Important role of temporal cues in speaker identification for simulated cochlear implants

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

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

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

Abstract

Speaker identification is still challenging issue for cochlear implant (CI) users due to the poor spectral cue provided by the CI device. To optimize CI systems for the users, it is important to understand the role of temporal modulation cues in speaker identification, as the CI device provides temporal modulation cues as primary cues. This study investigates the relative contributions of spectral and temporal cue on speaker identification by using noise-vocoded speech (NVS) as a CI simulation. In the experiment, speaker identification was conducted in normal-hearing listeners as a function of the number of channels (4, 8, and 16) and upper limitation of envelope frequency (0, 0.5, 1, 2, 4, 8, 16, 32, and 64 Hz) in NVS. The number of channels and upper limitation of envelope frequency present the spectral and temporal resolution of NVS separately. Results showed that the performance of speaker identification was not affected by spectral resolution significantly, at least in the limited set of stimuli in the present study. In addition, the results also showed that the performance was more sensitive to temporal resolution. It is suggested that temporal modulation cues contribute to speaker identification and have the potential to improve speaker identification if enhanced.

Index Terms: temporal modulation cue, speaker identification, noise-vocoded speech, cochlear implant

1. Introduction

The temporal envelope of speech has been proved to be an important cue in perceiving linguistic 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 [1]. The modulation frequency bands from 4 to 16 Hz have been shown to be important regions in speech recognition [2]. Also, cochlear-implant (CI) users can achieved good performance in speech recognition, as the CI device can provide sufficient temporal cues. However, human speech includes not only linguistic information, but also nonlinguistic information such as speaker individuality. CI users cannot accurately identify speakers as the CI device provides poor spectral cues [3].

To optimize CI systems for their users, the role of temporal modulation cues in speaker identification must be understood. It is necessary to know which aspects of the temporal modulation cues have the potential to improve speaker identification if enhanced. Luo and Fu successfully enhanced the tone recognition on the NVS scheme by manipulating the amplitude envelope to more closely resemble the F0 contour [4]. Their results showed the possibility of enhancing the recognition of nonlinguistic information by modifying the temporal envelope.

Traditional research about speaker identification by humans has focused on spectral cues based on speech production. The formant frequencies have been found to carry not only information about vowels but also information regarding speaker individuality [5]. Kitamura et al. indicated that speaker individuality exists mainly in the frequency bands higher than 2212 Hz of the speech spectral envelope [6]. The fundamental frequency contours are also shown to be important cues in speaker identification [7]. Generally, the speaker individualities related to fundamental frequency and spectral envelope can be thought of as results of the individual differences of vocal organs. Unfortunately, current CI devices cannot encode the spectral and fundamental frequency information of speech sufficiently for speaker identification.

As CI listeners using the temporal envelope of speech as a primary cue, Vongphoe and Zeng evaluated whether temporal cues are sufficient to support both speech recognition and speaker identification [3]. Their results showed a disassociation between speech and speaker recognition using primarily temporal cues: CI users performed well at vowel recognition but poorly at speaker recognition. On the other hand, Gonzalez and Oliver investigated speaker identification as a function of the number of channels in both noise and sin-wave vocoded speech as CI simulations [8]. The performance of speaker identification was shown to be poorer with fewer number of channels of noise-vocoded speech (NVS). However, Krull et al. showed that training resulted in improved identification of speakers in CI simulations [9]. Moreover, child CI users succeeded in differentiating their mothers' utterances from those of other people [10]. CI users's differentiation of speakers was facilitated by long-term familiarity. It is suggested that the temporal modulation information has possibility to be an effective cue for CI users to distinguish speakers.

In a previous study, the relative contributions of spectral and temporal cues in vocal emotion recognition for NVS is clarified by varying the the number of channels and upper limitation of envelope frequency systematically [11]. As the result, the temporal resolution of NVS affected the vocal emotion recognition significantly. Moreover, we examined word and speaker recognition using NVS while systematically varying the upper limit of the modulation frequency [12]. The results suggested that the temporal resolution of NVS should contribute to the speaker recognition. However, the role of temporal cues in speaker identification is still unknown.

This paper aims to clarify the role of temporal cues in speaker identification with NVS as a CI simulation. In the experiment, speaker identification was conducted by normal-hearing listeners as a function of the number of channels (4, 8, and 16) and upper limitation of envelope frequency (0, 0.5, 1, 2, 4, 8, 16, 32, and 64 Hz) in NVS. The number of channels and upper limitation of envelope frequency present the spectral and temporal resolutions of NVS separately. The experimental

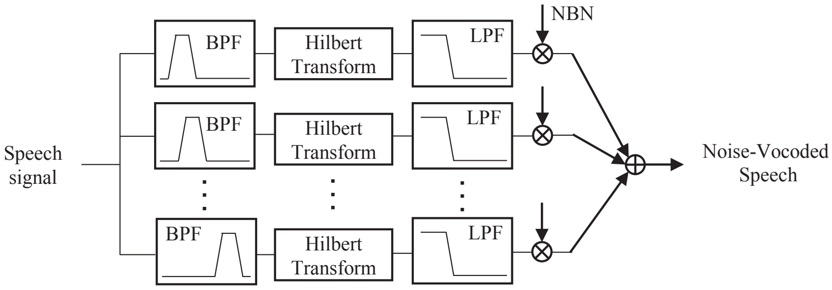


Figure 1: *Signal processing method for noise-vocoded speech. BPF: bandpass filter; LPF: low-pass filter; NBN: narrow-band noise*

paradigm used in this study can clarify the important modulation frequency band for speaker identification. The potential to improve speaker identification by enhance the temporal modulation cues is then discussed.

2. Speech data and signal processing

2.1. Speech data

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

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.

Kitamura et al. measured the perceptual similarity of speaker individualities of 20 female and 20 male Japanese speakers in ATR speech database set C [13]. 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 perceptual similarity of speakers is considerable to generate some undesirable bias in the XAB test. Therefore, in order to remove the impact of similarity, the speaker pairs of speech data used in this study have perceptual similarity closest to the average value of perceptual similarity (female: 1.87, and male: 1.99) measured by Kitamura et al. [13]. 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. 6 sentences of each speaker were used to generate the NVS stimuli.

Table 1: *Speaker pairs selected from ATR database and their average similarity index measured by Kitamura et al. [13]. Left and right halves show female and male speaker pairs, respectively.*

Speaker pair		Similarity	Speaker pair		Similarity
F407	F306	1.87	M509	M318	1.99
F611	F418	1.86	M603	M409	1.98
F606	F605	1.875	M508	M113	2.00
F720	F213	1.88	M519	M211	2.01
F709	F614	1.83	M520	M517	1.97

2.2. Signal processing

Figure 1 schematically illustrates a schematic diagram of the signal processing to generate NVS. First, to reduce the effect of the average intensity, the active speech levels of all speech signals were normalized to -26 dBov by using the P.56 speech voltmeter [14]. Speech signal was first divided into several frequency bands with a band-pass filterbank. The bandwidth and boundary frequencies of the band-pass filters (6th-order Butterworth Infinite Impulse Response (IIR) filter) were defined using $\mathrm{ERB_N}$ (Equivalent Rectangular Bandwidth) and $\mathrm{ERB_N}$-number scale [15]. The $\mathrm{ERB_N}$-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 $\mathrm{ERB_N}$-number. The relationship between $\mathrm{ERB_N}$-number and acoustic frequency is defined as follows:

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

where f is acoustic frequency in Hz. The boundary frequencies of the band-pass filters were defined from 3 to 35 $\mathrm{ERB_N}$-number with bandwidth as 2, 4, or 8 $\mathrm{ERB_N}$. Therefore, the numbers of channels of the band-pass filterbank were 16, 8, or 4. The number of channels presents the frequency resolution of NVS: higher frequency resolution is obtained with more number of channels.

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 determined the upper limit of envelope frequency that presents the temporal resolution of NVS. To investigate the role of temporal envelope cues for speaker identification, the conditions of the cut-off frequencies of the low-pass filter were 0.5, 1, 2, 4, 8, 16, 32, and 64 Hz. Moreover, there was an additional “0” Hz condition where only the direct current component of the Hilbert envelope was extracted.

Finally, the temporal envelope in each channel served to amplitude modulation with the band-limited noise 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. The NVS was widely used as a CI simulation, as the spectral cues of speech were reduced.

3. Experimental procedure

Nine native Japanese speakers (two female and seven male) 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. All the speaker pairs of A and B are shown in the Table 1.

A total of 3 different number of channels (4, 8, and 16) and 9 upper limits of envelope frequency (0, 0.5, 1, 2, 4, 8, 16, 32, and 64 Hz) created 18 NVS conditions. The original speech was also presented as a control condition. The participants were allowed to listen to each stimulus only once. Before the experiment, 10 stimuli were presented to the participants to familiarize participants with the CI simulation and the experimental environment. The stimuli used in the experiment were different from that used in the practice. The number of stimuli was 560 and all stimuli were presented totally randomized.

The experiment was conducted while the participants were in a sound-proof room. The sound pressure level of background noise was lower than 25.8 dB. The stimuli were simultaneously presented to both ears of a participant through a PC, audio interface (RME, Fireface UCX), and a set of headphones (SENNHEISER HDA 200). The sound pressure levels were calibrated to be the same among participants by using a head and torso simulator (B&K, type 4128) and sound level meter (B&K type 2231).

4. Results

Figure 2 shows the average value of speaker recognition rates, and the error bars indicate ±1 standard error of the mean. Under the original speech condition, the recognition rate was close to 95 %. Participants performed nearly perfectly in speaker identification with the original speech. The results of NVS stimuli showed that the performance of speaker identification improved as the upper limit of envelope frequency increased. The results for 4-band NVS were lower than 8 or 16-band NVS at some upper limits of envelope frequency. However, the performance was not obviously affected by the number of channels.

A repeated-measures analysis of variance (ANOVA) was conducted on the results with the number of channels and upper limit of envelope frequency as the factors. It is confirmed that there was a significant main effect of the upper limit of envelope frequency ($F(8,64) = 23.8631, p < 0.01$). However, there was no significant main effect of the number of bands ($F(2,16) = 3.3230, p = 0.29$) and there was also no significant interaction between the two factors ($F(16,128) = 1.1608, p = 0.16$). These results showed that the performance of speaker identification was significantly affected by the temporal resolution, which suggest that temporal modulation cues contribute to speaker identification. The performance was less sensitive to the spectral resolution, however, at least in the limited set of stimuli in the present study.

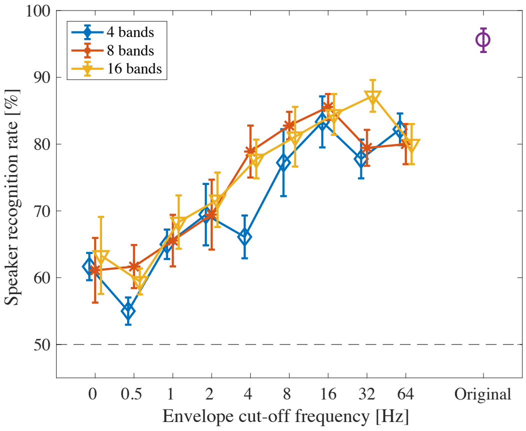


Figure 2: *Speaker recognition rates in all 27 NVS conditions and original speech condition. Error bars indicate* ±1 *standard error of the mean.*

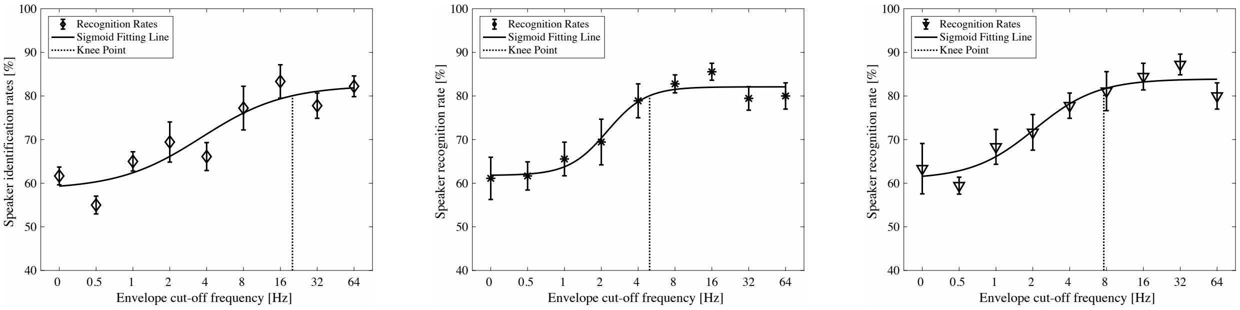
5. Discussion

5.1. Effect of spectral resolution

The speaker identification rates of 4-band NVS are lower in some conditions of the upper limit of envelope frequency. However, the number of channels did not affect the performance of speaker identification significantly. These results were different from the results of previous studies in which the performance was improved as the number of channels increased [3][8]. One difference between the present study and previous studies is that the upper limit of envelope frequencies in this study was lower. In previous studies, the cut-off frequencies of the low-pass filter were 500 Hz [3] and 160 or 400 Hz [8]. The modulation frequency bands between about 50 and 500 Hz are related to the periodicity information about fundamental frequency [16], which is not included in the stimuli used in the present study. One possible explanation may be that the temporal cues related to the periodicity information in higher modulation frequency bands are more sensitive to the number of channels. The main target of this study is to clarify the role of temporal cues in lower modulation frequency bands that include the information about variations of intensity, duration, attack, decay, and segmental cues of speech.

5.2. Effect of upper limit of envelope frequency

This study is intended to clarify the role of temporal modulation cues in speaker identification. Specifically, the important modulation frequency bands for speaker identification are investigated. To identify the important modulation frequency bands, a



(a) 4 bands

(b) 8 bands

(c) 16 bands

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

sigmoid function was used to fit the data of the experiment. The sigmoid function was mathematically defined as follows:

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

where x is the upper limit of envelope frequency and y is the percent-correct scores. The parameters $a, b, c,$ and d were calculated on the basis of the method of least squares. Moreover, the upper limit of envelope frequency at which 90% of the performance plateau was defined as a knee point. The results of fitting lines and knee points of each condition of the number of channels are shown in Fig. 3. The coefficients of determinations R^2 of the fitting results in 4, 8, and 16-band NVS were 0.86, 0.95, and 0.93.

The knee point of 4-band NVS was about 20.09 Hz which was higher than those of 8-band NVS (4.96 Hz) and 16-band NVS (7.60 Hz) . As the spectral cues provided by 4-band NVS was poor, participants may primarily use the temporal modulation cues to recognize the speaker rather than spectral cues. However, it still should be mentioned that there was no significant interaction between the number of channels and the upper limit of envelope frequency. More Xu and Pfingst measured both consonant and vowel recognition as a function of the number of channels (1 to 16) and upper limit of envelope frequency (1 to 512) [17]. The knee points of vowel recognition in different numbers of channels conditions are all below about 4 Hz. The knee points of consonant recognition are from 4 to 16 Hz, which are closer to the knee points for speaker identification in this study. Tachibana et al. conducted a experiment of NVS sentence recognition with various of upper limits of envelope frequency [18]. They found that an increase in the upper limit of envelope frequency from 4 to 8 Hz improved the correct response rate more that increasing the upper limit of envelope frequency from 8 to 16 Hz. Both studies showed that the duration and segmental cues included in such modulation frequency band below about 16 Hz are important in the perception of linguistic information. In this study, these duration and segmental cues of the temporal envelope are also suggested to be used in speaker identification. These segmental cues related to the rhythm, tempo, and the speaking style of the speaker which should be different with different speakers.

The results of this paper have shown that the temporal modulation cues contribute to speaker identification and that the temporal modulation information below about 20 Hz seems to be important. In the future, the modulation spectral features [19] related to speaker individuality and the effect of modifying such modulation spectral on speaker identification will be investigated. In a previous study, we confirmed that the vocal emotion of NVS can be converted by modifying the modulation spectrogram of temporal envelope [20]. Whether the speaker individuality information of NVS can be converted by modifying the modulation spectrogram should also be discussed further.

6. Summary

This study aimed to clarify the role of temporal cues in speaker identification with noise-vocoded speech (NVS) as a cochlear implant (CI) simulation. Speaker identification was conducted by normal-hearing listeners as a function of the number of channels (4, 8, and 16) and the upper limitation of envelope frequency (0, 0.5, 1, 2, 4, 8, 16, 32, and 64 Hz) in NVS. The result showed that speaker identification rates improved significantly as the upper limit of envelope frequency increased. However, the performance was not obviously affected by the number of channels. The modulation frequency bands below about 20 Hz were shown to be important in speaker identification with 4-band NVS. In conclusion, it is suggested that temporal modulation cues contribute to speaker identification and have the potential to improve speaker identification if enhanced. It is important to understand not only which parts but also exactly what kinds of features of temporal envelope have possibility to be important cues for speaker identification.

7. Acknowledgements

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.

8. 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, .

[3] M. Vongphoe, and F. G. Zeng, "Speaker recognition with temporal cues in acoustic and electric hearing," *Journal of the Acoustical Society of America*, vol. 118, no. 2, pp. 1055–1061, 2005.

[4] 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, 2004.

[5] R. E. Remez, J. M. Fellowes, and P. E. Rubin, "Talker identification based on phonetic information," *Journal of Experimental Psychology: Human Perception and Performance*, vol. 23, pp. 651–666, 1997.

[6] T. Kitamura, and M. Akagi, "Speaker individualities in speech spectral envelopes," *Journal of Acoustical Society of Japan (E)*, vol. 16, no. 5, pp. 283–289, 1995.

[7] M. Akagi, and T. Ienaga, "Speaker individuality in fundamental frequency contours and its control," *Journal of Acoustical Society of Japan (E)*, vol. 18, no. 2, pp. 73–80, 1997.

[8] J. Gonzalez, and J. Oliver, "Gender and speaker identification as a function of the number of channels in spectrally reduced speech," *Journal of the Acoustical Society of America*, vol. 118, no. 1, pp. 461–470, 2005.

[9] V. Krull, X. Luo and K. Kirk, "Talker-identification training using simulations of binaurally combined electric and acoustic hearing: generalization to speech and emotion recognition," *Journal of the Acoustical Society of America*, vol. 131, no. 4, pp. 3069–3078, 2012.

[10] T. Vongpaisal, S. E. Trehub, E. G. Schellenberg, P. Lieshout, and B. C. Papsin, "Children with cochlear implants recognize their mother's voice," *Ear and Hearing*, vol. 31, no. 4, pp. 555–566, 2010.

[11] 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*, 2017. (in press)

[12] Z. Zhu, Y. Nishino, R. Miyauchi, and M. Unoki, "Study on linguistic information and speaker individuality contained in temporal envelope of speech," *Acoustical Science & Technology*, vol. 37, no. 5, pp. 258–261, 2017.

[13] 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.

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

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

[16] S. Rosen, "Temporal Information in Speech: Acoustic, Auditory and Linguistic Aspects," *Philosophical Transactions: Biological Sciences*, vol. 336, no. 1278, pp. 367–373, 1992.

[17] L. Xu, and P. Pfingst, "Spectral and temporal cues for speech recognition: Implications for auditory prostheses," *Hearing Research*, vol. 242, pp. 132–140, 2008.

[18] R. Tachibana, Y. Sasaki, and H. Riquimaroux, "Relative contributions of spectral and temporal resolutions to the perception of syllables, words, and sentences in noise-vocoded speech," *Acoustical Science and Technology*, vol. 34, no. 4, pp. 263–270, 2013.

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

[20] Z. Zhu, R. Miyauchi, Y. Araki and M. Unoki, "Feasibility of Vocal Emotion Conversion on Modulation Spectrogram for Simulated Cochlear Implants," *EUSIPCO2017*, 2017. (in press)