Feasibility of Vocal Emotion Conversion on Modulation Spectrogram for Simulated Cochlear Implants

Zhi Zhu*, Ryota Miyauchi*, Yukiko Araki†, and Masashi Unoki*
*School of Information Science, Japan Advanced Institute of Technology and Science, Japan
†School of Humanitites, Kanazawa University, Japan

***Abstract*—Cochlear implant (CI) listeners were found to have great difficulty with vocal emotion recognition because of the limited spectral cues provided by CI devices. Previous studies have shown that the modulation spectral features of temporal envelopes may be important cues for vocal emotion recognition of noise-vocoded speech (NVS) as simulated CIs. In this paper, the feasibility of vocal emotion conversion on a modulation spectrogram for simulated CIs for correctly recognizing vocal emotion is confirmed. A method based on a linear prediction scheme is proposed to modify the modulation spectrogram and its features of neutral speech to match that of emotional speech. The logic of this approach is that if vocal emotion perception of NVS is based on the modulation spectral features, NVS with similar modulation spectral features of emotional speech will be recognized as the same emotion. As a result, it was found that the modulation spectrogram of neutral speech can be successfully converted to that of emotional speech. The results of the evaluation experiment showed the feasibility of vocal emotion conversion on the modulation spectrogram for simulated CIs. The vocal emotion enhancement on the modulation spectrogram was also further discussed.**

I. INTRODUCTION

High intelligibility of speech can be achieved by cochlear implant (CI) listeners. However, it was found that CI listeners' performance of vocal emotion recognize was lower than that of normal-hearing (NH) listeners. The main reason they failed is due to the limited spectral cues provided by CI devices because the temporal envelope cue is used as a primary cue. Research on speech perception by CI listeners has been conducted using acoustic simulations such as noise-vocoded speech (NVS) [2] with normal-hearing listeners. An NVS stimulus is generated by replacing the temporal fine structure of speech with a noise carrier while the temporal amplitude envelope is preserved. It is related to the fact that CI devices provide the temporal envelope information as a primary cue, and the temporal fine structure information is not effectively encoded [3].

Chatterjee *et al.* provided a comparison of the performance of vocal emotion recognition by both CI and NH listeners with NVS as CI simulations [1]. They also analyzed the mean intensity, intensity range, and duration of stimuli to clarify the acoustic features that contribute to the perception of vocal emotion. However, they found that the results of acoustic analyses cannot account for all of the perceptual data of experiments. For CI listeners, the temporal envelope was provided as a primary cue. The modulation spectral features extracted from the temporal envelope of speech should be considerable cues for vocal emotion recognition by CI listeners.

Modulation spectral features have been successfully applied in automatic vocal-emotion recognition system [4]. That means modulation spectral features can be used to represent the vocal emotional information. Zhu *et al.* investigated the relationship between the modulation spectral features of the temporal envelope and human perception of emotion with NVS [5]. The results showed that sadness and hot anger are more easily recognized than joy and cold anger with simulated CIs. Similar trends were also shown from experiments with CI listeners [6]. High correlations between modulation spectral features and the perception of vocal emotion based on the NVS scheme were found. These important studies suggested that the modulation spectrogram of speech should be an important cue for voice emotion recognition with simulated CIs.

This paper aims to study the feasibility of vocal emotion conversion on a modulation spectrogram for simulated CIs. Luo and Fu successfully enhanced the tone recognition on the NVS scheme by manipulating the amplitude envelope to more closely resemble the F0 contour [7]. Their results showed the possibility of enhancing the recognition of non-linguistic information by modifying the temporal envelope. It is also found that the sound texture can be converted successfully by modifying the modulation spectrogram [8].

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

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

ISBN 978-0-9928626-7-1 © EURASIP 2017

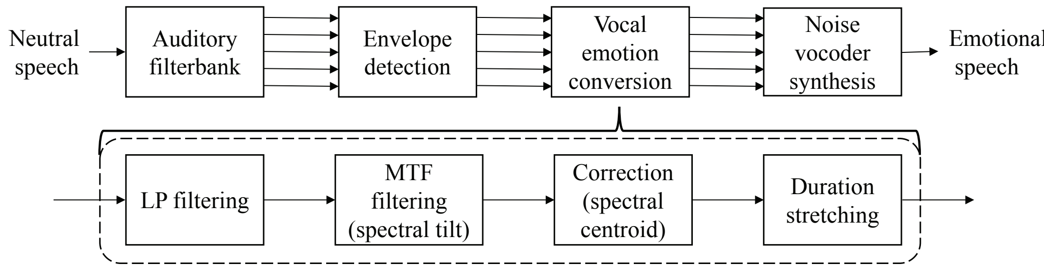

Fig. 1. Scheme of LP based vocal emotion conversion method.

TABLE I
BOUNDARY FREQUENCIES OF AUDITORY-INSPIRED BAND-PASS FILTERBANK.

Band number	$\mathrm{ERB_N}$-number	Boundary frequencies [Hz]
1	3 - 7	87.18 - 257.2
2	7 - 11	257.2 - 518.5
3	11 - 15	518.5 - 920.5
4	15 - 19	920.5 - 1539
5	19 - 23	1539 - 2489
6	23 - 27	2489 - 3951
7	27 - 31	3951 - 6200
8	31 - 35	6200 - 9657

neutral and emotional speech.

Finally, a vocal-emotion recognition experiment using NVS generated by the converted temporal envelope is carried out. The method for enhancing the vocal-emotion information of the modulation spectrogram is also discussed further. The final goal of this research is to propose a front-end processor for a CI device to improve the vocal emotion recognition by CI listeners. The novelty of this study is considering the conversion of the vocal emotion information on the modulation frequency domain and trying to enhance the modulation spectral features of vocal emotion to improve the vocal emotion recognition on the NVS scheme.

II. VOCAL EMOTION CONVERSION ON MODULATION SPECTROGRAM

In this section, the method of vocal emotion conversion on the modulation spectrogram as shown in Fig. 1 is described.

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

A. Auditory-inspired band-pass filterbank and temporal envelope extraction

The performance of vocal emotion recognition by CI listeners was found to be similar to that of NH listeners with 8-band NVS [1]. Therefore, in this study, the speech signal was divided into 8 bands by an auditory-inspired band-pass filterbank as follows:

$$s(k,n) = h_{\mathrm{BPF}}(k,n) * s(n) \tag{1}$$

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

The auditory filterbank was constructed by using 3rd-cascaded 2nd-order Butterworth IIR filters. The bandwidth of the filter was designed as $\mathrm{ERB_N}$ (equivalent rectangular bandwidth), and all filters were placed on the $\mathrm{ERB_N}$-number scale [11]. $\mathrm{ERB_N}$-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 the acoustic frequency in Hz. This scale is comparable to a scale of distance along the basilar membrane so that the frequency resolution of the auditory system can be faithfully replicated by dividing the frequency bands according to $\mathrm{ERB_N}$-number. In this study, the boundary frequencies of band-pass filters are spaced from 3 to 35 $\mathrm{ERB_N}$-numbers with 4 $\mathrm{ERB_N}$ as the bandwidth of the acoustic frequency region (8-bands). Table I shows the boundary frequencies of the band-pass filterbank in Hz.

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

$$e(k,n) = |s(k,n) + j\mathcal{H}[s(k,n)]| * h_{\mathrm{LPF}}(n) \tag{3}$$

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

B. Vocal emotion conversion based on LP scheme

In the previous study, it was found that modulation spectral features are suggested to be important cues for vocal emotion recognition with simulated CIs [5]. Table 1 in [5] showed that the discriminability indices of modulation spectral features (kurtosis, tilt, and centroid as higher order statistics) have high correlation with the perceptual data of experiments with NVS stimuli. Modulation spectral kurtosis gives a measure of the peakedness of the modulation spectrum. Modulation spectral

ISBN 978-0-9928626-7-1 © EURASIP 2017

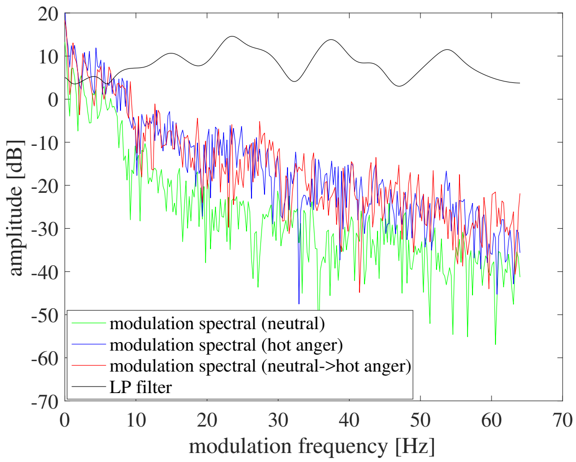

Fig. 2. Modulation spectrum of neutral, hot anger, and NE-HA converted speech on 3rd band and frequency characteristic of LP based conversion filter.

tilt is the linear regression coefficient obtained by fitting a first-degree polynomial to the modulation spectrum. Modulation spectral centroid indicates the center of spectral balance across acoustic frequency bands. If these modulation spectral features are important cues for vocal emotion recognition, converting the vocal emotion by modifying these modulation spectral features should be possible. In this study, we used three steps to modify the modulation spectrogram and these modulation spectral features of neutral speech close to the target emotion.

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

$$H_{\mathrm{LP}}(z) = \frac{\sum_{i=0}^{p} b_{NE,i} z^{-i}}{\sum_{i=0}^{p} a_{EM,i} z^{-i}} \tag{4}$$

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

$$\hat{e}_{\mathrm{LP}}(k,n) = e_{\mathrm{NE}}(k,n) * h_{\mathrm{LP}}(k,n) \tag{5}$$

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

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

$$\hat{e}_{\mathrm{MTF}}(k,n) = \hat{e}_{\mathrm{LP}}(k,n) * h_{\mathrm{MTF}}(k,n) \tag{6}$$

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

$$\hat{e}(k,n) = \hat{e}_{\mathrm{MTF}}(k,n) \frac{\bar{e}_{\mathrm{NE}}(k)}{\bar{e}_{\mathrm{EM}}(k)} \tag{7}$$

where $\bar{e}_{\mathrm{NE}}(k)$ and $\bar{e}_{\mathrm{EM}}(k)$ are the average amplitude of the envelope of neutral speech and the target emotional speech in the kth band. This process can modify the modulation spectrogram on the acoustic frequency domain to shift the spectral centroid close to the target emotion.

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

Figure 2 shows an example of the modulation spectrum of the converted temporal envelope. The target emotion is hot anger and the modulation spectrum in the 3rd channel is shown. The modulation spectrum is the amplitude spectrum of the temporal envelope calculated by the Fourier transform. The results show that the modulation spectrum of the converted temporal envelope (blue line) is very close to that of the target emotion (red line) from neutral speech (green line). Figure 3 shows the modulation spectrograms of neutral, emotional speech, and converted speech. As a result, the shape of the modulation spectrogram of converted speech is similar to that of hot anger speech. That means the modulation spectrogram of neutral speech was successfully converted to that of emotional speech.

III. Evaluation Experiment

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

A. Stimuli

To generate a stimulus in the 8–band NVS scheme, the envelope of each band was used to amplitude modulated with band-limited noise limited in the same band. Then, all amplitude modulated band-limited noises were summed to generate a stimulus. To confirm the effect of modifying the modulation spectrum with LP filtering, a condition with only amplitude correction and no modification of modulation

ISBN 978-0-9928626-7-1 © EURASIP 2017

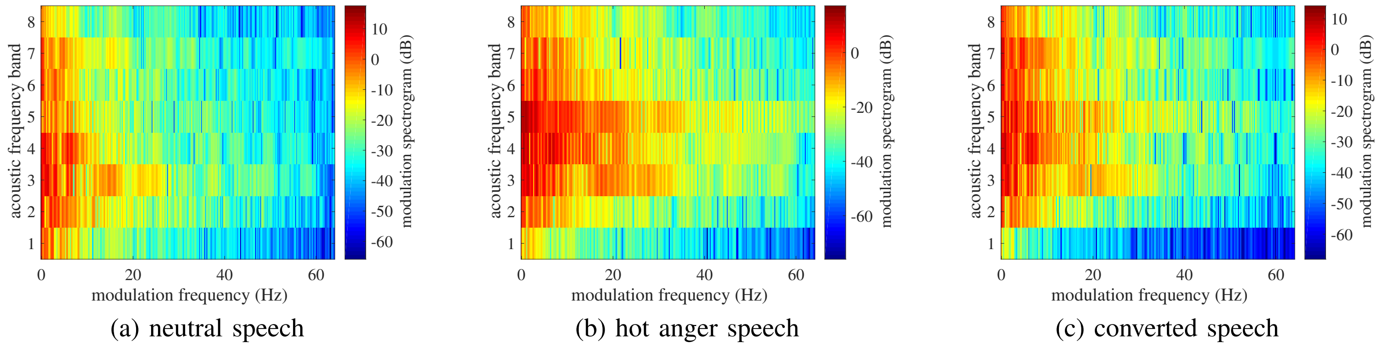

Fig. 3. Modulation spectrograms of (a) neutral, (b) hot anger, and (c) NE-HA converted speech.

spectrum by LP filtering was added. For joy, sadness, and hot anger, 10 sentences of vocal emotion conversion with the LP filter and vocal emotion conversion with only amplitude correction were generated. There were also 10 sentences of neutral NVS for the balance of stimuli.

B. Procedure

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

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

C. Results

Figure 4 shows the vocal emotion recognition rates of the experiment. The vocal emotion recognition rate was very low for joy. However, joy was found to be more difficult to recognize than the other emotions, even with the original joy NVS. The method of further enhancing the modulation spectral features to increase the recognition rate of joy is discussed in the next section. For sadness and hot anger, the results of vocal emotion conversion with the LP filter were higher than those without the LP filter. The results show that the process of LP filtering for modifying the modulation spectrogram is effective for the vocal emotion conversion of sadness and hot anger. Furthermore, the modulation spectrogram is confirmed to be an important cue for the perception of vocal emotion with simulated CIs. However, the results of repeatedly measured analyses of variance showed that there was no significant difference between the process method with and without the LP filter ($F(1, 3) = 4.84$). More experiments with more participants are necessary.

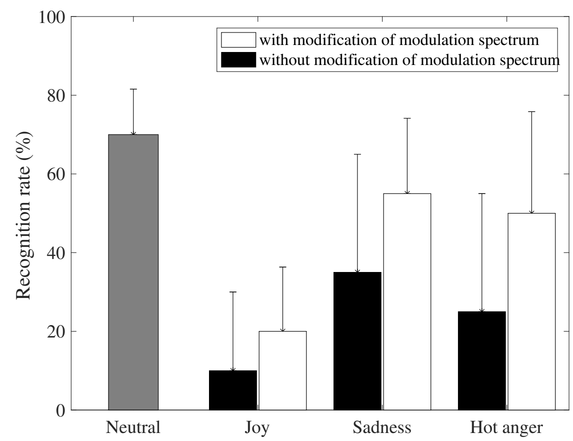

Fig. 4. Results of vocal-emotion recognition experiment.

IV. Discussion

McDermott *et al.* successfully converted the texture of sound by modifying the modulation spectrogram [8]. The method they used began with processing stages from the auditory periphery (auditory filterbank, envelope extraction, and modulation filterbank) to calculate the modulation spectrogram and culminated with the measurement of simple statistics of these stages. It was found that the synthetic textures will sound like another example of the corresponding real-world texture if the statistics of the modulation spectrogram used for synthesis are similar to those of the real-world texture. Their results suggested the importance of the modulation spectrogram in the timbre perception by humans and the possibility of converting sound signals by modifying the modulation spectrogram.

In the previous study, we investigated the relationship between the modulation spectral features of the temporal envelope and human perception of NVS [5]. These results suggested that the modulation spectral centroid, modulation

TABLE II

MEAN VALUE OF MODULATION SPECTRAL FEATURES OF ORIGINAL AND CONVERTED EMOTIONAL NVS OVER ALL MODULATION OR ACOUSTIC FREQUENCY BANDS. (MSCR: MODULATION SPECTRAL CENTROID, MSKT: MODULATION SPECTRAL KURTOSIS, MSTL: MODULATION SPECTRAL TILT. NE-EM (JO, SA, HA): VOCAL EMOTION CONVERTED NVS FROM NEUTRAL TO EMOTIONAL)

	NE	JO	NE-JO	SA	NE-SA	HA	NE-HA
$\overline{MSCR}$	3.55	3.71	3.74	2.34	2.42	4.39	4.72
$\overline{MSKT}$	5.06	5.92	6.67	8.46	6.31	6.60	6.43
$\overline{MSTL}$	-4.23	-4.04	-3.71	-3.10	-3.78	-3.79	-3.81

spectral kurtosis, and modulation spectral tilt are important cues for vocal emotion recognition with simulated CIs. These modulation spectral features of original and converted emotional NVS were calculated by using the same method in [5]. Table II shows the results of the modulation spectral features. It was confirmed that the modulation spectral features could be converted to the direction of the target emotion using the proposed method. In addition, the NVS with converted modulation spectral features should sound like the target emotional NVS.

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

In this paper, a vocal emotion conversion method for simulated CIs was proposed. The final goal of this research is to propose a signal process method for improving the vocal emotion recognition by CI listeners. We assumed that the target of vocal emotion is known (e.g., vocal-emotion recognition methods can be used to predict the target emotion via a dimension approach (V-A) [10]). In the future, the method to enhance the vocal emotion information of emotional NVS by modifying the modulation spectral features will be discussed further.

V. CONCLUSION

The aim of this paper was to study the feasibility of vocal emotion conversion on the modulation spectrogram for simulated CIs to recognize vocal emotion correctly. A method based on a LP scheme was proposed to modify the modulation spectrogram and its features of neutral speech to that of emotional speech. The results showed that the modulation spectrogram of neutral speech can be successfully converted to that of emotional speech by the proposed method. Then a vocal-emotion recognition experiment using NVS generated by the converted temporal envelope was carried out. The results of the evaluation experiment confirmed the feasibility of vocal emotion conversion on the modulation spectrogram for simulated CIs. The method for enhancing the vocal-emotion information of the modulation spectrogram was then further discussed. In the future, the proposed method will be used to enhance the vocal emotion information of emotional NVS and to improve the vocal emotion recognition by the CI listeners.

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] 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, April 2015.

[2] 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, October 1995.

[3] P. C. Loizou, "Mimicking the human ear," *IEEE Signal Processing Magazine*, vol. 98, pp. 101–130, Spetember 1998.

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

[5] Z. Zhu, R. Miyauchi, Y. Araki and M. Unoki, "Modulation spectral features for predicting vocal emotion recognition," *INTERSPEECH2016*, pp. 262–266, September 2016.

[6] Z. Zhu, R. Miyauchi, Y. Araki and M. Unoki, "Recognition of Vocal emotion in noise-vocoded speech by normal hearing and cochlear implant listeners," *5th Joint Meeting Acoustical Society of America and Acoustical Society of Japan*, pp. 3271, December 2016.

[7] X. Luo and Q. Fu, "Enhancing Chinese tone recognition by manipulating amplitude envelope: Implications for cochlear implants," *Journal of the Acoustical Society of America*, vol. 116, pp. 3659–3667, December 2004.

[8] J. H. McDermott, and E. P. Simoncelli, "Sound Texture Perception via Statistics of the Auditory Periphery: Evidence from Sound Synthesis," *Neuron*, vol. 71, pp. 926–940, September 2011.

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

[10] X. Li and M. Akagi, "Multilingual Speech Emotion Recognition System based on a Three-layer Model," *INTERSPEECH2016*, pp. 2608–2612, September 2016.

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