Study on the relationship between modulation spectral features and the perception of speaker individuality with 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, Ishikawa, 923-1292 Japan
[2)] Kanazawa University, Kakuma-machi, Kanazawa, Ishikawa, 920-1192 Japan
E-mail: [1)]{zhuzhi, ryota, unoki}@jaist.ac.jp, [2)]yukikoa@staff.kanazawa-u.ac.jp

Abstract Our previous study showed that the modulation spectral features play an important role in the perception of vocal emotion with noise-vocoded speech. In this paper, the relationship between the perception of speaker individuality and modulation spectral features of noise-vocoded speech was investigated. For human perception, a speaker recognition experiment using noise-vocoded speech was carried out to clarify whether temporal envelope cues can support speaker recognition. Modulation spectral features of noise-vocoded speech were then extracted by using modulation filterbank to account for the perceptual data. The results showed that there were positive correlations between the modulation spectral features and the perceptual data of speaker recognition experiment. It is suggested that modulation spectral features may also contribute to the perception of speaker individuality with noise-vocoded speech.

Keywords Noise-vocoded speech, Modulation spectral features, Speaker individuality, Temporal cue

1. Introduction

The temporal envelope of speech has been proved to be an important cue in perceiving linguistic and nonlinguistic information included in the speech. Shannon et al. showed that the presentation of a dynamic temporal pattern in only a few broad spectral regions is sufficient for listeners to the recognize of linguistic information using noise-vocoded speech (NVS) [1]. The modulation frequency bands from 4 to 16 Hz have been shown to be important regions in speech recognition [2].

For the perception of nonlinguistic information, Zhu et al. examined word and speaker recognition using NVS while systematically varying the upper limit of the modulation frequency [3]. The results suggested that the temporal resolution of NVS should contribute to the speaker recognition. They then investigated the relative contributions of the spectral and temporal cues on the perception of vocal emotion [4] and speaker individuality [5]. Vocal emotion and speaker recognition experiments were carried out by systematically varying the number channels and the upper limit of the modulation frequency of NVS stimuli. The results showed that the modulation components contained in the temporal envelope play an important role in both vocal emotion and speaker recognition.

Furthermore, to clarify the specific features of temporal envelope that contribution to the perception of vocal emotion, Zhu et al. investigated the modulation spectral features of emotional speech and its relationship between the perceptual data of vocal emotion recognition experiment [6]. It was found that there was a high correlation between the modulation spectral features and the perceptual data. Based on these results, Zhu et al. then proposed a vocal-emotion conversion method by modifying the modulation spectrogram and its features of neutral speech to match that of emotion speech [7]. As a result, it was found that the modulation spectrogram of neutral speech can be successfully converted to that of emotional speech. Therefore, modulation spectral features were proved to be important cues in the perception of vocal emotion with NVS. Because the modulation components of temporal envelope were shown to contribute to the perception of speaker individuality, the modulation spectral features may also be important in the speaker recognition with NVS.

In this paper, the relationship between the modulation spectral features and the perception of speaker individuality with NVS was investigated. At first, ten types of modulation spectral feature were extracted from the modulation spectrogram of the speech spoken by 20 different speakers. Then, for human perception, a speaker recognition experiment using NVS stimuli was carried out. Finally, the correlation between the modulation spectral features and the perceptual data was calculated to discuss whether the modulation spectral features will contribute to the perception of speaker individuality with NVS.

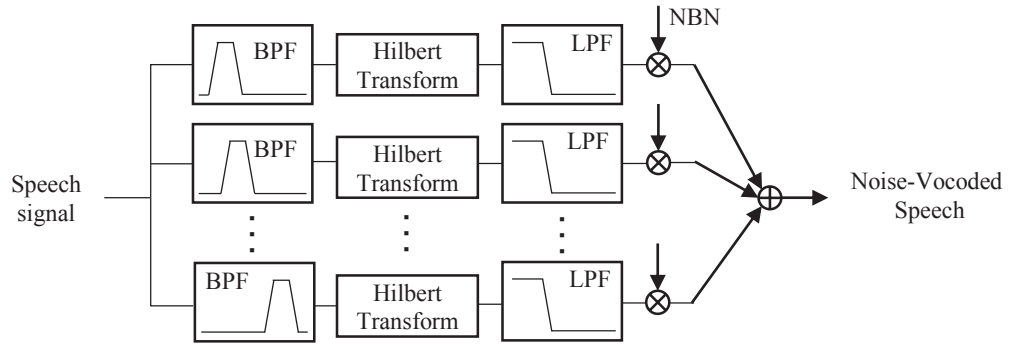

Figure 1: Schematic diagram of the noise-vocoder method used to generate the stimuli (BPF: band-pass filter; LPF: low-pass filter; NBN: narrow-band noise)

2. Speaker recognition experiment using NVS

2.1. Speech data

Speech data from the ATR Japanese speech database set C were used in this study. All speech data were recorded at a 20 kHz sampling frequency. Each sentence was uttered for about 4 to 5 s.

In this study, the XAB method was used in the speaker identification 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 also the 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 a speaker pair will affect the results of experiment. The speaker pair with high similarity may be difficult to be distinguish, even when the spectral and temporal cues were preserved. On the contrary, the speaker pair with low similarity may be still easy to be distinguish, even if the cues related to speaker identification were reduced. This kind of bias is not desirable.

Therefore, the speaker pairs used in this study were selected based on the perceptual similarity data measured be Kitamura et al. [8]. Kitamura et al. measured the perceptual similarity of speaker individualities of 20 female and 20 male Japanese speakers in ATR speech database set C. Two same sentences with different speakers were presented to normal-hearing listeners, and the listeners were asked to select the similarity of these two speakers from 1 to 5. The 5 female and 5 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. 12 sentences of each speaker were used to generate the NVS stimuli.

Table 1: Speaker pairs selected from ATR database and their average similarity index. Left and right halves show female and male speaker pairs, respectively.

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

2.2. Signal generation

Figure 1 shows the method to generate NVS stimuli used in this study was the same as the method used in [5]. 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 [9]. Speech signal was first divided into 8 or 16 frequency bands with a band-pass filterbank based on the ERB_N-number scale [10]. The ERB_N-number scale is comparable to a scale of distance along the basilar membrane so that the frequency resolution of the auditory system can be faithfully replicated by dividing frequency bands in accordance with the ERB_N-number.

Then, the temporal envelope of the output signal from each band-pass filter was extracted by using a Hilbert transformation and performing a low-pass filter (2nd-order Butterworth IIR filter). The cut-off frequency of the low-pass filter was 64 Hz.

Finally, the temporal envelope in each channel served to amplitude modulation with the band-limited noise which 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.

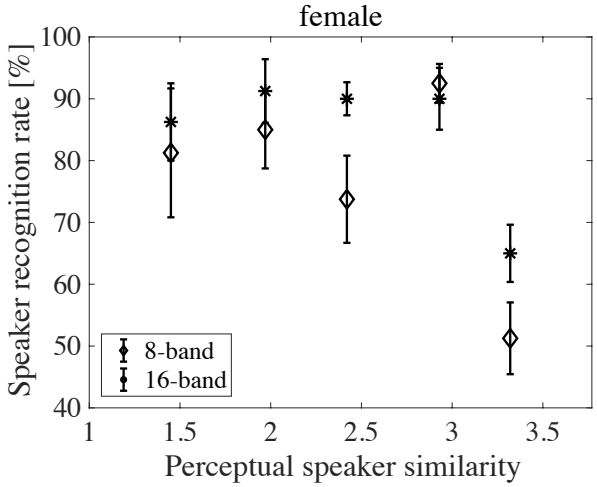

Figure 2: Results of speaker recognition rate for female speaker pairs.

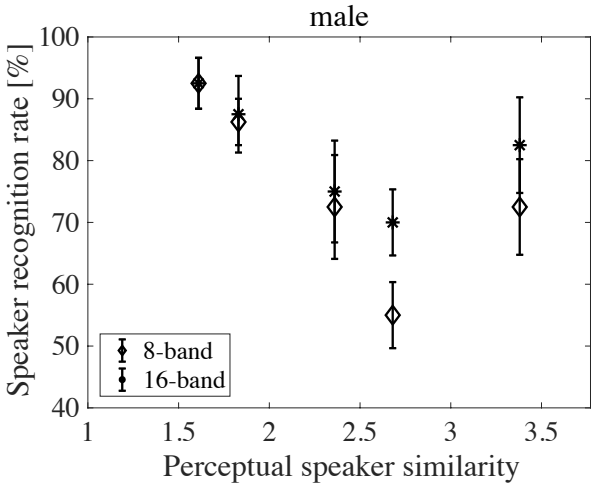

Figure 3: Results of speaker recognition rate for male speaker pairs.

2.3. Procedure

Eight native Japanese speakers (two females and six males) participated in this experiment. All participants had normal hearing (hearing losses of the participants were below 12 dB in the frequency range from 125 to 8000 Hz).

This experiment was carried out by 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: Noise-vocoded speech

A: Noise-vocoded speech with the same speaker as X

B: Noise-vocoded speech with a different speaker from X.

Participants were asked to compare the speakers of A and B with the speaker of X to select which speaker was more similar to the speaker of first speech X. Both stimulus with XAB and XBA orders were presented to counterbalance any effects due to the order of presentation.

2.4. Results

Figure 2 and 3 shows the results of speaker recognition rates of female and male speaker pairs. For female speaker pairs, the speaker recognition rate decreased dramatically when the speaker similarity was higher than 3. For male speaker pairs, when the speaker similarities were lower than 3, the speaker recognition rates decreased with the increasing of similarity. However, the speaker recognition rate was suddenly increased when the speaker similarity was higher than 3.

A 3-way repeated measures ANOVA was then conducted on the results with the gender of speaker pairs, speaker similarity, and the number of channels as the factors. The results of ANOVA show that the main effect of the gender of speaker pairs was not significant ($F(1,7) = 1.38$, $p = 0.28$). The main effect of the number of channels ($F(1,7) = 8.58$, $p < 0.05$) was significant. These results are different from the previous study [5]. The effect of the number of channels in speaker recognition was shown to be different when the speaker pairs and their similarity were different. Furthermore, the main effect of the speaker similarity ($F(4,28) = 9.59$, $p < 0.01$) was also significant. In the next section, the modulation spectral features were calculated to account for the perceptual data obtained in this experiment and the effect of speaker similarity.

3. Modulation spectral features

3.1. Modulation spectrogram analysis

To extract the modulation spectral features, the modulation spectrogram was calculated as the first step. Figure 4 shows the auditory-inspired process used to calculate the modulation spectrogram. The method used in this was the same as the method in [5].

Speech signals were first divided into 8 or 16 acoustic frequency bands using the same filterbank described in Section 2.2. The instantaneous amplitude of each band was then calculated using Hilbert transform.

The next step involved decomposing the instantaneous amplitude in to several modulation frequency bands by using a modulation filterbank. The modulation filterbank consisted of six filters (one low-pass filter and five band-pass filters). The low-pass filter was a 2nd order Butter-worth 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 time-

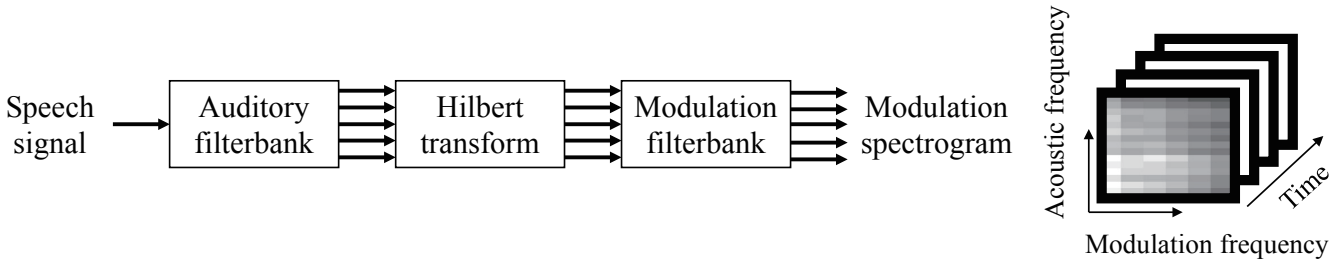

Figure 4: Block diagram overview of the process used to calculate modulation spectrogram.

averaged power of the output signal from each modulation filter was calculated to obtain the modulation spectrogram. The modulation spectrogram can describe the speaker individuality of speech signal not only on the acoustic frequency domain but also on the modulation frequency domain.

3.2. Modulation-spectral features extraction

Ten types of modulation spectral feature were then extracted from the modulation spectrogram. Two kinds of modulation spectral feature were calculated by analyzing the modulation spectrogram in the acoustic frequency domain and the modulation frequency domain. In the acoustic frequency domain, the first feature was the modulation spectral centroid MSCR_m, which can be defined as follows:

$$\mathrm{MSCR}_m = \frac{\sum_{k=1}^{K} kE_{k,m}}{\sum_{k=1}^{K} E_{k,m}}, \quad (1)$$

where k and m are the acoustic and modulation frequency band respectively, K is the number of channels which can be 8 or 16. $E_{k,m}$ is the modulation spectrogram which is also the time-averaged power in the k^{th} acoustic frequency band and the m^{th} modulation frequency band. The MSCR_m indicates the center of the spectral balance across acoustic frequency bands.

The modulation spectral spread (MSSP_m), was the calculated by:

$$\mathrm{MSSP}_m = \frac{\sum_{k=1}^{K}[k - \mathrm{MSCR}_m]^2 E_{k,m}}{\sum_{k=1}^{K} E_{k,m}}. \quad (2)$$

The MSSP_m can represent the spread of the spectrum around its MSCR_m as the 2^{nd} 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 3^{rd} order moment. The MSKT_m gives a measure of the peakedeness of the spectrum which was calculated from the 4^{th} order moment.

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

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). The other three higher order features of the modulation spectrogram $(\mathrm{MSSP}_k$, MSSK_k, and $\mathrm{MSKT}_k)$ and the modulation spectral tilt (MSTL_k) on the modulation frequency domain were also calculated.

A discriminability index (d') 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 each speaker pairs, divided by their average standard deviation. The average value of discriminability indices (taken across all the utterances used in the experiment) was computed as a measure of the net discriminability provided by this feature. The d' values of the perceptual data obtain in speaker recognition experiment were then calculated. The d' values of modulation spectral features present the physical distance of such features between two speakers. The calculated d' values of perceptual data present the psychological distance between two speakers based on the results of speaker recognition experiment. Finally, the correlation coefficients of the d' values of modulation spectral features were calculated.

3.3. Results

Table 2, 3, 4, and 5 show results of the d' values of perceptual data and modulation spectral features can their correlations for 8 and 16 bands NVS with female and male speaker pairs. As a result, the correlation coefficients in all

Table 2: The d' values of perceptual data and modulation spectral features and their correlations for 8 bands NVS and female speaker pairs.

	F507&F609	F407&F702	F213&F214	F611&F614	F606&F704	Correlation
Perceptual data	1.349	0.843	1.355	1.690	0.289	
$MSCR_m$	1.828	1.012	1.285	0.756	0.542	0.14
$MSSP_m$	3.508	1.318	1.598	0.882	0.617	0.17
$MSSK_m$	2.822	0.743	0.900	0.499	0.475	0.19
$MSKT_m$	5.530	1.163	1.700	0.609	0.381	0.14
$MSTL_m$	9.104	1.329	1.652	0.803	0.519	0.03
$MSCR_k$	1.612	0.636	0.713	0.744	0.274	0.21
$MSSP_k$	2.081	0.851	0.755	0.615	0.400	0.28
$MSSK_k$	2.696	1.323	0.881	0.762	0.485	0.27
$MSKT_k$	3.181	1.699	0.967	1.016	0.533	0.24
$MSTL_k$	1.998	1.428	0.675	1.233	0.586	0.47

Table 3: The d' values of perceptual data and modulation spectral features and their correlations for 16 bands NVS and female speaker pairs.

	F507&F609	F407&F702	F213&F214	F611&F614	F606&F704	Correlation
Perceptual data	1.825	2.195	1.272	2.926	0.063	
$MSCR_m$	2.181	2.165	1.072	0.812	0.583	0.33
$MSSP_m$	7.400	4.346	1.633	0.763	0.535	0.21
$MSSK_m$	2.974	2.529	1.043	1.078	0.655	0.45
$MSKT_m$	11.235	5.601	1.289	0.554	0.614	0.19
$MSTL_m$	11.555	10.914	1.207	0.759	0.471	0.28
$MSCR_k$	1.741	1.525	1.165	0.464	0.378	0.18
$MSSP_k$	2.017	1.455	1.396	0.518	0.418	0.10
$MSSK_k$	2.184	1.664	1.397	0.627	0.468	0.13
$MSKT_k$	2.341	1.919	1.266	0.713	0.483	0.17
$MSTL_k$	1.540	1.100	0.901	0.826	0.482	0.20

Table 4: The d' values of perceptual data and modulation spectral features and their correlations for 8 bands NVS and male speaker pairs.

	M504&M601	M614&M710	M214&M519	M509&M603	M409&M705	Correlation
Perceptual data	1.272	0.910	1.272	0.316	0.864	
$MSCR_m$	2.794	1.030	1.567	0.887	0.739	0.72
$MSSP_m$	4.426	1.387	1.434	0.678	0.430	0.65
$MSSK_m$	3.027	1.078	1.013	0.691	0.862	0.61
$MSKT_m$	6.808	1.273	1.194	0.502	0.240	0.63
$MSTL_m$	11.294	1.452	1.741	0.941	0.649	0.56
$MSCR_k$	1.780	0.748	0.765	0.567	0.245	0.41
$MSSP_k$	2.299	0.977	0.858	0.378	0.314	0.55
$MSSK_k$	2.827	1.243	0.895	0.444	0.427	0.52
$MSKT_k$	3.063	1.431	0.903	0.721	0.501	0.46
$MSTL_k$	1.343	1.151	0.809	1.120	0.604	0.12

Table 5: The d' values of perceptual data and modulation spectral features and their correlations for 16 bands NVS and male speaker pairs.

	M504&M601	M614&M710	M214&M519	M509&M603	M409&M705	Correlation
Perceptual data	2.879	2.187	1.199	0.252	1.209	
$MSCR_m$	3.396	2.683	1.178	0.925	0.633	0.72
$MSSP_m$	8.596	4.947	1.607	0.498	0.661	0.81
$MSSK_m$	3.279	2.801	0.810	1.230	0.626	0.82
$MSKT_m$	13.756	6.308	0.830	0.585	0.853	0.79
$MSTL_m$	16.702	13.176	1.269	0.824	0.626	0.76
$MSCR_k$	2.380	1.930	1.259	0.628	0.730	0.76
$MSSP_k$	2.565	1.683	1.382	0.638	0.808	0.87
$MSSK_k$	2.704	1.784	1.266	0.651	0.810	0.89
$MSKT_k$	2.769	1.969	1.104	0.649	0.787	0.86
$MSTL_k$	1.774	1.115	0.810	0.733	0.716	0.50

conditions are positive. However, the correlations of male speaker pairs are obviously higher than that of female speaker pairs. The correlations of 16 bands NVS are higher than that of 8 bands NVS. In conclusion, as the correlations are all positive, the results showed that the psychological distance of each speaker pair increases as the distance of modulation spectral features increases. Therefore, it is suggested that the modulation spectral features may contribute the perception of speaker individuality.

4. Discussion

The values of correlation coefficient were different with different conditions. The correlations for male speaker pairs are higher than that of female speaker pairs. The d' values of modulation spectral features roughly decreased with the increasing of perceptual speaker similarity of speaker pairs. However, the perceptual data showed that the speaker recognition rates did not decreased with the increasing of speaker similarity linearly. A possibility reason may be that the relationship between the modulation spectral feature and perceptual data is not linear. Moreover, the number of speaker pairs may not be large enough. Kitamura et al. measured the perceptual similarity of total 380 speaker pairs [8]. Speaker recognition experiments with more speaker pairs are necessary to obtain more general role of the modulation spectral features in the perception of speaker individuality.

In all conditions, the results showed that there are positive correlations between the modulation spectral features and perceptual data. These results have shown the potential possibility of modulation spectral features for speaker individuality analysis. The modification of modulation spectrogram and its features has been shown to be useful to convert vocal emotion of NVS [7]. In the future, the effect of modifying modulation spectral features on the speaker recognition will also be investigated to confirm whether these features contribute to the perception of speaker individuality.

5. Summary

The relationship between the perception of speaker individuality and modulation spectral features of noise-vocoded speech was investigated in this paper. At first, a speaker recognition experiment using 8 and 16 bands NVS was carried out. To account for the perceptual data obtained in speaker recognition experiment, the modulation spectral features of NVS stimuli were then extracted. Finally, the correlation coefficients of the discriminability index (d') of modulation spectral features and the perceptual data were calculated. The results showed that there were positive correlations between the modulation spectral features and the perceptual data of speaker recognition experiment. It was suggested that modulation spectral features should contribute to the perception of speaker individuality with NVS.

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*, vol. 270, no. 5234, pp. 303–304, 1995.

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

[3] Z. Zhu, R. Miyauchi, Y. Araki, and M. Unoki, "Study on linguistic information and speaker individuality contained in temporal envelope of speech," *Acoustical Science and Technology,* Vol. 37, No. 5, pp. 258-261, 2016.

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

[5] Z. Zhu, R. Miyauchi, Y. Araki, and M. Unoki, "Important role of temporal cues in speaker identification for simulated cochlear implants," *Proc. of the 1st International Workshop on Challenges in Hearing Assistive Technology*, pp. 51-55, 2017.

[6] Z. Zhu, R. Miyauchi, Y. Araki, and M. Unoki, "Modulation spectral features for predicting vocal emotion recognition by simulated cochlear implants," *INTERSPEECH 2016*, pp. 262-266, 2016.

[7] Z. Zhu, R. Miyauchi, Y. Araki, and M. Unoki, "Feasibility of Vocal Emotion Conversion on Modulation Spectrogram for Simulated Cochlear Implants," *EUSIPCO2017*, pp. 1884-1888, 2017.

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

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

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