The perceptual similarity of speakers is considered to generate some undesirable bias in the XAB test. Therefore, to remove the impact of similarity, the speaker pairs used in this study have perceptual similarity closest to the average value of perceptual similarity (female: 1.87, and male: 1.99) measured by Kitamura *et al.* [1]. The five female and five male speaker pairs used in this study and their perceptual similarities are shown in Table 3.2. All 20 speakers are different, and the speakers of each pair have the same gender.

3.3.2 Participants and Procedure

Nine native Japanese speakers (two females and seven males) participated in this experiment. All participants had normal hearing (hearing losses of the participants were below

Table 3.2: Speaker pairs selected from ATR database and their average similarity index measured by Kitamura *et al.* [1]. Left and right halves show female and male speaker pairs, respectively.

Speaker pair		Similarity	Speaker pair		Similarity
F407	F306	1.87	M509	M318	1.99
F611	F418	1.86	M603	M409	1.98
F606	F605	1.88	M508	M113	2.00
F720	F213	1.88	M519	M211	2.01
F709	F614	1.83	M520	M517	1.97

the mean hearing level of 12 dB in the frequency range from 125 to 8000 Hz).

This experiment was carried out using the XAB method. The contents of stimuli X, A, and B were as follows:

- X: NVS
- A: NVS with the same speaker as X
- B: NVS with a different speaker from X.

The sentences of X, A, and B were different. Participants were asked to compare the speakers of A and B with the speaker of X to select which one was more similar to the speaker of X. Both stimuli with XAB and XBA orders were presented to counterbalance any effects due to the order of presentation. All the speaker pairs of A and B are shown in the Table 3.2.

A total of 3 different numbers of channels (4, 8, and 16) and 9 upper limits of modulation frequency (0, 0.5, 1, 2, 4, 8, 16, 32, and 64 Hz) created 27 NVS conditions. The original speech was also presented as a control condition. The participants were allowed to listen to each stimulus only once. Before the experiment, 10 stimuli were presented to the participants to familiarize them with the NVS and the experimental environment. The stimuli used in the experiment were different from those used in the practice. In the experiment, all stimuli were presented randomly.

The experiment was conducted while the participants were in a sound-proof room. The sound pressure level of background noise was lower than 25.8 dB. The stimuli were

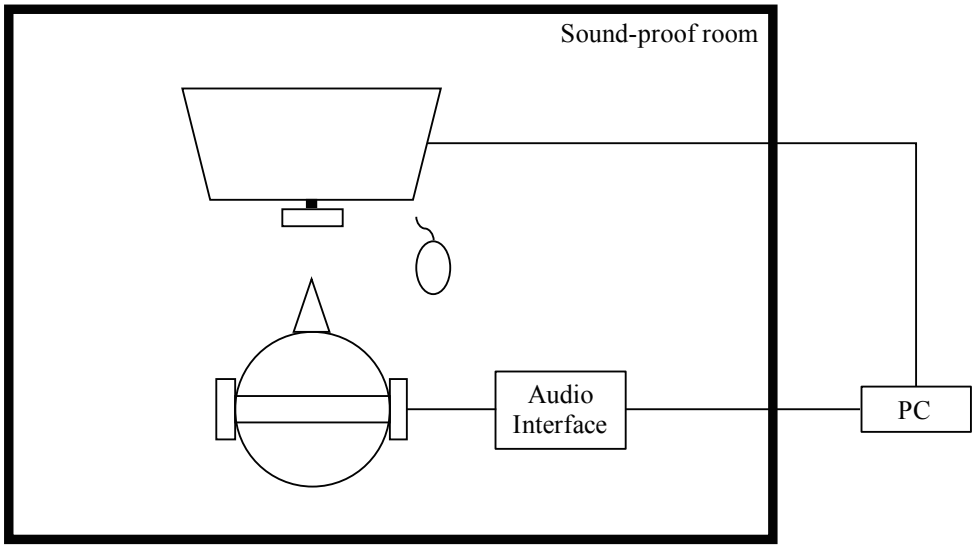


Figure 3.8: The experiment environment.

simultaneously presented to both ears of a participant through a PC, audio interface (RME, Fireface UCX), and a set of headphones (SENNHEISER HDA 200). The sound pressure levels were calibrated to be the same (65 dB SPL) for all participants by using a head and torso simulator (B&K, type 4128) and sound level meter (B&K type 2231).

3.3.3 Results

Figure 3.9 shows the average value of speaker distinction rates, and the error bars indicate a ± 1 standard error of the mean. Under the original speech condition, the mean recognition rate was close to 95%. Thus, participants were nearly perfect at speaker distinction with the original speech. The results for NVS stimuli showed that speaker distinction improved as the upper limit of modulation frequency increased. The results for 4-band NVS were lower than those for 8 or 16-band NVS at some upper limits (0.5, 4, 8, and 32 Hz). However, the performance was not obviously affected by the number of channels.

A three-way repeated-measures analysis of variance (ANOVA) was conducted on the results with the number of channels, upper limit of modulation frequency and speaker pairs as the factors. There was significant main effect of the speaker pairs ($F(9, 72) = 20.99, p < 0.01$). It was shown that, even the perceptual similarities of all speaker pairs were close, the results of different speaker pair were still different. It should be mentioned that the data of perceptual similarities were measured by using original speech signals

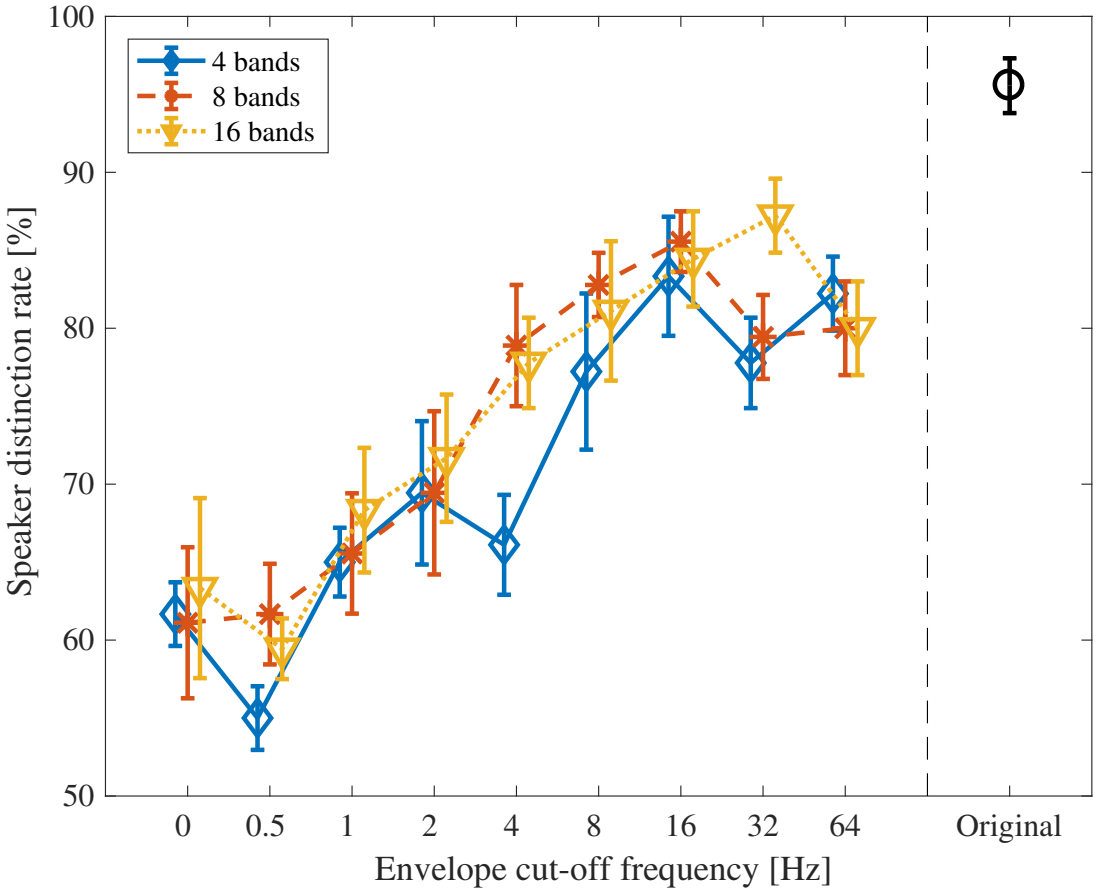


Figure 3.9: speaker distinction rates in all 27 NVS conditions and original speech condition. Error bars indicate ±1 standard error of mean.

and the stimuli used in this experiment were NVS. The reducing of spectral cues may be the reason of the difference between the perceptual similarities and the results of this expeirment.

There was a significant main effect of the upper limit of modulation frequency ($F(8, 64) = 23.86, p < 0.01$) but no significant main effect of the number of bands ($F(2, 16) = 3.32$) or significant interaction between the two factors ($F(16, 128) = 1.16$). These results showed that speaker distinction was significantly affected by the temporal resolution. Therefore, this suggests that temporal cues should contribute to speaker distinction with NVS. The speaker distinction performance was less sensitive to the spectral resolution, however, at least in the limited set of stimuli used in the present study.

3.3.4 Discussion

This experiment was intended to clarify the role of temporal cues in speaker distinction. Specifically, the important modulation frequency band for speaker distinction was investigated. To identify the important modulation frequency band, a sigmoid function was used to fit the data of the experiment. The sigmoid function was mathematically defined as follows:

$$y = \frac{a}{1 + e^{b(x-c)}} + d \tag{3.6}$$

where x is the upper limit of modulation frequency and y is the percent-correct scores. The values of parameters a, b, c, and d were calculated on the basis of the method of least squares. Moreover, the upper limit of modulation frequency at which 90% of the performances plateaued was defined as a knee point. The results of fitting lines and knee points for each number of channels are shown in figure 3.10, 3.11, and 3.12. The coefficients of determinations R^2 of the fitting results in 4, 8, and 16-band NVS were 0.86, 0.96, and 0.93.

The knee point of 4-band NVS was about 20.09 Hz, which was higher than those of 8-band NVS (4.96 Hz) and 16-band NVS (7.60 Hz). This result suggests that the temporal cue may contribute more to speaker distinction when the spectral resolution is limited further.

Note that the speaker distinction rates of 4-band NVS are lower at some upper limits of modulation frequency. However, the number of channels did not affect the performance of speaker distinction significantly. These results were different from those of previous studies [23] [24] in which the performance was improved as the number of channels increased. One difference between the present study and previous studies is that the upper limit of modulation frequency in this study was lower. In previous studies, the cut-off frequencies of the low-pass filter were 500 Hz [23] and 160 or 400 Hz [24]. The modulation frequency band between about 50 and 500 Hz is related to the periodicity information about F0 [21], which is not included in the stimuli used in the present study. One possible explanation may be that the temporal cue related to the periodicity information in the higher modulation frequency bands is more sensitive to the number of channels. The main target of this study is to clarify the role of temporal cues in the modulation frequency band below 64 Hz [21]. Such modulation frequency band includes the information about

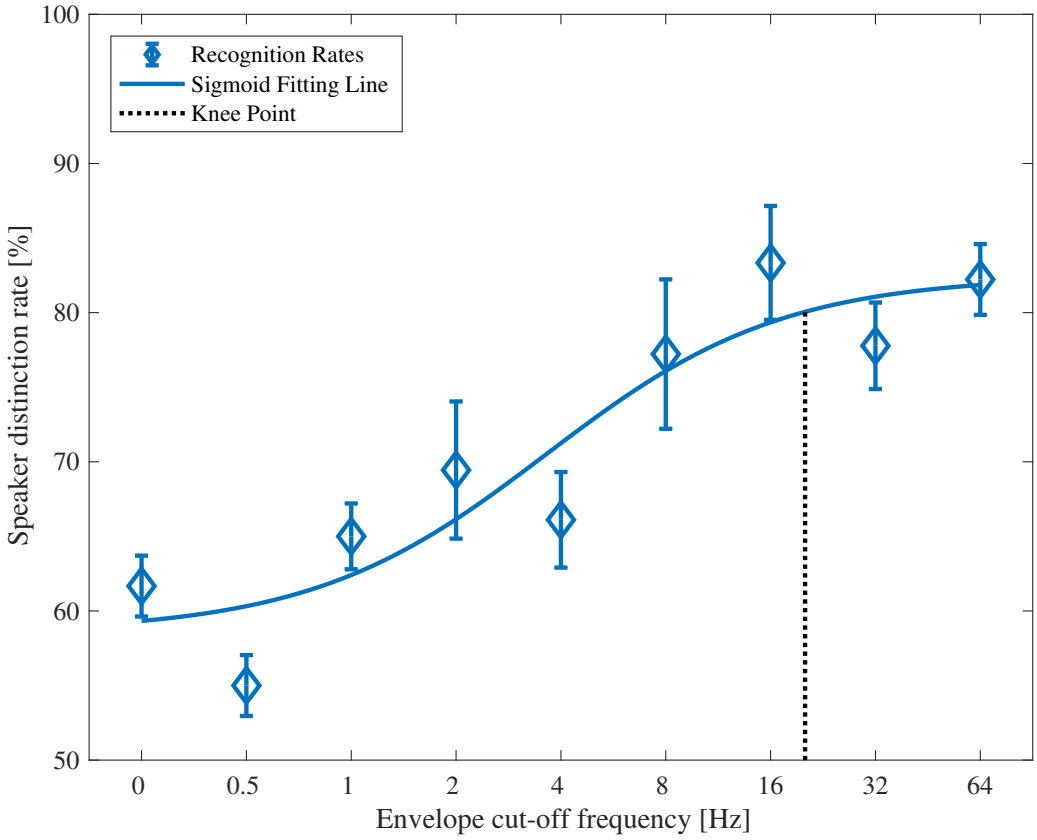


Figure 3.10: Speaker distinction rates in each condition of number of channels and their sigmoid fitting lines for 4-band NVS. Coefficients (95 % confidence interval): a = 23.84 (0.0342, 47.65), b = -0.8913 (-2.873, 1.091), c = 4.862 (2.509, 7.215), d = 58.59 (44.04, 73.14). Coefficient of determinations: $R^2 = 0.86$.

variations of intensity, duration, attack, decay, and segmental cues of speech.

As the spectral cue provided by 4-band NVS was reduced dramatically, participants may have primarily used the temporal cues rather than spectral cues to recognize the speakers. Even so, the average speaker distinction rate for 4-band NVS with a 64 Hz upper limit for modulation frequency was about 80%. Therefore, the temporal cue is showed to be important in the perception of speaker individuality.

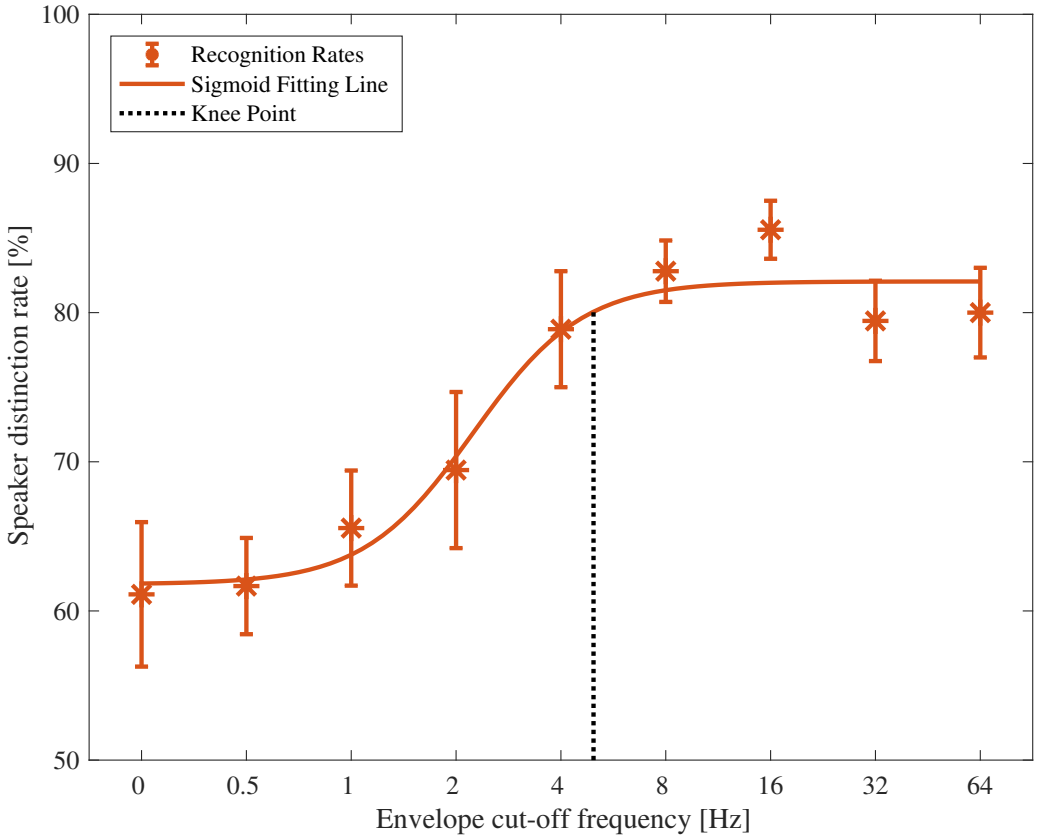


Figure 3.11: Speaker distinction rates in each condition of number of channels and their sigmoid fitting lines for 8-band NVS. Coefficients (95 % confidence interval): a = 20.3 (13.99, 26.61), b = -1.914 (-4.204, 0.3764), c = 4.163 (3.472, 4.854), d = 61.79 (57, 66.58). Coefficient of determinations: $R^2 = 0.96$.

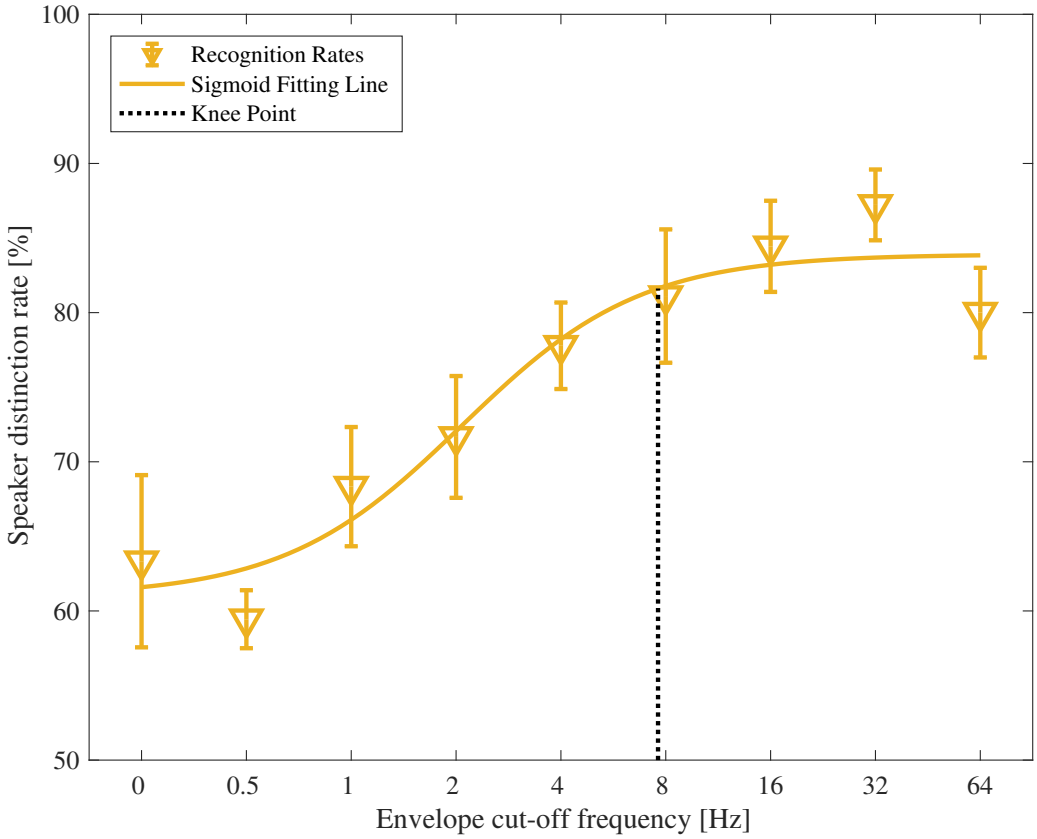


Figure 3.12: Speaker distinction rates in each condition of number of channels and their sigmoid fitting lines for 16-band NVS. Coefficients (95 % confidence interval): a = 23.02 (10.46, 35.58), b = -1.163 (-2.792, 0.4665), c = 4.054 (2.748, 5.359), d = 60.91 (51.28, 70.55) Coefficient of determinations: $R^2 = 0.93$.

3.4 Vocal emotion recognition experiment using noise-vocoded speech with different temporal resolution

3.4.1 Speech Data

The emotional speech data used in this study were selected from the Fujitsu Japanese Emotional Speech Database [9]. This database includes five emotions (*neutral, joy, cold anger, sadness, and hot anger*) expressed by one professional actress. The same sentence was spoken with five emotions. Ten utterances of each emotion were selected. The linguistic contents of each sentence were semantically emotion-neutral to minimize any biasing effect of context. The duration of each utterance was about 3 or 4 s. The sampling frequency and quantization bits were 22.05 kHz. and 16 bits.

3.4.2 Participants and Procedure

Eleven native Japanese speakers (seven males and four females) participated in this experiment. All participants had normal hearing (hearing levels of the participants were below hearing level of 12 dB in the frequency range from 125 to 8000 Hz).

The same as experiment I, there were 27 NVS conditions with 3 different numbers of channels (4, 8, and 16) and 9 upper limits of modulation frequency (0, 0.5, 1, 2, 4, 8, 16, 32, and 64 Hz). The original speech was also presented as a control condition. All stimuli were randomly presented to the participants during the experiment. Participants were asked to indicate which of the five emotions (*neutral, joy, cold anger, sadness, and hot anger*) he/she thought was associated with the stimulus. Each stimulus was presented only once. The experimental environment was as the same as that in speaker distinction experiment in section 3.3.

3.4.3 Results

Figure 3.13 shows the results of the vocal-emotion recognition experiment. First, the recognition rates of the original emotional speech are fixed to 100% for all participants. The Fujitsu database was determined to be a reliable emotional speech database.

The results also showed that vocal-emotion recognition improved as not only the upper limit of modulation frequency but also the number of channels increased. A three-way repeated-measures ANOVA was conducted on the results with the number of channels, upper limit of modulation frequency, and emotion as the factors. Results revealed significant main effects of the number of channels ($F(2, 20) = 79.83, p < 0.01$) and upper limit of modulation frequency ($F(8, 80) = 76.36, p < 0.01$). The interaction between the number of channels and upper limit of modulation frequency was also significant ($F(16, 160) = 8.61, p < 0.01$). The ANOVA also showed a significant main effect of emotion ($F(4, 40) = 31.16, p < 0.01$). Therefore, the emotion significantly affected results for recognition rates. The results for different emotions need to be analyzed separately.

Figure 3.14 - 3.18 shows the vocal-emotion recognition rates of different emotions. The results for different emotions are obviously different. The ANOVA showed significant interactions between the number of channels and emotion ($F(8, 80) = 19.11, p < 0.01$) and between the upper limit of modulation frequency and emotion ($F(32, 320) = 2.02, p < 0.01$). Following these significant interactions, as sub effect test, the analysis of simple main effect showed that for all emotions, there was a significant simple main effect ($p < 0.01$) of the upper limit of modulation. However, the simple main effect of the number of channels was significant ($p < 0.01$) for only neutral, joy, and cold anger. The simple main effect of the number of channels was not significant for sadness ($p = 0.20$) and hot anger ($p = 0.10$). Furthermore, there was a significant interaction between the number of channels, the upper limit of modulation frequency and emotion ($F(64, 640) = 2.59, p < 0.01$). The simple interactions between the number of channels and the upper limit of modulation frequency were significant ($p < 0.01$) for all emotions except sadness ($p = 0.86$).

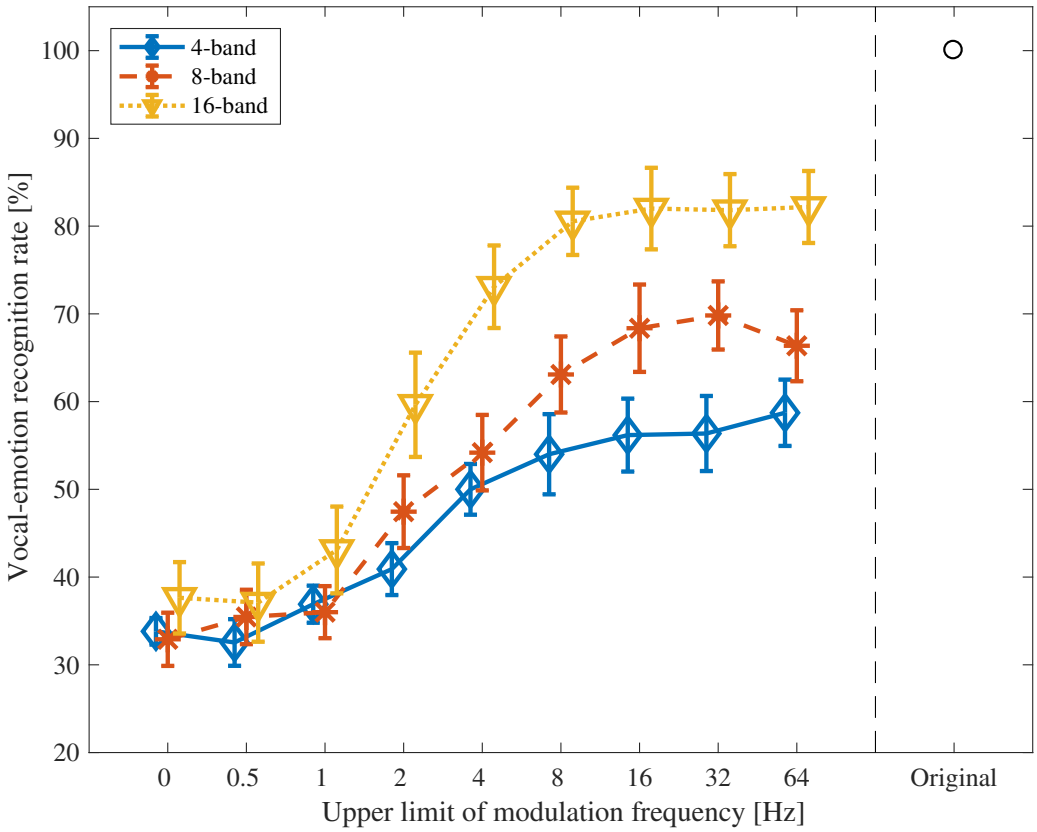


Figure 3.13: Vocal-emotion recognition rates in all 27 NVS conditions and original speech condition. Error bars indicate ±1 standard error of mean.

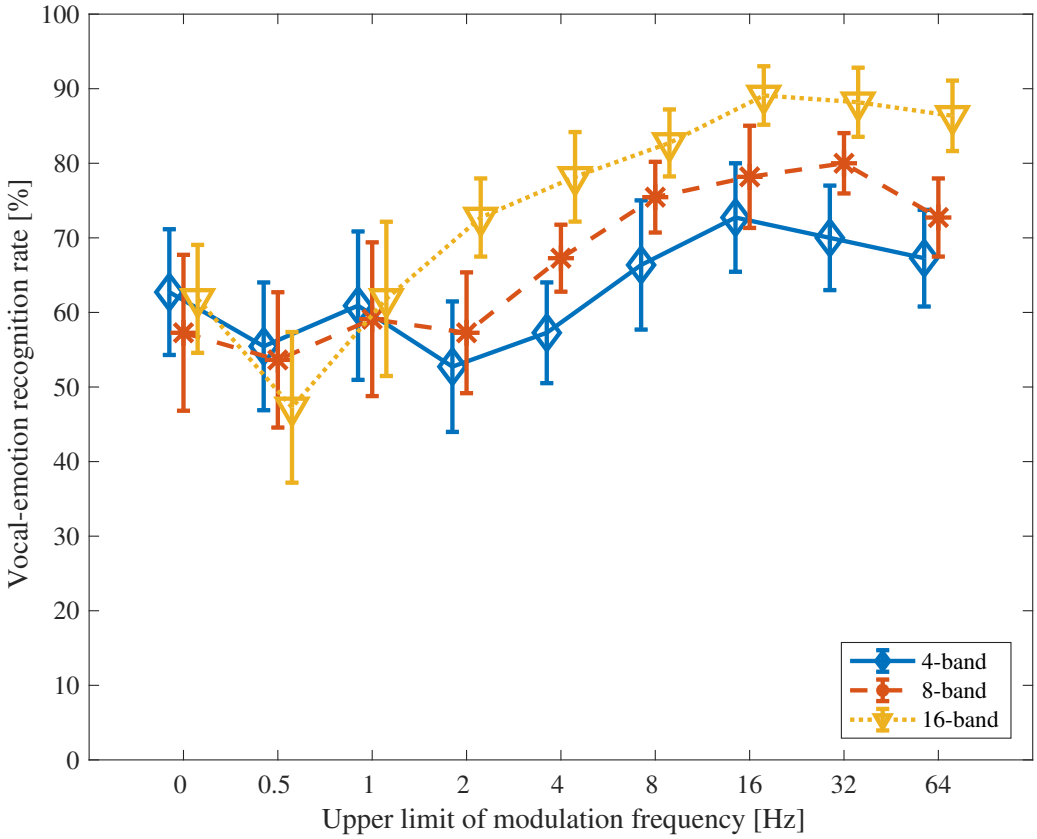


Figure 3.14: Vocal-emotion recognition rates of neutral speech. Error bars indicate ±1 standard error of mean.

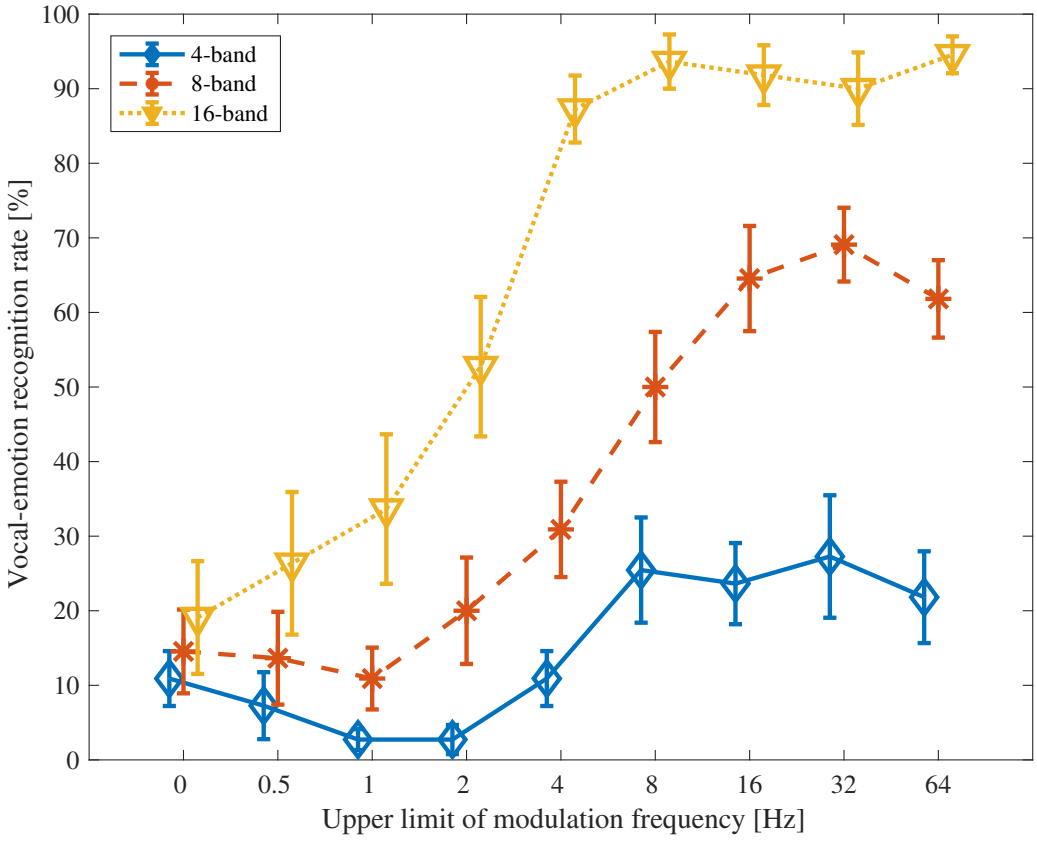


Figure 3.15: Vocal-emotion recognition rates of joy speech. Error bars indicate ±1 standard error of mean.

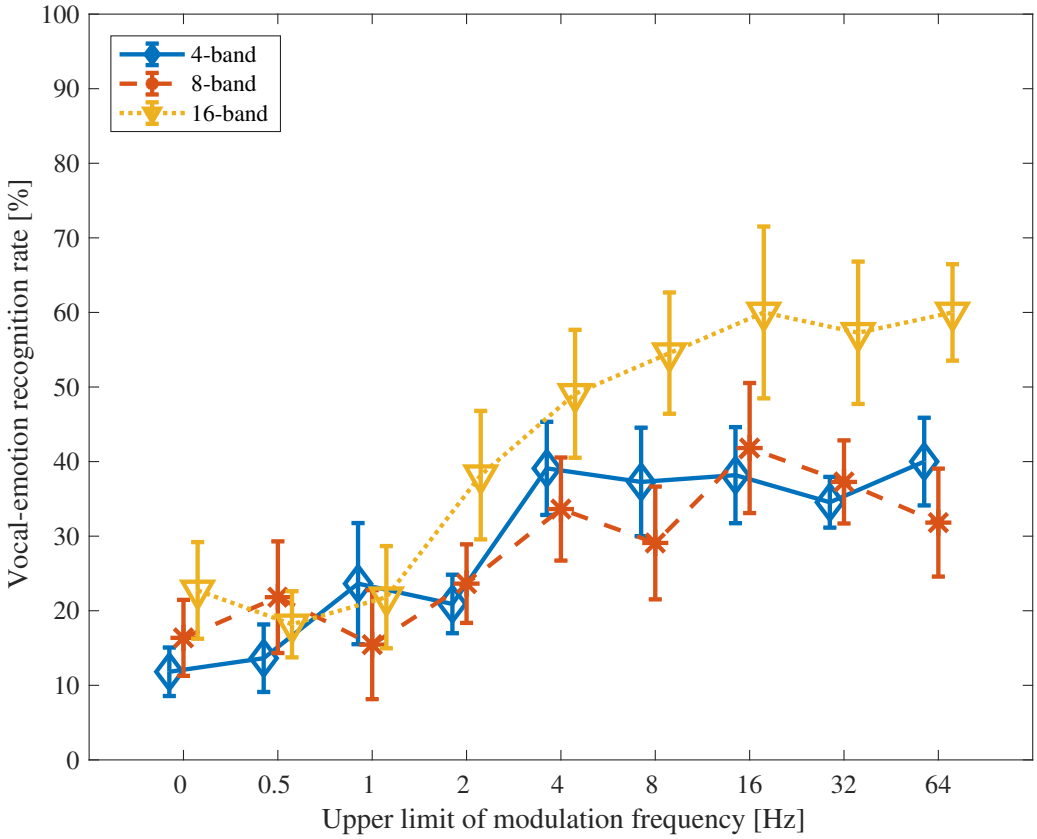


Figure 3.16: Vocal-emotion recognition rates of cold anger speech. Error bars indicate ±1 standard error of mean.

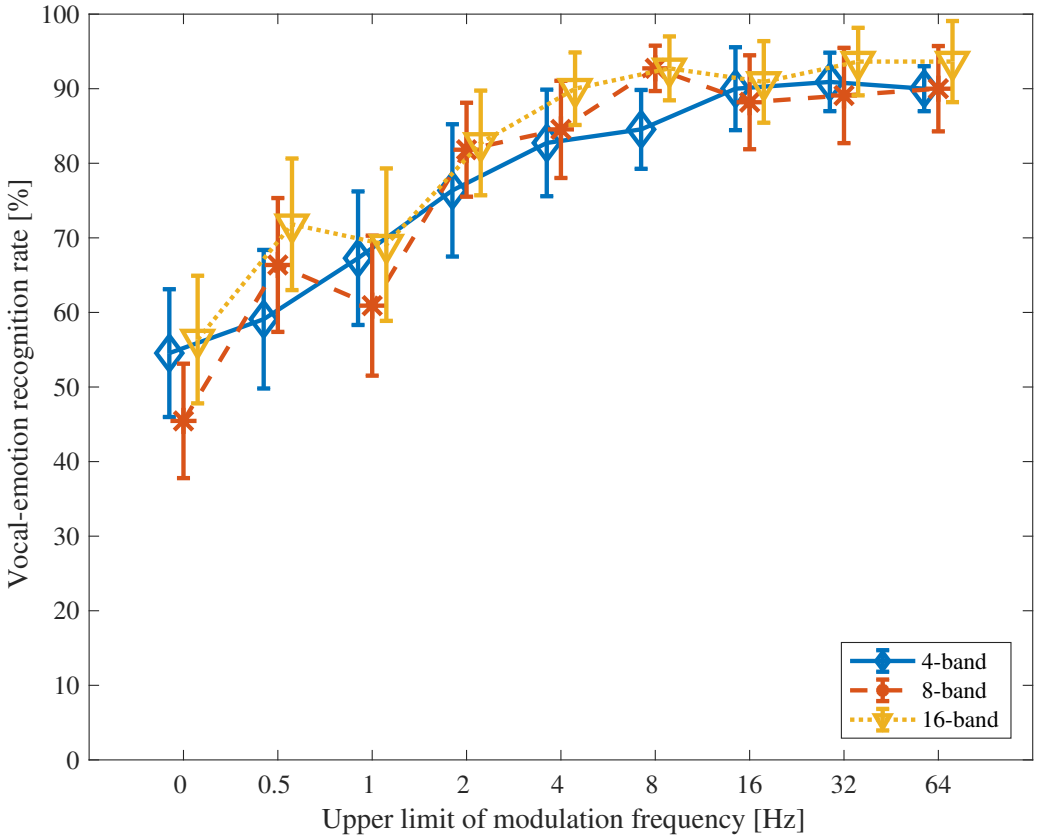


Figure 3.17: Vocal-emotion recognition rates of sadness speech. Error bars indicate ±1 standard error of mean.

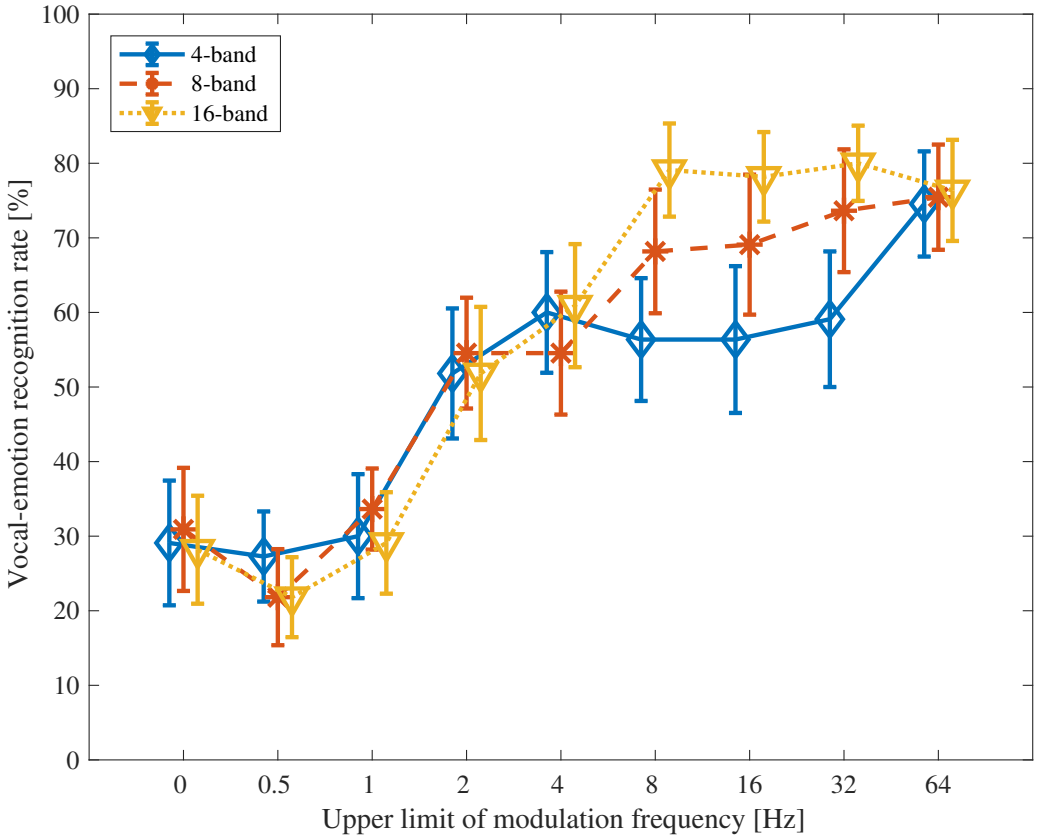


Figure 3.18: Vocal-emotion recognition rates of hot anger speech. Error bars indicate ±1 standard error of mean.

3.4.4 Discussion

Mean confusion matrices obtained with the results of vocal-emotion recognition experiments are shown in Appendices A. The results showed that the selection rates of Neutral were increased with the reducing of the upper limit of modulation frequency or the number of channels. Subjects may tend to select Neutral when they can not recognize the emotion of stimuli.

This experiment was intended to clarify the role of temporal cues in vocal-emotion recognition, so the same fitting method used in Experiment I was also used to the results of the vocal-emotion recognition experiment. Figure 3.19 - 3.21 shows the sigmoid fitting lines of the mean vocal-emotion recognition rates. The coefficients of determinations R^2 of the fitting results in 4, 8, and 16-band NVS were 0.9880, 0.9886, and 0.9986. The knee points of 4- and 8-band NVS were 9.16 and 10.62 Hz, which were higher than that of 16-band NVS (5.26 Hz). The same as for speaker distinction, if the spectral cue is limited further, the temporal cue may contribute more to vocal-emotion recognition. The relationship of the important modulation frequency bands for the perception of linguistic information, speaker individuality and vocal emotion will be discussed in the next section.

The results also showed that the effects of the number of channels and upper limit of modulation frequency were different for different emotions. For neutral (Fig. 3.14), the mean recognition rates was higher than that for other emotions, even for 4-band and 0 Hz upper limit of modulation frequency. Participants may select neutral when they could not recognize the emotion of the NVS stimuli.

For joy (Fig. 3.15), both the number of channels and upper limit of modulation frequency significantly affected the vocal-emotion recognition rates. Therefore, both spectral and temporal resolutions are important for the recognition of joy NVS stimuli. The recognition rates improved as the number of channels and upper limit of modulation frequency increased. At 64 Hz upper limit of modulation frequency, participants performed almost perfectly for the 16-band NVS stimuli. On the other hand, the mean recognition rates for 4-band NVS stimuli were close to or even below the chance level (20%). The fine structure of the spectrum and the temporal variation of amplitude envelope are shown to be important for the recognition of joy NVS.

For cold anger (Fig. 3.16), analyses of simple main effects also showed that both spec-

tral and temporal resolutions affected the results significantly. However, the recognition rates were lower than those of other emotions. Participants performed remarkably more poorly for cold anger when the spectral and temporal cues were reduced.

For sadness (Fig. 3.17) and hot anger (Fig. 3.18), only the upper limit of modulation frequency showed significant simple main effects. This indicates that the spectral solution seems to be unimportant for recognizing sadness and hot anger NVS.

The results showed that temporal solution significantly affected the recognition rates of all emotions. It is confirmed that the temporal cue plays an important role on the perception of vocal emotion. The results also showed that the contribution of spectral cue on the perception of vocal emotion is different for different emotion. The high recognition rates of sadness and hot anger with only 4-band NVS showed that only a rough shape of spectrum is enough for the participants to recognize such emotions. On the other hand, to recognize joy speech, more details of spectrum are necessary. The potential reason of the different contribution of spectral cue should be the different spectral structure of emotional speech.

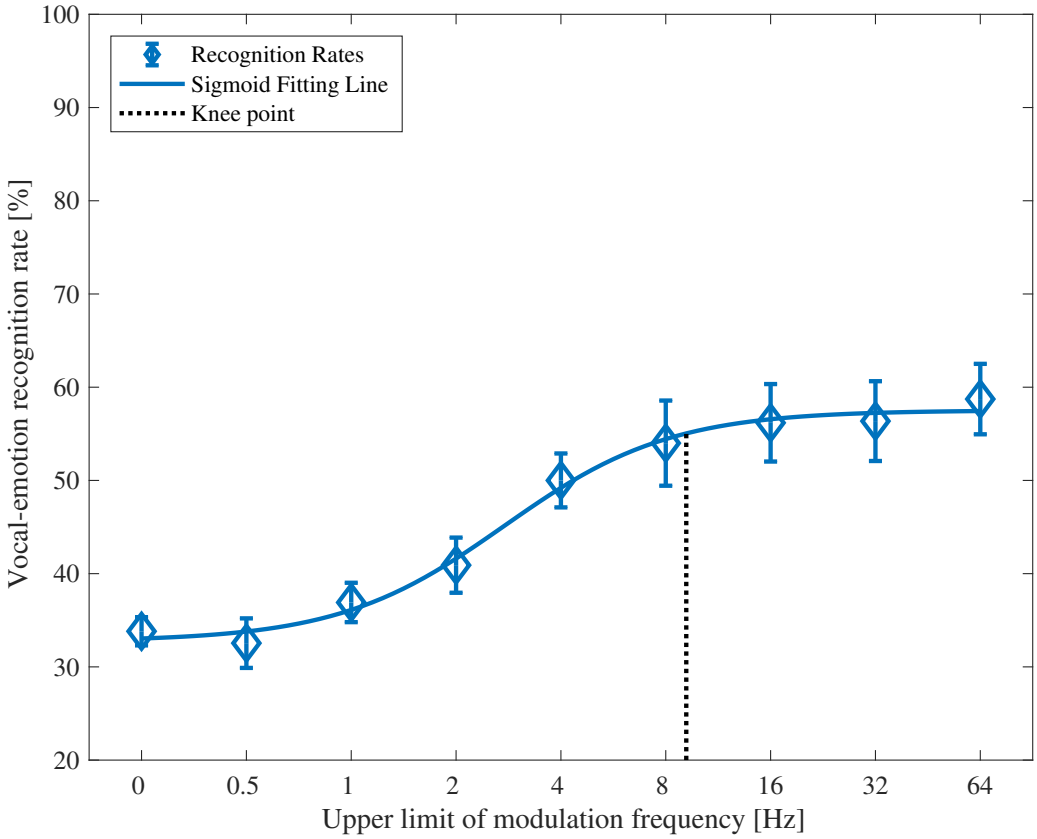


Figure 3.19: Vocal-emotion recognition rates in each condition of number of channels and their sigmoid fitting lines for 4-band NVS. Coefficients (95 % confidence interval): a = 24.78 (28.68, 20.89), b = -1.266 (-1.839, 0.6936), c = 4.461 (4.076, 4.845), d = 57.52 (55.44, 59.6). Coefficient of determinations: $R^2 = 0.9880$.

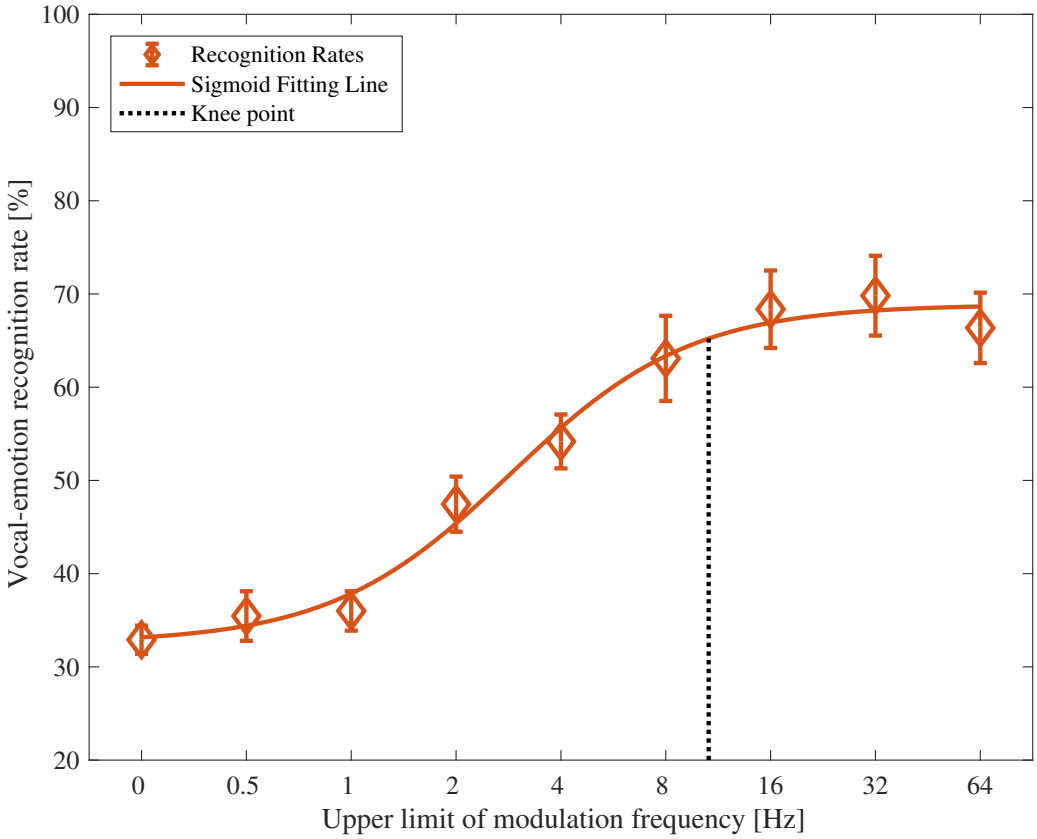


Figure 3.20: Vocal-emotion rates in each condition of number of channels and their sigmoid fitting lines for 8-band NVS. Coefficients (95 % confidence interval): a = 36.28 (28.76, 43.8), b = -1.163 (-1.815, -0.5106), c = 4.52 (4.012, 5.028), d = 32.57 (27.33, 37.82). Coefficient of determinations: $R^2 = 0.9886$.

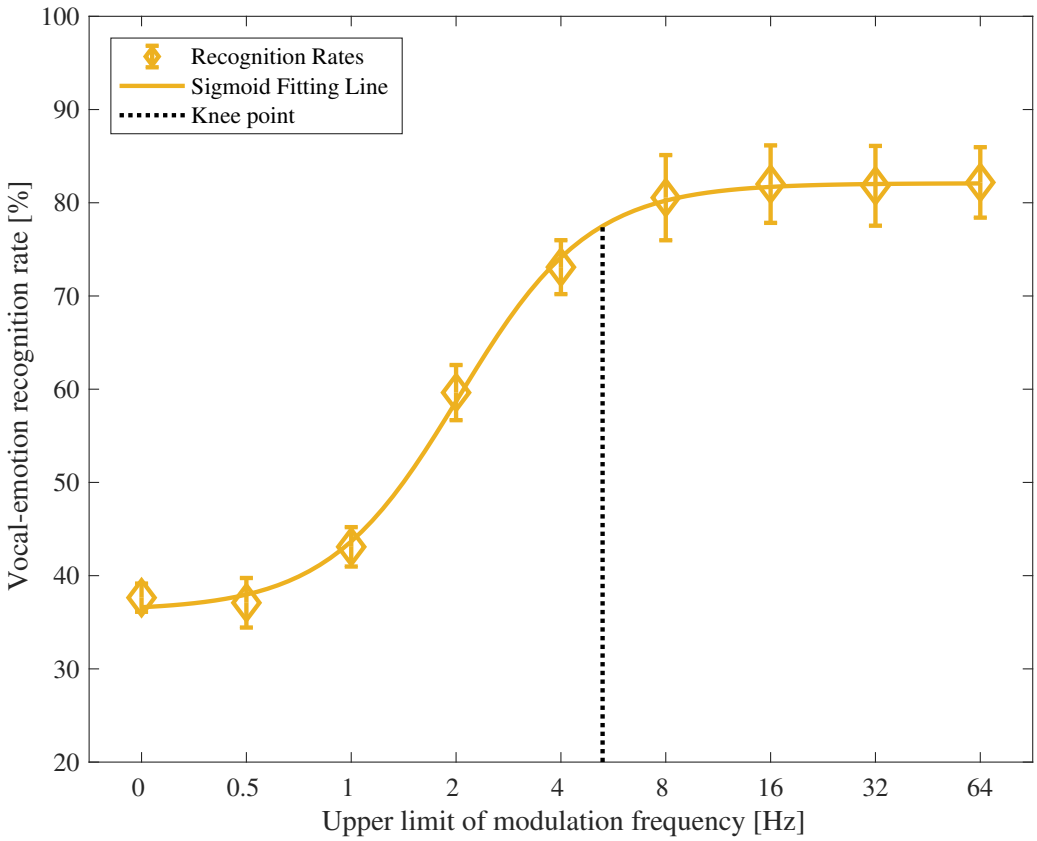


Figure 3.21: Vocal-emotion recognition rates in each condition of number of channels and their sigmoid fitting lines for 16-band NVS. Coefficients (95 % confidence interval): a = 45.83 (43.07, 48.59), b = -1.603 (-1.923, -1.283), c = 4.027 (3.889, 4.164), d = 36.26 (34.12, 38.4). Coefficient of determinations: $R^2 = 0.9986$.

3.5 General discussion

The temporal envelope of speech has been demonstrated to be an important cue for perceiving of linguistic information. The results obtained in this study demonstrated that the temporal cue is also important for the perceiving nonlinguistic information. However, the important modulation frequency bands for linguistic and nonlinguistic information are different.

Xu and Pfingst measured both consonant and vowel recognition as a function of the number of channels (1 to 16) and upper limit of modulation frequency (1 to 512 Hz) [16]. The knee points of vowel recognition for different numbers of channels are all below about 4 Hz. Tachibana *et al.* conducted an experiment of NVS sentence recognition with various upper limits of modulation frequency [14]. They found that increasing the upper limit from 4 to 8 Hz improved the correct response rate more that increasing the upper limit from 8 to 16 Hz. In previous study, the effect of controlling the upper limit of modulation frequency on the recognition of words and speakers while using a fixed number of channels was investigated [79]. The result of word intelligibility tests showed that the average correct number of morae decreased when the upper limit of modulation frequency was less than 5 Hz. The moraic syllable structure is suggested to contribute to the perception of speech. Houtgast and Steeneken demonstrated that the most important modulation frequencies for linguistic information are 3-4 Hz, reflecting the syllable rate in speech [80]. This result is consistent with Arai and Greenberg's previous study about the temporal properties of speech [81]. Their modulation spectral analysis of speech showed that there is a peak on the modulation spectrum at around 4 and 5 Hz. And such temporal characteristics of English and Japanese are remarkably similar.

In this study, the important modulation frequency bands for speaker and vocal-emotion recognition were investigated by using NVS with 3 different numbers of channels (4, 8, and 16). The knee points of 4-, 8-, and 16-band NVS were 20.09, 4.96, and 7.60 Hz for speaker distinction and 9.16, 10.62, and 5.26 Hz for vocal-emotion recognition. The knee points for speaker and vocal emotion recognition were all above 4 Hz. The duration and segmental cues below about 5 Hz for the temporal envelope are also suggested to be used in recognizing speakers and vocal emotions. These segmental cues related to the rhythm, tempo, and the speaking style of the speaker which should be different with dif-

ferent speaker and different emotion. Furthermore, the important modulation frequency bands for nonlinguistic information are suggested to be higher than those for linguistic information. The higher modulation frequency bands are considered to be related to the perception of voice quality.

It is necessary to clarify exactly what kinds of features of the temporal envelope are important for perceiving nonlinguistic information. One possible way to do this is to compare the results of speaker and vocal-emotion recognition experiments with modulation spectral features (MSFs). MSFs, which are the static features extracted from the modulation spectrum of speech, have been shown to be useful for automatic vocal-emotion recognition [77]. The relationship between MSFs and the response of humans will be investigated further.

3.6 Summary

In this chapter the role of temporal modulation cues in the perception of speaker individuality and vocal emotions was investigated. Speaker and vocal-emotion recognition experiments were carried out using NVS as stimuli. The temporal resolution was controlled by varying the upper limits of the modulation frequency. In addition, the role of temporal modulation cues in the different spectral resolution conditions was also investigated by varying the number of channels.

For both speaker and vocal emotion, the recognition rates were significantly decreased with lower upper limit of modulation frequency. Therefore, the results demonstrated temporal modulation cues contribute to the recognition of speakers and vocal emotions. However, the speaker distinction performance was not sensitive to the spectral resolution, at least in the limited set of stimuli in the present study. For vocal-emotion recognition, the spectral resolution was important for the recognition of only neutral, joy, and cold anger NVS, but not sadness or hot anger. Compared to the perception of linguistic information, the temporal modulation cues provided by higher modulation frequency bands were suggested to be important for the perception of speaker individuality and vocal emotion. It is confirmed that the temporal modulation cues contributes to the perception of not only linguistic information but also speaker individuality and vocal emotion.

Chapter 4

Contributions of modulation spectral features on the perception of speaker individuality and vocal emotion

4.1 Introduction

From the speaker and vocal-emotion recognition experiment with NVS, it has been confirmed that the temporal envelope of speech contributes to the perception of not only linguistic information but also speaker individuality and vocal emotion. The temporal modulation cue is suggested to play an important role in the auditory system to extract various information from speech. However, it is still unknown that exactly what kinds of features of the temporal modulation components contribute to the perception of speaker individuality and vocal-emotion.

On the other hand, the modulation spectrum of temporal envelope of speech has been proved to be important for many research fields such as auditory physiology, psychoacoustics, speech perception, and signal analysis and synthesis [65]. Moreover, modulation spectral features have been successfully applied in automatic speaker or vocal emotion recognition systems [69–74, 76, 77]. Therefore, modulation spectral features can represent speaker individuality and vocal emotion information. Such kinds of modulation spectral features are calculated based on the signal process of auditory system. However, it is still unclear whether modulation spectral features contribute to the perception of nonlinguistic

information in auditory system.

In this chapter, the relationship between the modulation spectral features and perceptual data is investigated to clarify the contribution of modulation spectral features on the perception of speaker individuality and vocal-emotion. At first, ten types of modulation spectral feature are extracted from the modulation spectrogram of speech data. The correlation between the modulation spectral features and the perceptual data was calculated to discuss whether the modulation spectral features will contribute to the perception of speaker individuality or vocal-emotion.

4.2 Method to analysis modulation spectral features

4.2.1 Modulation Spectrogram Analysis

A previous study suggested that the acoustic features of intensity and duration cannot account for the human perception of vocal emotion with noise-vocoded speech [27]. Moreover, for vocal emotion recognition by machine, it has been proved that the modulation spectral features perform better than the traditional acoustic features such as Mel frequency cepstrum coefficient (MFCC) and perceptual linear predictive (PLP) coefficient [77]. For these reasons, we only investigated the modulation spectral features for this study.

All emotional speech signals used in this study were selected from the Fujitsu Japanese Emotional Speech Database [9]. This database includes five emotions (*neutral, joy, cold anger, sadness, and hot anger*) spoken by one female speaker. Ten utterances of each emotion were used.

Figure 4.1 shows the auditory-inspired process used in this study to calculate the modulation spectrogram. The signal process until the temporal envelope extraction is as same as that of the signal process to generate NVS stimuli. Emotional speech signals s were first band-pass filtered using an auditory-inspired band-pass filterbank as follows:

$$s(k,n) = s(n) * h_{BPF}(k,n) \tag{4.1}$$

where $h(k,n)$ is the impulse response of the kth channel and n is sample number in the time domain. The same 4-,8-,16-band filterbank used in NVS stimuli generation (section 3.2) is also used here.

The instantaneous amplitude of kth channel signal $e(k,n)$ was then calculated using the Hilbert transform as follows:

$$e(k,n) = |s(k,n) + j\mathcal{H}[s(k,n)]|, \tag{4.2}$$

where $\mathcal{H}$ denotes the Hilbert transform. The next step involved decomposing the instantaneous amplitude into several modulation frequency bands by using a modulation filterbank. The modulation filterbank consisted of six filters, $g_m(n)$, (one low-pass filter and five band-pass filters). The low-pass filter was a 2nd order Butterworth IIR filter with a cut-off frequency of 2 Hz. The cut-off frequencies of the band-pass filters were equally spaced on a logarithm scale from 2 to 64 Hz. Figure 4.2 shows the frequency response of the modulation filterbank. Finally, the modulation spectrogram $E(k,m,n)$ was obtained by:

$$E^2(k,m,n) = |g(m,n) * e(k,m,n)|^2, \tag{4.3}$$

where m is the channel number of the modulation filter. Finally, the time averaged modulation spectrogram $\overline{E}(k,m)$ in dB was used to calculate the modulation spectral features.

$$\overline{E}(k,m) = 10\log_{10}\frac{1}{N}\sum_{n=0}^{N} E^2(k,m,n) \tag{4.4}$$

where N is the length of speech signal.

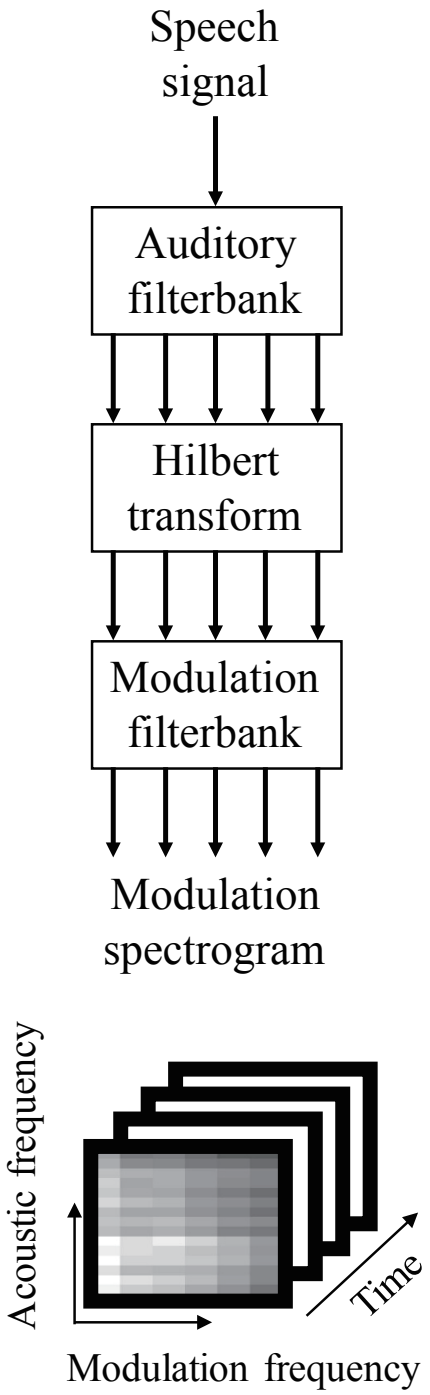


Figure 4.1: Schematic diagram of noise-vocoder method used to generate stimuli (BPF: band-pass filter; LPF: low-pass filter; and NBN: narrow-band noise).

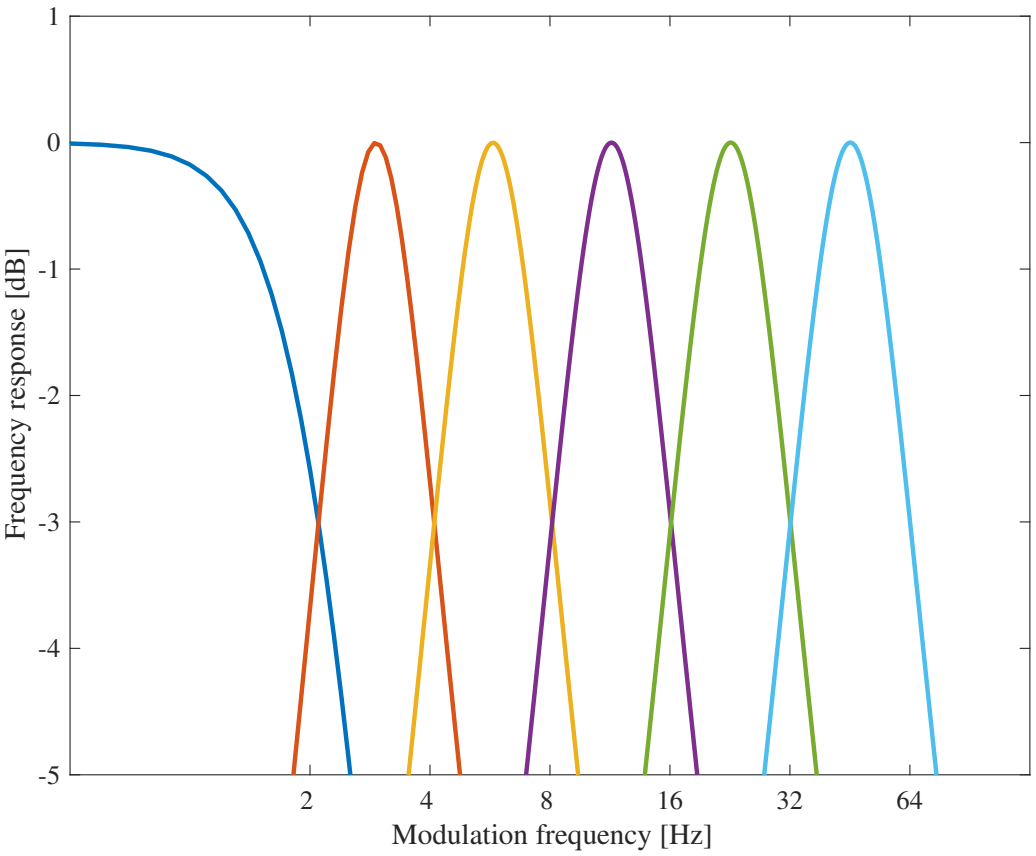


Figure 4.2: Frequency response of the modulation filterbank.

4.2.2 Modulation-Spectral Feature Extraction

We extracted ten types of modulation spectral feature to determine whether these features can be used to identify the corresponding vocal emotion with noise-vocoded speech. Two kinds of modulation spectral feature were calculated by analyzing the modulation spectrogram in the acoustic frequency domain and the modulation frequency domain.

In the acoustic frequency domain, the first feature was the modulation spectral centroid (MSCR_m), which can be defined as follows:

$$\mathrm{MSCR}_m = \frac{\Sigma_{k=1}^{K} k\overline{E}(k,m)}{\Sigma_{k=1}^{K} \overline{E}(k,m)}, \tag{4.5}$$

where K is the number of acoustic frequency bands (4, 8 or 16). The MSCR_m indicates the center of the spectral balance across acoustic frequency bands (k).

The modulation spectral spread (MSSP_m) was then calculated by:

$$\mathrm{MSSP}_m = \frac{\Sigma_{k=1}^{K} [k - \mathrm{MSCR}_m]^2 \overline{E}(k,m)}{\Sigma_{k=1}^{K} \overline{E}(k,m)}. \tag{4.6}$$

The MSSP_m can represent the spread of the spectrum around its MSCR_m as the 2nd moment.

Two other higher order features, modulation spectral skewness (MSSK_m) and kurtosis (MSKT_m), were also calculated. The MSSK_m describes the degree of asymmetry of the spectrum which was calculated from the 3rd order moment:

$$\mathrm{MSSK}_m = \frac{\Sigma_{k=1}^{K} [k - \mathrm{MSCR}_m]^3 \overline{E}(k,m)}{\Sigma_{k=1}^{K} \overline{E}(k,m)}. \tag{4.7}$$

The MSKT_m gives a measure of the peakedness of the spectrum which was calculated from the 4th order moment:

$$\mathrm{MSKT}_m = \frac{\Sigma_{k=1}^{K} [k - \mathrm{MSCR}_m]^4 \overline{E}(k,m)}{\Sigma_{k=1}^{K} \overline{E}(k,m)}. \tag{4.8}$$

On the modulation frequency domain, the first feature is the MSCR_k which is the barycenter of the modulation spectrum in each acoustic frequency band. Different from the MSCR_m which was calculated across the acoustic frequency bands (k), the MSCR_k was calculated across the modulation frequency bands (m). Then the other three higher order features of the modulation spectrogram on the modulation frequency domain (MSSP_k, MSSK_k, and MSKT_k) were also calculated as following.

$$\mathrm{MSCR}_k = \frac{\Sigma_{m=1}^{M} m\overline{E}(k,m)}{\Sigma_{m=1}^{M} \overline{E}(k,m)}, \tag{4.9}$$

$$\mathrm{MSSP}_k = \frac{\Sigma_{m=1}^{M}[m - \mathrm{MSCR}_k]^2 \overline{E}(k,m)}{\Sigma_{m=1}^{M} \overline{E}(k,m)}, \tag{4.10}$$

$$\mathrm{MSSK}_k = \frac{\Sigma_{m=1}^{M}[m - \mathrm{MSCR}_k]^3 \overline{E}(k,m)}{\Sigma_{m=1}^{M} \overline{E}(k,m)}, \tag{4.11}$$

$$\mathrm{MSKT}_k = \frac{\Sigma_{m=1}^{M}[m - \mathrm{MSCR}_k]^4 \overline{E}(k,m)}{\Sigma_{m=1}^{M} \overline{E}(k,m)}, \tag{4.12}$$

where the M is the number of channels of the modulation filterbank which is 6.

The last feature on the acoustic frequency domain was modulation spectral flatness (MSFT_m), which was computed from the ratio of the geometric mean to the arithmetic mean of the spectrum:

$$\mathrm{MSFT}_m = \frac{\sqrt[K]{\prod_{k=1}^{K} E^2(k,m)}}{\frac{1}{K}\Sigma_{k=1}^{K} E^2(k,m)}. \tag{4.13}$$

The MSFT_m is a measure of the noisiness of a spectrum.

The last modulation spectral feature on the modulation frequency domain was modulation spectral tilt (MSTL_k), which is the linear regression coefficient obtained by fitting a first-degree polynomial to the modulation spectrum in dB scale.

4.3 Modulation spectral features related to the perception of speaker individuality

In this section, at first, the relationship between modulation spectral features and perceptual speaker similarity is investigated to confirm whether the modulation spectral features could be possible cues in the perception of speaker individuality. In the next step, the correlation between modulation spectral features and the perceptual data obtained from speaker distinction experiments using NVS is investigated to clarify that whether the modulation spectral features contribute to the speaker distinction. In section 3.3, speaker distinction experiments using NVS was carried out. However, the perceptual speaker similarities of the speaker pairs are too close, so the perceptual data can not be used to investigated the relationship with modulation spectral features. Therefore, another speaker distinction experiment is carried out using the speaker pairs with different perceptual speaker similarity.

4.3.1 The relationship between modulation spectral features and perceptual speaker similarity

Kitamura et al. measured the perceptual similarity of speaker individualities of 20 female and 20 male Japanese speakers in ATR speech database set C [1]. Two same sentences with different speakers were presented to NH listeners, and the listeners were asked to select the similarity of these speaker spairs from 1 to 5. If the modulation spectral features contribute to the perception of speaker individuality, there must be a high correlation between the modulation spectral features and the perceptual speaker similarity. In this section, the relationship between modulation spectral features and perceptual speaker similarity is discussed. At first, the modulation spectral features of the speech data from 20 female and 20 male Japanese speakers used in [1] is calculated. Then, an discriminability index called d' (d-prime) to describe the separation of each modulation spectral feature between different speaker pairs. Finally, the correlation between the d' of modulation spectral features and perceptual speaker similarity is calculated.

Speech data

The speech data from 20 female and 20 male Japanese speakers used by Kitamura [1] was used in this study. The speaker numbers of all 40 speaker are F213, F214, F306, F308, F406, F407, F409, F418, F507, F509, F605, F606, F609, F611, F614, F702, F704, F709, F714, F720, M109, M113, M211, M214, M318, M409, M504, M508, M509, M510, M517, M519, M520, M601, M603, M614, M705, M710, M714, M718. The table 3 and 4 in [1] show the data of perceptual speaker similarity of the 20 male and 20 female speakers. 10 utterance spoken by each speaker was used to calculate the modulation spectral features.

The discriminability index (d') of modulation spectral features

The modulation spectral features of the speech data was calculated by the method described in section 4.2. Then, an discriminability index called d' (d-prime) to describe the separation of each modulation spectral feature between different speaker pairs. The discriminability index is defined as the absolute value of the difference between the mean values of the modulation spectral feature (taken across the 10 utterances) for two speakers,

divided by the root of their average variance as follows.

$$d' = \frac{\mu_{speaker1} + \mu_{speaker2}}{\sqrt{\frac{1}{2}(\sigma^2_{speaker1} + \sigma^2_{speaker1})}} \tag{4.14}$$

where μ and σ^2 are the mean value and variance of a modulation spectral feature taken across the 10 utterances. The average value of discriminability indices (taken across all the acoustic frequency or modulation frequency bands) was computed as a measure of the net discriminability provided by this feature. The d' can present the distance of such modulation spectral feature between two different speakers. Higher value of d' means the modulation spectral feature's distributions of two speakers are more separated. The correlation coefficients between the average value of the d' of modulation spectral features and the perceptual speaker similarity were then calculation to clarify the relationship between modulation spectral features and perceptual speaker similarity.

Results

Figure 4.3 shows the results of the correlation coefficients between the d' of 16-band modulation spectral features and the perceptual speaker similarity for female and male speakers. The scatterplots of the d' of MSFs and perceptual data of perceptual speaker similarity are shown in Appendices B. For each modulation spectral feature, there was a minus correlation between the d' and perceptual speaker similarity. The results showed that the distance of modulation spectral features of two speakers will be closer when the two speakers are more similar. It is suggested that the modulation spectral features should contribution to the perception of speaker individuality. The auditory system may take advantage of modulation spectral features to distinguish different speakers.

Figure 4.4 shows the correlation coefficients between the d' of MSFs. The correlation between each MSF were all significant. The results showed that there were high correlation between MSCR_m, MSSK_m, and MSFT_m. There were also high correlation between MSSP_m, and MSKT_m and the same trend was also appeared for the MSFs in modulation frequency domain. However, the correlations between the MSFs in acoustic frequency domain and modulation frequency domain were low.

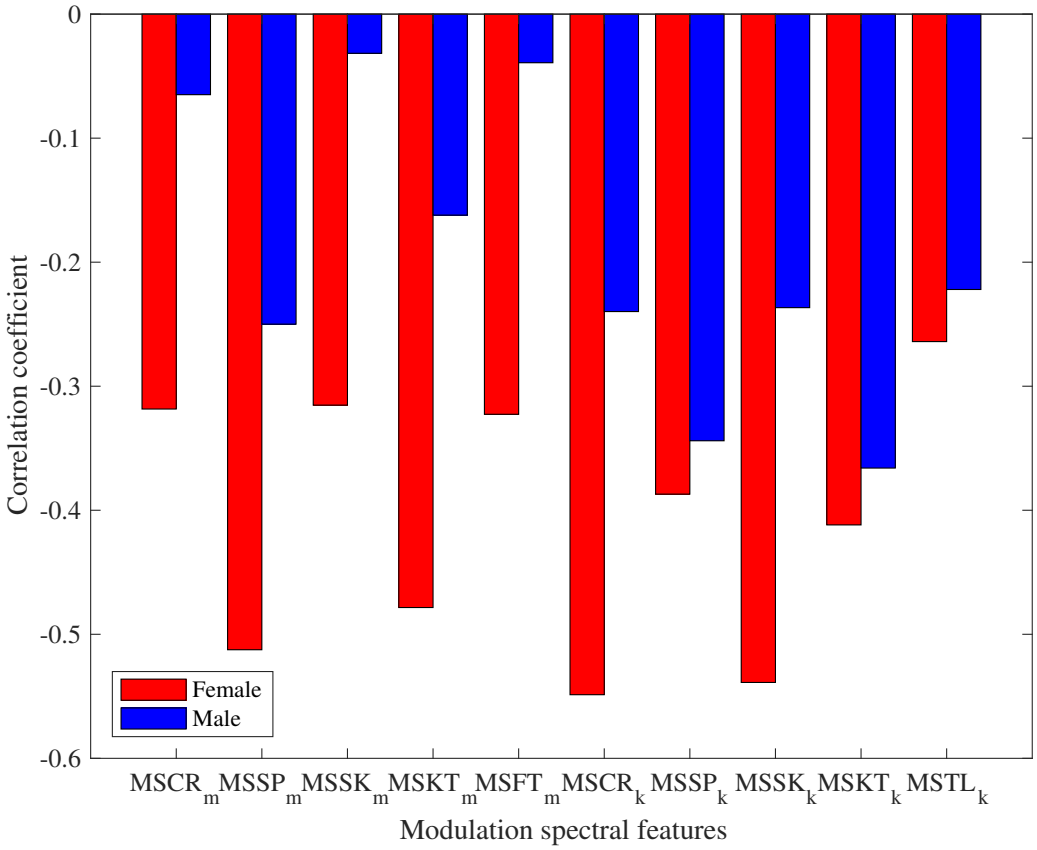


Figure 4.3: The correlation coefficients between the d' of modulation spectral features and the perceptual speaker similarity for female and male speaker pairs.

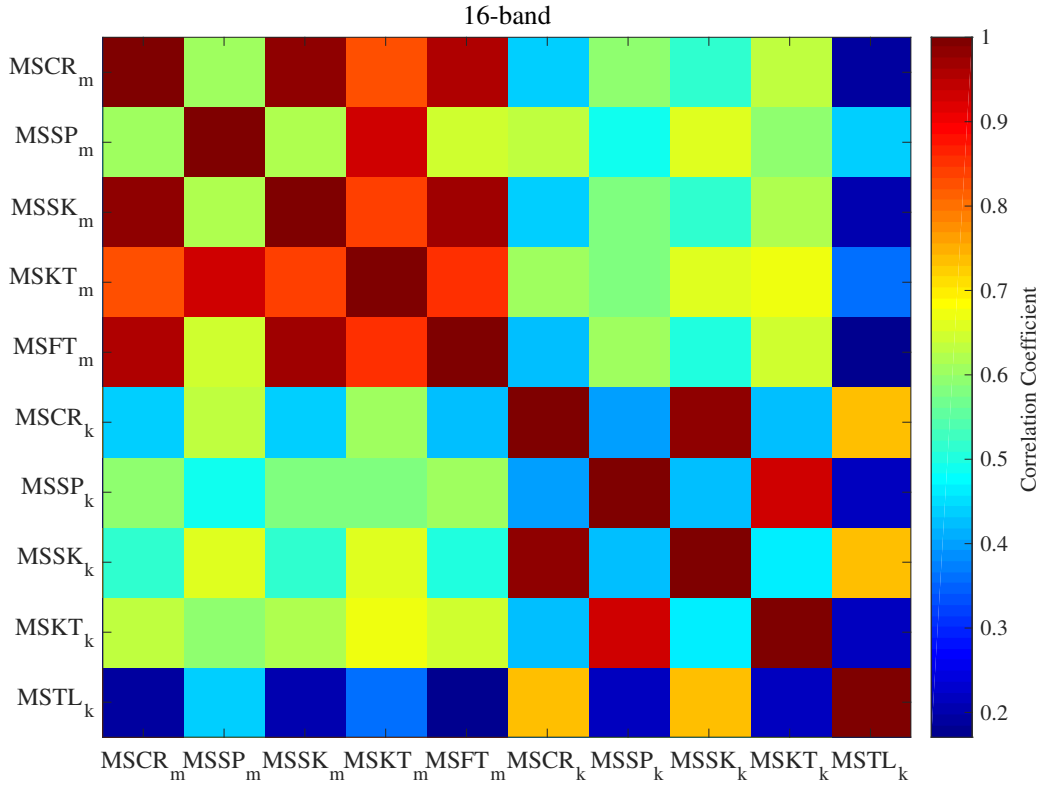


Figure 4.4: The correlation coefficients between the d' of modulation spectral features.

Table 4.1: Speaker pairs selected from ATR database and their average similarity index measured by Kitamura *et al.* [1]. Left and right halves show female and male speaker pairs, respectively.

Speaker pair		Similarity	Speaker pair		Similarity
F507	F609	1.45	M504	M601	1.61
F407	F702	1.97	M614	M710	1.83
F213	F214	2.42	M214	M519	2.36
F611	F614	2.93	M509	M603	2.68
F606	F704	3.32	M409	M705	3.38

4.3.2 Speaker distinction experiment using NVS

Speech data

In section 3.3, speaker distinction experiments using NVS were carried out. However, the perceptual speaker similarities of the speaker pairs are too close, so the perceptual data can not be used to investigated the relationship with modulation spectral features. Therefore, another speaker distinction experiment is carried out using the speaker pairs with different perceptual speaker similarity.

The speaker pairs used in this experiment were also selected based on the perceptual similarity data measured be Kitamura et al. [1]. The 5 female and 5 male speaker pairs used in this study and their perceptual similarities are shown in Table 4.1. All 20 speakers are different and the speakers of each pair have the same gender. 12 sentences of each speaker were used to generate the NVS stimuli.

Stimuli and procedure

NVS stimuli were used in this experiment. The number of channels of NVS stimuli was 8, or 16, and the upper limit of modulation frequency was only 64 Hz. Eight native Japanese speakers with NH (two females and six males) participated in this experiment. XAB method was also used in this experiment. The experimental environment was as same as that in section 3.3

Results

Figure 4.5 and 4.6 shows the results of speaker distinction rates of female and male speaker pairs. For female speaker pairs, the speaker distinction rate decreased dramatically when the speaker similarity was higher than 3. For male speaker pairs, when the speaker similarities were lower than 3, the speaker distinction rates decreased with the increasing of similarity. However, the speaker distinction rate was suddenly increased when the speaker similarity was higher than 3.

A 3-way repeated measures ANOVA was then conducted on the results with the gender of speaker pairs, speaker similarity, and the number of channels as the factors. The results of ANOVA show that the main effect of the gender of speaker pairs was not significant ($F(1,7) = 1.38, p = 0.28$). The main effect of the number of channels ($F(1,7) = 8.58, p < 0.05$) was significant. These results are different from the previous study [23]. The effect of the number of channels in speaker distinction was shown to be different when the speaker pairs and their similarity were different. Furthermore, the main effect of the speaker similarity ($F(4,28) = 9.59, p < 0.01$) was also significant. In the next section, the modulation spectral features were calculated to account for the perceptual data obtained in this experiment and the effect of speaker similarity.

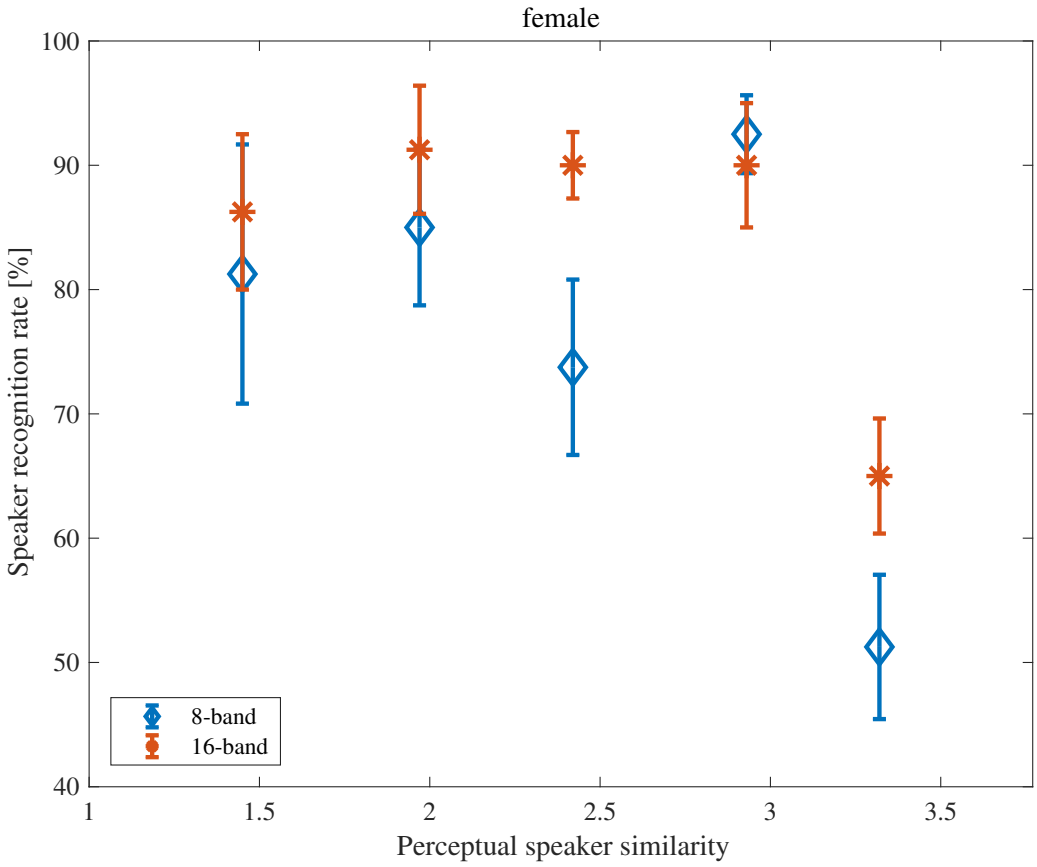


Figure 4.5: Results of speaker distinction rate for female speaker pairs.

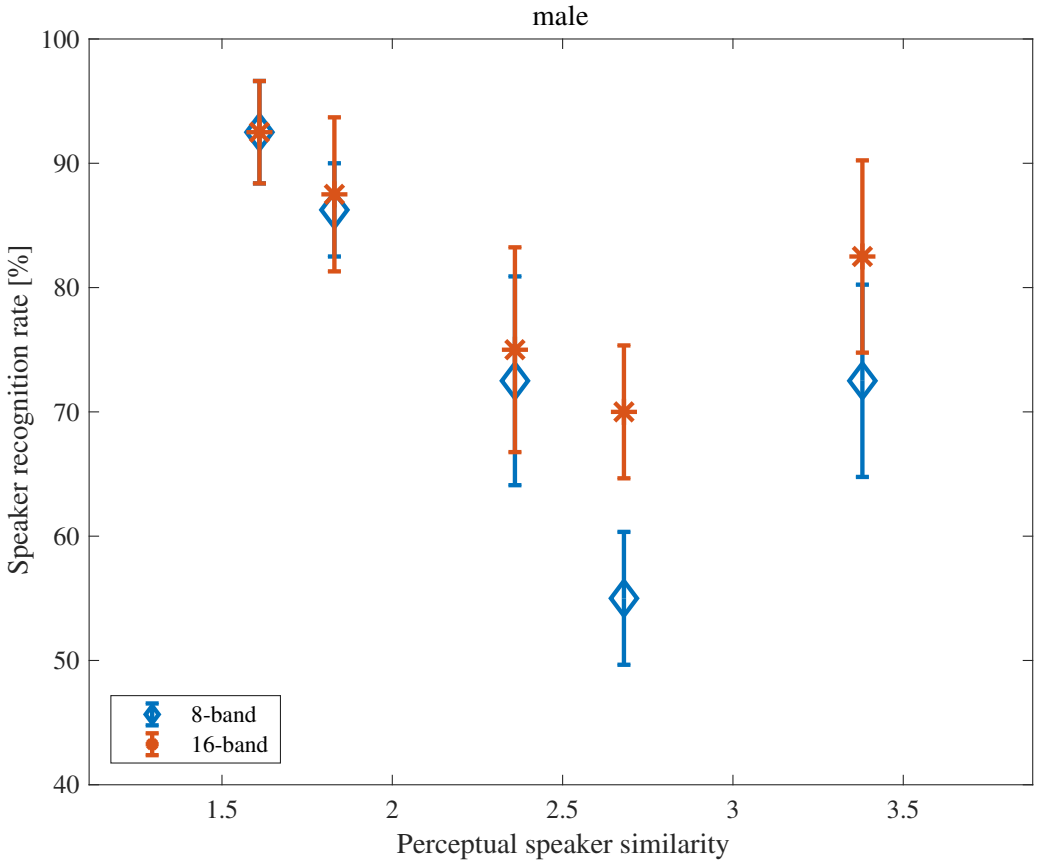


Figure 4.6: Results of speaker distinction rate for male speaker pairs.

4.3.3 The correlation between the perceptual data and modulation spectral features

To investigate the correlation between the perceptual data and modulation spectral features, the d' values of the perceptual data of the speaker distinction experiment were calculated as the method in [82]. The d' value was calculated as following:

$$d' = \phi^{-}1(Hit) - \phi^{-}1(FalseAlarm), \tag{4.15}$$

where $\phi^{-}1$ means the z score, Hit is the hit rate, and $FalseAlarm$ is the false alarm rate. Tabel 4.2 and 4.3 show the d' values of the perceptual data for female and male speakers. Figure 4.7 shows the results of the correlation coefficients between the d' of modulation spectral features and the perceptual data for all speakers. The scatterplots of the d' of MSFs and the perceptual data for all speakers are shown in Appendices C. For all modulation spectral features the correlation coefficients are positive. As the correlations are all positive, the results showed that the psychological distance of each speaker pair increases as the distance of modulation spectral features increases. Therefore, it is suggested that the modulation spectral features should contribute to the perception of speaker individuality.

Table 4.2: The d ' values of perceptual data for female speakers.

	F507&F609	F407&F702	F213&F214	F611&F614	F606&F704
8-band	1.825	2.195	1.272	2.926	0.063
16-band	2.187	2.795	2.590	2.563	0.818

Table 4.3: The d ' values of perceptual data for male speakers.

	M504&M601	M614&M710	M214&M519	M509&M603	M409&M705
8-band	2.879	2.187	1.199	0.252	1.209
16-band	2.879	2.318	1.353	1.095	1.906

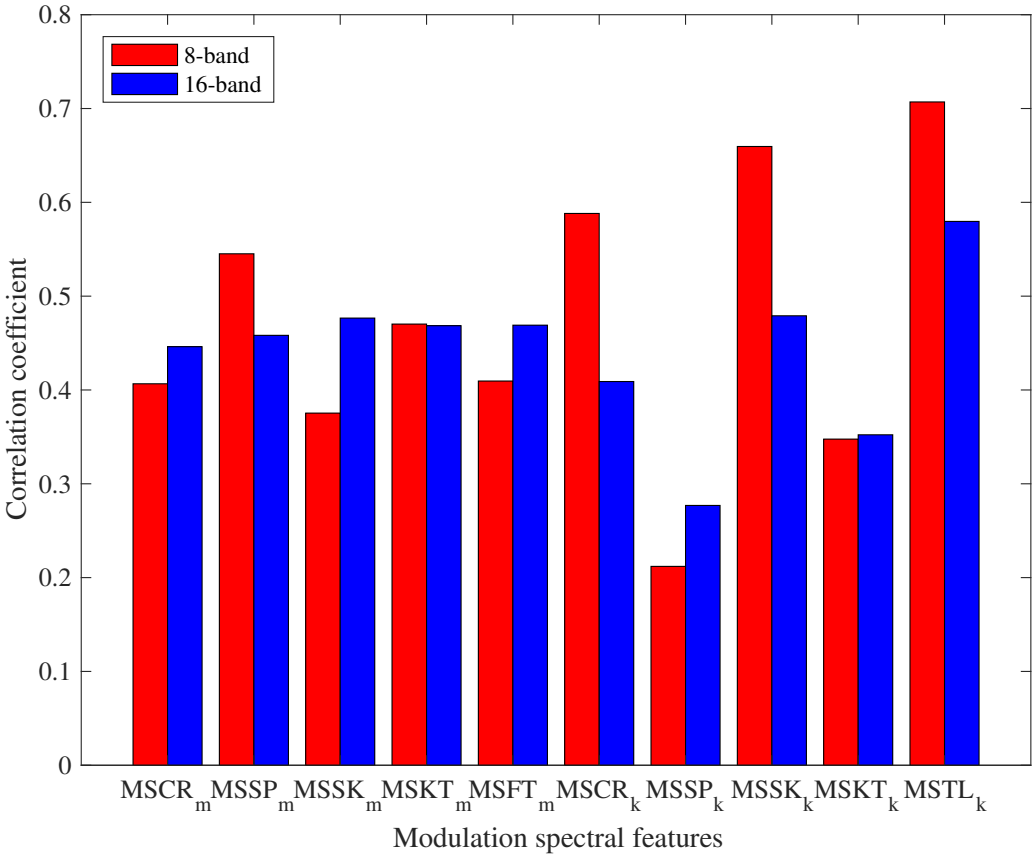


Figure 4.7: The correlation coefficients between the d' of modulation spectral features and the perceptual data for all speakers.

4.4 Modulation spectral features related to the perception of vocal emotion

4.4.1 The perceptual data of vocal-emotion recognition experiment

To discuss the relationship between modulation spectral features and the perception of vocal-emotion, the perceptual data collected in the vocal-emotion recognition experiments (section 3.4) were used. In this section, only the perceptual data on the condition that the upper limit of modulation frequency was 64 Hz were used. Figure 4.8 and table 4.4 show the perceptual used in this section and the d' values of the perceptual data. The d' values of the perceptual data were calculated based on the confusion matrices in Table A.9, A.18, and A.27. The average recognition rate decreased when the number of bands decreased, and the results of joy were mostly effected by the number of bands. In addition, the average recognition rates of sadness and hot anger were higher than the other three emotions.

Table 4.4: The d' values of the perceptual data on the condition that the upper limit of modulation frequency was 64 Hz.

	Neutral	Joy	Cold Anger	Sadness	Hot Anger
4-band	1.314	1.266	0.908	2.737	1.904
8-band	1.438	1.863	1.090	2.905	2.356
16-band	2.366	3.457	1.968	3.349	2.996

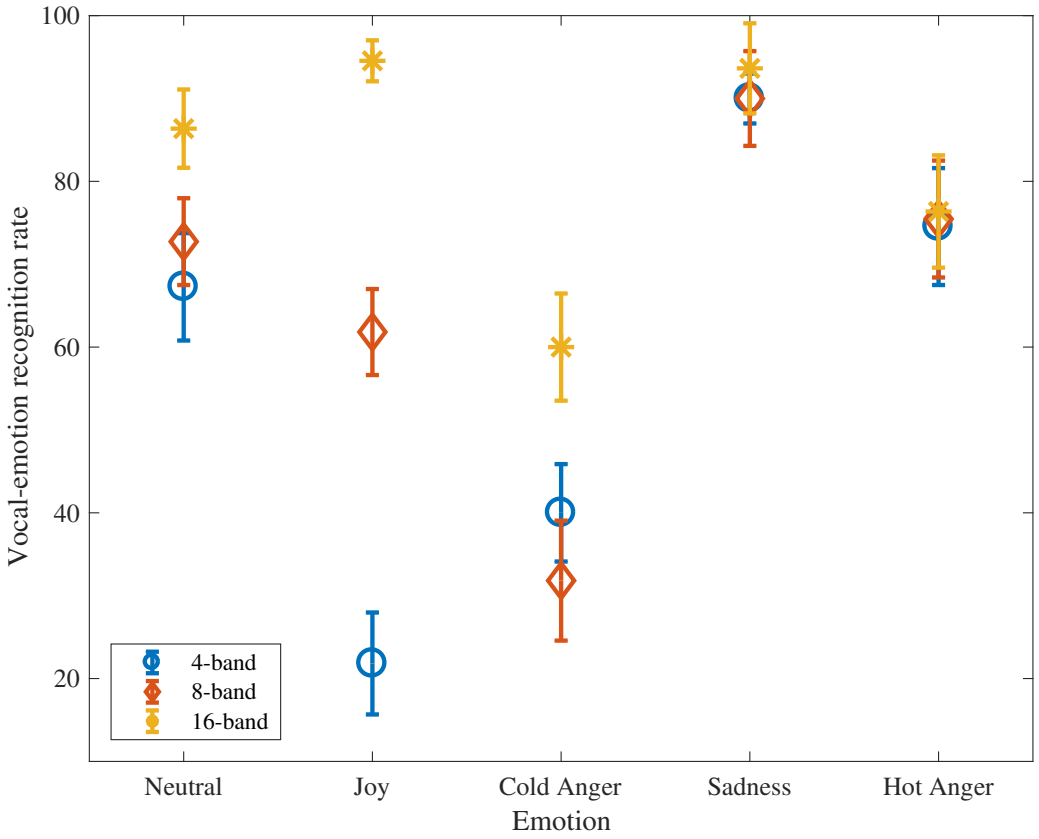


Figure 4.8: The results of vocal-emotion recognition experiment on the condition that the upper limit of modulation frequency was 64 Hz.

4.4.2 The modulation spectrogram of vocal-emotion speech

Before the discussion of modulation spectral features, in this section, the modulation spectrogram of emotional speech is discussed firstly. The modulation spectrogram of the emotional speech data used in section 3.4 was calculated. Figure 4.9-4.13 show the examples of the time-average modulation spectrogram of the speech with 5 different emotions in the Fujitsu database. The results show that different emotion has different characteristic on the time-averaged modulation spectrum, suggesting they could be well discriminated from each other. Compared to Neutral, Sadness has significantly more low acoustic frequency energy as Sadness should be a less expressive emotion. To the contrary, Hot Anger and Joy both have more high acoustic frequency energy. However in the higher acoustic frequency bands beyond about 25, Joy has less energy than Hot-Anger. For Cold-Anger, the distribution of modulation spectrum in acoustic frequency is significantly low. In the next step, the modulation spectral features of emotional speech is extracted to discuss whether these features can be used to account for the perceptual data obtained from the vocal-emotion recognition experiments in figure 4.8.

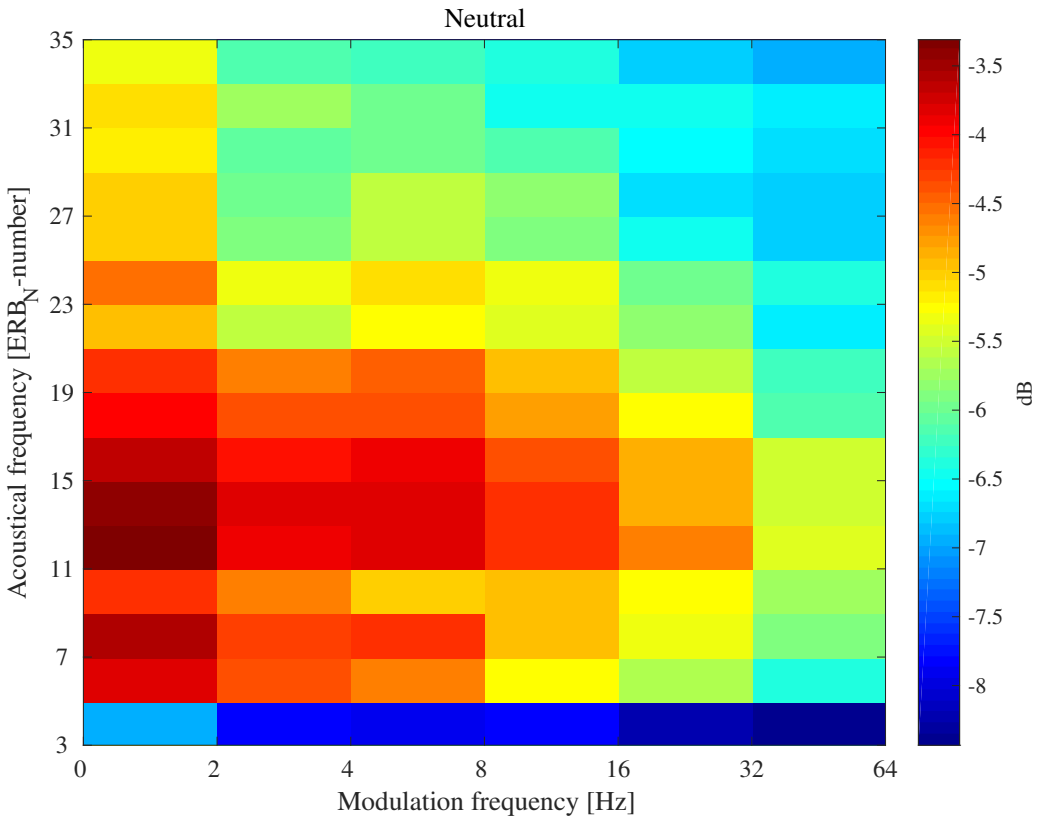


Figure 4.9: The time averaged modulation spectrogram of a neutral speech data with 16-bands.

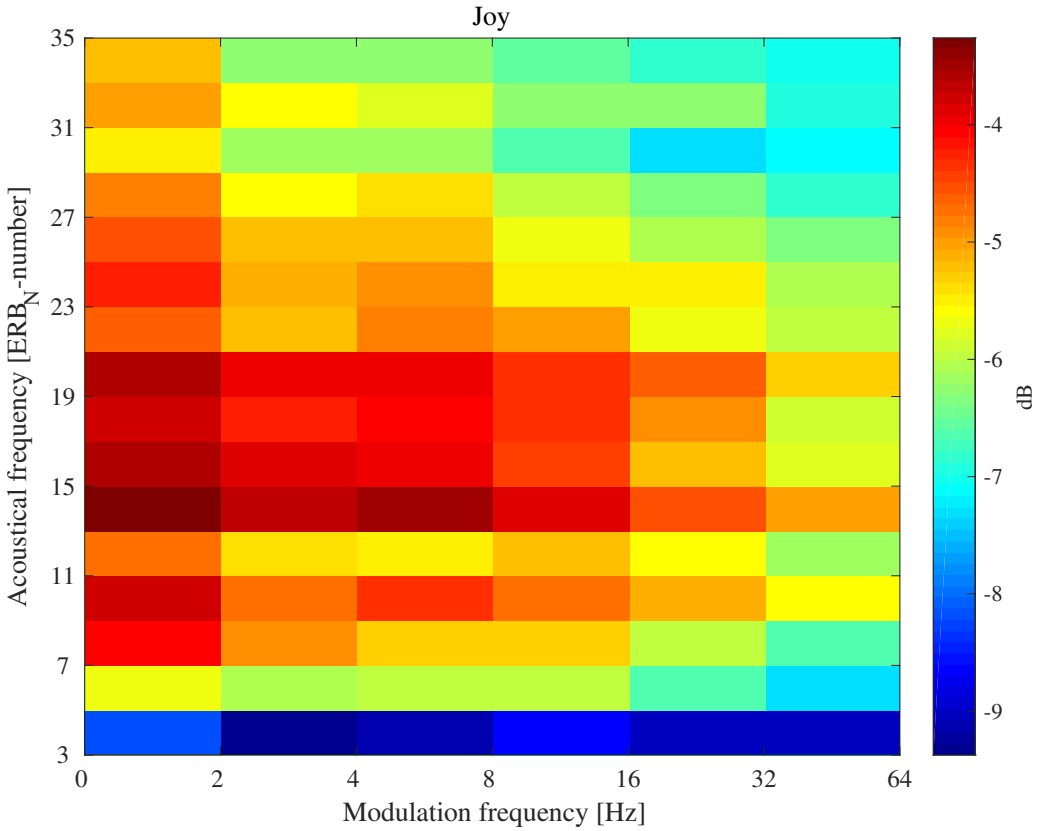


Figure 4.10: The time averaged modulation spectrogram of a joy speech data with 16-bands.

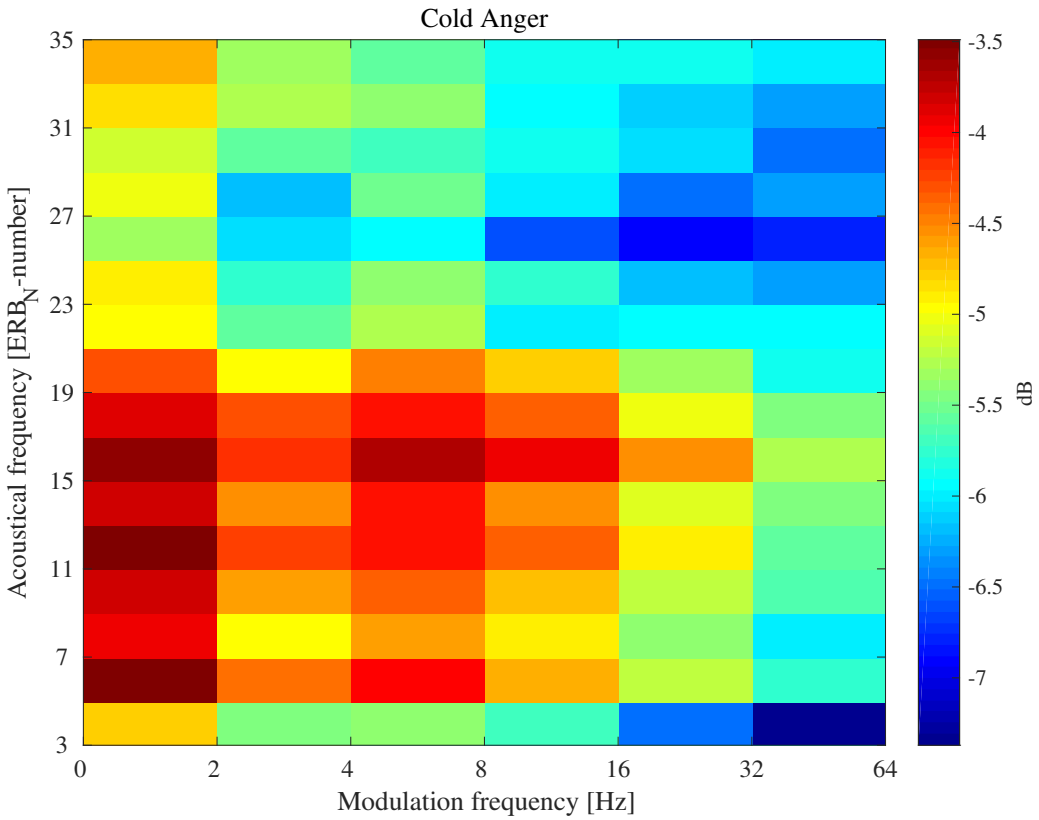


Figure 4.11: The time averaged modulation spectrogram of a cold anger speech data with 16-bands.

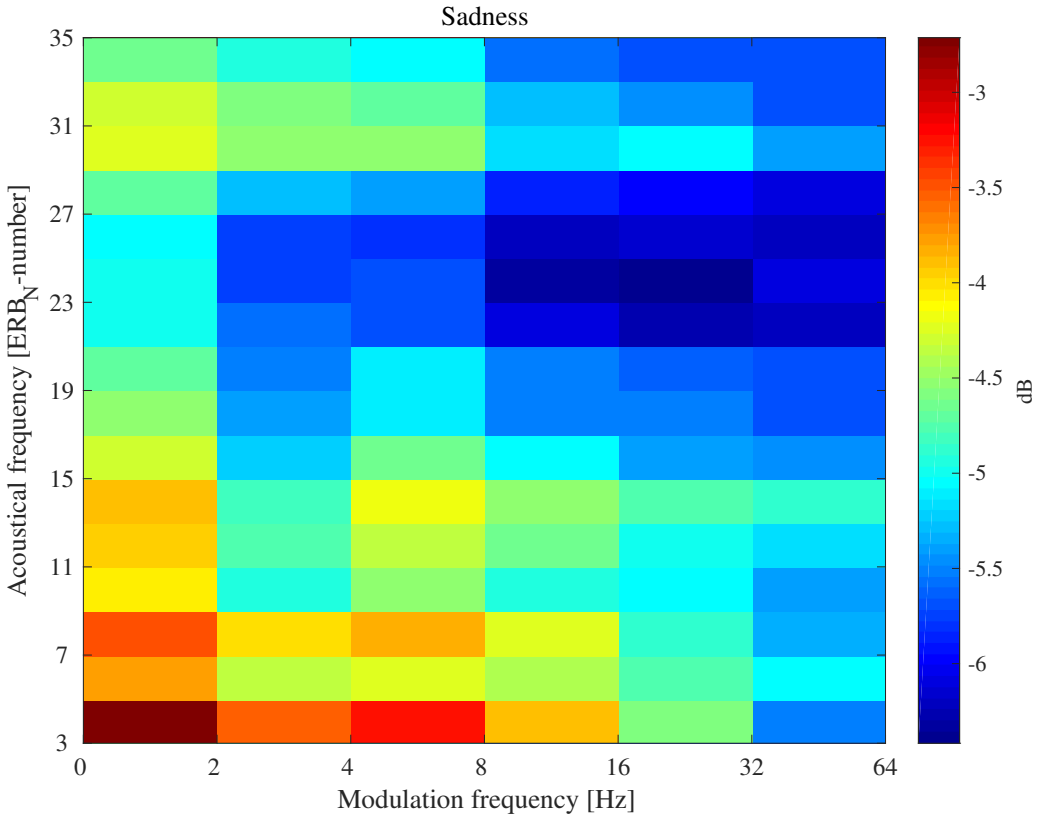


Figure 4.12: The time averaged modulation spectrogram of a sadness speech data with 16-bands.

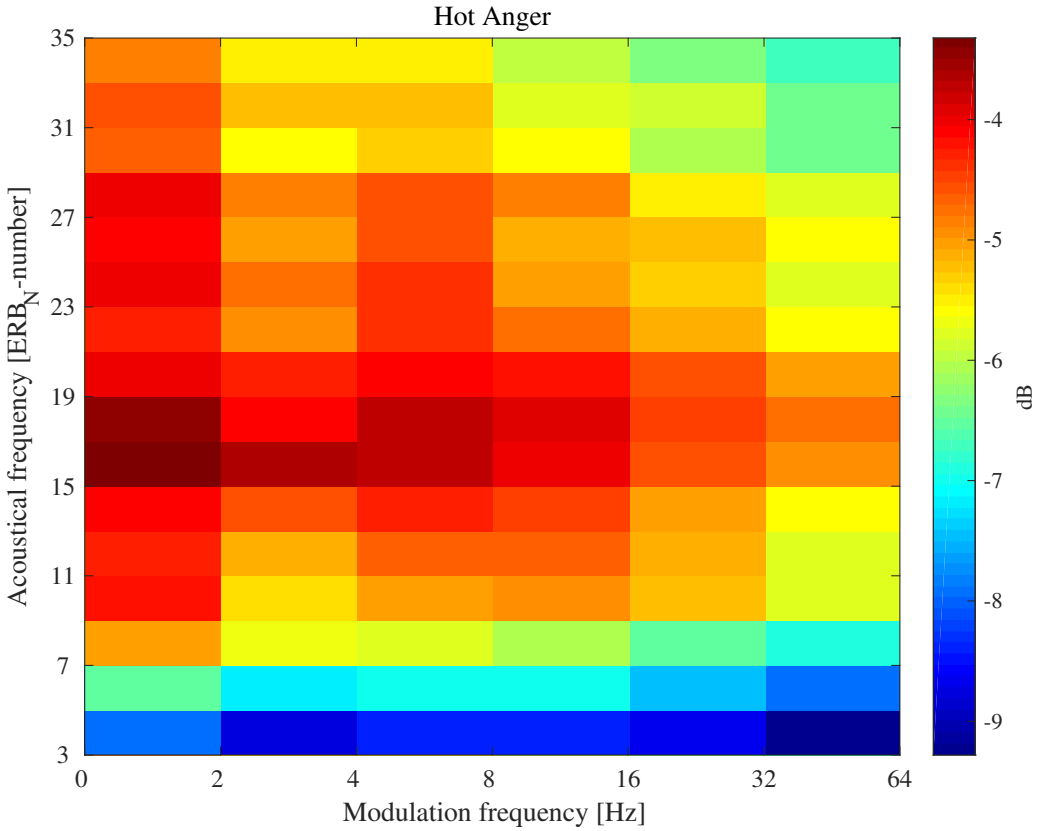


Figure 4.13: The time averaged modulation spectrogram of a hot anger speech data with 16-bands.

4.4.3 The correlation between the perceptual data and modulation spectral features

In this section the modulation spectral features of emotional speech data were calculated by the method described in section 4.2 to investigate those that may account for the perceptual data of human perception. The results of the vocal-emotion recognition experiment showed that participants achieved better performance with sadness and hot anger stimuli and there was a significant effect of the type of emotion on emotion recognition. It is necessary to discuss the modulation spectral features depend on different emotions. The averaged d' of modulation spectral features of different emotional speech were calculated.

Figure 4.14 shows the correlation coefficients between the d' of modulation spectral features and the perceptual data (table 4.4). The scatterplots of the d' of MSFs and perceptual data of vocal-emotion recognition experiments are shown in Appendices D. Except MSFT_m in the condition of 4-band and 8-band, the correlation of all other modulation spectral features are positive. In the condition of 4-band, the correlation coefficients of MSSP_m, MSKT_m, MSCR_k, and MSTL_k are close to 1. The modulation spectral features on the modulation frequency domain are higher than that on the acoustic frequency domain roughly. Moreover, the correlation increased with the decreasing of the number of channels. These results suggest that these modulation spectral features may be important cues for vocal emotion recognition with NVS. The contribution of modulation spectral features may increase with the decreasing of the number of channels of NVS.

Moreover, figure 4.15, 4.16, and 4.17 show the the correlation coefficients between the d' of modulation spectral features of emotional speech. The results showed that there were high correlation between MSCR_m, MSSK_m, and MSFT_m. Different from the results in figure 4.4, there were also high correlation between MSSP_m, MSKT_m and the MSFs in modulation frequency domain. The correlation of all the MSFs in modulation frequency domain were high.

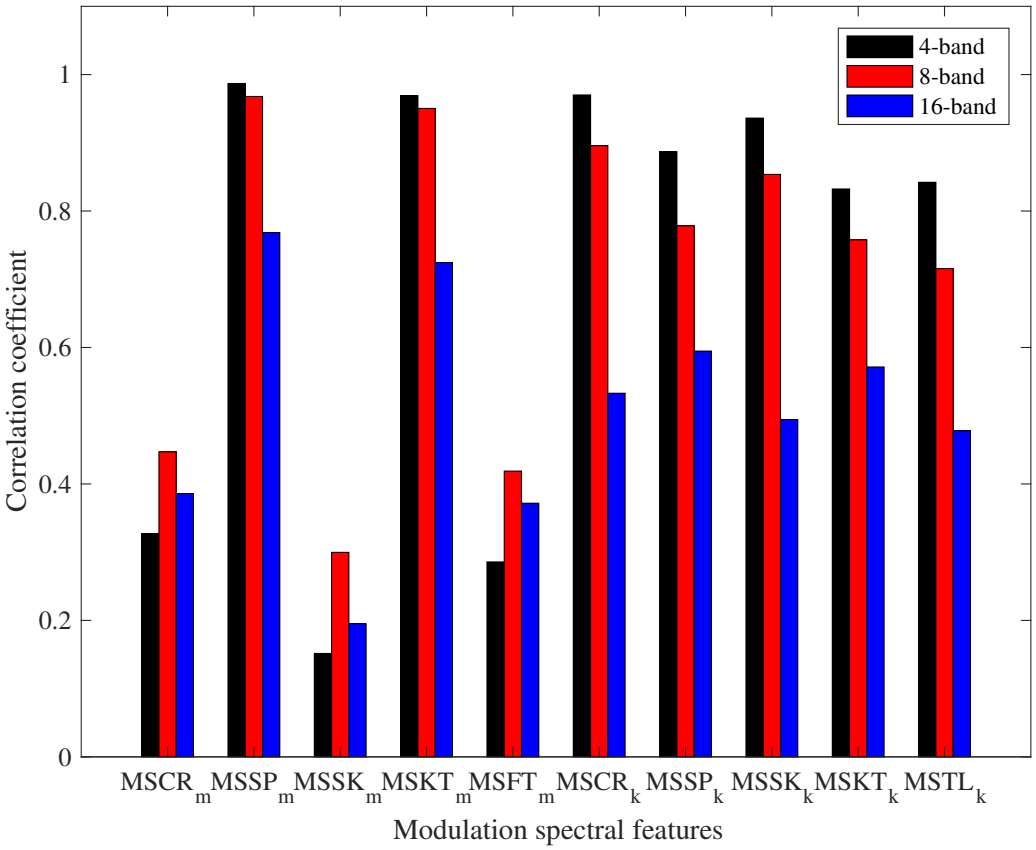


Figure 4.14: The correlation coefficients between modulation spectral features and the perceptual data of vocal-emotion recognition experiments.

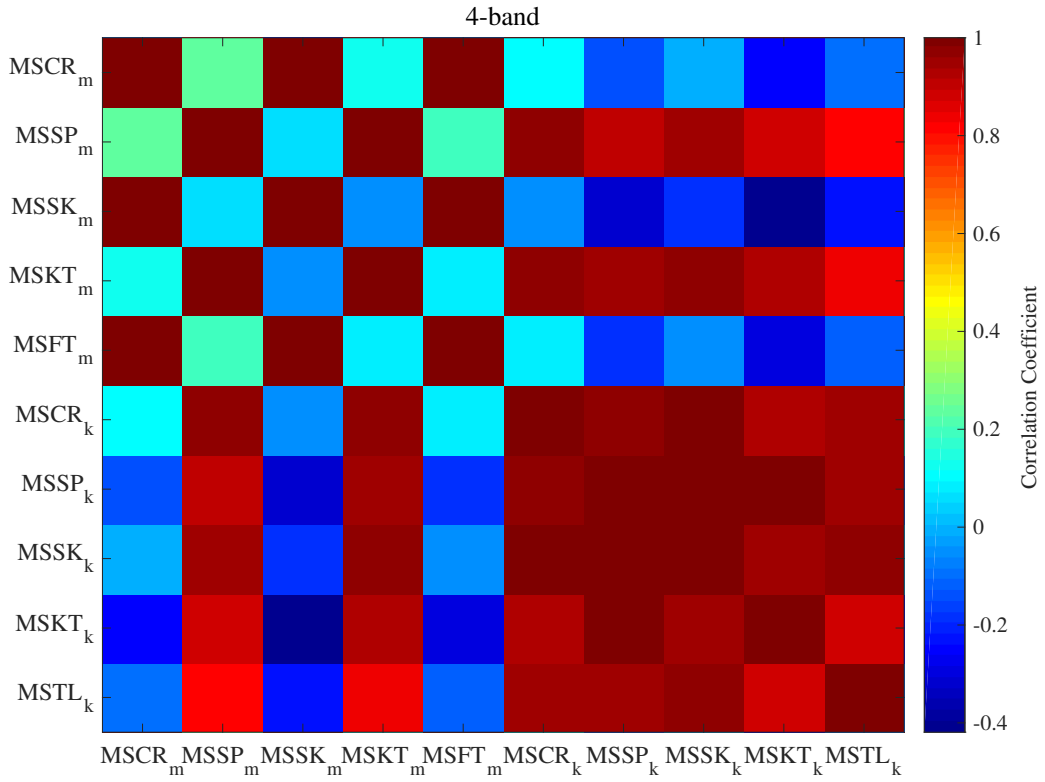


Figure 4.15: The correlation coefficients between the d' of 4-band modulation spectral features of emotional speech.

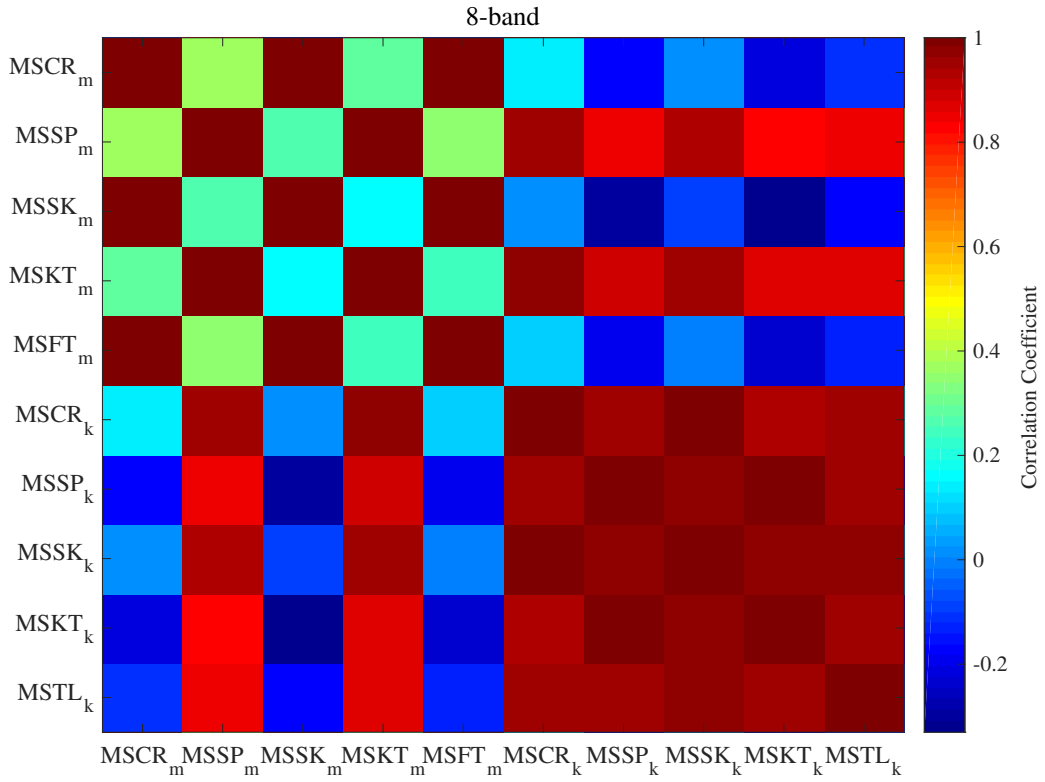


Figure 4.16: The correlation coefficients between the d' of 8-band modulation spectral features of emotional speech.

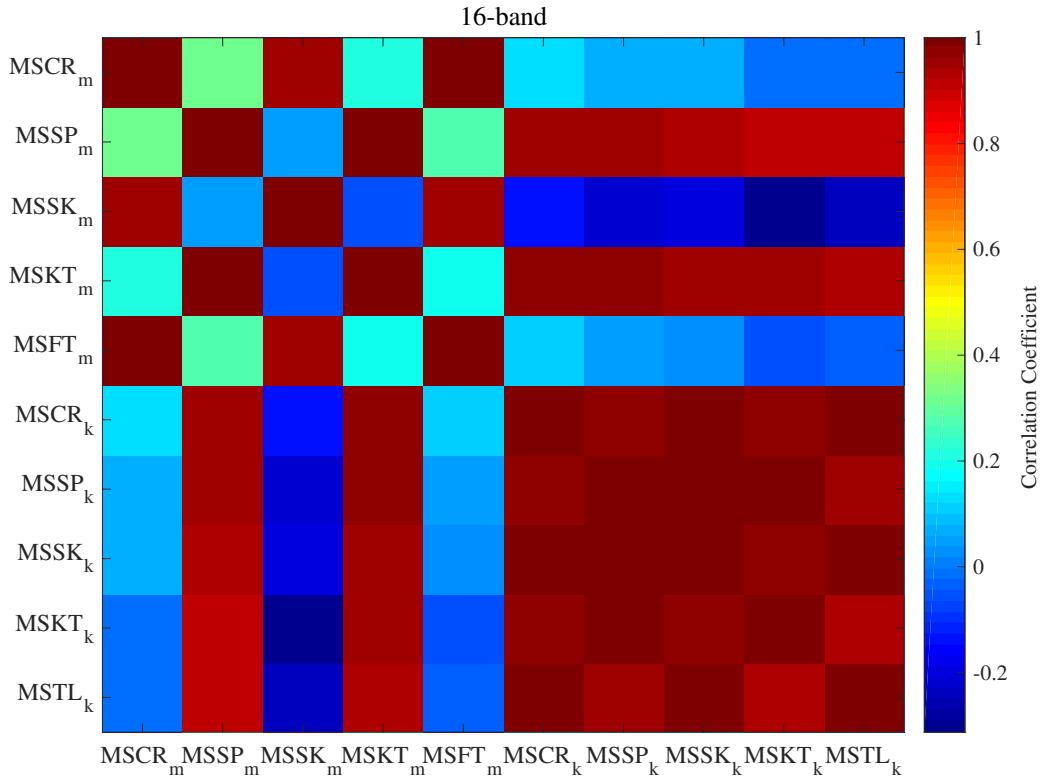


Figure 4.17: The correlation coefficients between the d' of 16-band modulation spectral features of emotional speech.

4.4.4 Discussion

Modulation spectral features on acoustic frequency domain

Both MSSP_m and MSKT_m had high correlation with the perceptual data. The MSSP_m can represent the spread of the spectrum around its MSCR_m as the 2nd moment. The MSKT_m gives a measure of the peakedness of the spectrum which was calculated from the 4th order moment. Therefore, there was also a high correlation between MSSP_m and MSKT_m.

From the scatter of MSSP_m for 4-band NVS (figure D.2), sadness stimuli had the highest d' value of both perceptual data and MSSP_m. Table A.9 shows the confusion matrix for 4-band and 64 Hz stimuli. The results showed that sadness had the highest hit rate and the false alarm rate is very low. Therefore the d' value of perceptual data of sadness stimuli was highest. From the figures of modulation spectrogram (figure 4.9 - 4.13), it is shown that the modulation spectrogram of sadness stimuli was so different from the other emotions. On the acoustic frequency domain, the power of sadness stimuli concentrated on the low frequency bands. The spread and peakedness of sadness stimuli were lower that other emotions.

Hot anger stimuli had the second highest d' value of both perceptual data and modulation spectral features. From the confusion matrix, hot anger stimuli also had the second highest hit rate. For the false alarm rates, 35 % of joy stimuli were recognized as hot anger. The modulation spectrogram of hot anger and joy were also similar, except that hot anger had more high frequency components.

Then neutral, joy, and cold anger stimuli had lower d' values. For modulation spectrogram these three emotions were more similar to each other than sadness and hot anger. For the perceptual data, the neutral stimuli had high hit rate, however the false alarm rate was also high. Joy stimuli had low false alarm rate, however the hit rate was also low. The hit rate and false alarm rate of cold anger stimuli were in the middle of neutral and joy stimuli. Such facts resulted the low d' values of neutral, joy and cold anger stimuli.

Modulation spectral features on modulation frequency domain

For the modulation spectral features on modulation frequency domain, all features had high correlation with the perceptual data. For neutral, joy, and cold anger, the distribution

of modulation spectrogram on modulation frequency domain were still very similar. The modulation spectrogram of sadness stimuli showed that the modulation components on high modulation frequency bands (8 - 64 Hz) were much lower than the other emotions. The reason should be that the speaker speaking sadness speech data very slow. Therefore the temporal envelope of sadness should be more smooth that the other emotions. Such facts resulted the highest d' values of sadness stimuli of the modulation spectral features on modulation frequency domain.

To the contrary, hot anger stimuli had much more high modulation frequency components. It should be related to that the speaker speaking hot anger speech data very fast. The speech rate was not the only reason, because the speech rate of joy stimuli was also high. However, comparing with hot anger, the high modulation frequency components of joy stimuli were lower. Another reason should be that the hot anger speech was spoken more roughly and the high modulation frequency components were related to the roughness of speech. Therefore, for the perception of hot anger speech, the high modulation frequency components were shown to be an important factor.

4.5 General discussion

For the results of speaker individuality (figure 4.5 and 4.6), the values of correlation coefficient were different with different conditions. The d ' values of modulation spectral features decreased with the increasing of perceptual speaker similarity of speaker pairs. However, the perceptual data showed that the speaker distinction rates did not decreased with the increasing of speaker similarity monotonically. A possibility reason may be that the relationship between the modulation spectral feature and perceptual data is not linear. Moreover, the number of speaker pairs may not be large enough. Kitamura et al. measured the perceptual similarity of total 380 speaker pairs [1]. speaker distinction experiments with more speaker pairs are necessary to obtain more general role of the modulation spectral features in the perception of speaker individuality. The results showed that there are positive correlations between the modulation spectral features and perceptual data. These results have shown the potential possibility of modulation spectral features for speaker individuality analysis.

For the results of vocal-emotion (figure 4.14), the correlations between the modulation

spectral features and perceptual data were higher than that of speaker distinction experiment (figure 4.5 and 4.6). The results suggested the potential of modulation spectral features for vocal emotion analysis. The effect of modifying modulation spectral features on vocal emotion recognition is needed to be investigated to clarify whether these features can contribute to the perception of vocal emotion. Moreover, the variation in modulation spectral features in the time domain should be discussed in detail. Since human perception of speaker individuality and vocal-emotion may not depend on just one single feature, the interaction of modulation spectral features should also be further discussed.

4.6 Summary

In this chapter, the relationship between the modulation spectral features and perceptual data was investigated to clarify the contribution of modulation spectral features on the perception of speaker individuality and vocal-emotion. Ten types of modulation spectral feature were extracted from the modulation spectrogram of speech data. The correlation between the modulation spectral features and the perceptual data obtained from speaker or vocal-emotion recognition experiments was calculated.

For speaker individuality, there were positive correlations between the modulation spectral features and the perceptual data of speaker distinction experiment. Similarity results were also obtained from the results of vocal emotion, however, the correlations were roughly higher than that of speaker distinction experiments. The results showed that the modulation spectral features were useful to account for the perceptual data of speaker and vocal-emotion recognition experiments. It was suggested that modulation spectral features could be important cues contribute to the perception of speaker individuality and vocal-emotion with NVS.

Chapter 5

Discussion of the application of temporal modulation information

So far, it is confirmed that the temporal envelope of speech contributes to the perception of nonlinguistic information. Moreover, modulation spectral features are suggested to be important cues contribute to the perception of nonlinguistic information. The temporal modulation information has been proved to play an important role in the perception of not only linguistic but also nonlinguistic information. In this chapter, two kinds of applications of the temporal modulation information are discussed.

At first, the feasibility of using NVS to simulate CI listeners' response in vocal emotion recognition is discussed. In section 4.4, the high correlation of the modulation spectral features and the perceptual data obtained from the vocal-emotion recognition experiments using NVS with NH listeners showed that the modulation spectral features could be used to account for the perceptual data of vocal-emotion perception. NVS is known as a CI simulation to simulate the poor spectral cue provided by CI device. The fact that modulation spectral features are useful to account for the perceptual data of vocal-emotion recognition experiments using NVS with NH listeners shows that the modulation spectral features could also be used to account for the perceptual data from CI listeners. A vocal-emotion recognition experiment was carried out to confirm this concept.

Then the effect of the modification of modulation spectrogram on the vocal emotion recognition with noise-vocoded speech is discussed. A method based on a linear prediction scheme is proposed to modify the modulation spectrogram and its features of neutral

speech to match that of emotional speech. The logic of this approach is that if vocal emotion perception of CI simulation is based on the modulation spectral features, NVS with similar modulation spectral features of emotional speech will be recognized as the same emotion.

5.1 Feasibility of using noise-vocoded speech to simulate cochlear implant listeners' response in vocal emotion recognition

5.1.1 Intoduction

It has been known that CI listeners' performances of vocal-emotion are poorer than normal-hearing (NH) listeners, as the poor spectral cues provided by CI device [23, 24, 27, 28]. Luo *et al.* showed that vocal-emotion recognition of NH listeners using NVS was significantly improved as the cut-off frequency of modulation low-pass filter was increased from 50 to 500 Hz [28]. The modulation frequency bands between 50 and 500 Hz mainly included the periodic information related to F0 [21]. However, the contribution of the temporal cue defined as the modulation frequency band below 50 Hz is still unknown. By comparing the performances of vocal-emotion recognition by CI listeners and HN listeners using NVS, Chatterjee *et al.* [27] found that the mean performance of CI listeners was similar to that of NH listeners with 8-band NVS. Chatterjee *et al.* then analyzed the F0, intensity, and duration of stimuli. However, it was found that, the acoustic analyses could not account for all of the perceptual data of the vocal-emotion recognition experiment with NVS.

In section 4.4, the high correlation of the modulation spectral features and the perceptual data obtained from the vocal-emotion recognition experiments using NVS with NH listeners showed that the modulation spectral features could be used to account for the perceptual data of vocal-emotion perception. The fact that modulation spectral features are useful to account for the perceptual data of vocal-emotion recognition experiments using NVS with NH listeners shows that the modulation spectral features could also be used to account for the perceptual data from CI listeners.

In this section, vocal-emotion recognition experiments using NVS with NH and CI listeners were carried to confirm that whether NVS can be used to simulate CI listeners' response in vocal emotion recognition. The feasibility of using modulation spectral features to account for the perceptual data from CI listeners is also discussed by comparing the results of NH and CI listeners.

5.1.2 Method

Stimuli

The same Fujitsu Japanese Emotional Speech Database used in section 3.4 was used in these experiments. This database includes five emotions (*neutral, joy, cold anger, sadness, and hot anger*) expressed by one professional actress. The same sentence was spoken with five emotions and ten utterances of each emotion were selected. The linguistic contents of each sentence were semantically emotion-neutral to minimize any biasing effect of context. The duration of each utterance was about 3 or 4 s. The sampling frequency and quantization bits were 22.05 kHz. and 16 bits.

The original emotional speech and the NVS generated from the emotional speech were used as stimuli. The method to generate NVS stimuli is described in section 3.2. However, the conditions of NVS are only 8- and 16-band with 64 Hz upper limit of modulation frequency.

Participants

3 CI listeners participated in this experiment. Tabel 5.1 shows the detailed information about the CI listeners. The averaged age of CI listeners was 18. Previous study showed that a strong developmental effect was observed in the NH listeners with NVS in vocal-emotion recognition [27]. For this reason, 9 NH high school students (5 males and 4 females) participated in this experiment. The averaged age of NH listeners was about 17.

Procedure

There were 2 NVS conditions with 2 different numbers of channels (8, and 16) and the upper limits of modulation frequency was 64 Hz. The original speech was also presented. All stimuli were randomly presented to the participants during the experiment. Participants

Table 5.1: Detailed information about the CI listeners, Mean ATH is the mean absolute threshold of hearing of the ear using CI.

ID	Age	Gender	Manufacturer/device	Mean ATH [dB]
CI01	17	Female	Cochlear Ltd./Nucleus 6	40
CI02	16	Male	Cochlear Ltd./Nucleus 6	30
CI03	21	Male	Cochlear Ltd./Nucleus Freedom	40

were asked to indicate which of the five emotions (*neutral, joy, cold anger, sadness, and hot anger*) he/she thought was associated with the stimulus. Each stimulus was presented only once.

For NH listeners, the experimental environment was as same as that in section 3.3 For CI listeners, the experiment was also conducted while the participants were in a sound-proof room. The sound pressure level of background noise was lower than 25.8 dB. The stimuli were simultaneously presented to a participant through a PC, audio interface (RME, Fireface UCX), a power amplifier (YAMAHA, A-U671), and two speakers (YAMAHA, NS-pf7). The sound pressure levels were calibrated to be the same for all participants by using a head and torso simulator (B&K, type 4128) and sound level meter (B&K type 2231).

5.1.3 Results

Figure 5.1 shows the results of vocal-emotion recognition experiment for NH listeners. For original emotional speech, the recognition rates are all 100 %. Therefore, it is confirmed that NH participants can perceive the vocal emotion information from the original emotional speech successfully. For NVS, similar to the results in 4.4, NH listeners performed better on the Sadness and Hot Anger NVS than the other emotions. The results of Joy NVS were also mostly affected by the number of channels.

Figure 5.2 shows the results of vocal-emotion recognition experiment for CI listeners. The results showed that CI listeners could not recognize the emotion of even original emotional speech successfully. The results for CI listeners revealed that they recognized sadness and hot anger more easily than joy and cold anger in both original emotional speech and NVS conditions.

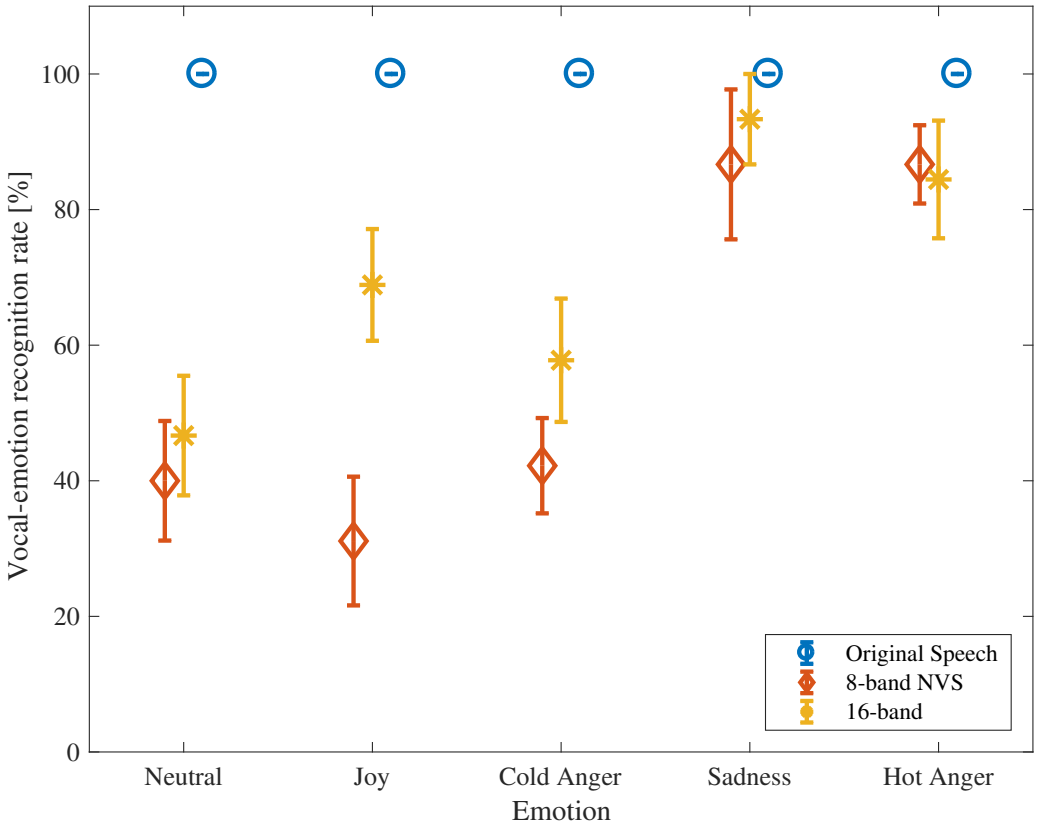


Figure 5.1: The results of vocal-emotion recognition experiment for NH listeners.

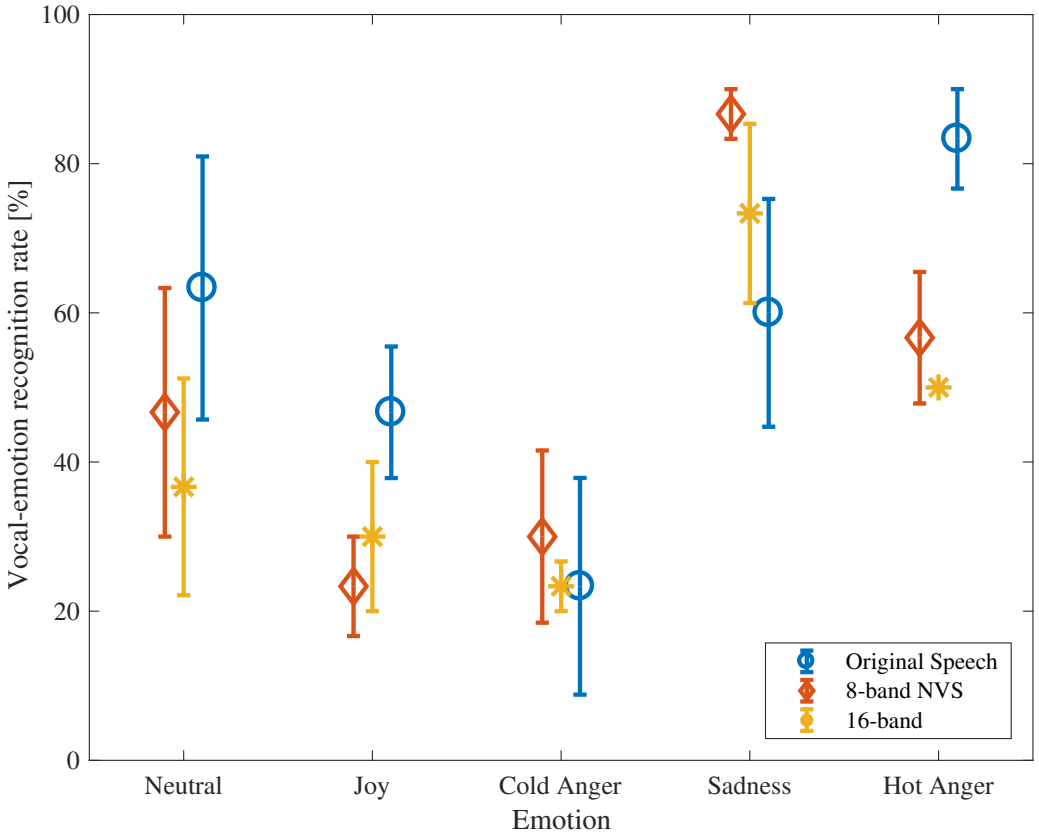


Figure 5.2: The results of vocal-emotion recognition experiment for CI listeners.

Table 5.2: Mean confusion matrix with 8-band NVS stimuli for CI listeners.

	Neutral	Joy	Cold Anger	Sadness	Hot Anger
Neutral	0.47	0.07	0.33	0.10	0.03
Joy	0.27	0.23	0.27	0.03	0.20
Cold Anger	0.53	0.03	0.30	0.13	0
Sadness	0.07	0.03	0.03	0.87	0
Hot Anger	0.03	0.07	0.33	0	0.57

Table 5.3: Mean confusion matrix with 16-band NVS stimuli for CI listeners.

	Neutral	Joy	Cold Anger	Sadness	Hot Anger
Neutral	0.37	0.17	0.33	0.10	0.03
Joy	0.17	0.30	0.33	0.07	0.13
Cold Anger	0.50	0.03	0.23	0.23	0
Sadness	0.17	0	0.10	0.73	0
Hot Anger	0	0.17	0.30	0.03	0.50

Table 5.4: Mean confusion matrix with original emotional speech for CI listeners.

	Neutral	Joy	Cold Anger	Sadness	Hot Anger
Neutral	0.63	0.03	0.27	0.07	0
Joy	0.20	0.47	0.13	0.10	0.10
Cold Anger	0.67	0	0.23	0.10	0
Sadness	0.10	0.07	0.23	0.60	0
Hot Anger	0.07	0.03	0.07	0	0.83

Table 5.5: Mean confusion matrix with 8-band NVS stimuli for NH listeners.

	Neutral	Joy	Cold Anger	Sadness	Hot Anger
Neutral	0.40	0.18	0.20	0.02	0.20
Joy	0.04	0.31	0.13	0.07	0.44
Cold Anger	0.27	0.04	0.42	0.20	0.07
Sadness	0	0.04	0.04	0.87	0.04
Hot Anger	0.02	0.07	0.02	0.02	0.87

Table 5.6: Mean confusion matrix with 16-band NVS stimuli for NH listeners.

	Neutral	Joy	Cold Anger	Sadness	Hot Anger
Neutral	0.47	0.13	0.16	0.13	0.11
Joy	0.04	0.69	0.04	0.07	0.16
Cold Anger	0.22	0.02	0.58	0.13	0.04
Sadness	0	0	0.02	0.93	0.04
Hot Anger	0.04	0	0.11	0	0.84

5.1.4 Discussion

Table 5.2 - 5.6 show the confusion matrices in each condition. There was a common trend that the Cold Anger stimuli were recognized as Neutral speech. For NH listeners, Joy stimuli were recognized as Hot Anger speech. However, for CI listeners, the selection for Joy stimuli was more random.

Figure 5.3 shows the results of the average vocal-emotion recognition rate for each CI listeners. By comparing the results with the Mean ATH of each CI listeners, it is suggested the CI listener with lower ATH will performs better at vocal emotion recognition.

The purpose of these experiments is to discuss whether NVS can be used to simulate CI listeners' response in vocal emotion recognition. Figure 5.4 shows the averaged vocal-emotion recognition rates of NH and CI listeners. At first, there was no remarkable difference of the results of CI listeners with different conditions. It is suggested that results of 8-band NVS with NH listeners can simulate the response of CI listeners better.

Moreover, in the condition of both 8-band NVS with NH listeners and original speech with CI listeners, there was a similar trend that the participants performed better on Sad-

ness and Hot Anger than the other emotions. This trend was also appeared in the results in section 4.4 which was the point to elucidate that the modulation spectral features can be used to account for the perceptual data obtained from the vocal-emotion recognition experiments with NVS. Therefore, the results showed that the modulation spectral features can also be used to account the performance of CI listeners in the vocal-emotion recognition. The modulation spectral features may also play an important role in the perception of vocal-emotion by CI listeners.

5.1.5 Summary

In this section, vocal-emotion recognition experiments with NH and CI listeners were carried out to clarify whether CI listeners can perceive vocal emotion the same way as NH listener with NVS do. The results for CI listeners revealed that they recognized sadness and hot anger more easily than joy and cold anger in both original emotional speech and NVS conditions. Moreover, the results for NH listeners with NVS showed the same trend. The results suggested that the vocal-emotion recognition paradigm using NVS can be used to investigate vocal emotion recognition by CI listeners. Therefore, the modulation spectral features can also be used to account the performance of CI listeners in the vocal-emotion recognition.

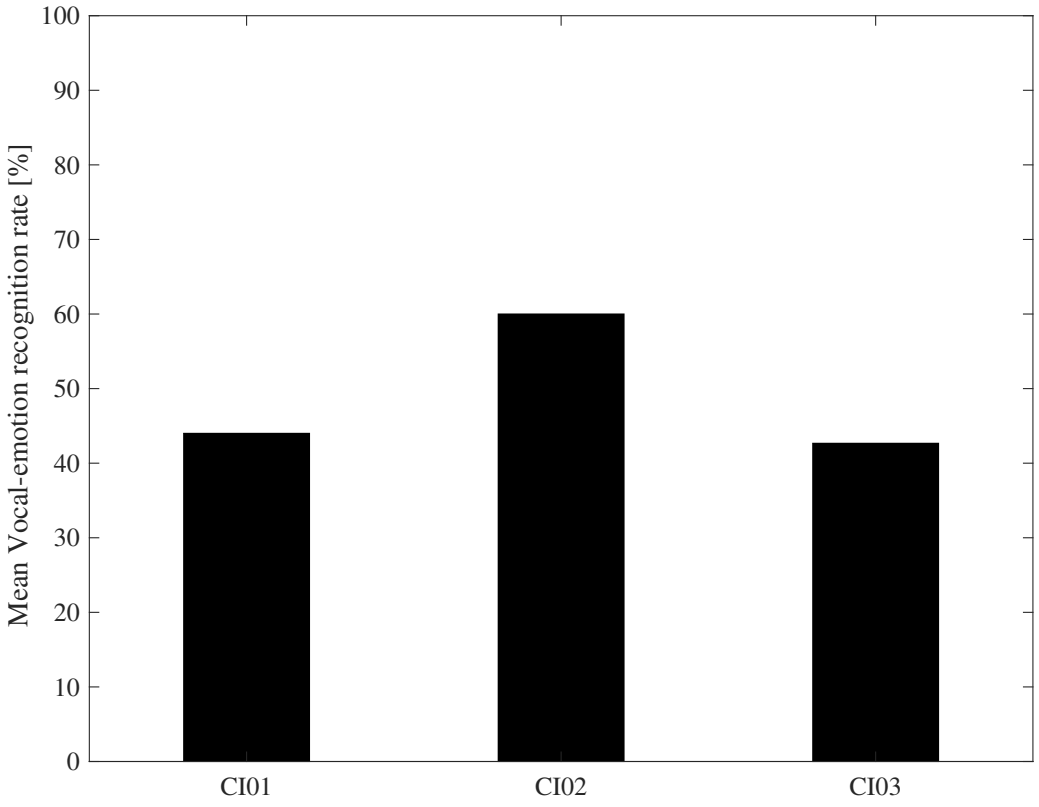


Figure 5.3: The average vocal-emotion recognition rate for each CI listener.

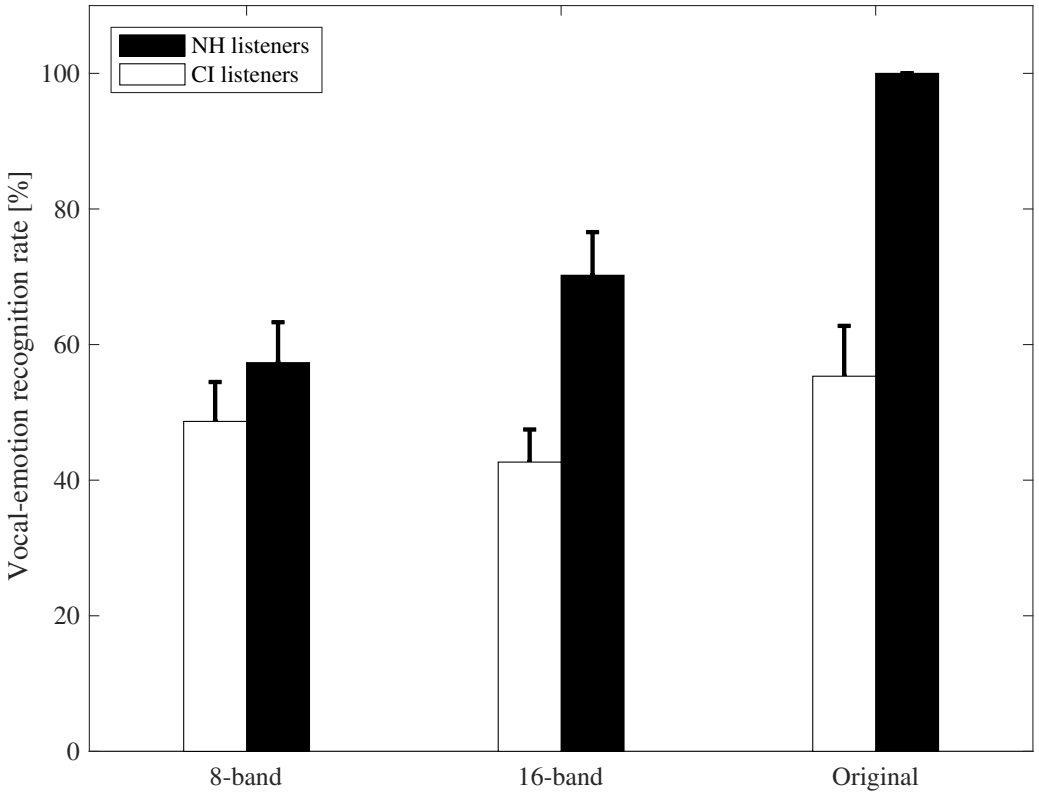


Figure 5.4: The averaged vocal-emotion recognition rates of NH and CI listeners.

5.2 Effect of the modification of modulation spectrogram on the vocal-emotion recognition with noise-vocoded speech

5.2.1 Introduction

The relationship between the modulation spectral features of the temporal envelope and human perception of vocal-emotion with NVS is discussed in section 4.4. The results showed that sadness and hot anger are more easily recognized than joy and cold anger with simulated CIs. Similar trends were also shown from experiments with CI listeners. High correlations between modulation spectral features and the perception of vocal emotion based on the NVS scheme were found. These results suggested that the modulation spectrogram of speech should be an important cue for voice emotion recognition with simulated CIs.

This section aims to study the feasibility of vocal emotion conversion on a modulation spectrogram for simulated CIs. Luo and Fu successfully enhanced the tone recognition on the NVS scheme by manipulating the amplitude envelope to more closely resemble the F0 contour [83]. Their results showed the possibility of enhancing the recognition of non-linguistic information by modifying the temporal envelope. It is also found that the sound texture can be converted successfully by modifying the modulation spectrogram [84].

In this section, a method based on a linear prediction (LP) scheme is proposed to modify the modulation spectrogram and its features of neutral speech to match that of emotional speech. The logic of this approach is that if vocal emotion perception of CI simulation is based on the modulation spectral features, NVS with similar modulation spectral features of emotional speech will be recognized as the same emotion.

In the process, the neutral speech is first divided into several bands using an auditory filterbank, and the temporal envelope of each band is extracted. Then, the temporal envelopes are modulation-filtered by using infinite impulse response (IIR) filters to modify the modulation spectrum from neutral to emotional speech. The IIR filters are derived from the relation of modulation characteristics of neutral and vocal emotions on a linear prediction scheme. On the acoustic frequency domain, the average amplitude of the

temporal envelope is corrected using the ratio of the average amplitude between neutral and emotional speech. Finally, a vocal-emotion recognition experiment using NVS generated by the converted temporal envelope is carried out. The method for enhancing the vocal-emotion information of the modulation spectrogram is also discussed further.

5.2.2 Vocal-emotion conversion on modulation spectrogram

In this section, the method of vocal emotion conversion on the modulation spectrogram as shown in Fig. 5.5 is described.

All emotional speech signals used in this study were selected from the Fujitsu Japanese Emotional Speech Database [9]. This database included five emotions (*neutral, joy, cold anger, sadness, and hot anger*) spoken by one female speaker. As the definition of cold anger is too ambiguous and not easily recognized, only neutral (NE), joy (JO), sadness (SA) and hot anger (HA) speech were used in this study.

Auditory-inspired band-pass filterbank and temporal envelope extraction

The performance of vocal emotion recognition by CI listeners was found to be similar to that of NH listeners with 8-band NVS [27]. Therefore, in this study, the speech signal was divided into 8 bands by an auditory-inspired band-pass filterbank as follows:

$$s(k,n) = h_{\mathrm{BPF}}(k,n) * s(n) \tag{5.1}$$

where $h_{\mathrm{BPF}}(k,n)$ is the impulse response of the band-pass filter in the kth band, "$*$" denotes the convolution operation, and n is the sample number in the time domain.

The auditory filterbank was constructed by using 3rd-cascaded 2nd-order Butterworth IIR filters. The bandwidth of the filter was designed as $\mathrm{ERB_N}$ (equivalent rectangular bandwidth), and all filters were placed on the $\mathrm{ERB_N}$-number scale [44]. $\mathrm{ERB_N}$-number is defined by the following equation,

$$\mathrm{ERB_N} - \mathrm{number} = 21.4 \log_{10}\left(\frac{4.37f}{1000} + 1\right) \tag{5.2}$$

where f is the acoustic frequency in Hz. This scale is comparable to a scale of distance along the basilar membrane so that the frequency resolution of the auditory system can be faithfully replicated by dividing the frequency bands according to $\mathrm{ERB_N}$-number. In

this study, the boundary frequencies of band-pass filters are spaced from 3 to 35 $\mathrm{ERB_N}$-numbers with 4 $\mathrm{ERB_N}$ as the bandwidth of the acoustic frequency region (8-bands).

Then, the temporal envelope of each band-limited signal was calculated by using the Hilbert transform and a low-pass filter.

$$e(k,n) = |s(k,n) + j\mathcal{H}[s(k,n)]| * h_{\mathrm{LPF}}(n) \tag{5.3}$$

where $\mathcal{H}$ denotes the Hilbert transform and $h_{\mathrm{LPF}}(n)$ is the impulse response of the low-pass filter. The low-pass filter was constructed by using a 2nd-order Butterworth IIR filter. The cut-off frequency of the low-pass filter was 64 Hz.

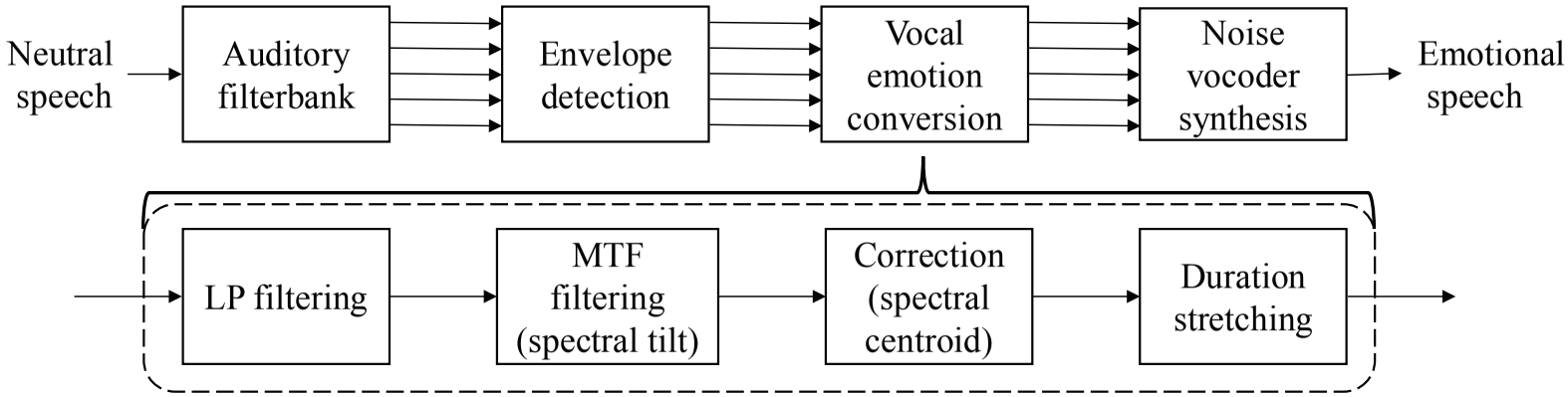


Figure 5.5: Scheme of LP based vocal emotion conversion method.

5.2.3 Vocal emotion conversion based on LP scheme

At first the temporal envelopes of the input signal were modulation-filtered by using IIR filters to modify the modulation spectrum from neutral to emotional speech. The transfer function of this IIR filter is represented as follows:

$$H_{\mathrm{LP}}(z) = \frac{\sum_{i=0}^{p} b_{NE,i} z^{-i}}{\sum_{i=0}^{p} a_{EM,i} z^{-i}} \tag{5.4}$$

where $b_{NE,i}$ and $a_{EM,i}$ are the linear prediction (LP) filter coefficients calculated from the envelope of neutral (NE) and target emotional (EM) speech and p is the order of filter. These LP coefficients are calculated by minimizing the linear prediction error in the least squares sense. The IIR filters were derived from the relation of modulation characteristics of neutral and vocal emotions on a LP scheme. From the preliminary experiments, the best performance of conversion was found when the order of LP filter p was 20. We found that the linguistic information will be destroyed when the order of the LP filter is higher than 20. But if the order is lower, the conversion of the modulation spectrum will not be enough. This process can also modify the modulation spectral kurtosis close to the target emotion. The process of LP filtering can be represented as follows:

$$\hat{e}_{\mathrm{LP}}(k,n) = e_{\mathrm{NE}}(k,n) * h_{\mathrm{LP}}(k,n) \tag{5.5}$$

where, $e_{\mathrm{NE}}(k,n)$ is the envelope of neutral speech, and $h_{\mathrm{LP}}(k,n)$ is the impulse response of the LP filter.

In the next step, we used a modulation transform function (MTF) filter (1st-order IIR filter) to modify the modulation spectral tilt of neutral speech close to the target emotion as follows:

$$\hat{e}_{\mathrm{MTF}}(k,n) = \hat{e}_{\mathrm{LP}}(k,n) * h_{\mathrm{MTF}}(k,n) \tag{5.6}$$

where $h_{\mathrm{MTF}}(k,n)$ is the impulse response of the 1st-order MTF filter. The frequency characteristics of this MTF filter are the best fits (in a least-squares sense) for the modulation spectrum of the target emotion. Then, the amplitude of the temporal envelope was corrected using the ratio of the average amplitude between emotional and neutral speech.

$$\hat{e}(k,n) = \hat{e}_{\mathrm{MTF}}(k,n) \frac{\bar{e}_{\mathrm{NE}}(k)}{\bar{e}_{\mathrm{EM}}(k)} \tag{5.7}$$

where $\bar{e}_{\mathrm{NE}}(k)$ and $\bar{e}_{\mathrm{EM}}(k)$ are the average amplitude of the envelope of neutral speech and the target emotional speech in the kth band. This process can modify the modulation

spectrogram on the acoustic frequency domain to shift the spectral centroid close to the target emotion.

Finally, a temporal stretching of the temporal envelopes based on the duration ratio of neutral to the target emotion was used to modify the duration. The amplitude of the converted temporal envelope in the interval in which the amplitude of the neutral speech is 40 dB smaller than the maximum was set to 0. This process aims to reduce the redundant components of the converted temporal envelope generated by the LP based conversion filtering. These redundant components will sound like reverberation of speech and destroy the linguistic information.

Figure 5.6 shows an example of the modulation spectrum of the converted temporal envelope. The target emotion is hot anger and the modulation spectrum in the 3rd channel is shown. The modulation spectrum is the amplitude spectrum of the temporal envelope calculated by the Fourier transform. The results show that the modulation spectrum of the converted temporal envelope (blue line) is very close to that of the target emotion (red line) from neutral speech (green line). Figure 5.7-5.9 show the modulation spectrograms of neutral, emotional speech, and converted speech. As a result, the shape of the modulation spectrogram of converted speech is similar to that of hot anger speech. That means the modulation spectrogram of neutral speech was successfully converted to that of emotional speech.

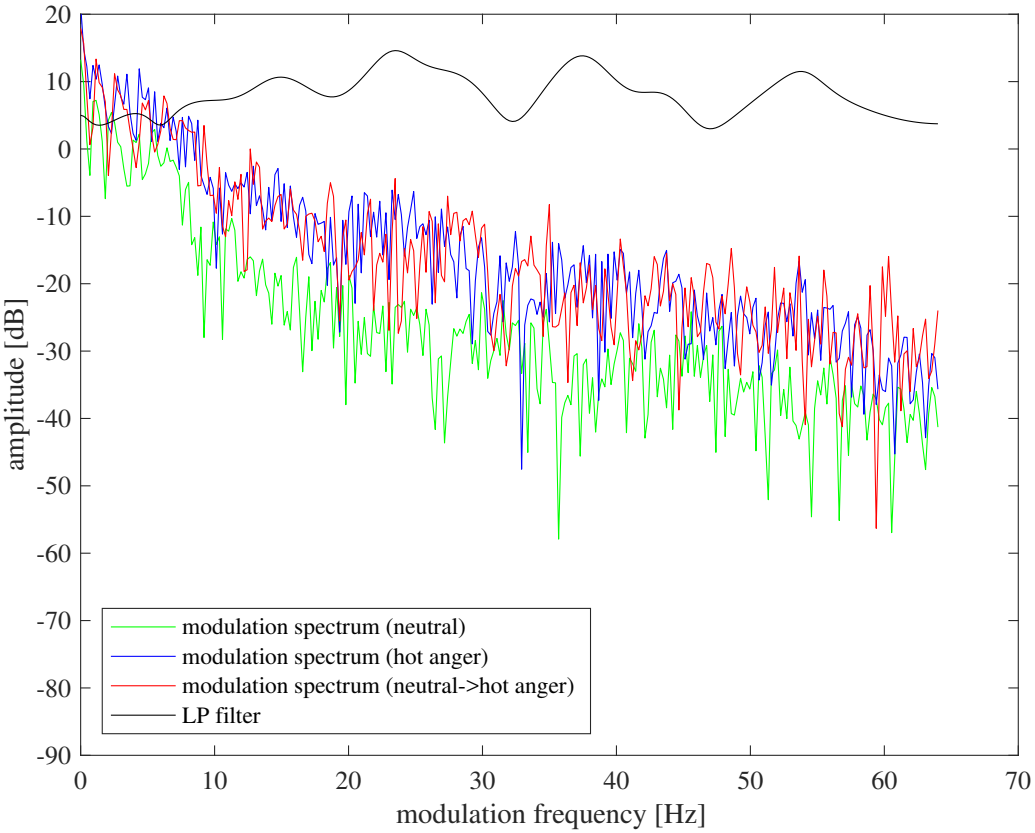
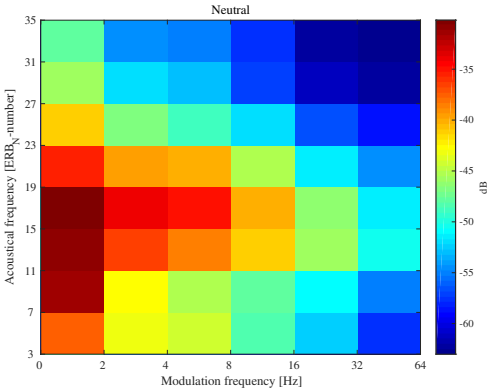
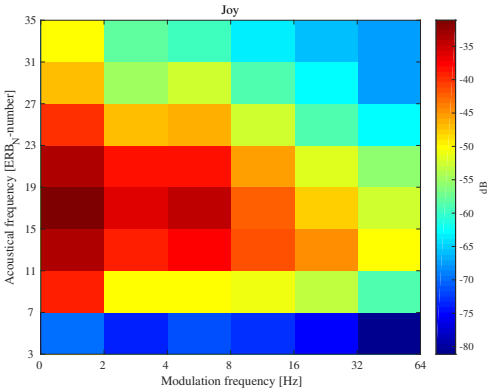


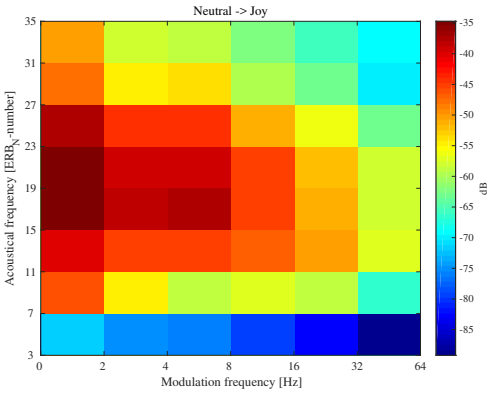
Figure 5.6: Modulation spectrum of neutral, hot anger, and NE-HA converted speech on 3rd band and frequency characteristic of LP based conversion filter.



(a) neutral speech

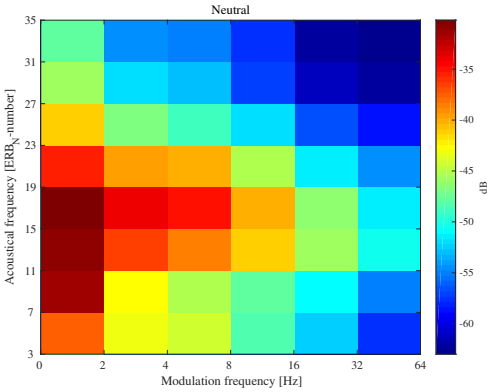


(b) joy speech

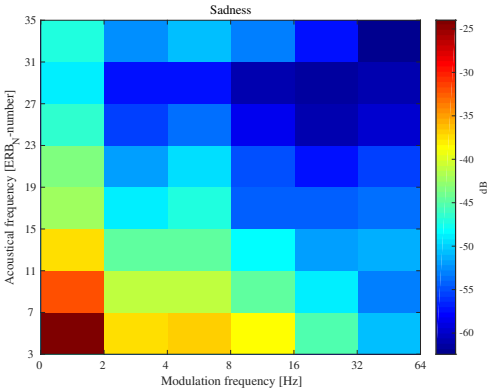


(c) converted speech

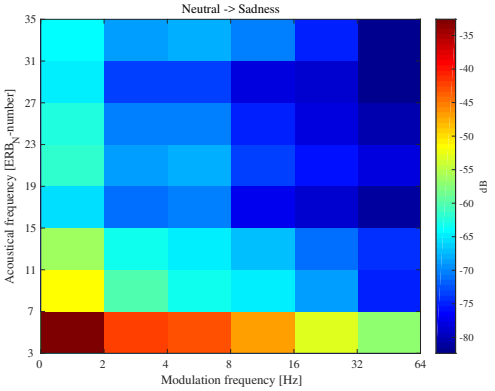
Figure 5.7: Modulation spectrograms of (a) neutral, (b) joy, and (c) neutral-joy converted speech.



(a) neutral speech

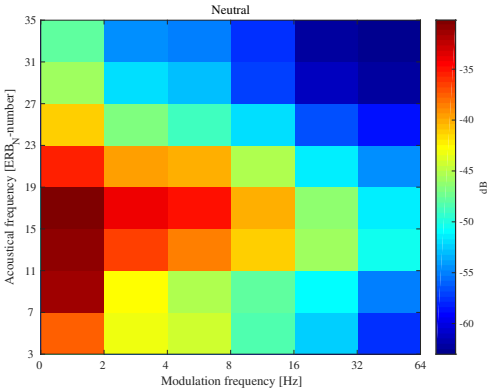


(b) sadness speech

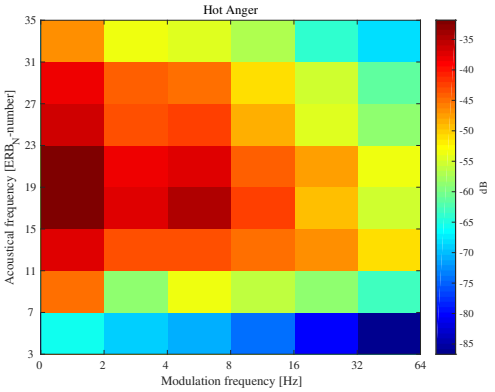


(c) converted speech

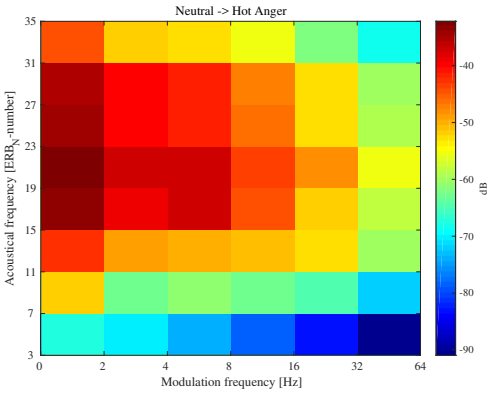
Figure 5.8: Modulation spectrograms of (a) neutral, (b) sadness, and (c) neutral-sadness converted speech.



(a) neutral speech



(b) hot anger speech



(c) converted speech

Figure 5.9: Modulation spectrograms of (a) neutral, (b) hot anger, and (c) neutral-hot anger converted speech.

5.2.4 Evaluation experiment

An experiment of vocal emotion recognition was carried out to confirm whether the vocal emotion of NVS can be converted successfully by using the proposed method.

Stimuli

To generate a stimulus in the 8–band NVS scheme, the envelope of each band was used to amplitude modulated with band-limited noise limited in the same band. Then, all amplitude modulated band-limited noises were summed to generate a stimulus. To confirm the effect of modifying the modulation spectrum with LP filtering, a condition with only amplitude correction and no modification of modulation spectrum by LP filtering was added. For joy, sadness, and hot anger, 10 sentences of vocal emotion conversion with the LP filter and vocal emotion conversion with only amplitude correction were generated. There were also 10 sentences of neutral NVS for the balance of stimuli.

Procedure

Four male native Japanese speakers participated in this experiment. All participants have normal hearing (hearing levels of the participants were below 12 dB in the frequency range from 125 to 8000 Hz). All participants were not familiar with NVS stimuli.

In this experiment, the NVS stimuli were presented to both ears of a participant through a PC, audio interface (RME, Fireface UCX), and a headphone (SENNHEISER HDA 200) in a sound-proof room. The sound pressure level of background noise was lower than 25.8 dB. The sound pressure level was calibrated to a comfortable level (about 65 dB) by using a head and torso simulator (B&K, type 4128) and sound level meter (B&K type 2231). All NVS stimuli were randomly presented to the participants. Participants were asked to indicate from all four kinds of emotions which emotion he/she thought was associated with the stimulus. Each stimulus was presented only once.

Results

Figure 5.10 shows the vocal emotion recognition rates of the experiment. The vocal emotion recognition rate was very low for joy. However, joy was found to be more difficult to recognize than the other emotions, even with the original joy NVS. The method of

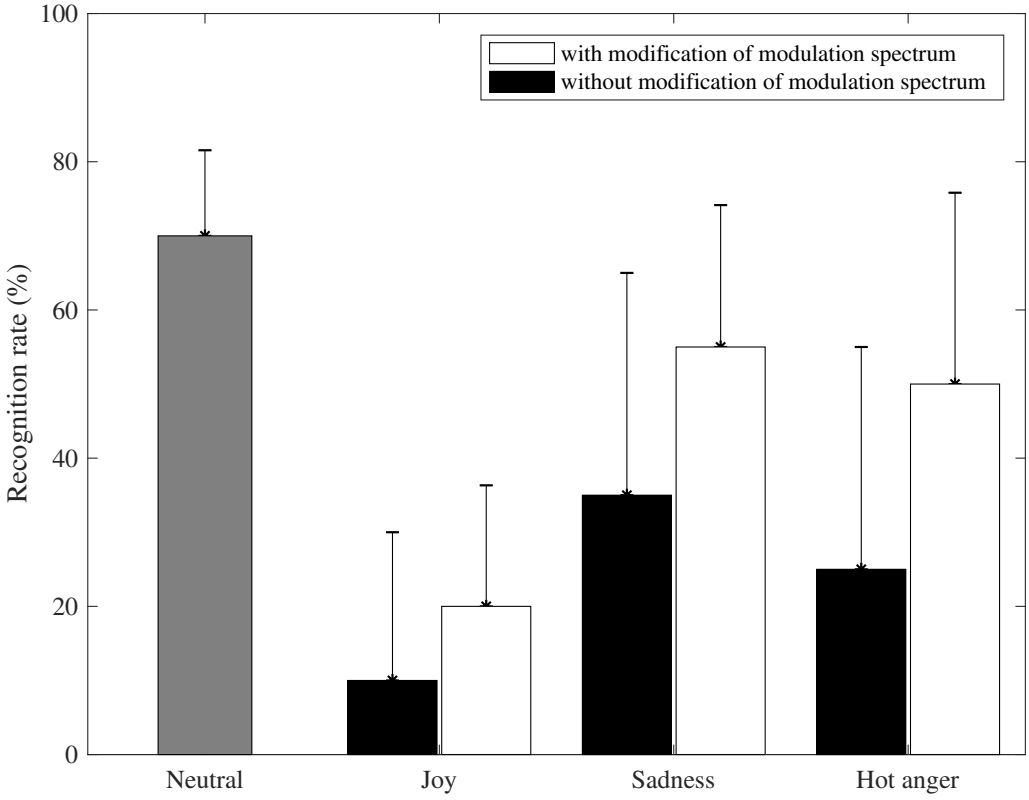


Figure 5.10: Results of vocal-emotion recognition experiment.

further enhancing the modulation spectral features to increase the recognition rate of joy is discussed in the next section. For sadness and hot anger, the results of vocal emotion conversion with the LP filter were higher than those without the LP filter. The results show that the process of LP filtering for modifying the modulation spectrogram is effective for the vocal emotion conversion of sadness and hot anger. Furthermore, the modulation spectrogram is confirmed to be an important cue for the perception of vocal emotion with simulated CIs. However, the results of repeatedly measured analyses of variance showed that there was no significant difference between the process method with and without the LP filter ($F(1,3) = 4.84$).

Discussion

McDermott *et al.* successfully converted the texture of sound by modifying the modulation spectrogram [84]. The method they used began with processing stages from the auditory periphery (auditory filterbank, envelope extraction, and modulation filterbank) to calculate the modulation spectrogram and culminated with the measurement of simple statistics of these stages. It was found that the synthetic textures will sound like another example of the corresponding real-world texture if the statistics of the modulation spectrogram used for synthesis are similar to those of the real-world texture. Their results suggested the importance of the modulation spectrogram in the timbre perception by humans and the possibility of converting sound signals by modifying the modulation spectrogram.

As a result of the evaluation experiment, modifying the modulation spectrogram using the LP filter was shown to be useful for the vocal emotion conversion of sadness and hot anger on the condition of simulated CIs. The results showed that the proposed method is not successful for joy on the NVS scheme. However, it should be mentioned that even the original joy NVS is difficult to be recognized. As the authors considered, by using the LP filtering and amplitude correction processes, the timbre of converted NVS is similar to the original emotional speech on the NVS scheme. However, this proposed method only focuses on the time averaged modulation spectrogram. The dynamic components of emotional speech such as accents are very important for the perception of vocal emotion. Therefore, a time varying modulation filtering process is considerably necessary as the next step in our future work.

5.2.5 Summary

In this section, a method based on a LP scheme was proposed to modify the modulation spectrogram and its features of neutral speech to that of emotional speech. The results showed that the modulation spectrogram of neutral speech can be successfully converted to that of emotional speech by the proposed method. Then a vocal-emotion recognition experiment using NVS generated by the converted temporal envelope was carried out. The results of the evaluation experiment confirmed the feasibility of vocal emotion conversion on the modulation spectrogram for simulated CIs.

Chapter 6

Conclusion

6.1 Summary

The purpose of this research is to clarify the contribution of temporal modulation cues to the perception of speaker individuality and vocal emotion. First of all, to confirm whether temporal modulation cues actually contribute to the perception of speaker individuality and vocal emotion, the role of temporal envelope and modulation frequency information in speaker and vocal emotion recognition was investigated. Speaker and vocal emotion recognition experiments using NVS were carried out to investigated the effects of different temporal and spectral resolutions of NVS on the perception of speaker individuality and vocal emotion. NVS is generated by dividing the speech signal into several band and replacing the carriers of each band with band-limited noise. The number of channels determines the spectral resolution of NVS: higher spectral resolution will be obtained with more channels. The upper limit of modulation frequency relates to the temporal resolution that higher temporal resolution will be provided with higher upper limit of modulation frequency. In the experiment, speaker distinction and vocal emotion recognition were conducted by NH listeners under different upper limit of modulation frequency (0, 0.5, 1, 2, 4, 8, 16, 32, and 64 Hz) of NVS. The role of temporal cues in the different spectral resolutions condition was also investigated by varying the number of channels (4, 8, and 16). The spectral and temporal modulation cues are reduced further when the number of channels and upper limit of modulation frequency decrease, respectively. If the temporal modulation cues contribute to the perception of nonlinguistic information, the

performance of speaker or vocal-emotion recognition will be poorer with lower temporal resolution of NVS. Therefore, this experimental paradigm can also clarify the important modulation frequency bands for speaker and vocal-emotion recognition.

For spectral cue, the speaker distinction performance was not sensitive to the spectral resolution, at least in the limited set of stimuli in the present study. For vocal-emotion recognition, the spectral resolution was important for the recognition of only neutral, joy, and cold anger NVS, but not sadness or hot anger NVS.

For temporal modulation cues, the results showed that the recognition rates were significantly decreased with lower upper limit of modulation frequency for both speaker and vocal emotion. On the other word, it was more difficult to recognize the speaker or vocal emotion from NVS if the temporal modulation cues provided by NVS were reduced. Therefore, it was confirmed that the temporal modulation cues contribute to speaker and vocal-emotion recognition. Compared to the perception of linguistic information, the temporal modulation cues provided by higher modulation frequency bands are suggested to be important for the perception of speaker individuality and vocal emotion.

At the next step, the relationship between the modulation spectral features and the perceptual data obtained from speaker and vocal-emotion recognition experiments was analyzed to clarify the exactly contribution of temporal modulation cues on the perception of speaker individuality and vocal-emotion. Modulation spectral features were extracted from the modulation spectrogram of speech data. The modulation spectrogram was calculated by the process of auditory filterbank, temporal envelope extraction and modulation filterbank. The modulation spectral centroid, spread, skewness, kurtosis, tilt and flatness were then extracted from the modulation spectrogram as modulation spectral features. In order to investigate the relationship between modulation spectral features and the perceptual data of speaker and vocal-emotion experiments, an discriminability index d' was used. The d' of each modulation spectral feature present the physical distance of the distributions of modulation spectral feature with different speakers or vocal emotions. On the other hand, the d' of the perceptual data present the psychological distance of different speakers or vocal emotions. The correlation between the d' of modulation spectral features and the perceptual data was calculated to demonstrate the relationship between modulation spectral features and the perception of speaker individuality and

vocal-emotion.

For speaker individuality, there were positive correlations between the modulation spectral features and the perceptual data of speaker distinction experiment. Similar results were also obtained from the results of vocal emotion, however, the correlations were roughly higher than that of speaker distinction experiments. The results showed that the modulation spectral features were useful to account for the perceptual data of speaker and vocal-emotion recognition experiments using NVS. It was suggested that modulation spectral features could be important cues contribute to the perception of speaker individuality and vocal emotion.

Finally, applications of the temporal modulation information in simulating CI listeners' response and vocal-emotion conversion of NVS were discussed. At first, the feasibility of using NVS to simulate CI listeners' response in vocal emotion recognition was discussed by carried out vocal-emotion recognition experiments using both NVS and original emotional speech with NH and CI listener. The results showed that the vocal-emotion recognition paradigm using NVS can be used to investigate vocal emotion recognition by CI listeners. Furthermore, it was suggested that the modulation spectral features can also be used to account the performance of CI listeners in the vocal-emotion recognition.

Effect of the modification of modulation spectrogram on the vocal-emotion recognition was also investigated. A method based on a linear prediction (LP) scheme was proposed to modify the modulation spectrogram and its features of neutral speech to match that of emotional speech. The logic of this approach is that if vocal emotion perception of CI simulation is based on the modulation spectral features, NVS with similar modulation spectral features of emotional speech will be recognized as the same emotion. The temporal envelopes were modulation-filtered by using IIR filters to modify the modulation spectrum from neutral to emotional speech. The IIR filters were derived from the relation of modulation characteristics of neutral and vocal emotions on a LP scheme. On the acoustic frequency domain, the average amplitude of the temporal envelope was corrected using the ratio of the average amplitude between neutral and emotional speech. Finally, a vocal-emotion recognition experiment using NVS generated by the converted temporal envelope was carried out. The results showed that the modulation spectrogram of neutral speech can be successfully converted to that of emotional speech by the pro-

posed method. The results of the evaluation experiment confirmed the feasibility of vocal emotion conversion on the modulation spectrogram for NVS.

In conclusion, the fact that the temporal modulation cues contribute to the perception of speaker individuality and vocal emotion was confirmed by the speaker and vocal-emotion recognition experiments using NVS. Furthermore, the investigation of modulation spectral features demonstrated that there were high correlations between modulation spectral features and the perceptual data obtained from speaker and vocal-emotion recognition experiments. Therefore, the modulation spectral features could be important cues contribute to the speaker and vocal-emotion recognition with NVS. These results further proved that the temporal modulation cues play an important role in the perception speaker individuality and vocal-emotion.

6.2 Contributions

The most important contribution is that the results of this study can help us to deepen our understanding of the relationship between the temporal modulation information of temporal envelope and the perception of nonlinguistic information. The previous studies about nonlinguistic information were almost based on the "classical" acoustical features such like F0, formant, spectral envelope, etc. However, the temporal modulation cues have been proved to be important cues in the speech perception. The results of this study demonstrated that the temporal modulation cues also contribute to the perception of nonlinguistic information. The modulation spectral features investigated in this study were analyzed based on the knowledge of human auditory system. Therefore, the temporal modulation information must play an important role in the perceptual process of various information in human auditory system directly.

The results of this study can also contribute to the development of cochlear implants (CI) device. CI system mimic the signal processing of the auditory peripheral system with four main steps: bandpass filterbank, envelope extraction, amplitude compression, and impulse signal generation. As the number of channels of the bandpass filterbank in CI system is so limited, CI device can only provide poor spectral cue. CI devices provide the temporal envelope information as a primary cue, and the temporal fine structure information is not effectively encoded. As the poor spectral cue, CI listeners have problem

of identifying the speaker or the emotion from only speech correctly. The method used in this study also mimic the signal processing of the auditory peripheral system. Therefore, the results of this study can be used to optimize the CI device. Particularly, the modulation spectrogram based vocal-emotion conversion method discussed in section 5.2 can be used in enhancing the modulation spectral features related to vocal-emotion to improve the performance of vocal-emotion recognition of CI listeners.

The modulation spectrogram and its features also have potential in modeling the perception of speaker individuality and vocal emotion. The modulation spectrogram is calculated based on the computational model of human peripheral auditory system. Therefore, the modulation spectrogram can be used in the physiology model of the perception of various information from not only speech but also other sounds. The modulation spectral features can also be used as acoustical features in the perceptual model of nonlinguistic information. The engineering applications of the modulation spectrogram and its features in the development of speaker or vocal-emotion recognition systems can also be expected.

6.3 Future works

1. **Analysis of modulation spectrogram in time domain**

   In this study, the modulation spectral features of time-averaged modulation spectrogram were analyzed. The modulation spectrogram is a 4-dimension data contained information about acoustic frequency, modulation frequency, amplitude and time. It is necessary to analysis the details of modulation spectrogram in time domain. However, as the modulation spectrogram is a 4-D data, it will be difficult extract the features related to nonlinguistic information from modulation spectrogram. Deep learning may be a good resolution for analyzing the modulation spectrogram in time domain.

2. **Modeling the perceptual process of nonlinguistic information based on modulation spectral features**

   The modulation spectrogram and its features has been proved to contribute the perception of nonlinguistic information. Therefore, the temporal modulation information can be used to modeling the perception of speaker individuality and vocal

emotion. For computational model such like the three-layer model [9], the modulation spectral features can be used as kinds of acoustical features. The method to calculate modulation spectrogram used in this study was based on the signal process in human peripheral auditory system. Therefore, the modulation spectrogram can be used in the physiology model. For example, the modulation spectrogram can be used as the input of a neural network based model instead of the traditional spectrogram calculated by short-time Fourier transformation.

3. **Details of modulation spectrogram related to the perception of nonlinguistic information**

   In this study, global features of modulation spectrogram were investigated. Such kinds of features may be used as cues in speaker and vocal-emotion recognition. However, the perceptual process of nonlinguistic information should not be that simple. It is undeniable that the local features such as the specific segmental cues are also used in the perception of nonlinguistic information. It is necessary to understand the contributions of the detailed information of the modulation spectrogram.

4. **Application of temporal modulation information in the development of CI device**

   As we known CI listeners have problem in speaker and vocal-emotion recognition as the poor spectral cue provided by CI device. Luo and Fu successfully enhanced the tone recognition on the NVS scheme by manipulating the amplitude envelope to more closely resemble the F0 contour [83]. Their results showed the possibility of enhancing the recognition of non-linguistic information by modifying the temporal envelope. However, as CI listeners using the temporal modulation cues as primarily cues, the results of this study can be used to optimize the CI device for improving the performance of speaker and vocal-emotion recognition by CI listeners. We can assumed that the target of vocal emotion is known (e.g., vocal-emotion recognition methods can be used to predict the target emotion via a dimension approach (V-A)) and enhance enhance the vocal emotion information of emotional NVS by modifying the modulation spectral features.

5. **Connect the temporal modulation information to the mechanism of speech**

production

This study demonstrated that the temporal modulation information contain the information related to speaker individuality and vocal-emotion. Such nonlinguistic information can be thought to be derived from human vocal organs. It is difficult to connect the temporal modulation information to the mechanism of speech production. However, it is still necessary to investigate the relationship between auditory-based modulation-spectral features and speech production.

6. **Contribution of temporal fine structure**

Speech signals can be represented as a sum of amplitude modulated frequency bands. The signal in each band can be regarded as a temporal amplitude envelope with a carrier (temporal fine structure). In this study, the temporal modulation cues contained in the temporal amplitude envelope has been proved to play an important role in the perception of speaker individuality and vocal-emotion. However, the temporal fine structure should also contribute to the speech perception of various information. It is necessary to understand the contribution of temporal fine structure further to complement the knowledge of the contributions of temporal information in speech perception.

Appendices

Appendix A

Confusion matrix of the results of vocal-emotion recognition experiments

Mean confusion matrices obtained with the results of vocal-emotion recognition experiments in section 3.4 are shown here. Confusion matrices are presented with the stimuli organized vertically and the response categories organized horizontally. Each cell shows the selection rate for that particular stimulus and response combination: the range is from 0 to 1.

Table A.1: Mean confusion matrix with 4-band, 0 Hz NVS stimuli.

	Neutral	Joy	Cold Anger	Sadness	Hot Anger
Neutral	0.63	0	0.15	0.06	0.16
Joy	0.61	0.11	0.10	0.05	0.14
Cold Anger	0.68	0	0.12	0.07	0.13
Sadness	0.35	0.01	0.08	0.55	0.01
Hot Anger	0.48	0.13	0.05	0.05	0.29

Table A.2: Mean confusion matrix with 4-band, 0.5 Hz NVS stimuli.

	Neutral	Joy	Cold Anger	Sadness	Hot Anger
Neutral	0.55	0.03	0.19	0.08	0.15
Joy	0.57	0.07	0.16	0.05	0.14
Cold Anger	0.52	0.02	0.14	0.15	0.17
Sadness	0.29	0.01	0.10	0.59	0.01
Hot Anger	0.43	0.15	0.10	0.05	0.27

Table A.3: Mean confusion matrix with 4-band, 1 Hz NVS stimuli.

	Neutral	Joy	Cold Anger	Sadness	Hot Anger
Neutral	0.61	0.02	0.16	0.06	0.15
Joy	0.53	0.03	0.16	0.04	0.25
Cold Anger	0.49	0.02	0.24	0.10	0.15
Sadness	0.29	0	0.03	0.67	0.01
Hot Anger	0.47	0.04	0.15	0.05	0.30

Table A.4: Mean confusion matrix with 4-band, 2 Hz NVS stimuli.

	Neutral	Joy	Cold Anger	Sadness	Hot Anger
Neutral	0.53	0.02	0.26	0.03	0.16
Joy	0.45	0.03	0.19	0.05	0.27
Cold Anger	0.43	0.02	0.21	0.10	0.25
Sadness	0.15	0	0.08	0.76	0
Hot Anger	0.31	0.03	0.14	0.01	0.52

Table A.5: Mean confusion matrix with 4-band, 4 Hz NVS stimuli.

	Neutral	Joy	Cold Anger	Sadness	Hot Anger
Neutral	0.57	0.02	0.24	0.07	0.10
Joy	0.33	0.11	0.12	0.01	0.44
Cold Anger	0.28	0.03	0.39	0.17	0.13
Sadness	0.06	0	0.11	0.83	0
Hot Anger	0.25	0.05	0.10	0.01	0.60

Table A.6: Mean confusion matrix with 4-band, 8 Hz NVS stimuli.

	Neutral	Joy	Cold Anger	Sadness	Hot Anger
Neutral	0.66	0.04	0.21	0.05	0.05
Joy	0.30	0.25	0.15	0.03	0.27
Cold Anger	0.36	0.04	0.37	0.14	0.09
Sadness	0.09	0	0.06	0.85	0
Hot Anger	0.25	0.10	0.09	0	0.56

Table A.7: Mean confusion matrix with 4-band, 16 Hz NVS stimuli.

	Neutral	Joy	Cold Anger	Sadness	Hot Anger
Neutral	0.73	0.03	0.15	0.05	0.04
Joy	0.28	0.24	0.12	0.02	0.35
Cold Anger	0.40	0	0.38	0.16	0.05
Sadness	0.06	0.01	0.02	0.90	0.01
Hot Anger	0.16	0.10	0.15	0.03	0.56

Table A.8: Mean confusion matrix with 4-band, 32 Hz NVS stimuli.

	Neutral	Joy	Cold Anger	Sadness	Hot Anger
Neutral	0.70	0.02	0.15	0.05	0.07
Joy	0.29	0.27	0.15	0.04	0.25
Cold Anger	0.43	0.01	0.35	0.15	0.06
Sadness	0.05	0	0.04	0.91	0
Hot Anger	0.25	0.05	0.10	0	0.59

Table A.9: Mean confusion matrix with 4-band, 64 Hz NVS stimuli.

	Neutral	Joy	Cold Anger	Sadness	Hot Anger
Neutral	0.67	0.04	0.20	0.06	0.03
Joy	0.23	0.22	0.18	0.02	0.35
Cold Anger	0.34	0.02	0.40	0.20	0.05
Sadness	0.05	0	0.05	0.90	0
Hot Anger	0.16	0.03	0.05	0.01	0.75

Table A.10: Mean confusion matrix with 8-band, 0 Hz NVS stimuli.

	Neutral	Joy	Cold Anger	Sadness	Hot Anger
Neutral	0.57	0.01	0.16	0.06	0.19
Joy	0.55	0.15	0.12	0.03	0.15
Cold Anger	0.66	0.02	0.16	0.08	0.07
Sadness	0.47	0.01	0.05	0.45	0.01
Hot Anger	0.39	0.15	0.10	0.05	0.31

Table A.11: Mean confusion matrix with 8-band, 0.5 Hz NVS stimuli.

	Neutral	Joy	Cold Anger	Sadness	Hot Anger
Neutral	0.54	0.06	0.17	0.09	0.14
Joy	0.53	0.14	0.09	0.13	0.12
Cold Anger	0.53	0.07	0.22	0.13	0.05
Sadness	0.22	0	0.12	0.66	0
Hot Anger	0.39	0.19	0.11	0.09	0.22

Table A.12: Mean confusion matrix with 8-band, 1 Hz NVS stimuli.

	Neutral	Joy	Cold Anger	Sadness	Hot Anger
Neutral	0.59	0.03	0.14	0.09	0.15
Joy	0.48	0.11	0.13	0.06	0.22
Cold Anger	0.57	0.02	0.15	0.21	0.05
Sadness	0.31	0	0.07	0.61	0.01
Hot Anger	0.41	0.14	0.07	0.05	0.34

Table A.13: Mean confusion matrix with 8-band, 2 Hz NVS stimuli.

	Neutral	Joy	Cold Anger	Sadness	Hot Anger
Neutral	0.57	0.18	0.10	0.06	0.08
Joy	0.46	0.20	0.09	0.04	0.21
Cold Anger	0.44	0.03	0.24	0.16	0.14
Sadness	0.11	0.01	0.06	0.82	0
Hot Anger	0.22	0.13	0.08	0.03	0.55

Table A.14: Mean confusion matrix with 8-band, 4 Hz NVS stimuli.

	Neutral	Joy	Cold Anger	Sadness	Hot Anger
Neutral	0.67	0.12	0.15	0.01	0.05
Joy	0.31	0.31	0.07	0.01	0.30
Cold Anger	0.41	0.03	0.34	0.21	0.02
Sadness	0.06	0.02	0.07	0.85	0
Hot Anger	0.17	0.20	0.06	0.02	0.55

Table A.15: Mean confusion matrix with 8-band, 8 Hz NVS stimuli.

	Neutral	Joy	Cold Anger	Sadness	Hot Anger
Neutral	0.75	0.13	0.05	0.05	0.03
Joy	0.21	0.50	0.04	0.02	0.24
Cold Anger	0.54	0	0.29	0.15	0.02
Sadness	0.05	0.01	0	0.93	0.01
Hot Anger	0.11	0.15	0.05	0	0.68

Table A.16: Mean confusion matrix with 8-band, 16 Hz NVS stimuli.

	Neutral	Joy	Cold Anger	Sadness	Hot Anger
Neutral	0.78	0.10	0.09	0.02	0.01
Joy	0.16	0.65	0.06	0	0.13
Cold Anger	0.42	0	0.42	0.15	0.01
Sadness	0.07	0	0.04	0.88	0.01
Hot Anger	0.08	0.15	0.06	0.01	0.69

Table A.17: Mean confusion matrix with 8-band, 32 Hz NVS stimuli.

	Neutral	Joy	Cold Anger	Sadness	Hot Anger
Neutral	0.80	0.10	0.07	0.01	0.02
Joy	0.13	0.69	0.03	0.01	0.15
Cold Anger	0.44	0.01	0.37	0.15	0.04
Sadness	0.07	0	0.04	0.89	0
Hot Anger	0.09	0.10	0.07	0	0.74

Table A.18: Mean confusion matrix with 8-band, 64 Hz NVS stimuli.

	Neutral	Joy	Cold Anger	Sadness	Hot Anger
Neutral	0.73	0.14	0.09	0.04	0.01
Joy	0.18	0.62	0.04	0.01	0.15
Cold Anger	0.49	0	0.32	0.16	0.03
Sadness	0.05	0	0.05	0.90	0
Hot Anger	0.08	0.10	0.06	0	0.75

Table A.19: Mean confusion matrix with 16-band, 0 Hz NVS stimuli.

	Neutral	Joy	Cold Anger	Sadness	Hot Anger
Neutral	0.62	0.02	0.21	0.04	0.12
Joy	0.45	0.19	0.11	0.08	0.16
Cold Anger	0.55	0	0.23	0.16	0.06
Sadness	0.35	0	0.08	0.56	0
Hot Anger	0.40	0.14	0.10	0.08	0.28

Table A.20: Mean confusion matrix with 16-band, 0.5 Hz NVS stimuli.

	Neutral	Joy	Cold Anger	Sadness	Hot Anger
Neutral	0.47	0.13	0.10	0.23	0.07
Joy	0.42	0.26	0.09	0.16	0.06
Cold Anger	0.51	0.02	0.18	0.25	0.05
Sadness	0.22	0.01	0.05	0.72	0
Hot Anger	0.36	0.26	0.06	0.09	0.22

Table A.21: Mean confusion matrix with 16-band, 1 Hz NVS stimuli.

	Neutral	Joy	Cold Anger	Sadness	Hot Anger
Neutral	0.62	0.06	0.14	0.05	0.13
Joy	0.35	0.34	0.10	0.08	0.13
Cold Anger	0.47	0.01	0.22	0.28	0.02
Sadness	0.24	0.01	0.05	0.69	0.01
Hot Anger	0.31	0.22	0.13	0.05	0.29

Table A.22: Mean confusion matrix with 16-band, 2 Hz NVS stimuli.

	Neutral	Joy	Cold Anger	Sadness	Hot Anger
Neutral	0.73	0.07	0.07	0.08	0.05
Joy	0.24	0.53	0.05	0.05	0.14
Cold Anger	0.32	0.03	0.38	0.22	0.05
Sadness	0.10	0.01	0.04	0.83	0.03
Hot Anger	0.18	0.25	0.03	0.02	0.52

Table A.23: Mean confusion matrix with 16-band, 4 Hz NVS stimuli.

	Neutral	Joy	Cold Anger	Sadness	Hot Anger
Neutral	0.78	0.08	0.10	0.03	0.01
Joy	0.05	0.87	0.03	0.01	0.04
Cold Anger	0.29	0.01	0.49	0.21	0
Sadness	0.03	0.01	0.06	0.90	0
Hot Anger	0.14	0.22	0.04	0	0.61

Table A.24: Mean confusion matrix with 16-band, 8 Hz NVS stimuli.

	Neutral	Joy	Cold Anger	Sadness	Hot Anger
Neutral	0.83	0.09	0.05	0.02	0.02
Joy	0.06	0.94	0	0	0
Cold Anger	0.33	0	0.55	0.12	0.01
Sadness	0.03	0	0.04	0.93	0.01
Hot Anger	0.06	0.08	0.06	0	0.79

Table A.25: Mean confusion matrix with 16-band, 16 Hz NVS stimuli.

	Neutral	Joy	Cold Anger	Sadness	Hot Anger
Neutral	0.89	0.03	0.05	0	0.03
Joy	0.05	0.92	0.01	0	0.02
Cold Anger	0.31	0.01	0.60	0.07	0.01
Sadness	0.03	0	0.05	0.91	0.01
Hot Anger	0.05	0.09	0.06	0.01	0.78

Table A.26: Mean confusion matrix with 16-band, 32 Hz NVS stimuli.

	Neutral	Joy	Cold Anger	Sadness	Hot Anger
Neutral	0.88	0.03	0.06	0.03	0
Joy	0.06	0.90	0.01	0.01	0.02
Cold Anger	0.31	0	0.57	0.11	0.01
Sadness	0.04	0	0.03	0.94	0
Hot Anger	0.06	0.04	0.10	0	0.80

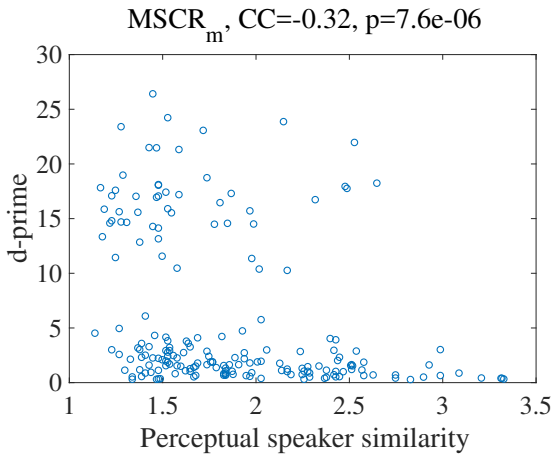
Table A.27: Mean confusion matrix with 16-band, 64 Hz NVS stimuli.

	Neutral	Joy	Cold Anger	Sadness	Hot Anger
Neutral	0.86	0.04	0.07	0.01	0.02
Joy	0.03	0.95	0	0.01	0.02
Cold Anger	0.26	0.01	0.60	0.12	0.01
Sadness	0.04	0	0.03	0.94	0
Hot Anger	0.08	0.08	0.07	0	0.76

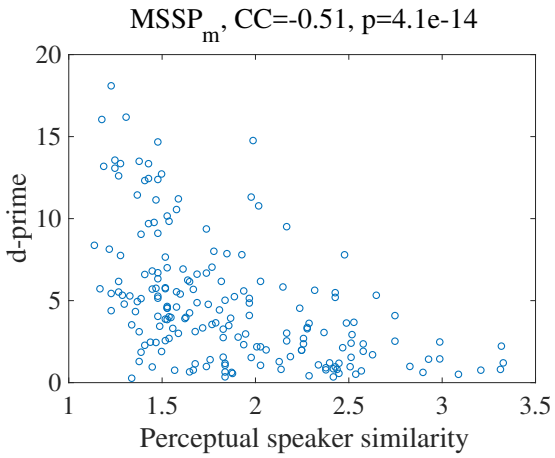
Appendix B

Scatterplots of perceptual speaker similarity and the d' of MSFs

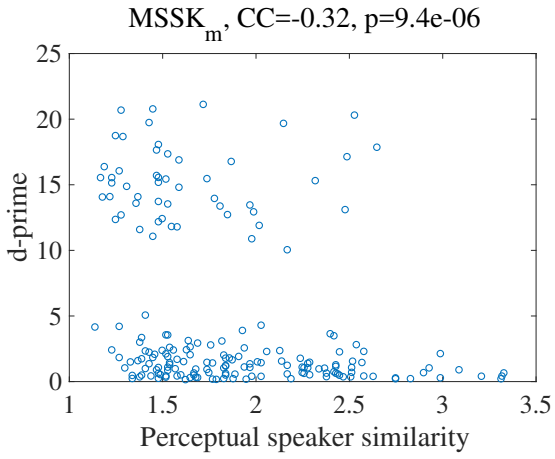
The scatterplots of perceptual speaker similarity and d' of MSFs are shown here. The horizontal axis is the perceptual speaker similarity of each speaker pairs measured by Kitamura *et al.* [1]. The vertical axis is the d' value of MSFs. The name of MSF, the correlation coefficient (CC), and the p-value for testing the hypothesis of no correlation are shown on the top of each figure. These results are related to the figure 4.3.



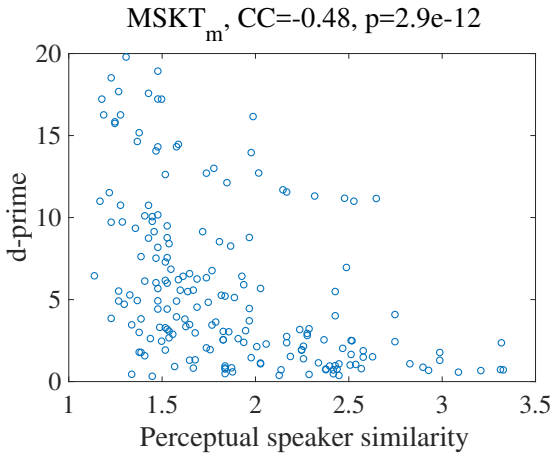
(a) MSCR$_m$



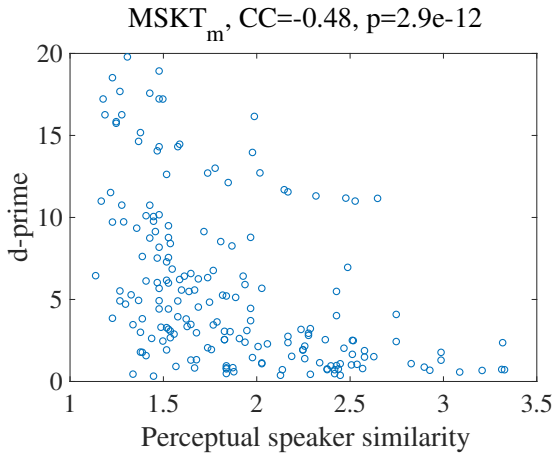
(b) MSSP$_m$



(c) MSSK$_m$

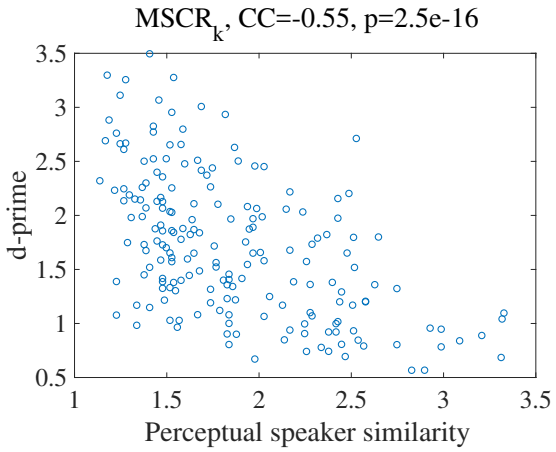


(d) MSKT$_m$

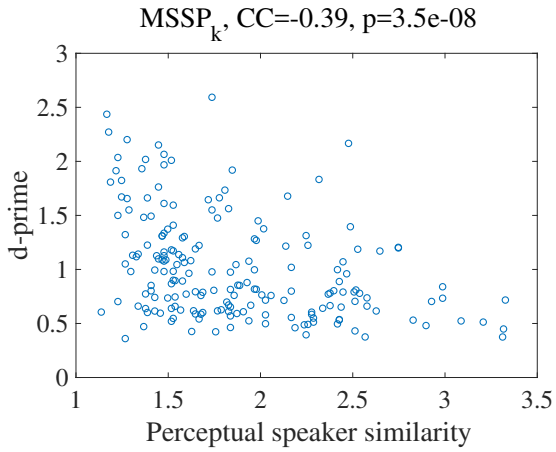


(e) MSKT$_m$

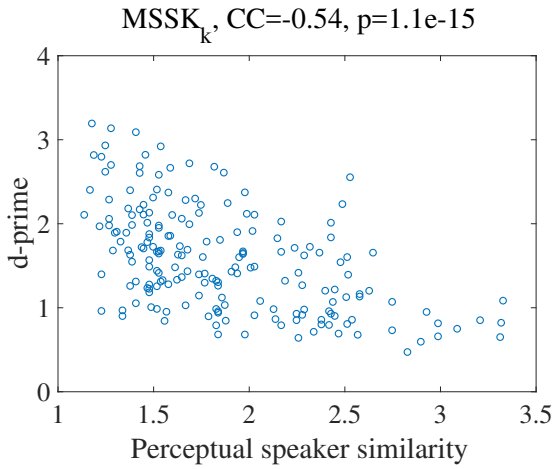
Figure B.1: The scatterplot of perceptual speaker similarity and d' of modulation spectral features on acoustic frequency domain for female speakers.



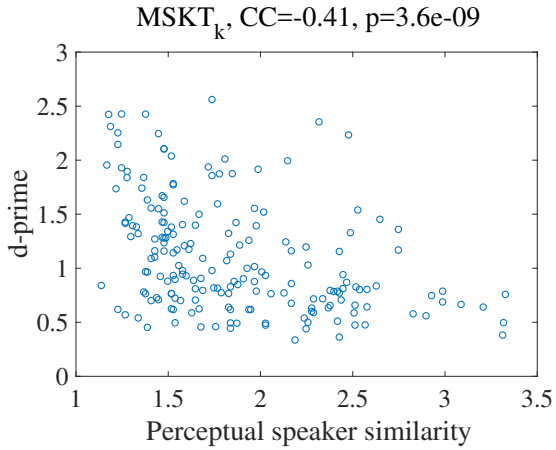
(a) MSCR_k



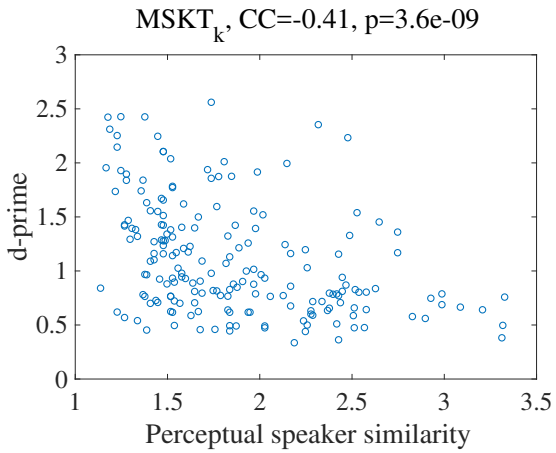
(b) MSSP_k



(c) MSSK_k

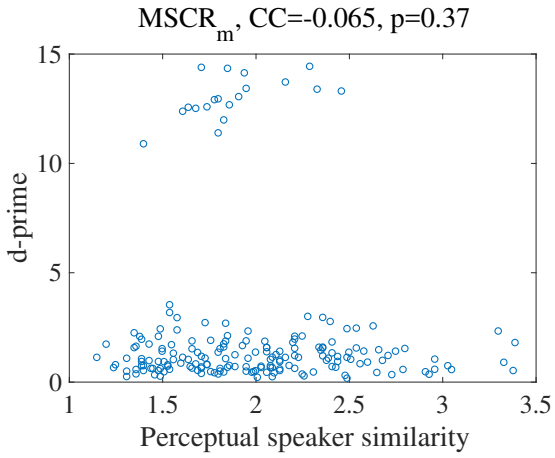


(d) MSKT_k

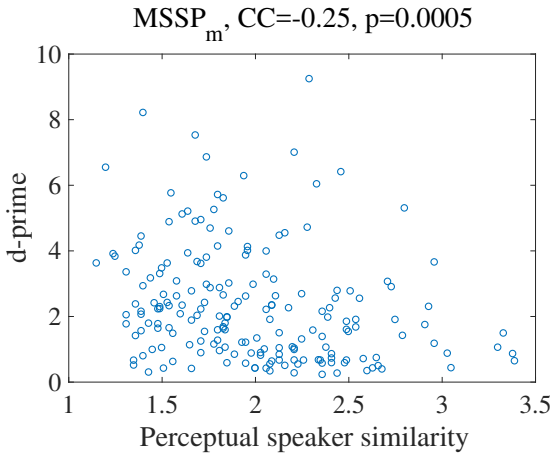


(e) MSKT_k

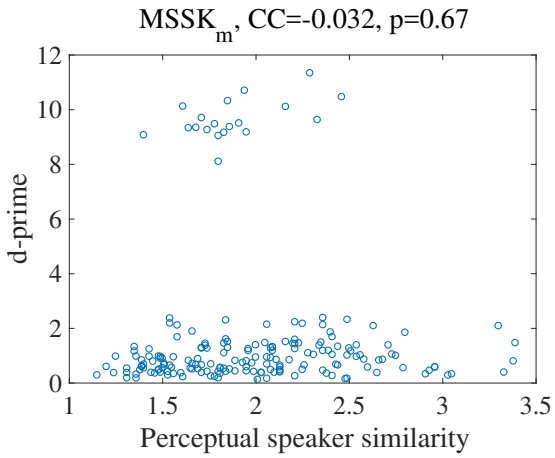
Figure B.2: The scatterplot of perceptual speaker similarity and d' of modulation spectral features on modulation frequency domain for female speakers.



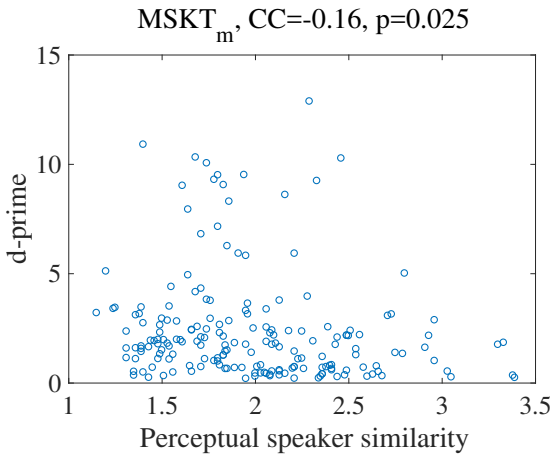
(a) MSCR$_m$



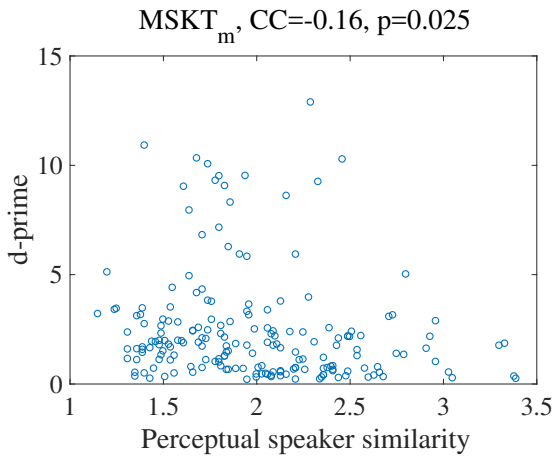
(b) MSSP$_m$



(c) MSSK$_m$

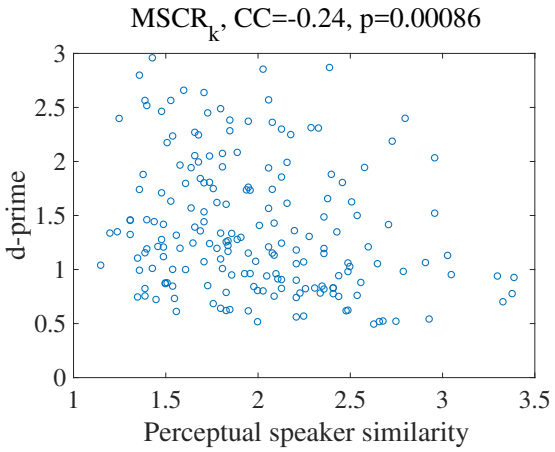


(d) MSKT$_m$

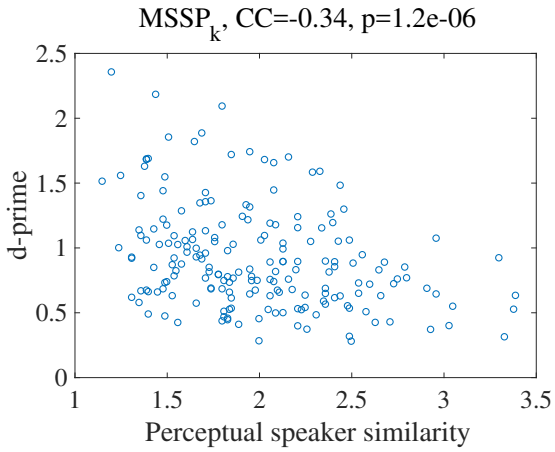


(e) MSKT$_m$

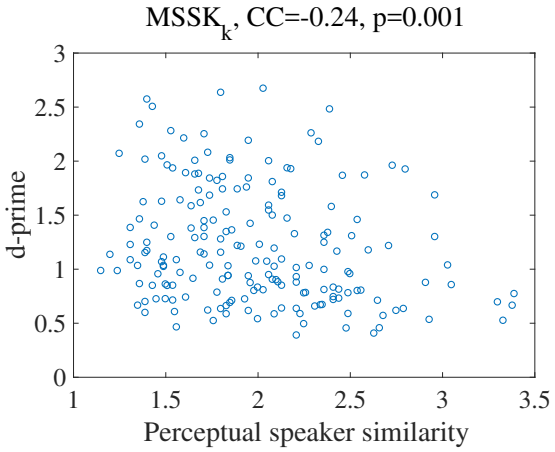
Figure B.3: The scatterplot of perceptual speaker similarity and d' of modulation spectral features on acoustic frequency domain for male speakers.



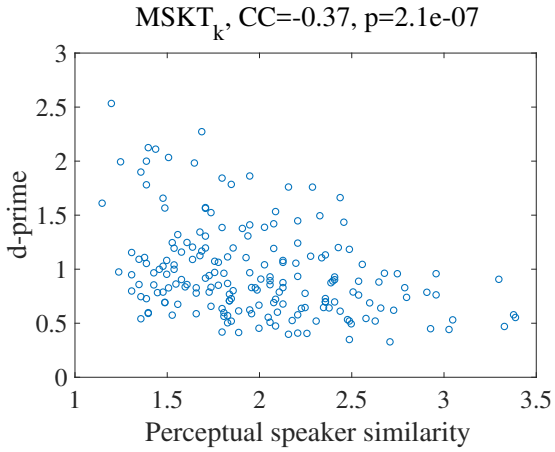
(a) MSCR$_k$



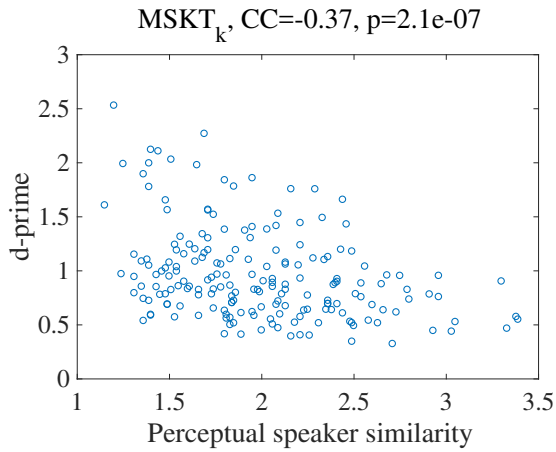
(b) MSSP$_k$



(c) MSSK$_k$



(d) MSKT$_k$



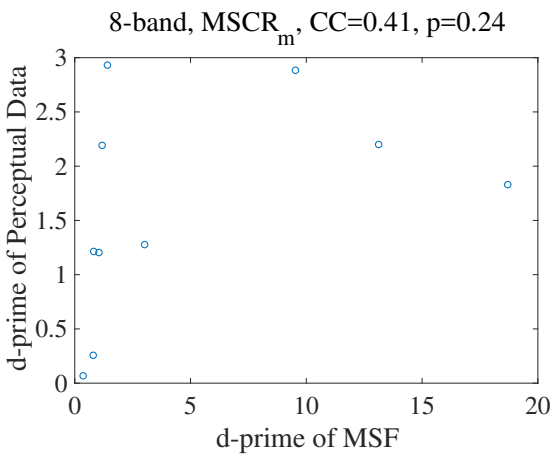
(e) MSKT$_k$

Figure B.4: The scatterplot of perceptual speaker similarity and d' of modulation spectral features on modulation frequency domain for male speakers.

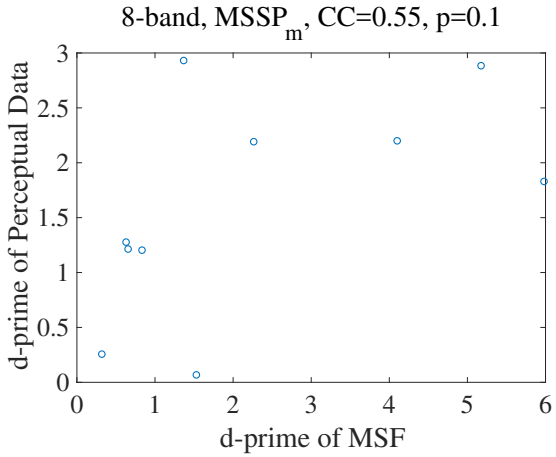
Appendix C

Scatterplots of the d' of MSFs and the results of speaker distinction experiments

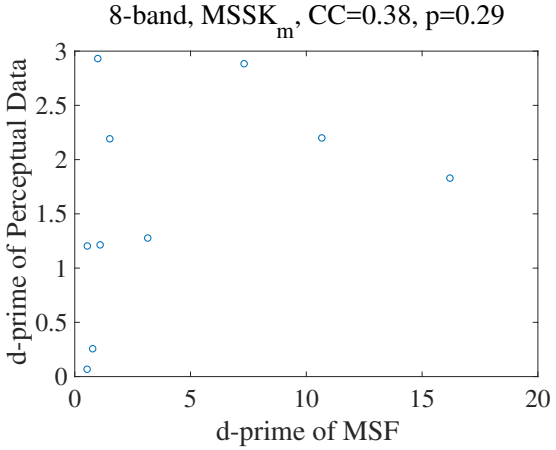
The scatterplots of the d' of MSFs and perceptual data of speaker distinction experiments are shown here. The horizontal axis is the d' of the perceptual data of speaker distinction experiment (Table 4.2 and 4.3). The vertical axis is the d' value of MSFs. The name of MSF, the correlation coefficient (CC), and the p-value for testing the hypothesis of no correlation are shown on the top of each figure. These results are related to the figure 4.7.



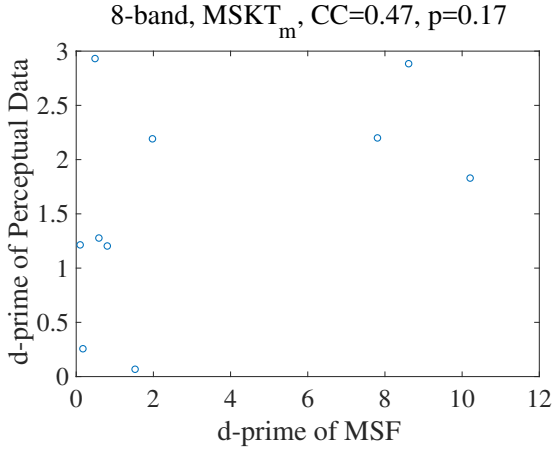
(a) MSCR$_m$



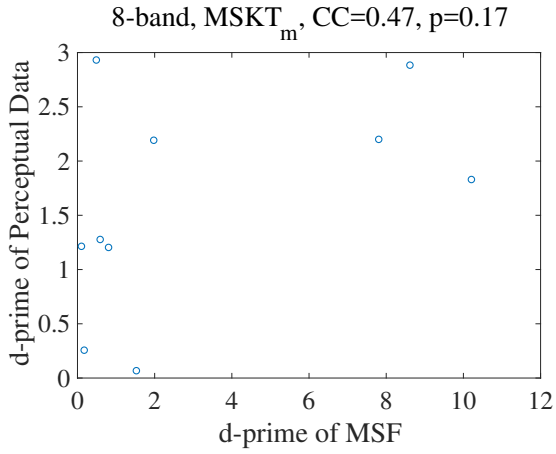
(b) MSSP$_m$



(c) MSSK$_m$

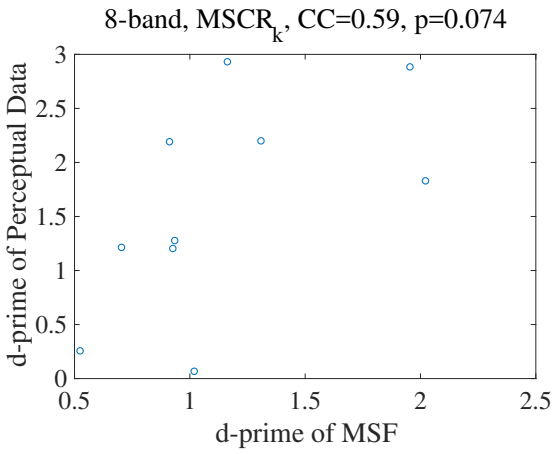


(d) MSKT$_m$

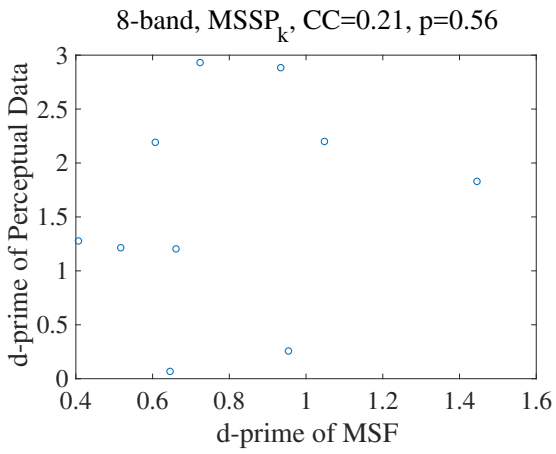


(e) MSKT$_m$

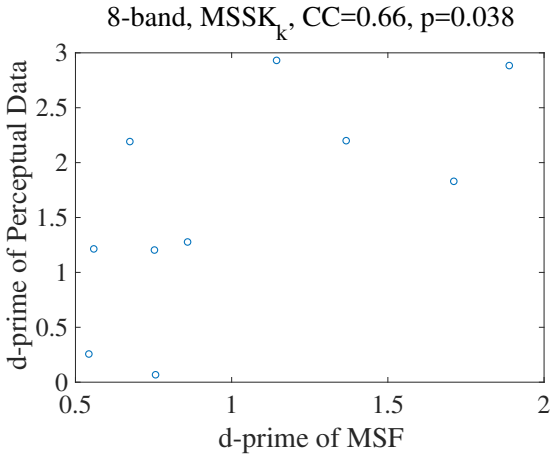
Figure C.1: The scatterplot of the d' of the perceptual data of speaker distinction experiment and modulation spectral features on acoustic frequency domain for for 8-band NVS.



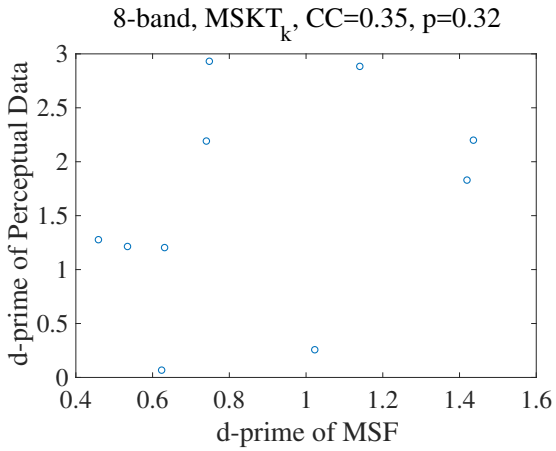
(a) MSCR$_k$



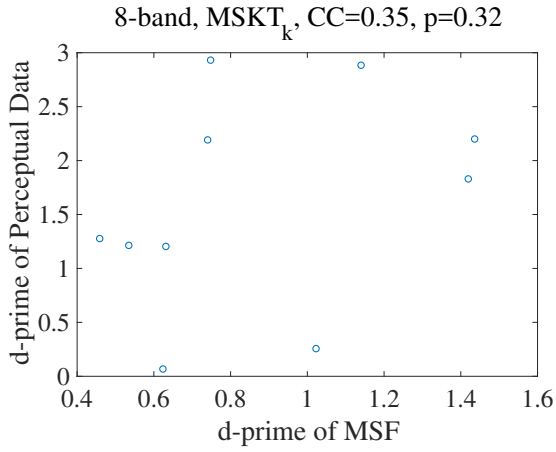
(b) MSSP$_k$



(c) MSSK$_k$

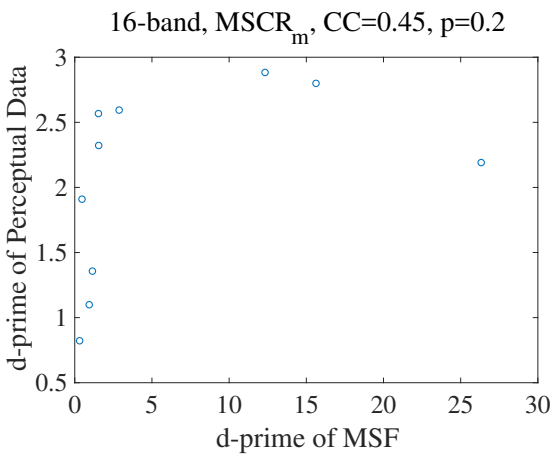


(d) MSKT$_k$

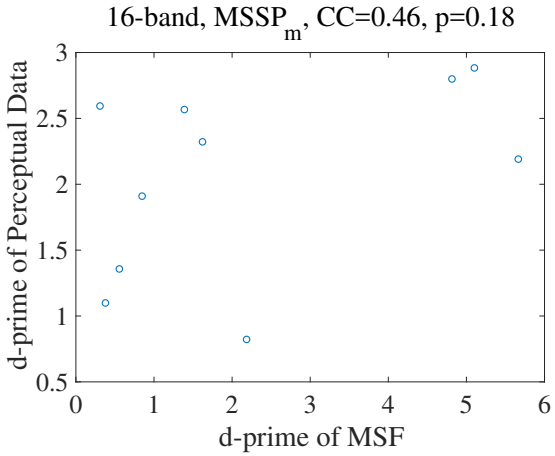


(e) MSKT$_k$

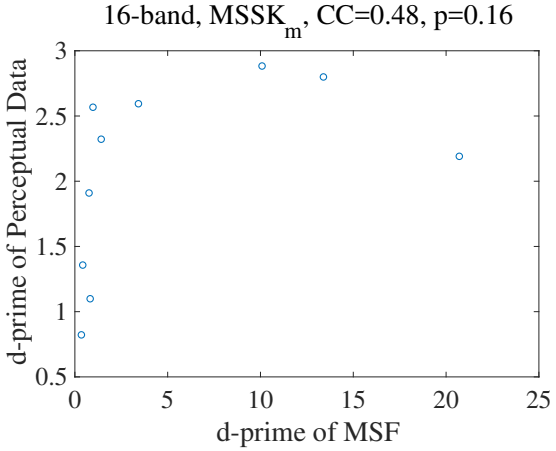
Figure C.2: The scatterplot of the d' of the perceptual data of speaker distinction experiment and modulation spectral features on modulation frequency domain for for 8-band NVS.



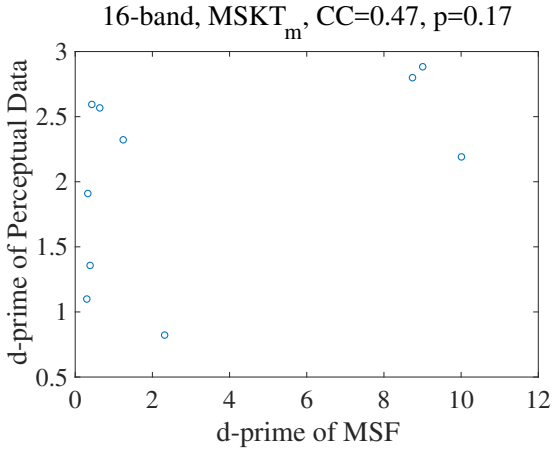
(a) MSCR$_m$



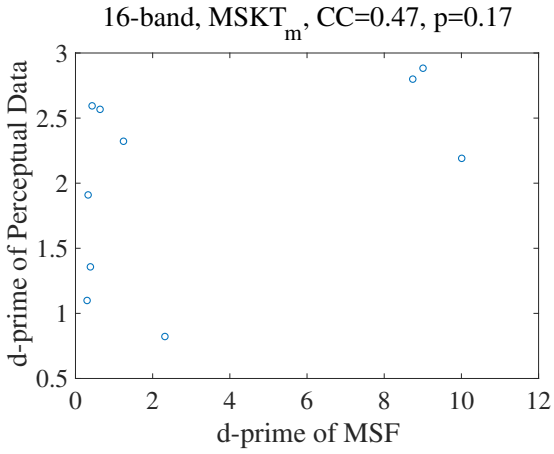
(b) MSSP$_m$



(c) MSSK$_m$

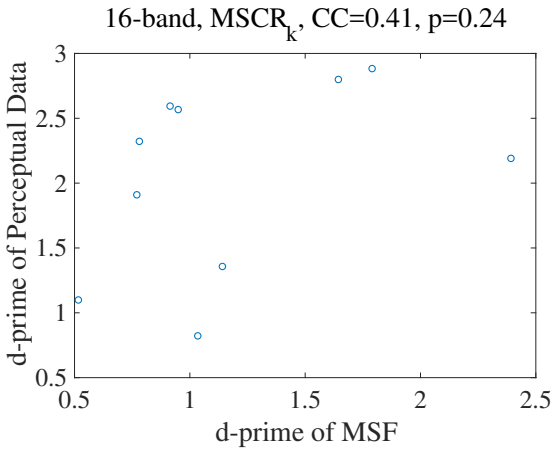


(d) MSKT$_m$

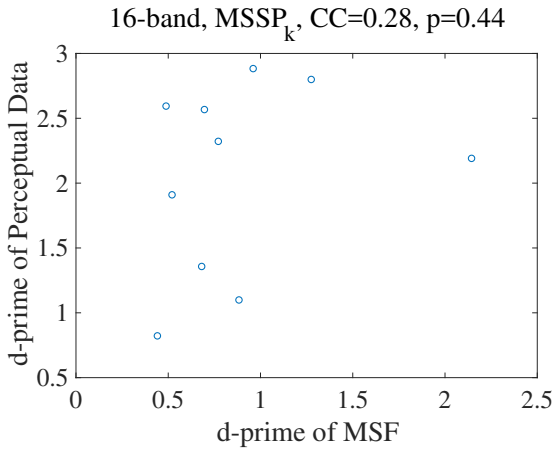


(e) MSKT$_m$

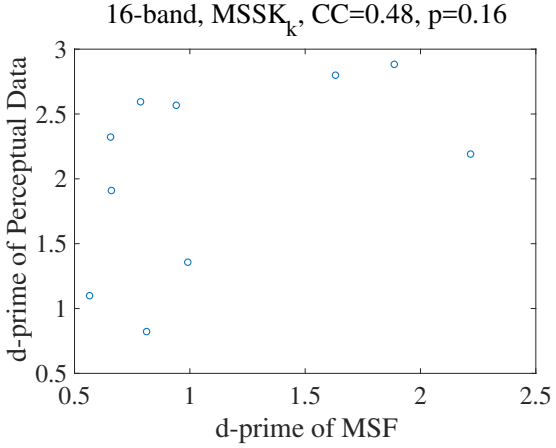
Figure C.3: The scatterplot of the d' of the perceptual data of speaker distinction experiment and modulation spectral features on acoustic frequency domain for for 16-band NVS.



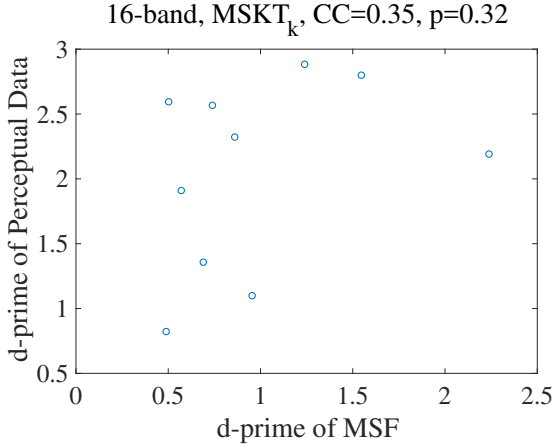
(a) MSCR$_k$



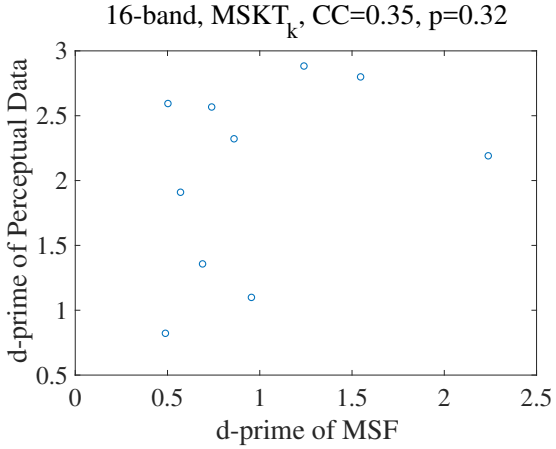
(b) MSSP$_k$



(c) MSSK$_k$



(d) MSKT$_k$



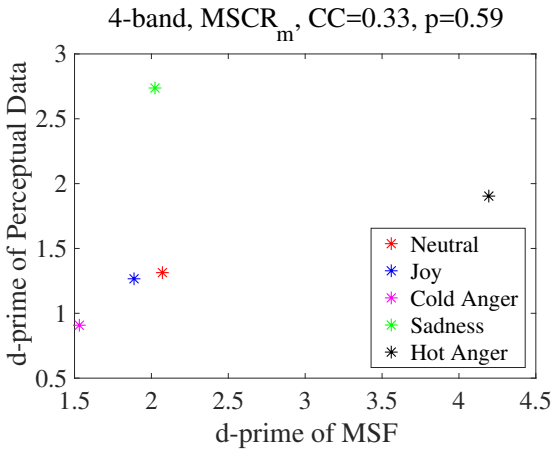
(e) MSKT$_k$

Figure C.4: The scatterplot of the d' of the perceptual data of speaker distinction experiment and modulation spectral features on modulation frequency domain for for 16-band NVS.

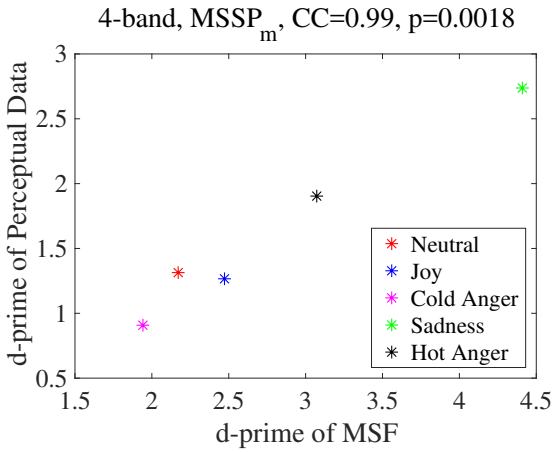
Appendix D

Scatterplots of the d' of MSFs and the results of vocal-emotion recognition experiments

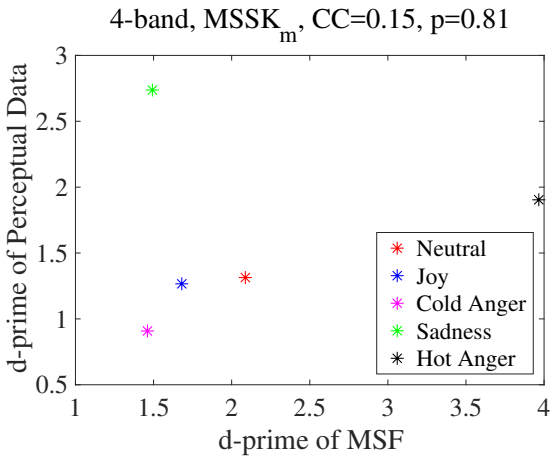
The scatterplots of the d' of MSFs and perceptual data of vocal-emotion recognition experiments are shown here. The horizontal axis is the d' of the perceptual data of vocal-emotion recognition experiment (Table 4.4). The vertical axis is the d' value of MSFs. The name of MSF, the correlation coefficient (CC), and the p-value for testing the hypothesis of no correlation are shown on the top of each figure. These results are related to the figure 4.8.



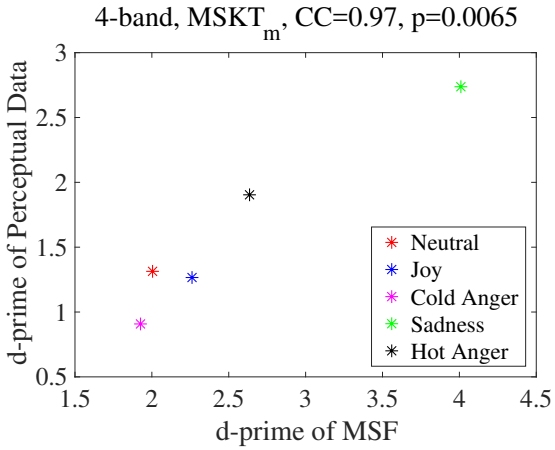
(a) MSCR$_m$



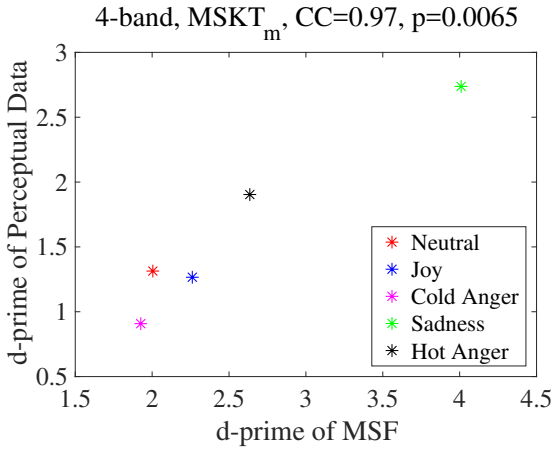
(b) MSSP$_m$



(c) MSSK$_m$

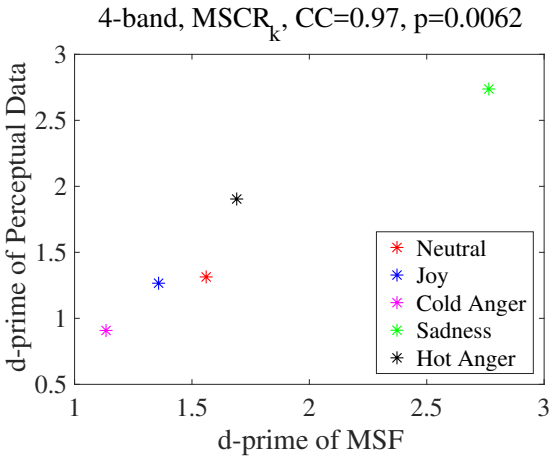


(d) MSKT$_m$

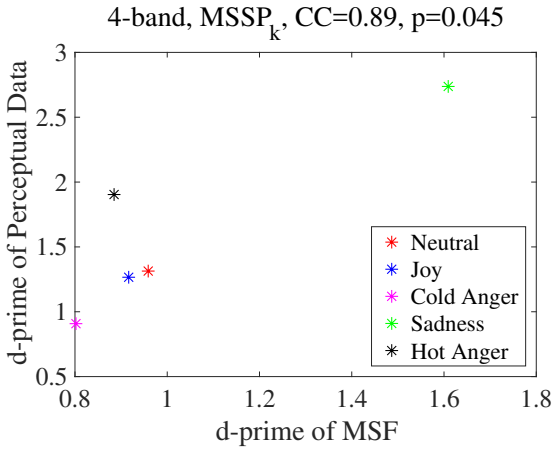


(e) MSKT$_m$

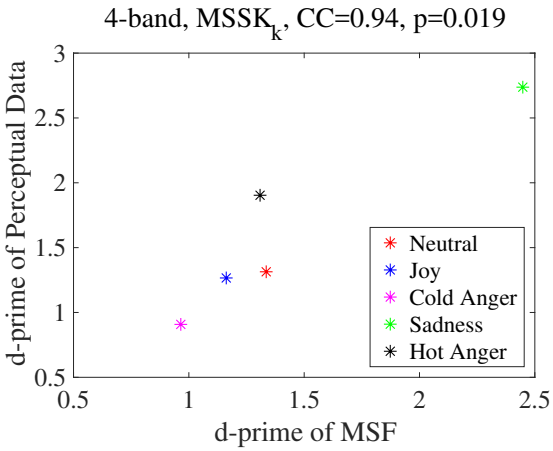
Figure D.1: The scatterplot of the d' of the perceptual data of vocal-emotion recognition experiment and modulation spectral features on acoustic frequency domain for for 4-band NVS.



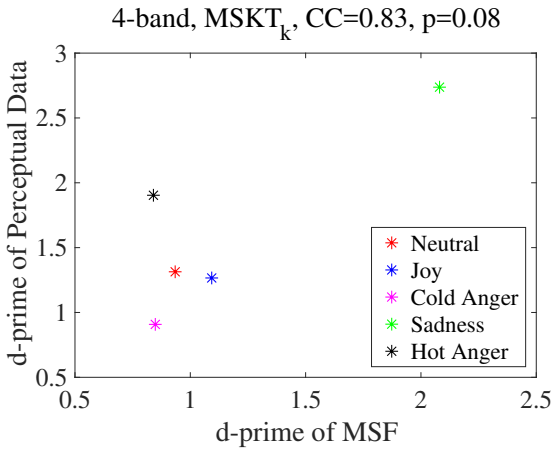
(a) MSCR$_k$



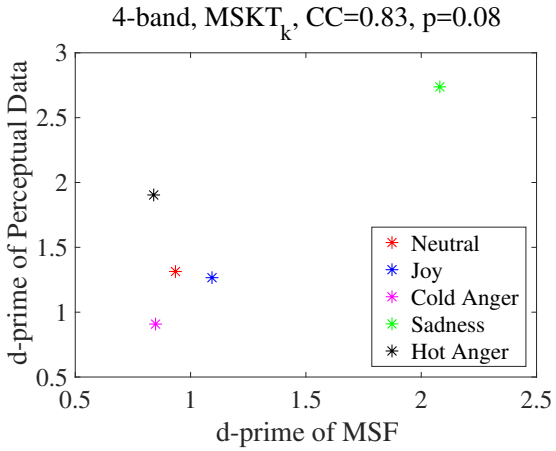
(b) MSSP$_k$



(c) MSSK$_k$

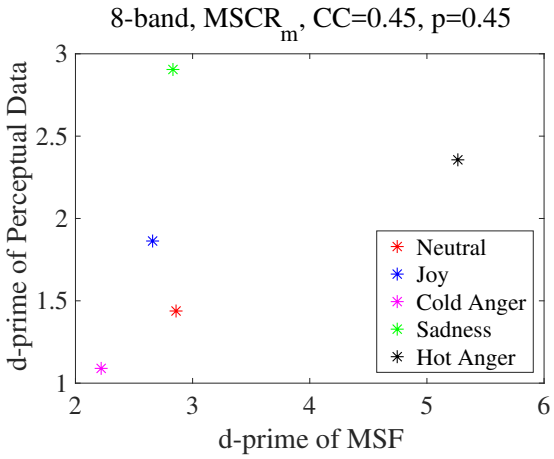


(d) MSKT$_k$

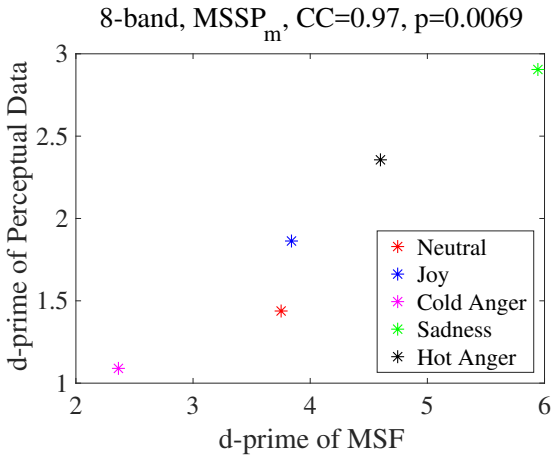


(e) MSKT$_k$

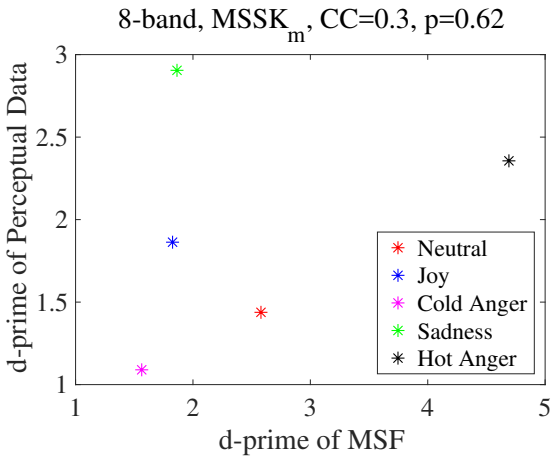
Figure D.2: The scatterplot of the d' of the perceptual data of vocal-emotion recognition experiment and modulation spectral features on modulation frequency domain for for 4-band NVS.



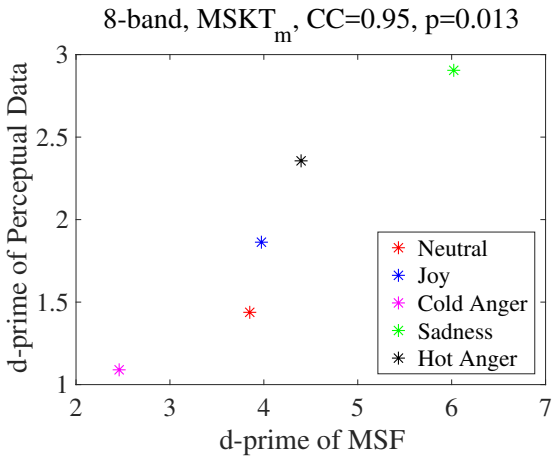
(a) MSCR$_m$



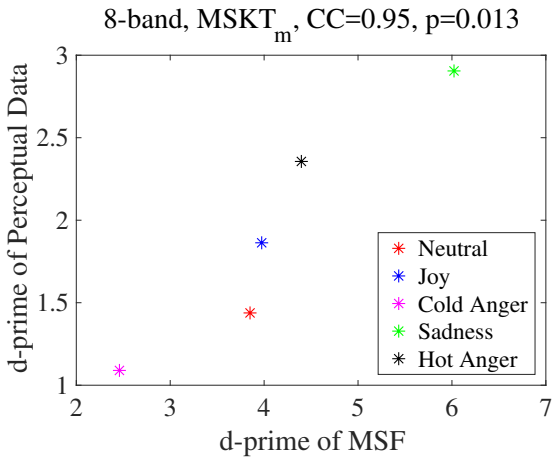
(b) MSSP$_m$



(c) MSSK$_m$

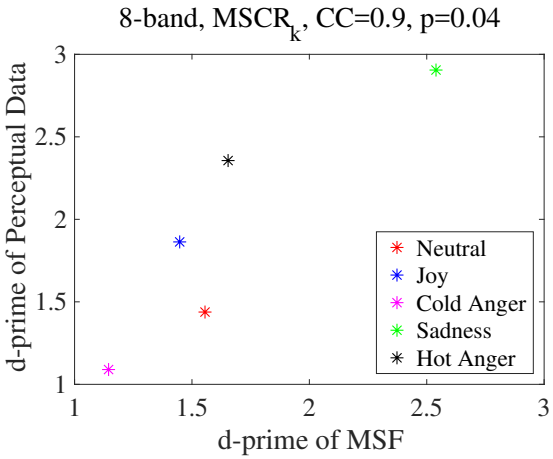


(d) MSKT$_m$

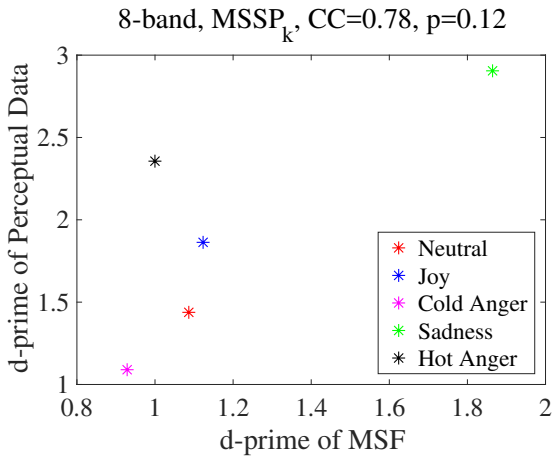


(e) MSKT$_m$

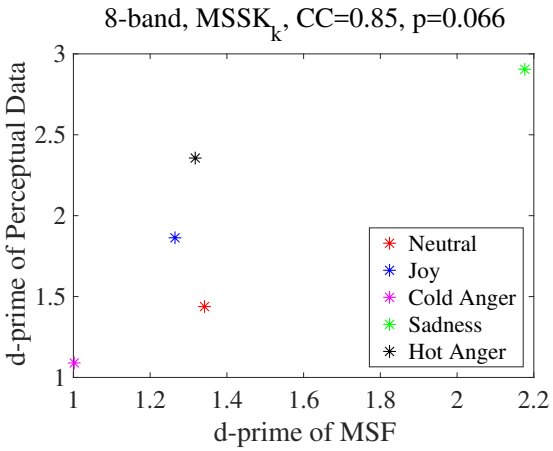
Figure D.3: The scatterplot of the d' of the perceptual data of vocal-emotion recognition experiment and modulation spectral features on acoustic frequency domain for for 8-band NVS.



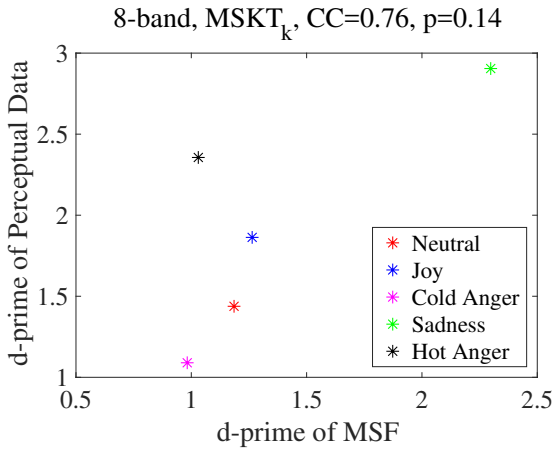
(a) MSCR$_k$



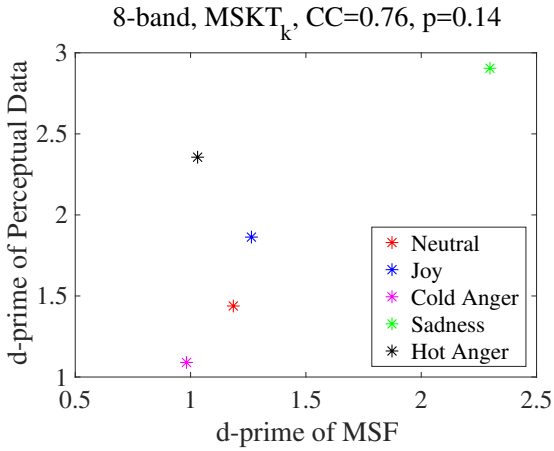
(b) MSSP$_k$



(c) MSSK$_k$

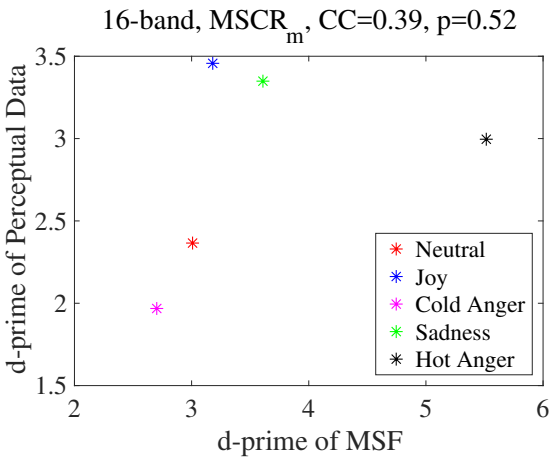


(d) MSKT$_k$

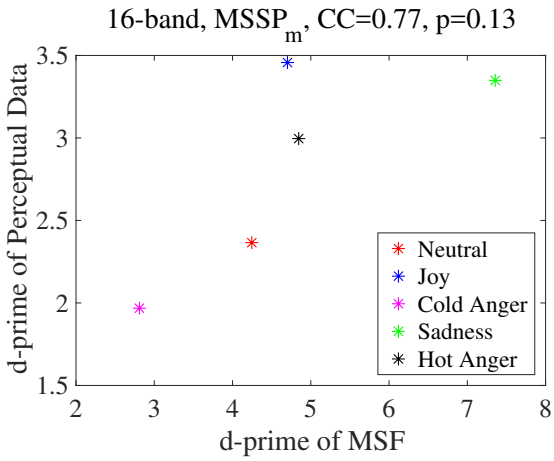


(e) MSKT$_k$

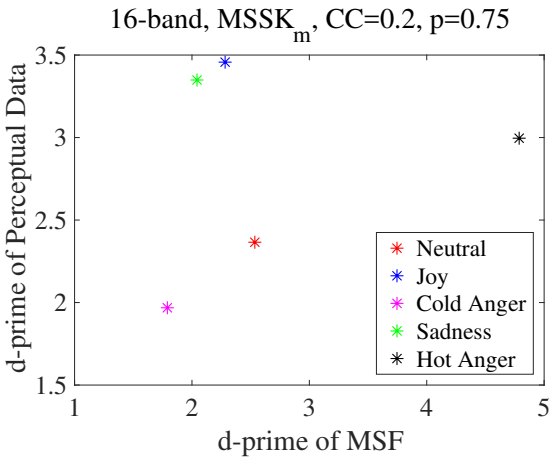
Figure D.4: The scatterplot of the d' of the perceptual data of vocal-emotion recognition experiment and modulation spectral features on modulation frequency domain for for 8-band NVS.



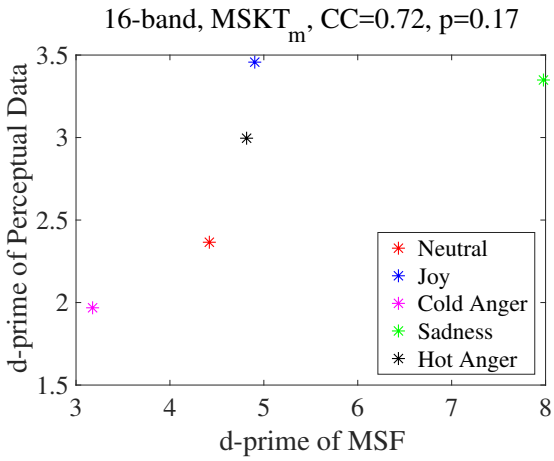
(a) MSCR_m



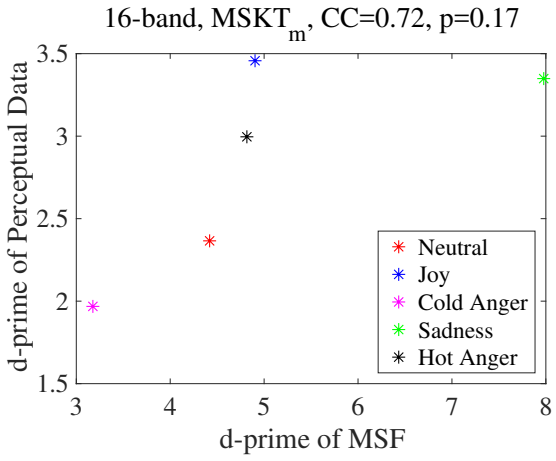
(b) MSSP_m



(c) MSSK_m

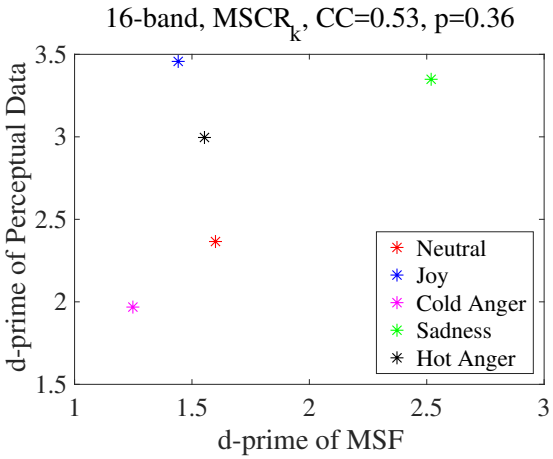


(d) MSKT_m

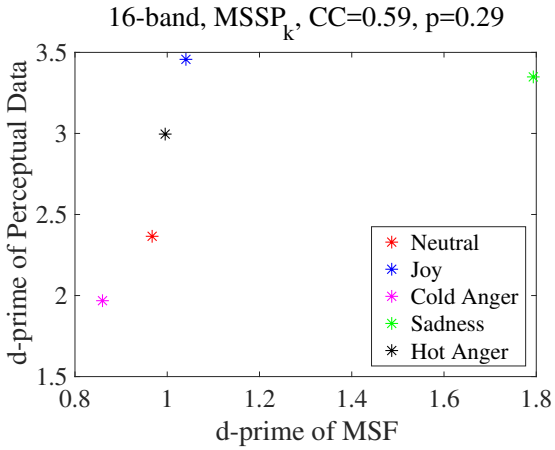


(e) MSKT_m

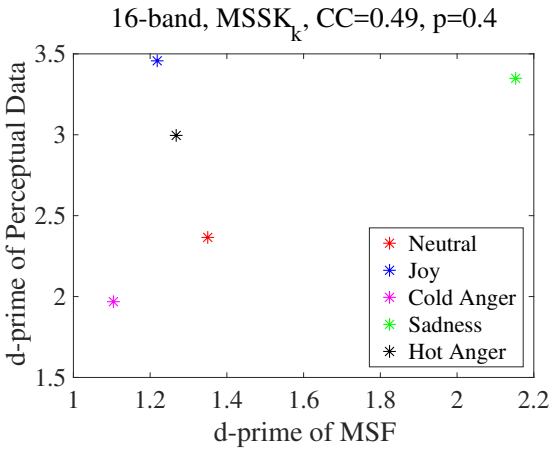
Figure D.5: The scatterplot of the d' of the perceptual data of vocal-emotion recognition experiment and modulation spectral features on acoustic frequency domain for for 16-band NVS.



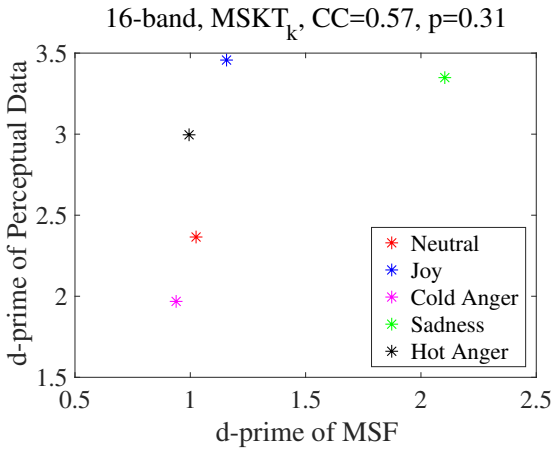
(a) MSCR$_k$



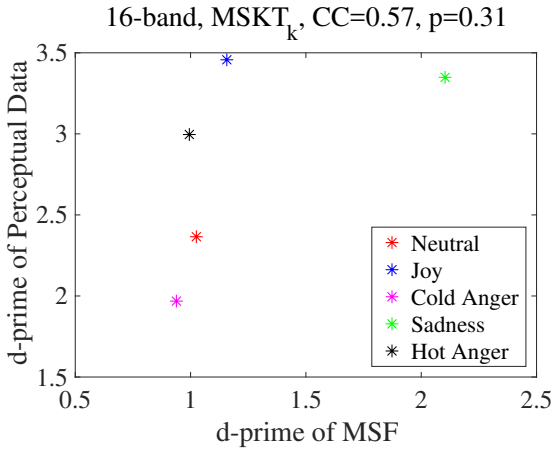
(b) MSSP$_k$



(c) MSSK$_k$



(d) MSKT$_k$



(e) MSKT$_k$

Figure D.6: The scatterplot of the d' of the perceptual data of vocal-emotion recognition experiment and modulation spectral features on modulation frequency domain for for 16-band NVS.

Bibliography

[1] T. Kitamura, T. Nakama, H. Ohmura, and H. Kawamoto, "Measurement of perceptual speaker similarity for sentence speech in atr speech database," *Journal of Acoustical Society of Japan (J)*, vol. 71, no. 10, pp. 516–525, 2015.

[2] H. Fujisaki, *Prosody, Models and Spontaneous Speech*, pp. 27–42. Computing Prosody, Springer, 1996.

[3] M. Akagi and T. Ienaga, "Speaker individuality in fundamental frequency contours and its control," *Journal of Acoustical Society of Japan (E)*, vol. 18, no. 2, pp. 73–80, 1997.

[4] R. E. Remez, J. M. Fellowes, and P. E. Rubin, "Talker identification based on phonetic information," *Journal of American Physiological Society*, vol. 23, no. 3, pp. 651–666, 1997.

[5] T. Kitamura and M. Akagi, "Speaker individualities in speech spectral envelopes," *Journal of Acoustical Society of Japan (E)*, vol. 16, no. 5, pp. 283–289, 1995.

[6] T. Kitamura, K. Honda, and H. Takemoto, "Individual variation of the hypopharyngeal cavities and its acoustic effects," *Acoustic Science and Technology*, vol. 26, no. 1, pp. 16–26, 2005.

[7] K. R. Scherer, "Vocal communication of emotion: A review of research paradigms," *Speech Communication*, vol. 40, pp. 227–256, 2003.

[8] R. Banse and K. R. Scherer, "Acoustic profiles in vocal emotion expression," *Journal of Personality and Social Psychology*, vol. 70, no. 3, pp. 614–636, 1996.

[9] C.-F. Huang and M. Akagi, "A three–layered model for expressive speech perception," *Speech Communication*, vol. 50, pp. 810–828, 2008.

[10] T. Dau, D. Puschel, and A. Kohlrausch, "A quantitative model of the "effective" signal processing in the auditory system. i. model structure," *Journal of Acoustical Society of America*, vol. 99, no. 6, pp. 3615–3622, 1996.

[11] T. Dau, D. Puschel, and A. Kohlrausch, "A quantitative model of the "effective" signal processing in the auditory system. ii. simulations and measurements," *Journal of Acoustical Society of America*, vol. 99, no. 6, pp. 3623–3631, 1996.

[12] S. D. Ewert and T. Dau, "Characterizing frequency selectivity for envelope fluctuations," *Journal of Acoustical Society of America*, vol. 108, no. 3, pp. 1181–1196, 2000.

[13] R. V. Shannon, F.-G. Zeng, V. Kamath, J. Wygonski, and M. Ekelid, "Speech recognition with primarily temopral cues," *Science*, vol. 270, no. 5234, pp. 303–304, 1995.

[14] R. O. Tachibana, Y. Sasaki, and H. Riquimaroux, "Relative contributions of spectral and temporal resolutions to the perception of syllables, words and sentences in noise–vocoded speech," *Acoustical Science and Technology*, vol. 34, no. 4, pp. 263–270, 2013.

[15] P. C. Loizou, M. Dorman, and Z. Tu, "On the number of channels needed to understand speech," *Journal of Acoustical Society of America*, vol. 106, no. 4, pp. 2097–2103, 1999.

[16] L. Xu and B. E. Pfingst, "Spectral and temporal cues for speech recognition: Implications for auditory prostheses," *Hearing Research*, vol. 242, pp. 132–140, 2008.

[17] H. Riquimaroux, "Perception of noise–vocoded speech sounds," *Journal of Acoustical Society of Japan (J)*, vol. 61, no. 5, pp. 273–278, 2005.

[18] R. Drullman, J. M. Festen, and R. Plomp, "Effect of temporal envelope smearing on speech reception," *Journal of Acoustical Society of America*, vol. 95, no. 2, pp. 1053–1064, 1994.

[19] R. Drullman, J. M. Festen, and R. Plomp, "Effect of reducing slow temporal modulations on speech reception," *Journal of Acoustical Society of America*, vol. 95, no. 5, pp. 2670–2680, 1994.

[20] L. Xu, C. S. Thompson, and B. E. Pfingst, "Relative contributions of spectral and temporal cues for phoneme recognition," *Journal of Acoustical Society of America*, vol. 117, no. 5, pp. 3255–3267, 2005.

[21] S. Rosen, "Temporal information in speech: Acoustic, auditory and linguistic aspects," *Philosophical Transactions: Biological Sciences*, vol. 336, no. 1278, pp. 367–373, 1992.

[22] F.-G. Zeng, S. Rebscher, W. Harrison, X. Sun, and H. Feng, "Cochlear implants: system design, integration and evaluation," *IEEE Reviews in Biomedical Engineering*, vol. 1, pp. 115–142, 2008.

[23] M. Vongphoe and F.-G. Zeng, "Speaker recognition with temporal cues in acoustic and electric hearing," *Journal of Acoustical Society of America*, vol. 118, no. 2, pp. 1055–1061, 2005.

[24] J. Gonzalez and J. C. Oliver, "Gender and speaker identification as a function of the number of channels in spectrally reduced speech," *Journal of Acoustical Society of America*, vol. 118, no. 1, pp. 461–470, 2005.

[25] V. Krull, X. Luo, and K. I. Kirk, "Talker–identification training using simulations of binaurally combined electric and acoustic hearing: Generalization to speech and emotion recognition," *Journal of Acoustical Society of America*, vol. 131, no. 4, pp. 3069–378, 2012.

[26] T. Vongpaisal, S. E. Trehub, E. G. Schellenberg, P. van Lieshout, and B. C. Papsin, "Children with cochlear implants recognize their mother's voice," *Ear Hearing*, vol. 31, no. 4, pp. 555–566, 2010.

[27] M. Chatterjee, D. J. Zion, M. L. Deroche, B. A. Burianek, C. J. Limb, A. P. Goren, A. M. Kulkarni, and J. A. Christensen, "Voice emotion recognition by cochlear–implanted children and their normally–hearing peers," *Hearing Research*, vol. 322, pp. 151–162, 2015.

[28] X. Luo, Q.-J. Fu, and J. J. G. III, "Vocal emotion recognition by normal–hearing listeners and cochlear implant users," *Trends in Amplification*, vol. 11, no. 4, pp. 301–315, 2007.

[29] T. Chiba and M. Kajiyama, *The vowel : its nature and structure*. Tokyo–Kaiseikan, 1941.

[30] K. Itoh and S. Saito, "Effects of acoustical feature parameters of speech on perceptual identification of speaker," *The IEICE Transactions*, vol. J65–A, no. 1, pp. 101–108, 1982.

[31] M. Hashimoto, S. Katagawa, and N. Higuchi, "Quantitative analysis of acoustic features affecting speaker identification," *Journal of Acoustical Society of Japan (J)*, vol. 54, no. 3, pp. 169–178, 1998.

[32] W. Zhu and H. Kasuya, "Study of perceptual contribution of static and dynamic features of vocal tract charateristics to speaker individuality," *The Jounral of Information Processing Society of Japan*, vol. 19, no. 13, pp. 69–65, 1997.

[33] T. Kitamura, M. Akagi, and S. Kitazawa, "Perceptual effect of spectral trajectory patterns for speaker identification," *Transactions of the Technical Committee of Psychological and Physiological Acoustics*, vol. H–98–97, 1998.

[34] H. Kuwabara and K. Ohgushi, "The role of formant frequencies and bandwidths in the perception of speaker," *The IEICE Transactions*, vol. J69–A, no. 4, pp. 509–517, 1986.

[35] T. Kitamura and M. Akagi, "Significant cues in spectral envelope of isolated vowels for speaker identification," *Journal of Acoustical Society of Japan (J)*, vol. 53, no. 3, pp. 185–191, 1997.

[36] T. Kitamura and T. Saitou, "Effects of acoustic modification on perception of speaker characteristics for sustained vowels," *Acoustic Science and Technology*, vol. 28, no. 6, pp. 434–437, 2007.

[37] K. Amino, T. Sugawara, and T. Arai, "Idiosyncrasy of nasal sounds in human speaker identification and their acoustic properties," *Acoustical Science and Technology*, vol. 27, no. 4, pp. 233–235, 2006.

[38] J. Kim, A. Toutios, S. Lee, and S. S. Narayanan, "A kinematic study of critical and non–critical articulators in emotional speech production," *Journal of Acoustical Society of America*, vol. 137, no. 3, pp. 1411–1429, 2015.

[39] D. Erickson, C. Menezes, and A. Fujino, "Some articulatory measurements of real sadness," in *INTERSPEECH 2004*, 2004.

[40] J.-A. Bachorowski, "Vocal expression and perception of emotion," *Current Directions in Psychological Science*, vol. 8, no. 2, pp. 53–57, 1999.

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

[42] Y. Xue, Y. Hamada, and M. Akagi, "Emotional speech synthesis system based on a three-layered model using a dimensional approach," in *2015 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, pp. 505–514, 2015.

[43] Y. Xue, Y. Hamada, and M. Akagi, "Voice conversion to emotional speech based on three-layered model in dimensional approach and parameterization of dynamic features in prosody," in *2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, pp. 1–6, 2016.

[44] B. C. J. Moore, *An introduction to the psychology of hearing.* London, Elsevier, 6th ed., 2013.

[45] M. Slaney, "An efficient implementation of the patterson–holdsworth auditory filter bank," tech. rep., Apple Computer Technical Report, 1993.

[46] M. Unoki, T. Irino, B. Glasber, B. C. J. Moore, and R. D. Patterson, "Comparison of the roex and gammachirp filters as representations of the audiotry filter," *Journal of Acoustical Society of America*, vol. 120, no. 3, pp. 1474–1492, 2006.

[47] B. R. Glasberg and B. C. Moore, "Derivation of auditory filter shapes from notched-noise data derivation of auditory filter shapes from notched-noise data derivation of audiotry filter shapes from notched–noise data," *Hearing Research*, vol. 47, pp. 103–138, 1990.

[48] E. Zwicker and E. Terhardt, "Analytical expressions for critical band rate and critical bandwidth as a function of frequency," *Journal of Acoustical Society of America*, vol. 68, pp. 1523–1525, 1980.

[49] P. C. Loizou, "Mimicking the human ear," *IEEE Signal Processing Magazine*, pp. 101–130, 1998.

[50] T. Dau and B. Kollmeier, "Modeling auditory processing of amplitude modulation. i. detection and masking with narrow-band carriers," *Journal of Acoustical Society of America*, vol. 102, no. 5, pp. 2892–2905, 1997.

[51] T. Dau and B. Kollmeier, "Modeling auditory processing of amplitude modulation. ii. spectral and temporal integration," *Journal of Acoustical Society of America*, vol. 102, no. 5, pp. 2906–2919, 1997.

[52] M. L. Jepsen, S. D. Ewert, and T. Dau, "A computational model of human auditory signal processing and perception," *Journal of Acoustical Society of America*, vol. 124, no. 1, pp. 422–438, 2008.

[53] S. Jorgensen, S. D. Ewert, and T. Dau, "A multi–resolution envelope–power based model for speech intelligibility," *Journal of Acoustical Society of America*, vol. 134, no. 1, pp. 436–446, 2013.

[54] I. C. Bruce, "Physiologically based predictors of speech intelligibility," *Acoustics Today*, vol. 13, no. 1, pp. 28–35, 2017.

[55] J. Xiang, D. Poeppel, and J. Z. Simon, "Physiological evidence for auditory modulation filterbanks: Cortical responses to concurrent modulations," *Journal of Acoustical Society of America*, vol. 133, no. 1, pp. 7–12, 2013.

[56] P. C. Loizou and O. Poroy, "Minimun spectral contrast needed for vowel identification by normal hearing and cochlear implant listeners," *Journal of Acoustical Society of America*, vol. 110, no. 3, pp. 1619–1627, 2001.

[57] Q.-J. Fu, S. Chinchilla, and J. J. Galvin, "The role of spectral and cues in voice gender discrimination by normal–hearing listeners and cochlear implant users," *Journal of the Association for Research in Otolaryngology*, vol. 5, pp. 253–260, 2004.

[58] S. Gilbers, C. Fuller, D. Gilbers, M. Broersma, M. Goudbeek, R. Free, and D. Baskent, "Normal–hearing listeners' and cochlear implant users' perception of pitch cues in emotional speech," *i–Perception*, vol. 6, no. 5, pp. 1–19, 2015.

[59] R. V. Shannon, F.-G. Zeng, and J. Wygonski, "Speech recognition with altered spectral distribution of envelope cues," *Journal of Acoustical Society of America*, vol. 104, no. 4, pp. 2467–2476, 1998.

[60] T. Araki, K. Ueda, and Y. Nakajima, "The effects of amplitude envelope coherence across frequency bands of noise-vocoded speech," *Transactions of the Technical Committee of Psychological and Physiological Acoustics*, vol. 38, no. 8, pp. 797–802, 2008.

[61] T. Araki, K. Ueda, and Y. Nakajima, "Effects of exchanging amplitude fluctuations between frequency bands of noise–vocoded speech," *Transactions of the Technical Committee of Psychological and Physiological Acoustics*, vol. 39, no. 8, pp. 573–578, 2009.

[62] K. Doumoto, K. Ueda, Y. Nakajima, W. Ellermeier, and F. Kattner, "Disruptive effect of unattended noise–vocoded speech on recall of visually presented digits: Interaction between the number of frequency bands and languages," *Transactions of the Technical Committee of Psychological and Physiological Acoustics*, vol. 41, no. 9, pp. 663–670, 2011.

[63] K. Doumoto, K. Ueda, Y. Nakajima, and W. Ellermeier, "The effects of boundary frequency differences of japanese noise–vocoded speech on intelligibility: the differences between japanese and german boundaries are negligible," *Transactions of the*

Technical Committee of Psychological and Physiological Acoustics, vol. 40, no. 10, pp. 783–788, 2010.

[64] S. Isaji, K. Ueda, and Y. Nakajima, "Effects of frequency–band elimination on identification of noise–vocoded japanese syllables," *Transactions of the Technical Committee of Psychological and Physiological Acoustics*, vol. 44, no. 1, pp. 45–51, 2014.

[65] L. Atlas and S. A. Shamma, "Joint acoustic and modulation frequency," *Journal on Applied Signal Processing*, vol. 7, pp. 668–675, 2003.

[66] R. Drullman, "Temporal envelope and fine structure cues for speech intelligibility," *Journal of Acoustical Society of America*, vol. 97, no. 1, pp. 585–592, 1995.

[67] T. M. Elliott and F. E. Theunissen, "The modulation transfer function for speech intelligibility," *PLoS Comutational Biology*, vol. 5, no. 3, pp. 1–14, 2009.

[68] R. Decorsiere, P. L. Sondergaard, E. N. MacDonald, and T. Dau, "Inversion of audiotry spectrograms, traditional spectrograms and other envleope representations," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 23, no. 1, pp. 46–56, 2015.

[69] T. H. Falk and W.-Y. Chan, "Modulation spectral features for robust far–field speaker identification," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 18, no. 1, pp. 90–100, 2010.

[70] T. Kinnunen, "Jount acoustic–modulation frequnecy for speaker recognition," in *ICASSP 2006*, pp. 665–668, 2006.

[71] M. Kazama, M. Tohyama, and Y. Ymasaki, "Speaker characteristics represented by narrow–band temporal–envelope correlation matrices," *The IEICE Transactions*, vol. J92–A, no. 4, pp. 205–215, 2009.

[72] H. Lei, B. T. Meyer, and N. Mirghafori, "Spectro–temporal gabor features for speaker recognition," in *ICASSP 2012*, pp. 4241–4244, 2012.

[73] S. M. Schimmel, L. E. Atlas, and K. Nie, "Feasibility of single channel speaker separation based on modulation frequency analysis," in *ICASSP 2007*, pp. 605–608, 2007.

[74] S. van Vuuren and H. Hermansky, "On the importance of components of the modulation spectrum for speaker verification," in *ICSLP 1998*, 1998.

[75] T. Chaspari, D. Dimitriadis, and P. Maragos, "Emotion classification of speech using modulation features," in *EUSIPCO 2014*, pp. 1552–1556, 2014.

[76] T.-S. Chi, L.-Y. Yeh, and C.-C. Hsu, "Robust emotion recognition by spectro–temporal modulation statistic features," *Journal of Ambient Intelligence and Humanized Computing*, vol. 3, no. 1, pp. 47–60, 2012.

[77] S. Wu, T. H. Falk, and W.-Y. Chan, "Automatic speech emotion recognition using modulation spectral features," *Speech Communication*, vol. 53, pp. 768–785, 2011.

[78] I. T. Union, "Objective measurement of active speech level," *ITU-T*, vol. P.56, no. Switzerland, 1993.

[79] Z. Zhu, R. Miyauchi, Y. Araki, and M. Unoki, "Study on linguistic information and speaker individuality contained in temporal envelope of speech," *Acoustic Science and Technology*, vol. 37, pp. 258–261, 2016.

[80] T. Houtgast and H. J. M. Steeneken, "A review of the mtf concept in room acoustics and its use for estimating speech intelligibility in auditoria," *Journal of Acoustical Society of America*, vol. 77, pp. 1069–1077, 1985.

[81] T. Arai and S. Greenberg, "The temporal properties of spoken japanese are similar to those of english," in *Proceedings of Eurospeech*, pp. 1011–1014, 1997.

[82] H. Stanislaw and N. Todorov, "Calculation of signal detection theory measures.," *Behavior research methods, instruments, & computers*, vol. 31, no. 1, pp. 137–149, 1999.

[83] X. Luo and Q.-J. Fu, "Enhancing chinese tone recognition by manipulating amplitude envelope: Implications for cochlear implants," *Journal of Acoustical Society of America*, vol. 116, no. 6, pp. 3659–3667, 2004.

[84] J. H. McDermott and E. P. Simoncelli, "Sound texture perception via statistics of the audiotry periphery: Evidence from sound synthesis," *Neuron*, vol. 71, pp. 926–940, 2011.

Publications

Publications related to the present study

Journal Paper

[1] Zhi Zhu, Ryota Miyauchi, Yukiko Araki, and Masashi Unoki, "Contributions of Temporal Cue on the Perception of Speaker Individuality and Vocal Emotion for Noise-vocoded Speech," Acoustical Science and Technology. (Conditional accepted)

Journal Letter

[2] Zhi Zhu, Ryota Miyauchi, Yukiko Araki, and Masashi Unoki, "Study on linguistic information and speaker individuality contained in temporal envelope of speech," Acoustical Science and Technology, Vol. 37, No. 5, pp. 258–261, 2016.

International Conference

[3] Zhi Zhu, Ryota Miyauchi, Yukiko Araki, and Masashi Unoki, "Feasibility of Vocal Emotion Conversion on Modulation Spectrogram for Simulated Cochlear Implants," 25th European Signal Processing Conference (EUSIPCO2017), pp. 1884–1888, 2017.

[4] Zhi Zhu, Ryota Miyauchi, Yukiko Araki, and Masashi Unoki, "Important role of temporal cues in speaker identification for simulated cochlear implants," Proc. of the

1st International Workshop on Challenges in Hearing Assistive Technology (CHAT-2017), pp. 51–55, 2017.

[5] Zhi Zhu, Ryota Miyauchi, Yukiko Araki, and Masashi Unoki, "The role of spectral and temporal cues for vocal emotion recognition by cochlear implant simulations," Acoustics ' 17, Journal of Acoustic Society of America, Vol. 141, No. 5, Pt. 2, pp. 3816, 2017.

[6] Zhi Zhu, Ryota Miyauchi, Yukiko Araki, and Masashi Unoki, "Recognition of Vocal Emotion in Noise-vocoded Speech by Normal Hearing and Cochlear Implant Listeners," 5th Joint Meeting Acoustical Society of America and Acoustical Society of Japan, Journal of Acoustic Society of America, Vol. 140, No. 4, Pt. 2, pp. 3271, 2016.

[7] Zhi Zhu, Ryota Miyauchi, Yukiko Araki, and Masashi Unoki, "Modulation spectral features for predicting vocal emotion recognition by simulated cochlear implants," INTERSPEECH 2016, pp. 262–266, 2016.

[8] Zhi Zhu, Ryota Miyauchi, and Masashi Unoki, "Analysis of Modulation-Spectral Features Extracted from Japanese Emotional Speech," The Taiwan/Japan Joint Research Meeting on Psychological & Physiological Acoustics and Electroacoustics, Proc. Auditory Research Meeting, The Acoustical Society of Japan, Vol. 45, No. 7, pp. 589–594, 2015.

[9] Zhi Zhu, Ryota Miyauchi, and Masashi Unoki, "Analysis of Speaker Individual Differences on Modulation Spectrum," 2015 RISP International Workshop on Nonlinear Circuits, Communications and Signal Processing (NCSP15), pp. 17–20, 2015.

Domestic Conference

[10] 朱治, 宮内良太, 荒木友希子, 鵜木祐史, "雑音駆動音声の個人性知覚に寄与する変調周波数成分の検討," 2017 年日本音響学会秋季研究発表会, pp. 359–362, 2017.

[11] 朱治, 宮内良太, 荒木友希子, 鵜木祐史, "変調周波数帯域の制御が雑音駆動音声の感情知覚に与える影響の検討," 2017 年日本音響学会春季研究発表会, pp. 1491–1494,

2017.

[12] Xinfeng Li, Zhi Zhu, and Masato Akagi, “Acoustic feature selection for improving estimation of emotions using a three layer model,” 2017 Spring Meeting of Acoustical Society of Japan, pp. 117–120, 2017.

[13] Zhi Zhu, Ryota Miyauchi, Yukiko Araki, and Masashi Unoki, “Effect of modifying modulation spectrogram on vocal emotion perception for noise-vocoded speech,” Proc. Auditory Research Meeting, The Acoustical Society of Japan, Vol. 46, No. 9, pp. 571–576, 2016.

[14] 朱治, 宮内良太, 荒木友希子, 鵜木祐史, “雑音駆動音声の感情知覚と振幅包絡線情報の関係に関する検討,” 日本音響学会聴覚研究会資料, Vol. 46, No. 2, pp. 73–76, 2016.

[15] 朱治, 宮内良太, 荒木友希子, 鵜木祐史, “変調フィルタバンクを用いた感情音声における変調スペクトルの特徴解析,” 2016 年日本音響学会春季研究発表会, pp. 507–510, 2016.

[16] 朱治, 宮内良太, 鵜木祐史, “雑音駆動音声の個人性知覚に関する基礎的検討,” 平成 27 年度電気関係学会北陸支部連合大会, 2015.

[17] 朱治, 宮内良太, 鵜木祐史, “変調スペクトルの帯域を制限した雑音駆動音声の個人性知覚に関する研究,” 2015 年日本音響学会春季研究発表会, pp. 491–494, 2015.

[18] 朱治, 宮内良太, 鵜木祐史, “音声の変調スペクトルに現れる個人差の分析,” 日本音響学会聴覚研究会資料, Vol. 44, No. 7, pp. 457–460, 2014.

Other publications

Journal Letter

[19] Zhi Zhu, Katsuhiko Yamamoto, Masashi Unoki, and Naofumi Aoki, "Study on Scramble Method for Speech Signal by Using Random–Bit Shift of Quantization," Journal of Signal Processing, Vol. 18, No. 6, pp. 303–307, 2014.

[20] Katsuhiko Yamamoto, Zhi Zhu, Masashi Unoki, and Naofumi Aoki, "Study on Semi–scramble Method for Speech Signals Based on Phonemic Restoration," Journal of Signal Processing, Vol. 18, No. 4, pp. 205–208, 2014.

International Conference

[21] Zhichao Peng, Zhi Zhu, Masashi Unoki, Jianwu Dang, and Masato Akagi, "Speech emotion recognition using MPCRNN based on Gammatone auditory filterbank," Asia-Pacific Signal and Information Processing Association Annual Summit and Conference 2017 (APSIPA 2017), 2017.

[22] Zhi Zhu, Katsuhiko Yamamoto, Masashi Unoki, and Naofumi Aoki, "Study on Scramble Method for Speech Signal by Using Random–Bit Shift of Quantization," 2014 RISP International Workshop on Nonlinear Circuits, Communications and Signal Processing (NCSP14), pp. 109–112, 2014.

[23] Katsuhiko Yamamoto, Zhi Zhu, Masashi Unoki, and Naofumi Aoki, "Study on Semi–scramble Method for Speech Signals Based on Phonemic Restoration," 2014 RISP International Workshop on Nonlinear Circuits, Communications and Signal Processing (NCSP14), pp. 201–204, 2014.

Domestic Conference

[24] 関谷伸一, 朱治, 鵜木祐史, "雑音駆動音声の言語・非言語知覚と室内音響特性による影響の検討," 日本音響学会聴覚研究会資料, Vol. 47, No. 7, pp. 551–556, 2017.

[25] 朱治, 山本克彦, 鵜木祐史, 青木直史, “量子化ビットのランダムシフトを利用した音声スクランブル法,” 信学技報, EMM2013-109, Vol. 113, No. 480, pp. 57–62, 2014.

[26] 山本克彦, 朱治, 鵜木祐史, 青木直史, “音韻修復現象に着目した半開示音声スクランブル法の検討,” 信学技報, EMM2013-78, Vol. 113, No. 291, pp. 59–64, 2013.

[27] 朱治, 山本克彦, 鵜木祐史, 青木直史, “量子化ビットのランダムシフトによる音声スクランブル法の検討,” 平成 25 年度電気関係学会北陸支部連合大会, 2013.

[28] 山本克彦, 朱治, 鵜木祐史, 青木直史, “音韻修復現象に着目した音声半開示スクランブル法の検討,” 平成 25 年度電気関係学会北陸支部連合大会, 2013.