INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Modulation spectral features for predicting vocal emotion recognition by simulated cochlear implants

Zhi Zhu[1], *Ryota Miyauchi*[1], *Yukiko Araki*[2], *Masashi Unoki*[1]

[1]School of Information Science, Japan Advanced Institute of Science and Technology, Japan
[2]School of Humanities, Kanazawa University, Japan

[1]{zhuzhi, ryota, unoki}@jaist.ac.jp, [2]yukikoa@staff.kanazawa-u.ac.jp

Abstract

It has been reported that vocal emotion recognition is challenging for cochlear implant (CI) listeners due to the limited spectral cues with CI devices. As the mechanism of CI, modulation information is provided as a primarily cue. Previous studies have revealed that the modulation components of speech are important for speech intelligibility. However, it is unclear whether modulation information can contribute to vocal emotion recognition. We investigated the relationship between human perception of vocal emotion and the modulation spectral features of emotional speech. For human perception, we carried out a vocal-emotion recognition experiment using noise-vocoder simulations with normal-hearing listeners to predict the response from CI listeners. For modulation spectral features, we used auditory-inspired processing (auditory filterbank, temporal envelope extraction, modulation filterbank) to obtain the modulation spectrogram of emotional speech signals. Ten types of modulation spectral feature were then extracted from the modulation spectrogram. As a result, modulation spectral centroid, modulation spectral kurtosis, and modulation spectral tilt exhibited similar trends with the results of human perception. This suggests that these modulation spectral features may be important cues for voice emotion recognition with noise-vocoded speech.

Index Terms: vocal emotion, cochlear implant, noise-vocoded speech, modulation spectral feature

1. Introduction

A cochlear implant (CI) is one of the most successful artificial organs. High levels of speech intelligibility can be achieved by CI listeners with clean speech. However, human voice includes not only linguistic information but also nonlinguistic information such as gender, age, and vocal emotion. Vocal emotion in particular plays an important role in human speech communication. Unfortunately, it has been found that CI listeners have great difficulty in perceiving vocal emotion due to the limited spectral cues provided by CI devices [1][2]. Research on speech perception by CI listeners has been conducted using acoustic simulations with normal-hearing (NH) listeners. Noise-vocoded speech, which is generated by replacing the temporal fine structure of speech with a noise carrier while the temporal amplitude envelope is preserved, is widely used in CI simulations [3]. It is related to the fact that CI devices provide the temporal envelope information as a primarily cue, and the temporal fine structure information is normally not effectively encoded [4].

Chatterjee *et al.* compared the performance of vocal emotion recognition by CI listeners and NH listeners with noise-vocoded speech as CI simulations [1]. The mean performance of CI listeners was similar to that of NH listeners with 8-channels noise-vocoded speech. The mean intensity, duration, and intensity range of stimuli were analyzed to clarify the acoustic features that contribute to the perception of vocal emotion. However, the results of acoustic analyses could not account for all the perceptual data of the experiments.

Luo *et al.* investigated the ability of NH and CI listeners to recognize vocal emotions [2]. The results showed that the performance of CI listeners was significantly lower than that of NH listeners. They also carried out vocal-emotion recognition experiments by using CI simulations with NH listeners. As a result, the emotion recognition rates significantly improved as the cut-off frequency of the modulation low-pass filter was increased from 50 to 500 Hz. Their results suggest that temporal envelope cues can contribute to vocal emotion recognition.

As the temporal envelope of speech is provided as a primarily cue in CI devices, the modulation spectral features extracted from the temporal envelope of speech should be considerable cues for vocal emotion recognition by CI listeners. The modulation spectrum of speech has been proved to be important for many research fields such as auditory physiology, psychoacoustics, speech perception, and signal analysis and synthesis [5]. Moreover, modulation spectral features have been successfully applied in automatic vocal emotion recognition systems [6][7]. Therefore, modulation spectral features can represent vocal emotion information. However, it is still unclear whether modulation spectral features can be used to account for vocal emotion recognition by humans.

For this study, we investigated the relationship between the modulation spectral features and human perception of noise-vocoded emotional speech. We first analyzed ten types of modulation spectral feature extracted from the modulation spectrogram of emotional speech. Then, for human perception, we carried out a vocal-emotion recognition experiment by using noise-vocoded speech with NH listeners to predict the response from CI listeners. Finally, we compared the modulation spectral features and perceptual data of the vocal-emotion recognition experiment to discuss which features can contribute to the perception of vocal emotion with noise-vocoded speech.

2. Modulation Spectral Features of Emotional Speech

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 [1]. Moreover, for

http://dx.doi.org/10.21437/Interspeech.2016-737

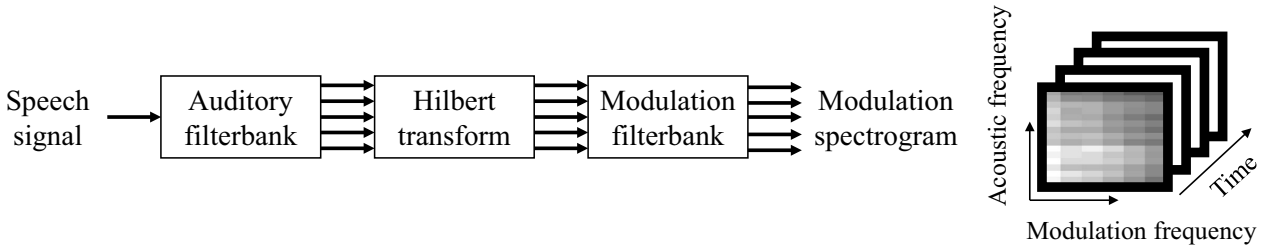

Figure 1: *Block diagram overview of the process used to calculate modulation spectrogram.*

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 [6]. 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 [8]. 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 1 shows the auditory-inspired process used in this study to calculate the modulation spectrogram. Emotional speech signals $s(n)$ were first band-pass filtered using an auditory-inspired band-pass filterbank as follows:

$$s_k(n) = h_k(n) * s(n), \tag{1}$$

where $h_k(n)$ is the impulse response of the kth channel and n is sample number in the time domain.

This auditory-inspired band-pass filterbank was constructed as an auditory filterbank by using 3rd-cascaded 2nd-order Butterworth infinite impulse response (IIR) filters. The bandwidth of each filter was designed using the equivalent rectangular bandwidth of auditory filter ($\mathrm{ERB_N}$) and all filters were placed according to the $\mathrm{ERB_N}$-number scale [9]. $\mathrm{ERB_N}$-number is defined by the following equation,

$$\mathrm{ERB_N} - \mathrm{number} = 21.4 \log_{10}\left(\frac{4.37f}{1000} + 1\right), \tag{2}$$

where f is acoustic frequency in Hz. This scale is comparable to that of the distance along the basilar membrane, so that the frequency resolution of the auditory system can be faithfully replicated by dividing frequency bands according to the $\mathrm{ERB_N}$-number. In this study, we carried out the signal process on the assumption of 32 auditory filters (from 3 to 35 $\mathrm{ERB_N}$-numbers). Since there are co-modulation characteristics of the amplitude envelopes in certain frequency bands, we considered the appropriate numbers of analysis bands from 32 $\mathrm{ERB_N}$-numbers. Thus, 16 bands (each 2 $\mathrm{ERB_N}$ selection) and 8 bands (each 4 $\mathrm{ERB_N}$ selection) were used as appropriated co-modulated bands.

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

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

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. Finally, the modulation spectrogram $E^2_{k,m}(n)$ was obtained by:

$$E^2_{k,m}(n) = |g_m(n) * e_k(n)|^2, \tag{4}$$

where m is the channel number of the modulation filter.

The modulation spectrogram can describe emotional speech on not only the acoustic frequency domain but also the modulation frequency domain. Modulation frequency can represent the fluctuation in the temporal envelope. For example, the hot anger speech has more high modulation frequency energy, since has faster fluctuation in the temporal envelope. On the contrary, sadness speech has slower fluctuation in the temporal envelope; thus it has lower high modulation frequency energy.

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:

$$\mathrm{MSCR}_m(n) = \frac{\Sigma_{k=1}^{K} k E^2_{k,m}(n)}{\Sigma_{k=1}^{K} E^2_{k,m}(n)}, \tag{5}$$

where K is the number of acoustic frequency bands (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:

$$\mathrm{MSSP}_m(n) = \frac{\Sigma_{k=1}^{K} [k - \mathrm{MSCR}_m(n)]^2 E^2_{k,m}(n)}{\Sigma_{k=1}^{K} E^2_{k,m}(n)}. \tag{6}$$

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:

$$\mathrm{MSSK}_m(n) = \frac{\Sigma_{k=1}^{K} [k - \mathrm{MSCR}_m(n)]^3 E^2_{k,m}(n)}{\Sigma_{k=1}^{K} E^2_{k,m}(n)}. \tag{7}$$

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

$$\mathrm{MSKT}_m(n) = \frac{\Sigma_{k=1}^{K} [k - \mathrm{MSCR}_m(n)]^4 E^2_{k,m}(n)}{\Sigma_{k=1}^{K} E^2_{k,m}(n)}. \tag{8}$$

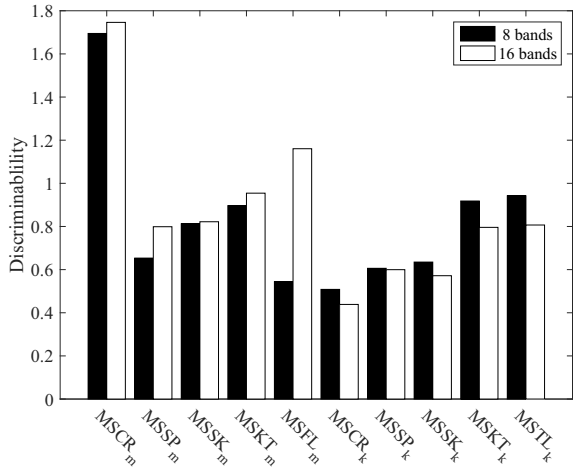

Figure 2: *The value of discriminability index for each modulation spectral feature.*

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:

$$\mathrm{MSFT}_m(n) = \frac{\sqrt[K]{\prod_{k=1}^{K} E_{k,m}^2(n)}}{\frac{1}{K}\Sigma_{k=1}^{K} E_{k,m}^2(n)}. \quad (9)$$

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

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. 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.

For this study, we analyzed only the time average values of all modulation spectral features. A discriminability index was then used to describe the separation of each modulation spectral feature between different emotion 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 emotions, divided by their average standard deviation. The average value of discriminability indices (taken across all the emotions and bands) was computed as a measure of the net discriminability provided by this feature.

Figure 2 shows the average discriminability indices of all ten types of modulation spectral feature under the conditions of 8 and 16 bands. As a result, the MSCR_m carried greater weight of the discriminability index than the other features. This might suggest that the MSCR_m of each modulation frequency band is an important cue for vocal emotion recognition. The average discriminability indices of modulation spectral features on the acoustic frequency domain decreased when the number of bands decreased from 16 to 8. On the contrary, for the modulation spectral features on the modulation frequency domain, the average discriminability indices increased. The average discriminability index of the MSFT_m was mostly affected by the number of bands. In addition, in the group of features based on the modulation frequency bands, the MSTL_k was highest.

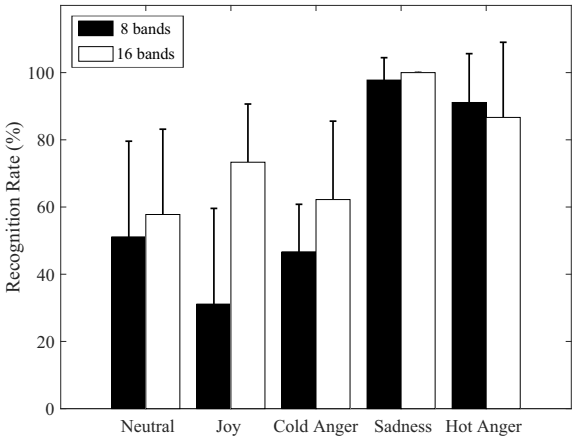

Figure 3: *Vocal emotion recognition rates of noise-vocoded emotional speech.*

3. Vocal-Emotion Recognition Experiment

3.1. Signal Generation

All emotional speech signals discussed in Section 2 were processed using a noise-vocoder method to generate the CI-simulation stimuli. Speech signals were first divided into 8 and 16 acoustic frequency bands using the same band-pass filterbank described in Section 2.1. Then, the temporal envelope was extracted using Hilbert transformation and a low-pass filter (2nd-order Butterworth IIR filter). The cut-off frequency of the low-pass filter was 64 Hz. The temporal envelope of each band was used to amplitude modulate the noise limited in the same band. Finally, all amplitude-modulated band-limited noises were summed to generate the noise-vocoded speech stimuli.

3.2. Procedure

Nine native Japanese speakers (5 males and 4 females) participated in this experiment. All participants had normal hearing (hearing levels of the participants were below 12 dB in the frequency range from 125 to 8000 Hz).

The noise-vocoded speech 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 levels of background noise was lower than 25.8 dB. The sound pressure levels of the output from headphone were 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).

Before the experiment, we confirmed that all participants could correctly recognize the emotion of the original speech. All noise-vocoded speech 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.

3.3. Results

Figure 3 shows the mean value and standard deviation of the results of vocal emotion recognition rates. The average recognition rate decreased when the number of bands decreased from 16 to 8, and the results of joy were mostly effected by the num-

Table 1: Modulation spectral features which have larger average discriminability index than 0.9 of 8 bands noise-vocoded speech and their correlation coefficients with the calculated discriminability index of the results of experiment.

	Neutral	Joy	Cold anger	Sadness	Hot anger	Correlation
MSCR_m	2.514	2.256	2.013	3.122	3.897	0.74
MSKT_k	1.849	1.769	1.572	2.958	1.797	0.86
MSTL_k	1.972	1.878	1.660	2.971	1.787	0.81

Table 2: Modulation spectral features which have larger average discriminability index than 0.9 of 16 bands noise-vocoded speech and their correlation coefficients with the calculated discriminability index of the results of experiment.

	Neutral	Joy	Cold anger	Sadness	Hot anger	Correlation
MSCR_m	3.511	3.161	2.648	4.340	5.298	0.45
MSKT_m	1.686	1.971	1.675	2.809	2.314	0.95
MSFT_m	1.993	3.096	2.175	3.368	2.008	0.78

ber of bands. In addition, the average recognition rates of sadness and hot anger were higher than the other three emotions under both conditions of 8 and 16 bands.

We conducted a repeated measures analysis of variance on the results with the number of bands and emotion as the factors. There were significant main effects of the number of bands ($F(1,8) = 11.4, p < 0.05$) and emotion ($F(4,32) = 16.2, p < 0.05$). There was also a significant interaction between the two factors ($F(4,32) = 4.34, p < 0.05$). Analyses of simple main effects of the number of bands showed that there was only a significant difference in the results of joy between 8 and 16 bands ($F(1,40) = 24.0, p < 0.05$). Moreover, there were significant simple effects of emotion under the both conditions of 8 bands ($F(4,64) = 17.2, p < 0.05$) and 16 bands ($F(4,64) = 6.13, p < 0.05$). The results suggest that the perception of vocal emotion with noise-vocoded speech significantly differs depending on the emotion.

4. Discussion

We calculated the modulation spectral features 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 modulation spectral features that had a larger average discriminability index than 0.9 were selected and their correlation coefficients with the calculated discriminability indices from the results of the vocal-emotion recognition experiment were computed. Tables 1 and 2 list the results under conditions of 8 and 16 bands. For the MSCR_m, sadness and hot anger exhibited a higher average discriminability index than the other three emotions. Thus, the distributions of the MSCR_m of sadness and hot anger were different to those of the other three emotions. This is consistent with the experimental results showing that the recognition rates of sadness and hot anger were higher than those of the other three emotions. However, the correlation coefficient of the MSCR_m was lower than the other features, even though it had the highest average discriminability index. Under the conditions of 8 and 16 bands, the MSKT_k, MSTL_k, MSKT_m, and MSFT_m highly correlated with the experimental results. These results suggest that it is not sufficient to analysis the value of discriminability of modulation spectral features to account for the results of human perception. The trend of the modulation spectral features for different emotions is also important.

The results of this paper suggest the potential of modulation spectral features for vocal emotion analysis. In the future, the effect of modifying modulation spectral features on vocal emotion recognition will 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 emotion with noise-vocoded speech may not depend on just one single feature, the interaction of modulation spectral features should also be further discussed.

5. Summary

We investigated the relationship between the modulation spectral features and human perception of noise-vocoded emotional speech. Ten types of modulation spectral feature of emotional speech were analyzed. Then, a vocal-emotion recognition experiment was carried out using noise-vocoder simulations with NH listeners to predict the response from CI listeners. As a result, the average recognition rate decreased when the number of bands decreased from 16 to 8. Moreover, participants achieved better performance with sadness and hot anger stimuli than the other emotions. Through comparing the modulation spectral features and perceptual data of the vocal-emotion recognition experiment, the modulation spectral centroid in each modulation frequency band and modulation spectral tilt had a similar trend with the results of human perception. This suggests that these modulation spectral features may be important cues for vocal emotion recognition with noise-vocoded speech. In the future, the effect of modifying modulation spectral features on vocal emotion recognition will be investigated to clarify whether these features can contribute to the perception of vocal emotion.

6. Acknowledgments

This work was supported by a Grant in Aid for Scientific Research (A) (No. 25240026), Young Scientists (A) (No. 24683026), and Innovative Areas (No. 16H01669) from MEXT, Japan. This work was also supported by the ICT Global Leader Program of the JAIST, an A3 foresight program made available by the JSPS and Mitani Foundation for Research and Development.

7. References

[1] M. C. 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.

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

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

[4] P. C. Loizou, "Mimicking the human ear," *IEEE Signal Processing Magazine*, vol. 98, pp. 1053–5888, 1998.

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

[6] 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.

[7] 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, pp. 47–60, 2012.

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

[9] B. C. J. Moore, *An introduction to the psychology of hearing*, 6th Edition, London, Elsevier, pp. 74–80, 2013.