©2018 The Acoustical Society of Japan

PAPER

Contribution of modulation spectral features on the perception of vocal-emotion using 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 15 February 2018, Accepted for publication 29 June 2018*)

Abstract: Previous studies on noise-vocoded speech showed that the temporal modulation cues provided by the temporal envelope play an important role in the perception of vocal emotion. However, the exact role that the temporal envelope and its modulation components play in the perceptual processing of vocal emotion is still unknown. To clarify the exact features that the temporal envelope contributes to the perception of vocal emotion, a method based on the mechanism of modulation frequency analysis in the auditory system is necessary. In this study, auditory-based modulation spectral features were used to account for the perceptual data collected from vocal-emotion recognition experiments using noise-vocoded speech. An auditory-based modulation filterbank was used to calculate the modulation spectrogram of noise-vocoded speech stimuli, and ten types of modulation spectral features were then extracted from the modulation spectrograms. The results showed that there were high similarities between modulation spectral features and the perceptual data of vocal-emotion recognition experiments. It was shown that the modulation spectral features are useful for accounting for the perceptual processing of vocal emotion with noise-vocoded speech.

Keywords: Temporal modulation cue, Modulation spectral feature, Vocal emotion, Noise-vocoded speech, Speech perception

PACS number: 43.71.Bp [doi:10.1250/ast.39.379]

1. INTRODUCTION

Speech waves are highly complex signals that transmit both linguistic information and various nonlinguistic information, such as vocal emotion. The human auditory system can ingeniously decode the emotional information included in speech signals to perceive the emotional state of speakers. Emotional expression in speech plays an important role in our daily lives; however, the perceptual processing of vocal emotion is still not fully clarified at present.

Previous studies related to the perception of vocal emotion focused on the acoustic features and sound patterns of speech signals. Banse and Scherer presented speech stimuli that contained 14 different emotions to listeners with normal hearing and asked them to label the emotion of each stimulus [1]. At the same time, they also extracted 29 different acoustic features (fundamental frequency (F0), intensity, speaking rate, duration, time-averaged spectrum, etc.) for each emotional speech stimulus. An emotion classification model was constructed using multiple regression analysis which analyzed the contributions of each acoustic feature. The results of discriminant analysis on the basis of this model showed that the confusion patterns were close to those of human responses. Huang and Akagi proposed a three-layered model with semantic primitives as a middle layer between vocal emotion and acoustic features [2]. In the previous studies, only the acoustic features based on the source-filter model (F0 and spectral envelope) and speech waveforms (intensity and duration) were investigated, regardless of what kinds of model were used.

In a study on vocal emotion perception by listeners with cochlear implants and its simulations, it was shown that such typical acoustic features have difficulty to account for the human response from cochlear-implant listeners [3].

[*] 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

Chatterjee *et al.* carried out vocal-emotion recognition experiments for cochlear implant listeners and normal hearing listeners using noise-vocoded speech as a cochlear implant simulation. They then analyzed the F0, intensity, and duration of the stimuli to clarify how cochlear implant listeners process the vocal emotion information included in speech. As cochlear implants only present a poor spectral resolution, the acoustic features related to the spectral envelope (formants, etc.) were not used. The results showed that the acoustic analyses could not account for all of the perceptual data from the vocal-emotion recognition experiments. An probable reason is that, for cochlear implant 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 envelope of sound signals has been proven to be important in auditory system. The signal processes in the peripheral auditory system can be computationally modeled as a bandpass filterbank, envelope extraction and amplitude compression [4,5]. Furthermore, Dau *et al.* proposed a computational model of human auditory signal processing and perception using a modulation filterbank after the process of temporal envelope extraction [6,7]. There are both physiological [8] and psychological [9] evidences that suggest the existence of a modulation filterbank in the auditory system. The auditory system has a modulation frequency analyzer which analyzes the modulation frequency components of the temporal envelope. On the other hand, Wu *et al.* proposed an automatic speech emotion recognition system using an auditory-based modulation analysis to extract the modulation spectral features of emotional speech [10]. The results showed that the modulation spectral features can be used to represent the features of temporal envelope related to vocal emotion better than the typical acoustic features.

In our previous study, we investigated the contribution of temporal modulation cues on the perception of non-linguistic information using noise-vocoded speech [11]. The results showed that the temporal modulation cues play an important role in the perception of vocal emotion. However, the role that temporal envelope is playing in the perceptual processing of vocal emotion is still unknown. As there is no harmonic structure, noise-vocoded speech does not contain the temporal fine structure of original speech, that is, the information related to F0. The intensity of noise-vocoded speech stimuli in the experiments was also normalized. Therefore, similar to the results in [3], the typical acoustic features cannot be used to account for the perceptual data collected from the experiments using noise-vocoded speech. An analysis based on the modulation frequency analysis mechanism of the auditory system is necessary. It has been shown that auditory-based modulation spectral features have the potential to account for the perceptual data of vocal-emotion recognition experiments [12]. However, the specific relationship between the modulation spectral features and the perception of vocal emotion is still unknown.

In this study, the relationship between the modulation spectral features and the perceptual data of vocal-emotion recognition experiments [11] was investigated to clarify the contribution of the modulation spectral features on the perception of vocal emotion. An auditory-based modulation filterbank was used to calculate the modulation spectrograms of the temporal envelope of noise-vocoded speech. Then, ten types of modulation spectral feature extracted from the modulation spectrograms were analyzed. Finally, the modulation spectral features and the perceptual data of vocal emotion recognition were compared to investigate the contribution of temporal modulation cues to the perception of vocal emotion with noise-vocoded speech. The originality of this study is that we considered the problem of vocal emotion perception from the viewpoint of auditory with the use of auditory-based modulation spectral features rather than from the viewpoint of speech production with the use of typical acoustic features.

This paper is organized as follows: Section 2 analyzes the perceptual data of vocal-emotion recognition experiments in [11]. Section 3 introduces the method for calculating the modulation spectral features from the noise-vocoded speech stimuli. Section 4 discusses the relationship between the modulation spectral features and the perceptual data. Section 5 summarizes the results and the discussion.

2. PERCEPTUAL DATA OF VOCAL-EMOTION RECOGNITION EXPERIMENTS

In our previous study, in order to study the contribution of temporal modulation cues on vocal-emotion recognition, we varied the spectral and temporal resolution of noise-vocoded speech stimuli presented to normal hearing listeners. The detailed method of signal processing to generate noise-vocoded speech can be found in [11]. The Fujitsu Japanese Emotional Speech Database was used. This database includes five emotions (*neutral, joy, cold anger, sadness, and hot anger*) expressed by a professional actress. The spectral resolution of the noise-vocoded speech stimuli was manipulated by varying the number of channels from 4 to 16. The temporal resolution was manipulated by varying the upper limits of the modulation frequency from 0 to 64 Hz. The results demonstrated that the vocal-emotion recognition rates significantly decreased as the upper limit of the modulation frequency decreased. Therefore, it was confirmed that the temporal modulation cues provided by the temporal envelope (in other words,

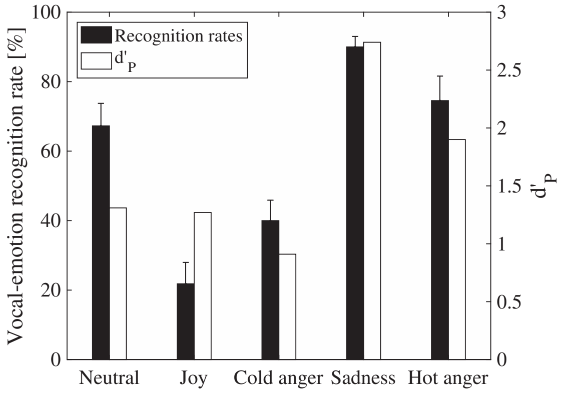


Fig. 1 The results of vocal-emotion recognition experiment in [11] on the condition that the upper limit of modulation frequency was 64 Hz and the number of channels was 4.

the information contained in the modulation frequency band below 64 Hz) contribute to the perception of vocal emotion.

To clarify the exact features that the temporal envelope contributes to the perception of vocal emotion, the current results regarding the condition of the 64-Hz upper limit of the modulation frequency and 4 channels noise-vocoded speech were used as the perceptual data of vocal-emotion recognition experiments. In this condition, the noise-vocoded speech stimuli contain all the information in the modulation frequency band below 64 Hz. Furthermore, the spectral cues were reduced mostly because we want to focus on the temporal modulation cues. Figure 1 shows the vocal-emotion recognition rates of the perceptual data that was used in this study. The results showed that joy was the most difficult to recognize and that the mean recognition rate was close to the chance level (20%). On the contrary, the recognition rates of sadness and hot anger were higher than that of the other emotions. The recognition rates of neutral emotion and cold anger were in the middle of the other three emotions, however, the recognition rate of cold anger was much lower than that of neutral.

To better understand the perceptual data, the discriminability index (d'_{P}) of each emotion was calculated from the mean confusion matrix of the perceptual data (Table 1). The d'_{P} values shown in Fig. 1 were based on the hit rates and false alarm rates derived from the confusion matrix, as follow:

$$d'_{\mathrm{P}} = \mathcal{Z}(H) - \mathcal{Z}(F) \tag{1}$$

where H and F are the hit rate and false alarm rate. $\mathcal{Z}(\cdot)$ is the inverse of the normal distribution function. Generally, high d'_{P} values are derived from high hit rates and low false alarm rates.

Because of the relatively higher hit rates and lower false alarm rates, the d'_{P} values of sadness and hot anger were much higher than those of the other emotions, as seen in the results for the recognition rates. The d'_{P} value of cold anger was lowest, due to the low hit rate and high false alarm rate. The hit rate of joy was the lowest, however, as it had a low false alarm rate, the d'_{P} value was higher than that of cold anger. For neutral emotion, the high hit rate and false alarm rate led to a low d'_{P} value.

Table 1 Mean confusion matrix of the perceptual data (in percent). Confusion matrix is presented as percentage with the stimuli organized vertically and the response categories organized horizontally.

	Neutral	Joy	Cold anger	Sadness	Hot anger
Neutral	67.27	3.646	20.00	6.364	2.727
Joy	22.73	21.82	18.18	1.818	35.45
Cold anger	33.64	1.818	40.00	20.00	4.546
Sadness	4.546	0	5.455	90.00	0
Hot anger	16.36	2.727	5.455	0.9091	74.55

For the perception of vocal emotion with noise-vocoded speech, the temporal modulation cues provided by the temporal envelope were used as primary cues. Therefore, in the next section, the modulation spectral features extracted directly from the modulation spectrograms of the temporal envelope were used to account for the perceptual data.

3. ANALYSIS OF THE MODULATION SPECTRAL FEATURES

3.1. Modulation Spectrogram

Figure 2 shows the auditory-based process used in this study to calculate the modulation spectrograms. Emotional speech signal s was first band-pass filtered using an auditory-based band-pass filterbank as follows:

$$s_k(n) = s(n) * h_k(n) \tag{2}$$

where $*$ denotes the convolution operation, $h_k(n)$ is the impulse response of the kth channel and n is the sample number in the time domain. The bandwidth and boundary frequencies of the band-pass filters (6th-order Butterworth infinite impulse response (IIR) filters) were defined using $\mathrm{ERB_N}$ (Equivalent Rectangular Bandwidth) and $\mathrm{ERB_N}$-number scales [13]. The boundary frequencies of the band-pass filters were defined as 3 to 35 $\mathrm{ERB_N}$-number with an 8 $\mathrm{ERB_N}$ bandwidth, and the number of channels was 4.

The temporal envelope of the output signal from each band-pass filter $s_k(n)$ was extracted using the Hilbert transformation, and a low-pass filter (2nd-order Butterworth IIR filter, cut-off frequency: 64 Hz) was performed as follows:

$$e_k(n) = |s_k(n) + j\mathcal{H}[s_k(n)]| * g(n), \tag{3}$$

where $\mathcal{H}$ denotes the Hilbert transform and, $g(n)$ denotes the impulse response of the low-pass filter. The signal

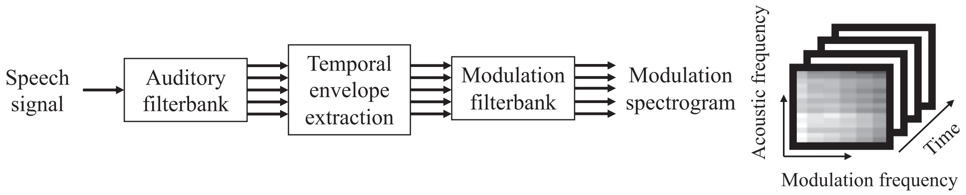


Fig. 2 Schematic diagram of the auditory-based process to calculate the modulation spectrogram.

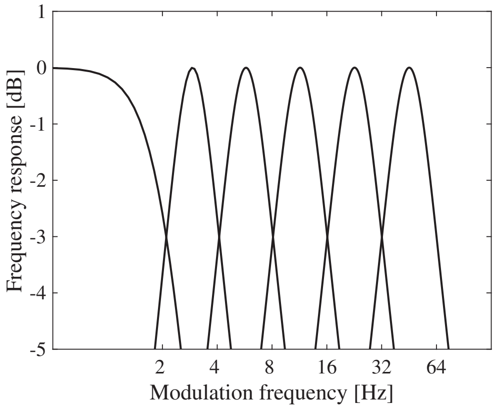


Fig. 3 The frequency response of the modulation filterbank.

processing methods of bandpass filterbank and temporal envelope extraction was as same as the methods used in [11].

The next step involved decomposing $e_k(n)$ into several modulation frequency bands by using a modulation filterbank:

$$E_{k,m}(n) = f_m(n) * (e_k(n) - \overline{e_k(n)}), \tag{4}$$

where m is the channel number of the modulation filter, $f_m(n)$ is the impulse response of the modulation filterbank, and $\overline{e_k(n)}$ is the time-averaged amplitude of $e_k(n)$. The 0 Hz component was removed because we only focused on the dynamic components of the temporal envelope. The modulation filterbank consisted of six filters (one low-pass filter and five band-pass filters). The boundary frequencies of the filters were spaced on an octave frequency band from 2 to 64 Hz. Figure 3 shows the frequency responses of the modulation filterbank. Finally, the root mean square of $E_{k,m}(n)$ calculated as the modulation spectrogram,

$$\overline{E}_{k,m} = \sqrt{\frac{1}{N}\sum_{n=0}^{N} E_{k,m}^2(n)}, \tag{5}$$

where N is the length of the speech signal $s(n)$. $\overline{E}_{k,m}$ was then used to calculate modulation spectral features.

Figure 4 shows examples of the modulation spectrograms of the speech with five different emotions from the Fujitsu database. The results show that each different emotion had different characteristics in the modulation spectrograms. The modulation spectrogram for sadness speech had significantly more low acoustic and modulation frequency energy. Contrarily, the modulation spectrogram of hot anger speech had more high acoustic and modulation frequency energy. These results should be related to the facts showed in [14,15] that sadness speech has lower high frequency energy and speech rate and anger speech has higher high frequency energy and speech rate. These results should be consistent with the perceptual data showing that sadness and hot anger had relatively higher d'_{P} values. However, it is difficult to directly connect the results of the modulation spectrograms and the perceptual data. Therefore, to quantitatively investigate the contributions of the modulation spectrogram to the perception of vocal emotion, the modulation spectral features extracted from the modulation spectrograms were then analyzed.

3.2. Modulation Spectral Features

Two kinds of modulation spectral feature were calculated by analyzing the modulation spectrograms in the acoustic frequency and modulation frequency domains. In the acoustic frequency domain, the first feature was the modulation spectral centroid (MSCR_m), which is defined as follows:

$$\mathrm{MSCR}_m = \frac{\Sigma_{k=1}^{K} k\overline{E}_{k,m}}{\Sigma_{k=1}^{K} \overline{E}_{k,m}}, \tag{6}$$

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

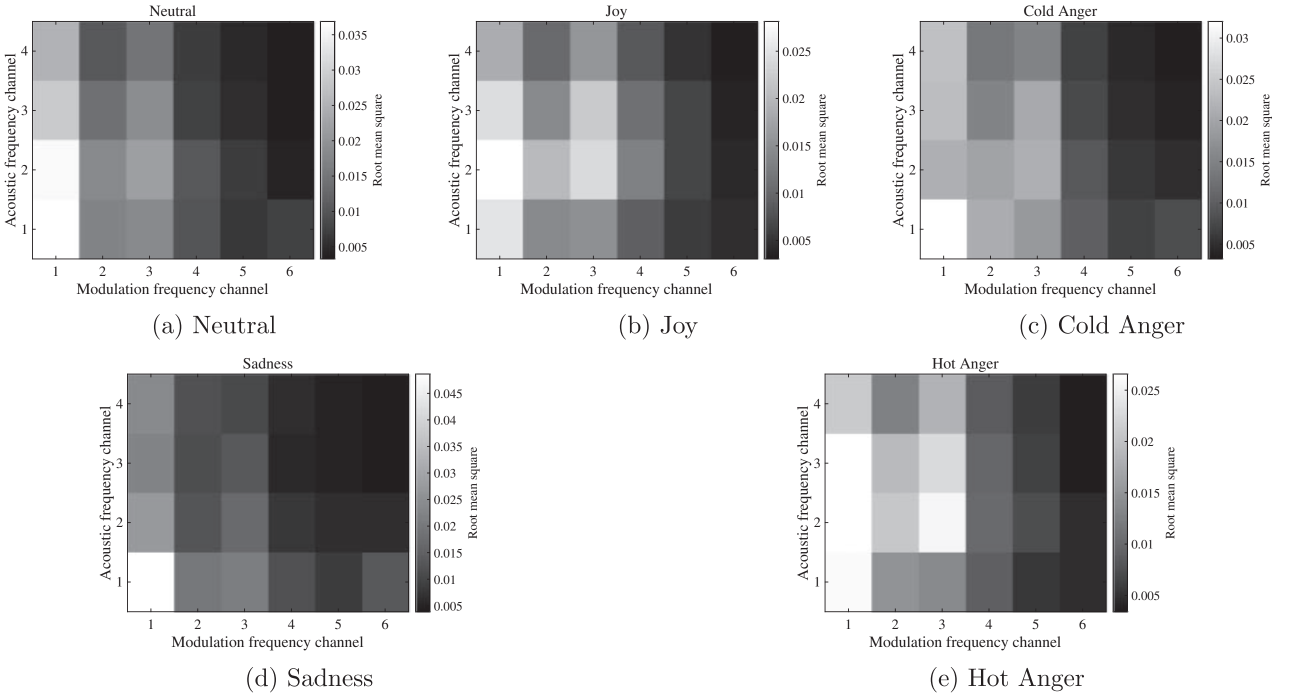
The modulation spectral spread (MSSP_m) was then calculated by:

$$\mathrm{MSSP}_m = \frac{\Sigma_{k=1}^{K} [k - \mathrm{MSCR}_m]^2 \overline{E}_{k,m}}{\Sigma_{k=1}^{K} \overline{E}_{k,m}}. \tag{7}$$

The MSSP_m represents the spread of the spectrum around its MSCR_m as the 2nd order 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 modulation spectrogram, which was calculated from the 3rd order moment:

$$\mathrm{MSSK}_m = \frac{\Sigma_{k=1}^{K} [k - \mathrm{MSCR}_m]^3 \overline{E}_{k,m}}{\Sigma_{k=1}^{K} \overline{E}_{k,m}}. \tag{8}$$



(a) Neutral

(b) Joy

(c) Cold Anger

(d) Sadness

(e) Hot Anger

Fig. 4 Examples of the time-average modulation spectrogram of different emotional speech.

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

$$\text{MSKT}_m = \frac{\Sigma_{k=1}^{K}[k - \text{MSCR}_m]^4 \overline{E}_{k,m}}{\Sigma_{k=1}^{K}\overline{E}_{k,m}}. \tag{9}$$

In the modulation frequency domain, the first feature was the MSCR_k which was 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).

$$\text{MSCR}_k = \frac{\Sigma_{m=1}^{M} m\overline{E}_{k,m}}{\Sigma_{m=1}^{M}\overline{E}_{k,m}}. \tag{10}$$

Then, the other three higher-order features of the modulation spectrograms in the modulation frequency domain (MSSP_k, MSSK_k, and MSKT_k) were calculated as follows:

$$\text{MSSP}_k = \frac{\Sigma_{m=1}^{M}[m - \text{MSCR}_k]^2 \overline{E}_{k,m}}{\Sigma_{m=1}^{M}\overline{E}_{k,m}}, \tag{11}$$

$$\text{MSSK}_k = \frac{\Sigma_{m=1}^{M}[m - \text{MSCR}_k]^3 \overline{E}_{k,m}}{\Sigma_{m=1}^{M}\overline{E}_{k,m}}, \tag{12}$$

$$\text{MSKT}_k = \frac{\Sigma_{m=1}^{M}[m - \text{MSCR}_k]^4 \overline{E}_{k,m}}{\Sigma_{m=1}^{M}\overline{E}_{k,m}}, \tag{13}$$

where M is the number of channels in the modulation filterbank which is six. Figure 5 shows an example of calculating the modulation spectral centroid in the acoustic frequency domain (MSCR_m) and the modulation frequency domain (MSCR_k). For the modulation spectral features in the acoustic frequency domain (modulation spectral features with subscript m), the modulation frequency channel was fixed, and the features were calculated on the basis of the acoustic frequency axis. On the contrary, for the modulation spectral features in the modulation frequency domain (modulation spectral features with subscript k), the acoustic frequency channel was fixed, and the features were calculated based on the modulation frequency axis.

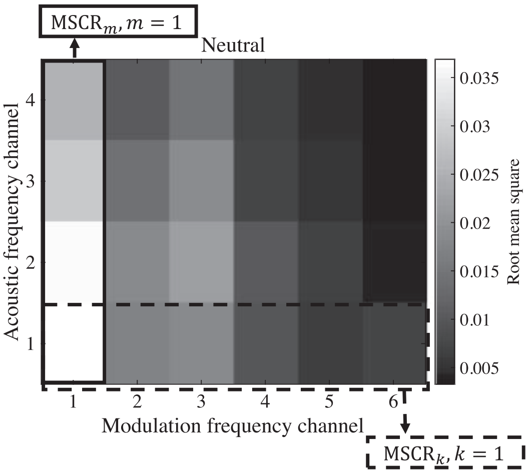


Fig. 5 An example of calculate the modulation spectral centroid of modulation spectrogram on the acoustic frequency domain (MSCR_m) and the modulation frequency domain (MSCR_k).

The last two modulation spectral features in the acoustic frequency and modulation frequency domains

Table 2 An example of the $\hat{d}'_{\mathrm{MSF}}$ value of each emotion for MSCR$_m$, $m = 1$ on acoustic frequency domains.

	Neutral	Joy	Cold anger	Sadness	Hot anger
Neutral	0	2.5793	1.0874	1.6855	6.0974
Joy	2.5793	0	0.2151	2.9507	4.1397
Cold anger	1.0874	0.2151	0	2.1366	2.1852
Sadness	1.6855	2.9507	2.1366	0	4.7163
Hot anger	6.0974	4.1397	2.1852	4.7163	0
$\hat{d}'_{\mathrm{MSF}}$	2.8624	2.4712	1.4061	2.8723	4.2846

Table 3 The similarities between modulation spectral features on the acoustic frequency domain and the perceptual data.

m	1	2	3	4	5	6
MSCR$_m$	0.5381	0.6287	0.7096	0.8854	0.7079	0.1662
MSSP$_m$	−0.1745	0.8341	0.8451	0.8805	0.8817	0.7908
MSSK$_m$	0.9874	0.9448	0.9601	0.9254	0.9454	0.6092
MSKT$_m$	0.8858	0.9104	0.9059	0.9006	0.9405	0.5531
MSTL$_m$	0.6619	0.8734	0.5452	0.8742	0.9050	0.5531

were modulation spectral tilt (MSTL$_m$ and MSTL$_k$), which are the linear regression coefficient obtained by fitting a first-degree polynomial to the modulation spectrograms.

Finally, to investigate the relationship between the modulation spectral features and the perceptual data, the discriminability index of the modulation spectral features (d'_{MSF}) were also calculated by the following equation:

$$d'_{\mathrm{MSF}} = \frac{|\mu_{emotion1} - \mu_{emotion2}|}{\sqrt{\frac{1}{2}(\sigma^2_{emotion1} + \sigma^2_{emotion2})}}, \tag{14}$$

where, μ and σ^2 are the mean value and variance of a modulation spectral feature (taken across the 10 utterances of each emotion). The mean value of all the d'_{MSF} values for each emotion was computed as an approximate measure of the net discriminability of the modulation spectral features (see Table 2). This $\hat{d}'_{\mathrm{MSF}}$ value represents the mean distance of a modulation spectral feature between different emotions.

3.3. Similarities between the Perceptual Data and Modulation Spectral Features

The centered cosine similarity between d'_{P} (Fig. 1) and $\hat{d}'_{\mathrm{MSF}}$ were calculated to investigate the relationship between modulation spectral features and the perception of vocal emotion with noise-vocoded speech. The similarity was defined as follow:

$$A(em) = d'_{\mathrm{P}}(em) - \frac{1}{5}\,\Sigma^{5}_{em=1} d'_{\mathrm{P}}(em), \tag{15}$$

Table 4 The similarities between modulation spectral features on the modulation frequency domain and the perceptual data.

k	1	2	3	4
MSCR$_k$	−0.3996	0.3628	0.7564	−0.5762
MSSP$_k$	−0.2670	0.8252	0.8632	0.8900
MSSK$_k$	−0.2696	0.7603	0.8068	−0.6825
MSKT$_k$	0.1213	0.8402	0.8405	0.9191
MSTL$_k$	0.9949	0.9557	0.9992	−0.3721

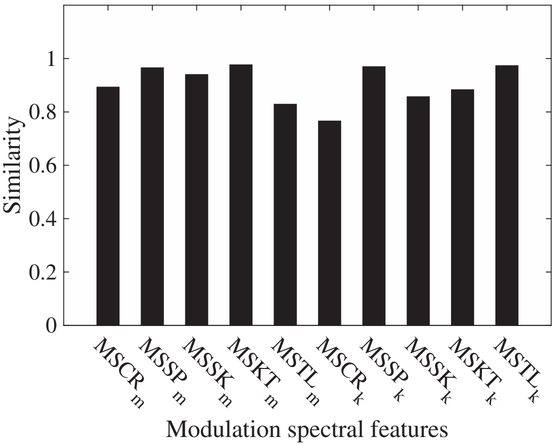


Fig. 6 The highest similarity of each modulation spectral feature (taken across all the acoustic or modulation frequency channels).

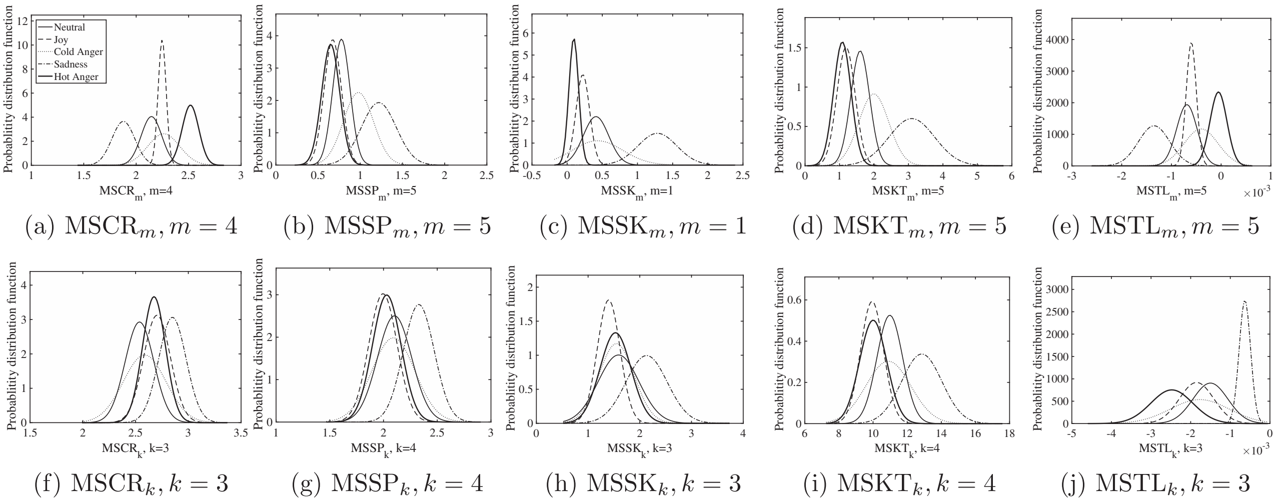
$$B(em) = \hat{d}'_{\mathrm{MSF}}(em) - \frac{1}{5}\,\Sigma^{5}_{em=1} \hat{d}'_{\mathrm{MSF}}(em), \tag{16}$$

$$Similarity = \frac{\Sigma^{5}_{em=1} A(em)B(em)}{\sqrt{\Sigma^{5}_{em=1} A(em)^2}\sqrt{\Sigma^{5}_{em=1} B(em)^2}}, \tag{17}$$

where, em is the emotion which could be *neutral, joy, cold anger, sadness, and hot anger*. Tables 3 and 4 show the results of the similarities of the modulation spectral features in the acoustic frequency and modulation frequency domains, respectively. Figure 6 shows the highest similarity of each modulation spectral feature (taken across all the acoustic frequency or modulation frequency channels). The results showed that there were high similarities between the modulation spectral features and the perceptual data. For some modulation spectral features, the similarities were close to 1. These results suggest that the modulation spectral features are useful in accounting for the perceptual data of vocal-emotion recognition experiment using noise-vocoded speech.

4. DISCUSSION

The d'_{P} values of the perceptual data obtained from vocal-emotion recognition experiments represent the psychological distance between the emotions for participants.



(a) $\mathrm{MSCR}_m, m = 4$ (b) $\mathrm{MSSP}_m, m = 5$ (c) $\mathrm{MSSK}_m, m = 1$ (d) $\mathrm{MSKT}_m, m = 5$ (e) $\mathrm{MSTL}_m, m = 5$

(f) $\mathrm{MSCR}_k, k = 3$ (g) $\mathrm{MSSP}_k, k = 4$ (h) $\mathrm{MSSK}_k, k = 3$ (i) $\mathrm{MSKT}_k, k = 4$ (j) $\mathrm{MSTL}_k, k = 3$

Fig. 7 Estimated probability distribution function of modulation spectral features for each emotion. Only the modulation spectral features with the highest similarity showed in Fig. 6 are demonstrated here.

The $\hat{d}'_{\mathrm{MSF}}$ values of the modulation spectral features represent the physical distance of the modulation spectral features between different emotions. The probability distribution functions (PDFs) of the modulation spectral features with the highest similarity showed in Fig. 6 were estimated to discuss the reason for the high similarity between the modulation spectral features and the perceptual data (Fig. 7).

Figure 7(a) shows that the MSCR$_m$ for hot anger speech was highest, and the MSCR$_m$ of sadness speech was lowest in the 4th modulation frequency channel. In addition, the distributions of the other emotions (neutral, joy, and cold anger) overlapped. Similar phenomenon also appeared in the distribution of MSTL$_m$. The reason for this is that hot anger speech had more high-acoustic frequency energy, and sadness speech had more low-acoustic frequency energy. The distributions of neutral, joy and cold anger speech on the acoustic frequency domain were similar. These results were consistent with the perceptual data that sadness and hot anger stimuli had higher d'_{P} values and that the d'_{P} values of other emotions were much lower.

On the contrary, for the other high-order features MSSP$_m$ (Fig. 7(b)), MSSK$_m$ (Fig. 7(c)), and MSKT$_m$ (Fig. 7(d)) in the acoustic frequency domain, the PDFs of hot anger speech were lowest and the PDFs of sadness speech were highest. The high-order features in the modulation frequency domain MSSP$_k$ (Fig. 7(g)) and MSKT$_k$ (Fig. 7(i)), also showed a similar trend. These results showed that the spread and peakedness of sadness speech in both the acoustic frequency and modulation frequency domains were higher than those of the other emotions. Moreover, the PDFs of joy and hot anger speech overlapped, which were consistent with the results of confusion matrix (Table 1) that nearly 35% of the joy stimuli were recognized as hot anger.

It was also shown that the similarities of the modulation spectral features in the acoustic frequency domain (Table 3) in the 4th and 5th modulation frequency channel (from 8 to 32 Hz) were much higher. Similar to the results in [11], high-modulation frequency band was shown to be more important to the perception of vocal emotion with noise-vocoded speech. The high-modulation frequency components are related to auditory roughness, which should affect the speech quality of noise-vocoded speech.

The high-modulation frequency components should also affect the modulation spectral features in the modulation frequency domain. Hot anger speech had much more high-modulation frequency components that resulted in higher MSCR$_k$. On the contrary, the MSCR$_k$ of sadness speech should be lower because sadness speech had much less high-modulation frequency components. The degree of asymmetry (MSSK$_k$) of sadness speech should be higher than that of hot anger speech as the modulation spectrogram for sadness speech was central in the low-modulation frequency band. The shapes of the modulation spectrograms of the other three emotions in the modulation frequency domain were similar.

To summarize the results: hot anger speech has more high acoustic frequency and modulation frequency components; sadness speech has less high acoustic frequency and modulation frequency components; regarding hot anger and sadness speech, the distributions of the modulation spectrogram for neutral, joy, and cold anger speech are similar. These physical characteristics are consistent with the perceptual data which showed that sadness and hot anger stimuli had higher $\hat{d}'_{\mathrm{MSF}}$ values and the $\hat{d}'_{\mathrm{MSF}}$ values of neutral, joy, and cold anger stimuli were much lower. Therefore, there were high similarities between the modulation spectral features and the perceptual data.

Modulation spectral features have been shown to be useful in accounting for the perception of vocal emotion. In this study, the modulation spectral features of time-averaged modulation spectrograms were analyzed. The modulation spectrograms were 4-dimensional data containing information on acoustic frequency, modulation frequency, amplitude, and time. It is necessary to analyze the details regarding modulation spectrograms in time domain. However, as the modulation spectrograms were 4-dimensional data, it would be difficult to extract the features related to nonlinguistic information from them. Deep learning may be a good solution for analyzing the modulation spectrogram in the time domain. The modulation spectral features should be derived from human vocal organs. It is also necessary to connect the auditory-based modulation spectral features to the mechanism of speech production to investigate the relationship between modulation spectral features and the perception of not only noise-vocoded speech but also normal speech.

5. SUMMARY

In this study, the relationship between the auditory-based modulation spectral features and perceptual data of vocal-emotion experiments using noise-vocoded speech was investigated to clarify the exact features that the temporal envelope contributes to the perception of vocal emotion. The discriminability indices (d'_{P} and $\hat{d}'_{\mathrm{MSF}}$) of each emotion were calculated from the modulation spectral features and the mean confusion matrix of the perceptual data. It was shown that for both the modulation spectral features and the perceptual data, the d'_{P} and $\hat{d}'_{\mathrm{MSF}}$ values of sadness and hot anger speech were higher than those of neutral, joy, and cold anger speech. These results led to high similarities between the modulation spectral features and the perceptual data. This suggests that the modulation spectral features play an important role in the perception of vocal emotion with noise-vocoded speech. The modulation spectral features have shown to be useful in accounting for the perceptual processing of the temporal modulation cues provided by the temporal envelope of speech.

ACKNOWLEDGMENTS

This work was supported by a Grant in Aid for Scientific Research (A) (No. 25240026), Innovative Areas (No. 16H01669) and (No. 18H05004) from MEXT, Japan, and the Mitsubishi Research Foundation. This work was also supported by JSPS KAKENHI Grant Number JP. 17J08312.

REFERENCES

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

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

[3] 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).

[4] 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).

[5] 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).

[6] T. Dau, B. Kollmeier and A. Kohlrausch, "Modeling auditory processing of amplitude modulation. I. Detection and masking with narrow-band carriers," *J. Acoust. Soc. Am.*, **102**, 2892–2905 (1997).

[7] T. Dau, B. Kollmeier and A. Kohlrausch, "Modeling auditory processing of amplitude modulation. II. Spectral and temporal integration," *J. Acoust. Soc. Am.*, **102**, 2906–2919 (1997).

[8] J. Xiang, D. Poeppel and J. Z. Simon, "Physiological evidence for auditory modulation filterbanks: Cortical responses to concurrent modulations," *J. Acoust. Soc. Am.*, **133**, 7–12 (2013).

[9] S. D. Ewert and T. Dau, "Characterizing frequency selectivity for envelope fluctuations," *J. Acoust. Soc. Am.*, **108**, 1181–1196 (2000).

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

[11] Z. Zhu, R. Miyauchi, Y. Araki and M. Unoki, "Contributions of temporal cue on the perception of speaker individuality and vocal emotion for noise-vocoded speech," *Acoust. Sci. & Tech.*, **39**, 234–242 (2018).

[12] Z. Zhu, R. Miyauchi, Y. Araki and M. Unoki, "Contributions of temporal cue on the perception of speaker individuality and vocal emotion for noise-vocoded speech," *INTERSPEECH2016*, pp. 262–266 (2016).

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

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

[15] T. Johnstone and K. R. Scherer, *Handbook of Emotion*, 2nd ed. (Guilford, New York, 2000), pp. 220–235.