PAPER

Contributions of temporal cue on the perception of speaker individuality and vocal emotion for noise-vocoded speech

Zhi Zhu[1,*], Ryota Miyauchi[1,†], Yukiko Araki[2,‡] and Masashi Unoki[1,§]

[1]*Japan Advanced Institute of Science and Technology,*
1–1 Asahidai, Nomi, 923–1292 Japan
[2]*Kanazawa University,*
Kakuma-machi, Kanazawa, 920–1192 Japan

(*Received 16 May 2017, Accepted for publication 22 November 2017*)

Abstract: This paper investigates the importance of temporal cues in the perception of speaker individuality and vocal emotion. Experiments of speaker and vocal-emotion recognition were carried out using an analysis/synthesis method of noise-vocoded speech (NVS). The temporal resolution of NVS was controlled by varying the upper limit of modulation frequency (0, 0.5, 1, 2, 4, 8, 16, 32, and 64 Hz). In addition, the role of temporal cue in the different spectral resolution condition was also investigated by varying the number of channels (4, 8, and 16). The results demonstrated that temporal resolution contributes to the recognition of both speaker and vocal emotion. Therefore, temporal cues are found to be important for the perception of not only linguistic information but also speaker individuality and vocal emotion. On the other hand, the performance of speaker recognition was less 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 shown to be important for recognizing only neutral, joy, and cold anger, but not sadness or hot anger. The important modulation frequency band for the perception of nonlinguistic information was suggested to be higher than that of linguistic information.

Keywords: Temporal cue, Speaker individuality, Vocal emotion, Noise-vocoded speech, Speech perception

PACS number: 43.71.Bp [doi:10.1250/ast.39.234]

1. INTRODUCTION

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). The temporal envelope of speech has been proved to be an important cue for speech perception from the studies using noise-vocoded speech (NVS) [1–4]. NVS is generated by replacing the carriers with band-limited noise, so the spectral cue is reduced dramatically and the temporal cue is preserved. Shannon *et al.* showed that NVS with only four bands is sufficient to achieve good vowel, consonant, and sentence recognition [1]. Therefore, human can successfully perceive linguistic information using the temporal envelope of speech signal as a primary cue.

From the viewpoint of auditory perception, the temporal cue provided by the temporal envelope of speech is very important. The signal processing in the auditory peripheral system can be computationally modeled as band-pass filtering and envelope extracting [5,6]. Cochlear implant (CI) mimic such signal processing in the auditory peripheral system [7]. As the number of channels in CI device is limited, a poor spectral cue is provided, and the temporal envelope is used as a primary cue by CI listeners. NVS was usually used as a CI simulation to simulate the poor spectral cue provided by CI device. CI listeners can accurately recognize linguistic information, as CI device can provide enough temporal cues.

The importance of temporal cues in the perception of linguistic information has been studied by many researchers. Drullman *et al.* investigated the important modulation frequency bands for speech perception by low- and high-pass filtering on the temporal envelope [8,9]. They showed that the modulation frequency bands from 4 to 16 Hz contained important cues related to linguistic information.

[*]e-mail: zhuzhi@jaist.ac.jp
[†]e-mail: ryota@jaist.ac.jp
[‡]e-mail: yukikoa@staff.kanazawa-u.ac.jp
[§]e-mail: unoki@jaist.ac.jp

Xu *et al.* attempted to elucidate the importance of temporal cues for phoneme recognition using NVS [10]. The results showed that vowel recognition plateaued at the 4 Hz upper limit of modulation frequency. Tachibana *et al.* used a similar experimental paradigm to demonstrate that the modulation frequency band below 8 Hz is important for sentence recognition [2]. However, human speech includes not only linguistic information but also nonlinguistic information such as speaker individuality and vocal emotion.

Previous studies about the perception of nonlinguistic information were always based on the source-filter theory from the viewpoint of speech production. For speaker individuality, spectral envelope and formants of speech have been proved to contribute speaker recognition [11–13]. For vocal emotion, previous works focused on the acoustic features conveyed in speech, such as F0, spectral envelope, intensity, and speech rate [14–16]. For both speaker individuality and vocal emotion, the time-averaged acoustic features were investigated. However, the temporal cues provided by the dynamic components of speech are also considered to be important for perceiving nonlinguistic information.

As CI listeners using the temporal cue as a primary cue, it is important to clarify the contributions of temporal cue on the perception of nonlinguistic information. It has been known that CI listeners' performances of vocal-emotion and speaker recognition are poorer than normal-hearing (NH) listeners, as the poor spectral cues provided by CI device [17–20]. Nonetheless, for speaker recognition, Krull *et al.* showed that training results in improved recognition rates of speaker in CI simulations [21]. Moreover, child CI listeners succeeded in differentiating their mothers' utterances from those of other people [22]. For vocal-emotion recognition, Chatterjee *et al.* [19] found that the mean performance of CI listeners was similar to that of NH listeners with 8-band NVS which was much better than the chance level. The temporal envelope of speech has been suggested to potentially be an important cue in the perception of nonlinguistic information.

However, the role of such temporal cues is still not clear. To clarify the mechanism of nonlinguistic information perception, the role of temporal cues in speaker and vocal-emotion recognition needs to be clarified. The elucidation of temporal cues can also be used to optimize CI systems for their users to improve their recognition of speakers and vocal emotions. Speaker recognition and vocal-emotion recognition need to be investigated together to discuss the common contribution of temporal cues for nonlinguistic information and how it differs from that for linguistic information. In a previous study, we examined word and speaker recognition using NVS while systematically varying the upper limit of modulation frequency [23]. Results suggested that the modulation frequency band below 5 Hz is important for word recognition.

In the present study, we extended this work to examine the perception of both speaker individuality and vocal emotion. This paper aims to clarify the role of temporal modulation cue in speaker and vocal-emotion recognition using NVS. Furthermore, the effects of different spectral resolutions are also investigated. In the experiment, speaker recognition 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. In addition, 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 cues are reduced further when the number of channels and upper limit of modulation frequency decrease, respectively. The experimental paradigm used in this study can clarify the important modulation frequency band for speaker and vocal-emotion recognition.

This paper is organized as follows. Section 2 introduces the signal processing method to generate NVS. Section 3 describes the speaker recognition experiment using NVS. Section 4 details the vocal-emotion recognition experiment. Section 5 discusses the contribution of temporal cues to the perception of speaker individuality and vocal emotion. Section 6 summarizes the results and discussion.

2. SIGNAL PROCESSING: NOISE-VOCODER METHOD

Figure 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 [24]. Speech signals were then divided into several frequency bands by using a band-pass filterbank. The bandwidth and boundary frequencies of the band-pass filters (6th-order Butterworth infinite impulse response (IIR) filter) were defined using $\mathrm{ERB_N}$ (Equivalent Rectangular Bandwidth) and $\mathrm{ERB_N}$-number scale [25]. The $\mathrm{ERB_N}$-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 $\mathrm{ERB_N}$-number. The relationship between $\mathrm{ERB_N}$-number and acoustic frequency is defined as follows:

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

where f is acoustic frequency in Hz. The boundary frequencies of the band-pass filters were defined from 3 to 35 $\mathrm{ERB_N}$-number with bandwidth as 2, 4, or 8 $\mathrm{ERB_N}$. Therefore, the numbers of channels of the band-pass

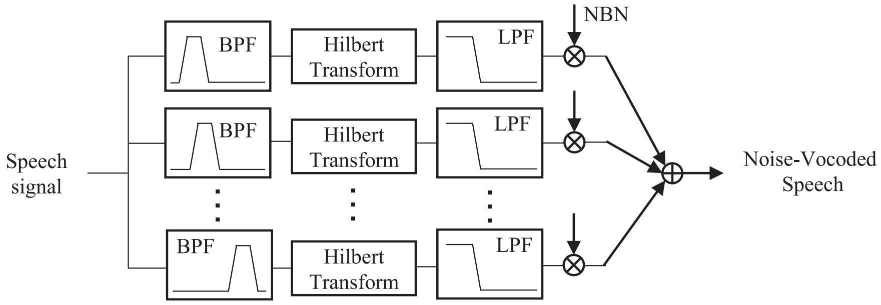


Fig. 1 Schematic diagram of noise-vocoder method used to generate stimuli (BPF: band-pass filter; LPF: low-pass filter; and NBN: narrow-band noise).

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.

Then, the temporal envelope 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). 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.

3. EXPERIMENT I: SPEAKER RECOGNITION

3.1. Speech Data

The speech data used in this study 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 recognition 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 recognition 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.

Table 1 Speaker pairs selected from ATR database and their average similarity index measured by Kitamura *et al.* [26]. 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

Kitamura *et al.* measured the perceptual similarity of 20 female and 20 male Japanese speakers in ATR speech database set C [26]. 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.* [26]. The five female and five male speaker pairs used in this study and their perceptual similarities are shown in Table 1. All 20 speakers are different, and the speakers of each pair have the same gender.

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 the hearing level of 12 dB in the frequency range from 125 to 8,000 Hz).

This experiment was carried out using the XAB method. One trial consisted of three different speech signals (X, A, and B). 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 1.

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 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 for all participants by using a head and torso simulator (B&K, type 4128) and sound level meter (B&K type 2231).

3.3. Results

Figure 2 shows the average value of speaker recognition 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 recognition with the original speech. The results for NVS stimuli showed that speaker recognition 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 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 experiment.

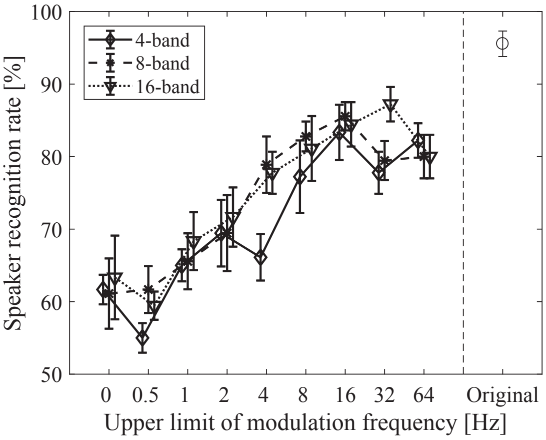


Fig. 2 Speaker recognition rates in all 27 NVS conditions and original speech condition. Error bars indicate ±1 standard error of mean.

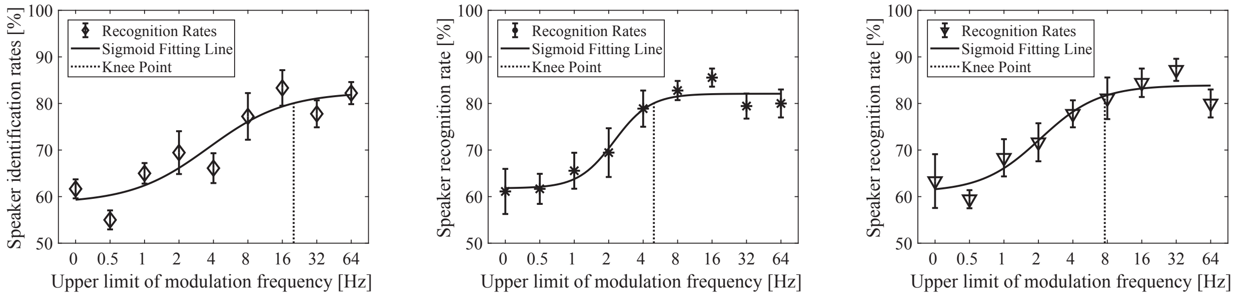
There was a significant main effect of the upper limit of modulation frequency ($F(8,64) = 23.86, p < 0.01$) but no significant main effect of the number of bands ($F(2,16) = 3.32$) or significant interaction between the two factors ($F(16,128) = 1.16$). These results showed that speaker recognition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

3.4. Discussion

This experiment was intended to clarify the role of temporal cues in speaker recognition. Specifically, the important modulation frequency band for speaker recognition 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:

$$y = \frac{a}{1 + e^{b(x-c)}} + d \qquad (2)$$

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 Fig. 3. The coefficients of determinations R^2 of the fitting results in 4, 8, and 16-band NVS were 0.86, 0.95, and 0.93.



(a) 4-band

(b) 8-band

(c) 16-band

Fig. 3 Speaker recognition rates in each condition of number of channels and their sigmoid fitting lines.

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 recognition when the spectral resolution is limited further.

Note that the speaker recognition rates of 4-band NVS were lower at some upper limits of modulation frequency. However, the number of channels did not affect the performance of speaker recognition significantly. These results were different from those of previous studies [17,18] 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 [17] and 160 or 400 Hz [18]. The modulation frequency band between about 50 and 500 Hz is related to the periodicity information about F0 [27], 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 [27]. Such modulation frequency band includes the information about 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 recognition 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.

4. EXPERIMENT II: VOCAL-EMOTION RECOGNITION

4.1. Speech Data

The emotional speech data used in this study were selected from the Fujitsu Japanese Emotional Speech Database [16]. 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.

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 8,000 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 the same as in experiment I.

4.3. Results

Figure 4 shows the results of the vocal-emotion recognition experiment. First, the recognition rates of the original emotional speech were 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

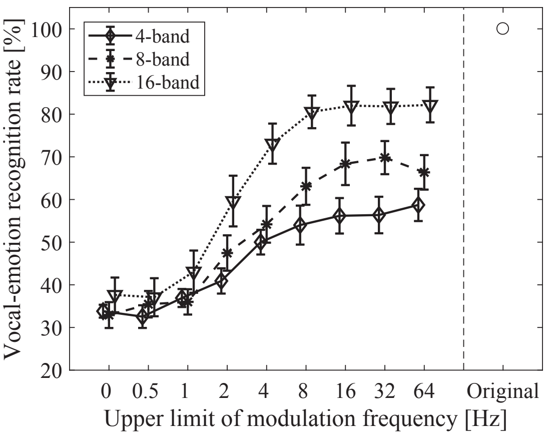
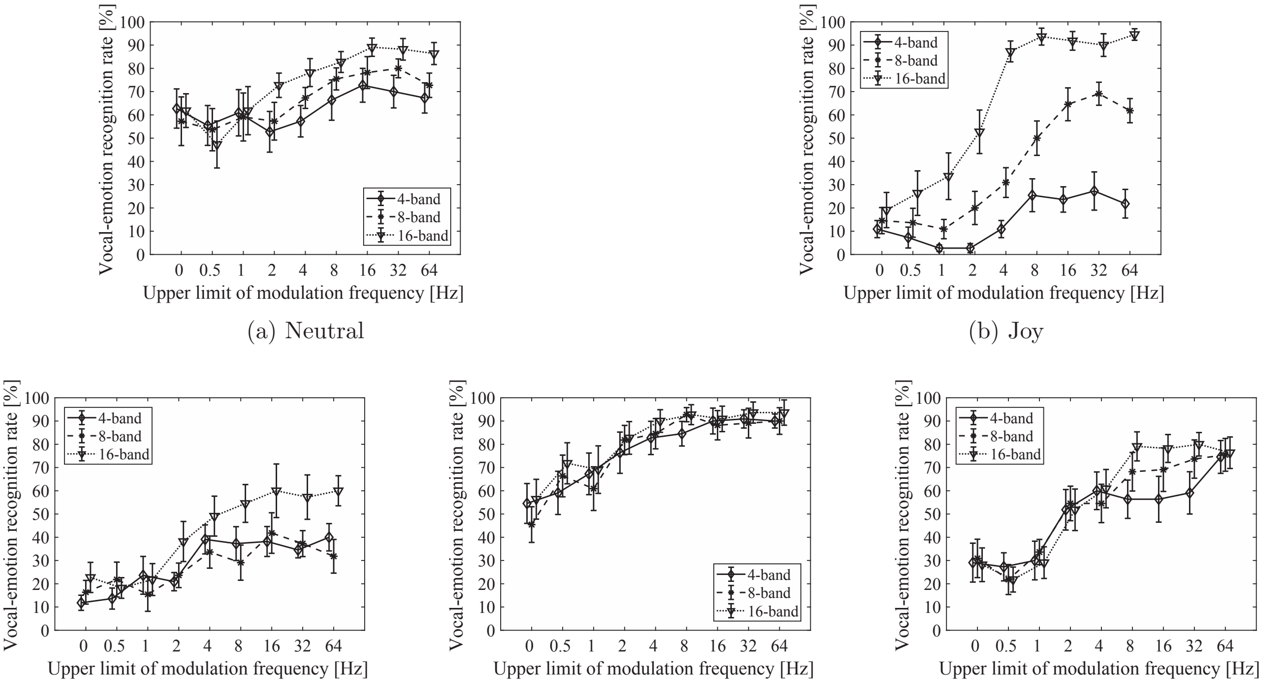


Fig. 4 Vocal-emotion recognition rates in all 27 NVS conditions and original speech condition. Error bars indicate ±1 standard error of mean.

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. Moreover, the 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 ($F(32,320) = 2.02, p < 0.01$) were both significant. There was also a significant interaction between the number of channels, the upper limit of modulation frequency and emotion ($F(64,640) = 2.59, p < 0.01$). Therefore, the effects of the number of channels, the upper limit of modulation frequency and their interaction on vocal emotion recognition were different with different emotions. The results for different emotions need to be analyzed separately.

Figure 5 shows the vocal-emotion recognition rates of different emotions. The results for different emotions are obviously different. Following the significant interactions, the analysis of simple main effect showed that there was a significant simple main effects of the upper limit of modulation for all emotions: neutral ($F(8,400) = 5.38, p < 0.01$), joy ($F(8,400) = 24.32, p < 0.01$), cold anger ($F(8,400) = 8.15, p < 0.01$), sadness ($F(8.400) = 11.29, p < 0.01$), hot anger ($F(8,400) = 21.87, p < 0.01$). However, the simple main effects of the number of channels were significant for only neutral ($F(2,100) = 6.93, p < 0.01$), joy ($F(2,100) = 132.99, p < 0.01$), and cold anger ($F(2,100) = 13.68, p < 0.01$). The simple main effect of the number of channels was not significant for sadness ($F(2,100) = 1.62, p = 0.20$) and hot anger ($F(2,100) = 2.40, p = 0.10$). The simple interactions between the number of channels and the upper limit of modulation frequency were significant for neutral ($F(2,100) = 2.02, p < 0.05$), joy ($F(2,100) = 11.49, p < 0.01$), cold anger ($F(2,100) = 2.20, p < 0.01$) and hot



(a) Neutral

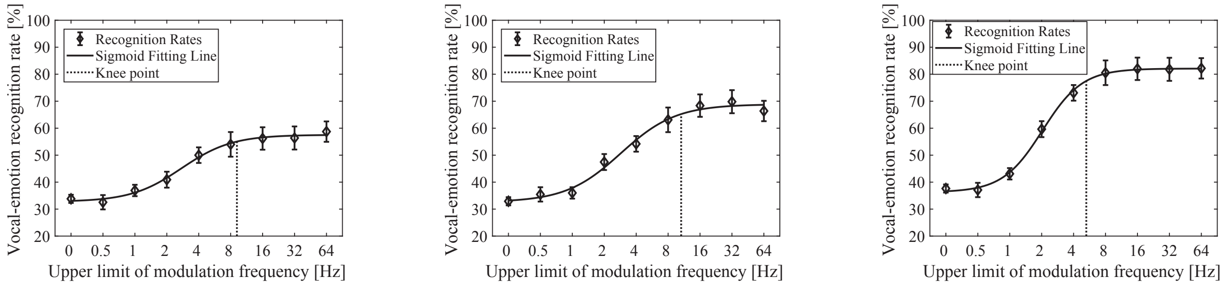
(b) Joy

(c) Cold Anger

(d) Sadness

(e) Hot Anger

Fig. 5 Vocal-emotion recognition rates of different emotions. Error bars indicate ±1 standard error of mean.



(a) 4-band (b) 8-band (c) 16-band

Fig. 6 Vocal-emotion recognition rates for each number of channels and their sigmoid fitting lines.

anger ($F(2, 100) = 2.26, p < 0.01$), but not significant for sadness ($F(2, 100) = 0.64, p = 0.86$).

4.4. Discussion

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 6 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 speaker recognition, 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. 5(a)), 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. 5(b)), 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. 5(c)), analyses of simple main effects also showed that both spectral 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. 5(d)) and hot anger (Fig. 5(e)), 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.

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) [4]. 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 [2]. They found that increasing the upper limit from 4 to 8 Hz improved the correct response rate more than increasing the upper limit from 8 to 16 Hz. In our previous study, we investigated the effect of controlling the upper limit of modulation frequency on the recognition of words and speakers while using a fixed number of channels [23]. 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. This result is consistent with Arai and Greenberg's previous study about the temporal properties of speech [28]. 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 recognition 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. 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.

In the future, 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 [29]. It is also suggested that the MSFs can potentially predict the vocal-emotion recognition by CI simulation [30]. The relationship between MSFs and the response of humans should be investigated further.

6. SUMMARY

This paper investigated the role of temporal cues in the perception of speaker individuality and vocal emotions. Speaker and vocal-emotion recognition experiments were carried out using noise-vocoded speech (NVS) as stimuli. The temporal resolution of NVS was controlled by varying the upper limits of the modulation frequency (0, 0.5, 1, 2, 4, 8, 16, 32, and 64 Hz). In addition, the role of temporal cues in the different spectral resolution conditions was also investigated by varying the number of channels (4, 8, and 16). The results demonstrated temporal resolution contributes to the recognition of speakers and vocal emotions. However, the speaker recognition 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.

It is confirmed that the temporal envelope of speech contributes to the perception of not only linguistic information but also speaker individuality and vocal emotion. The important modulation frequency bands for the perception of nonlinguistic information were suggested to be higher than that for linguistic information. The temporal modulation cue is suggested to play an important role in the auditory system to extract various information from speech.

ACKNOWLEDGMENTS

This work was supported by a Grant in Aid for Scientific Research (A) (No. 25240026), Innovative Areas (No. 16H01669) from MEXT, Japan, and the Mitsubishi Research Foundation. This work was also supported by JSPS KAKENHI Grant Number JP. 17J08312.

REFERENCES

[1] R. V. Shannon, F. G. Zeng, V. Kamath, J. Wygonski and M. Ekelid, "Speech recognition with primarily temporal cues," *Science*, **270**, 303–304 (1995).

[2] 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," *Acoust. Sci. & Tech.*, **34**, 263–270 (2013).

[3] P. C. Loizou, M. Dorman and Z. Tu, "On the number of channels needed to understand speech," *J. Acoust. Soc. Am.*, **106**, 2097–2103 (1999).

[4] L. Xu and B. E. Pfingst, "Spectral and temporal cues for speech recognition: Implications for auditory prostheses," *Hear. Res.*, **242**, 132–140 (2008).

[5] T. Dau, D. Puschel and A. Kohlrausch, "A quantitative model of the "effective" signal processing in the auditory system. I. Model structure," *J. Acoust. Soc. Am.*, **99**, 3615–3622 (1996).

[6] T. Dau, D. Puschel and A. Kohlrausch, "A quantitative model of the "effective" signal processing in the auditory system. II. Simulations and measurements," *J. Acoust. Soc. Am.*, **99**, 3623–3631 (1996).

[7] P. Loizou, "Mimicking the human ear," *IEEE Signal Process. Mag.*, **15**, 101–130 (1998).

[8] R. Drullman, J. M. Festen and R. Plomp, "Effect of temporal envelope smearing on speech reception," *J. Acoust. Soc. Am.*, **95**, 1053–1064 (1994).

[9] R. Drullman, J. M. Festen and R. Plomp, "Effect of reducing slow temporal modulations on speech reception," *J. Acoust. Soc. Am.*, **95**, 2570–2680 (1994).

[10] L. Xu, C. S. Thompson and B. E. Pfingst, "Relative contributions of spectral and temporal cues for phoneme

recognition," *J. Acoust. Soc. Am.*, **117**, 3255–3267 (2005).

[11] R. E. Remez, J. M. Fellowes and P. E. Rubin, "Talker recognition based on phonetic information," *J. Exp. Psychol. - Hum. Percept. Perform.*, **23**, 651–666 (1997).

[12] T. Kitamura and M. Akagi, "Speaker individualities in speech spectral envelopes," *J. Acoust. Soc. Jpn.* (*E*), **16**, 283–289 (1995).

[13] T. Kitamura, K. Honda and H. Takemoto, "Individual variation of the hypopharyngeal cavities and its acoustic effects," *Acoust. Sci. & Tech.*, **26**, 16–26 (2005).

[14] K. R. Scherer, "Vocal communication of emotion: A review of research paradigms," *Speech Commun.*, **40**, 227–256 (2003).

[15] R. Banse and K. R. Scherer, "Acoustic profiles in vocal emotion expression," *J. Pers. Soc. Psychol.*, **70**, 614–636 (1996).

[16] C. F. Huang and M. Akagi, "A three-layered model for expressive speech perception," *Speech Commun.*, **50**, 810–828 (2008).

[17] M. Vongphoe and F. G. Zeng, "Speaker recognition with temporal cues in acoustic and electric hearing," *J. Acoust. Soc. Am.*, **118**, 1055–1061 (2005).

[18] J. Gonzalez and J. Oliver, "Gender and speaker recognition as a function of the number of channels in spectrally reduced speech," *J. Acoust. Soc. Am.*, **118**, 461–470 (2005).

[19] 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," *Speech Commun.*, **322**, 151–162 (2015).

[20] X. Luo, Q. J. Fu and J. J. Galvin III, "Vocal emotion recognition by normal-hearing listeners and cochlear implant users," *Trends In Amplif.*, **11**, 301–315 (2007).

[21] V. Krull, X. Luo and K. Kirk, "Talker–recognition training using simulations of binaurally combined electric and acoustic hearing: Generalization to speech and emotion recognition," *J. Acoust. Soc. Am.*, **131**, 3069–3078 (2012).

[22] T. Vongpaisal, S. E. Trehub, E. G. Schellenberg, P. Lieshout and B. C. Papsin, "Children with cochlear implants recognize their mother's voice," *Ear Hear.*, **31**, 555–566 (2010).

[23] Z. Zhu, Y. Nishino, R. Miyauchi and M. Unoki, "Study on linguistic information and speaker individuality contained in temporal envelope of speech," *Acoust. Sci. & Tech.,* **37**, 258–261 (2016).

[24] Intl. Telecom. Union, "Objective measurement of active speech level," *ITU-T*, **P.56**, Switzerland (1993).

[25] B. C. J. Moore, *An Introduction to the Psychology of Hearing*, 6th Ed. (Elsevier, London, 2013), pp. 74–80.

[26] T. Kitamura, T. Nakama, H. Ohmura and H. Kawamoto, "Measurement of perceptual speaker similarity for sentence speech in ATR speech database," *J. Acoust. Soc. Jpn.* (*J*), **71**, 516–525 (2015) (in Japanese).

[27] S. Rosen, "Temporal information in speech: Acoustic, auditory and linguistic aspects," *Philos. Trans. S. Soc. Lond. Ser. B*, **336**, 367–373 (1992).

[28] T. Arai and S. Greenberg, "The temporal properties of spoken Japanese are similar to those of English," *Proc. EUROSPEECH 1997*, pp. 1011–1014 (1997).

[29] S. Wu, T. H. Falk and W. Chan, "Automatic speech emotion recognition using modulation spectral features," *Speech Commun.*, **53**, 768–785 (2011).

[30] Z. Zhu, R. Miyauchi, Y. Araki and M. Unoki, "Modulation spectral features for predicting vocal emotion recognition by simulated cochlear implants," *Proc. Interspeech 2016*, pp. 262–266 (2016).