

ACOUSTICAL LETTER

Study on linguistic information and speaker individuality contained in temporal envelope of speech

Zhi Zhu*, Yasutaka Nishino†, Ryota Miyauchi‡ and Masashi Unoki§

School of Information Science, Japan Advanced Institute of Science and Technology, 1–1 Asahidai, Nomi, Ishikawa, 923–1292 Japan

(*Received 15 April 2016, Accepted for publication 17 May 2016*)

Keywords: Noise-vocoded speech, Temporal envelope, Speech perception, Speaker individuality
PACS number: 43.71.Bp [doi:10.1250/ast.37.258]

1. Introduction

Speech signals can be represented as a sum of amplitude modulated frequency bands. This sum can also be regarded as a temporal amplitude envelope with carrier (temporal fine structure). A study using noise-vocoded speech showed that the temporal amplitude envelope of speech plays an important role in the perception of linguistic information [1]. Human speech includes not only linguistic information, but also nonlinguistic information such as speaker individuality. Kazama *et al.* [2] reported that it is possible to express speaker individuality information in connection with the difference in narrow-band temporal envelope correlation matrices. They confirmed that the temporal amplitude envelope of speech contains not only linguistic information, but also speaker individuality information.

Modern psychophysical models of temporal modulation processing suggest that the temporal amplitude envelope is processed by a modulation filterbank [3]. Therefore, in the auditory system, modulation frequency analysis should be used to extract the linguistic and nonlinguistic information from the temporal amplitude envelope of speech. The modulation frequency bands in the range between 4 and 8 Hz are reported to be important for the perception of linguistic information in noise-vocoded speech [4]. For speaker individuality, the modulation frequency bands ranging from 3 to 15 Hz have been shown to be robust enough for a machine to identify the speaker [5]. However, whether these modulation frequency bands account for the perception of speaker individuality remains unknown.

The purpose of the present study is to clarify which modulation frequency bands can contribute to the perception of linguistic and speaker individuality information in the temporal amplitude envelope of speech. Two separate experiments were conducted to investigate the effect of controlling the highest modulation frequency of the speech signal on the recognition of words and speaker by using noise-vocoded speech. The originality of this study is that it investigated important cues for perception of linguistic and speaker individuality information by systematically controlling the modulation component of the temporal amplitude envelope of speech.

*e-mail: zhuzhi@jaist.ac.jp
†e-mail: yasutaka.westfield@jaist.ac.jp
‡e-mail: ryota@jaist.ac.jp
§e-mail: unoki@jaist.ac.jp

2. Experiment 1: Word intelligibility tests

Two experiments were carried out in this paper. In the first experiment, accurate modulation frequency bands related to speech perception were confirmed. In order to focus on temporal envelope cues, noise-vocoded speech was used as the stimuli. Noise-vocoded speech was generated by transforming the temporal fine structure of speech to noise.

2.1. Test materials

The test materials in this experiment were chosen from the Familiarity-controlled Word-lists (FW03) [6]. The speech signal of the word list had a sampling frequency of 48 kHz. The words were composed of four morae. To eliminate the effect of familiarity on recognizing linguistic information, the familiarities of all words was at the lowest rate, i.e. between 1.0 and 2.5.

2.2. Signal generation

Figure 1 illustrates a schematic diagram of the noise-vocoder method used to generate the stimuli. Speech signal was first divided into several frequency bands. The bandwidth and boundary frequencies of the band-pass filter (sixth-order Butterworth IIR filter) were defined using ERB_N (equivalent rectangular bandwidth) and ERB_N-number scale [7]. 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 according to the ERB_N-number. However, if the bandwidth is narrow such as only 1 ERB_N, the linguistic information contained in the frequency domain may be used as robust cues rather than the temporal envelope. To provide suitable temporal envelopes as robust cues, the relative wide-bands according to ERB_N-numbers were used under the assumption that the signal of adjacent channels should be co-modulated. Thus, the boundary frequencies of the band-pass filter were defined from 2 to 32 ERB_N-number with 3 ERB_N.

The temporal envelope of the signal was extracted by making the Hilbert transformation and performing low-pass filtering (second-order Butterworth IIR filter). To investigate the upper limit of modulation frequency band relating to speech perception, the cut-off frequency of the low-pass filter was varied from 1 to 30 Hz in increments of 1 Hz. The temporal amplitude envelope of the signal in each channel served to amplitude modulate (AM) the band-limited noise which was generated by band-pass filtering white noise with

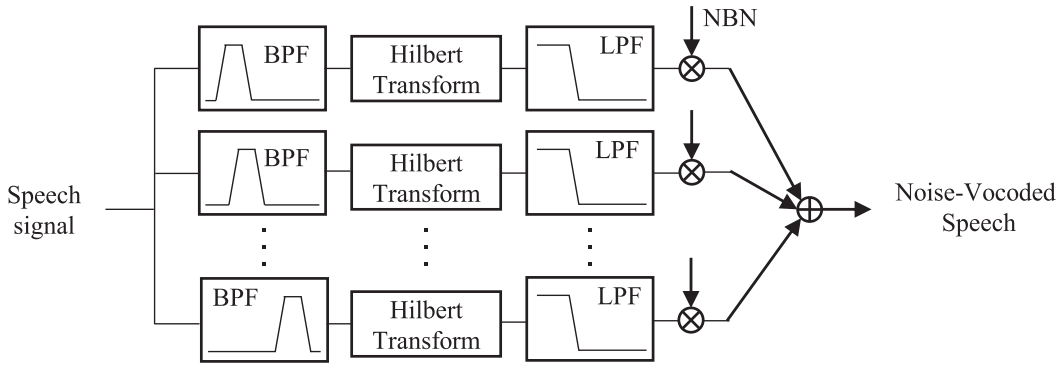

Fig. 1 Schematic diagram of the noise-vocoder method used to generate the stimuli (BPF: band-pass filter, LPF: low-pass filter, and NBN: narrow-band noise).

the same boundary frequency. Finally, all AM band-limited noises were summed to generate the noise-vocoded speech.

2.3. Procedure

Four male native Japanese speakers 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 8,000 Hz).

There were 180 trials in a session (30 cut-off frequencies $\times$ 6 words). The words for all trials were different. The trials were conducted in random order. Participants could replay the words as any time as they wanted. The task of the participants was to input a word of four morae as they understood it by using a keyboard. The participants had a break once every 45 trials. 45 practice trials were carried out before the actual trials.

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 headphone (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).

2.4. Results

The number of correct mora was counted at each cut-off frequency of the low-pass filter. As only four-mora words were used, the upper limit of the number of correct mora should be four. Figure 2 shows the results of averaged correct number of mora related to the cutoff frequency of the low-pass filter. The results of the ANOVA (analysis of variance) indicated there was a significant main effect of the cut-off frequency of the low-pass filter ($F(29, 87) = 18.98$, $p < 0.05$). Moreover, as a result of a *post hoc* Turkey's HSD (honestly significant different) test, there were significant differences between the conditions of 1 and 3–30 Hz. There were also significant differences between the conditions of 2–4 Hz and 5–30 Hz. When the cut-off frequency was more than or equal to 5 Hz, the results of averaged correct number of mora were almost the same. However, when the cut-off frequency was less than 5 Hz, the average correct number of mora suddenly decreased and was almost 0 at 1 Hz. This result indicates that the modulation frequency components of less than 5 Hz are important for accurately recognizing linguistic information.

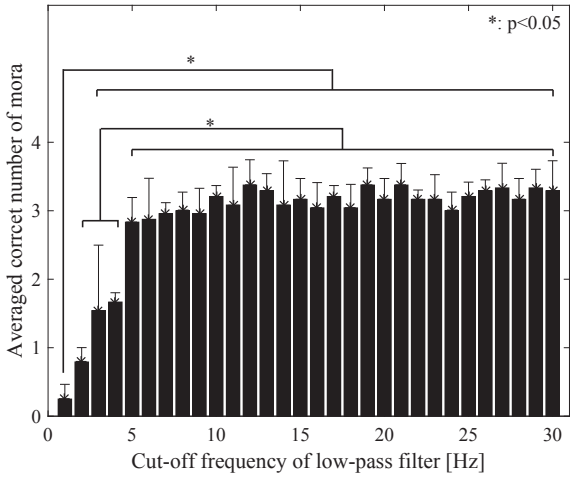

Fig. 2 Average correct number of mora related to cutoff frequency of low-pass filter.

3. Experiment 2: Speaker identification

The second experiment investigated the effect of controlling the highest modulation frequency of the speech signal on speaker identification.

3.1. Test materials

The test materials in this experiment were chosen from the Advanced Telecommunications Research (ATR) speech database set C. Speech data of 10 female speakers, wherein each speaker spoke 10 sentences were used. The length of each sentence ranged from 4 to 7 seconds. All speech signals had a sampling frequency of 20 kHz.

3.2. Signal generation

The signal generation method was the same as in Experiment 1, with the exception that the boundary frequencies of the band-pass filter and low-pass filter were changed. In the experiment 1, it was found that if the bandwidth is too narrow, the frequency domain may be used as robust cues rather than the temporal envelope. For the same reason, to provide suitable temporal envelopes as robust cues for speaker identification, the bandwidth was set to 2 $\mathrm{ERB_N}$. Moreover, it

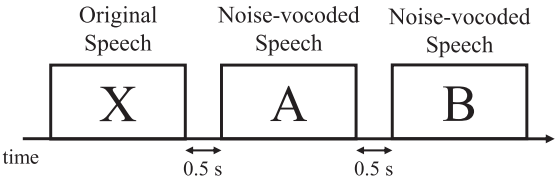

Fig. 3 XAB test procedure in experiment 2.

is known that speaker individuality information exists mainly in the frequency bands higher than 22 ERB_N-number [8]. Therefore, in this experiment, the upper limit of the frequency band extended to 35 ERB_N-number. The boundary frequencies of the band-pass filter were defined from 1 to 35 ERB_N-number with 2 ERB_N. To investigated the modulation frequency bands accounting for the speaker identification, the cut-off frequency of the low-pass filter was set to 1, 2, 4, 8, 16, 32, or 64 Hz.

3.3. Procedure

Six native Japanese speakers (five males and one female) participated in this experiment. All participants had normal hearing (the hearing losses of the participants were below 12 dB in the octave frequency range from 125 to 8,000 Hz).

This experiment was carried out by using XAB test procedure, which is illustrated in Fig. 3. One trial consisted of three different speech signals (X, A, and B). The contents of stimuli X, A, and B are shown below.

X: Original speech signal
A: Noise-vocoded speech with the same speaker of X
B: Noise-vocoded speech with a different speaker of X

Participants were asked to select which speaker, A or B, was similar to the speaker of X. Stimuli were presented in both XAB and XBA orders to counterbalance any effects due to the order of presentation. The number of stimuli was 140, and the participants were allowed to listen to each stimulus only once. The equipment of this experiment was the same as the experiment 1.

3.4. Results

Figure 4 shows the mean value and standard deviation of the speaker identification rate. The speaker identification rate increased as the cut-off frequency of the low-pass filter increased from 1 to 16 Hz. The results of the ANOVA indicated that there was a significant effect of the cut-off frequency ($F(6, 30) = 8.309$, $p < 0.05$) on the speaker identification. In addition, as a result of a *post hoc* Turkey's test, there were significant differences between the conditions of 1 Hz and 8–64 Hz. The results suggest that the modulation frequency components less than about 16 Hz should be important for speaker identification. The upper limit of the modulation frequency band that contains individuality information should be between 8 to 16 Hz.

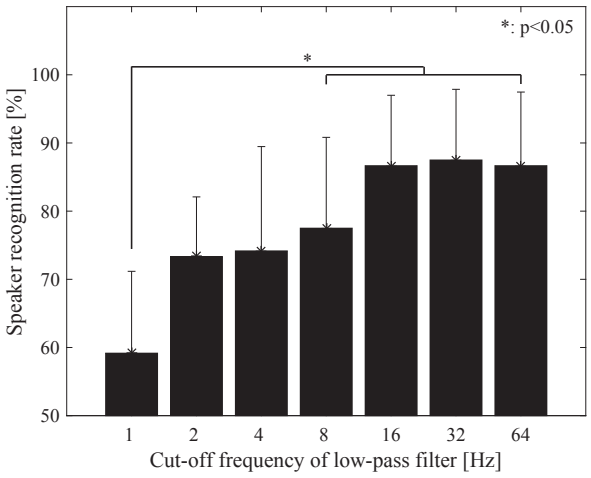

Fig. 4 Results of speaker recognition rate in the experiment 2.

4. Discussion

The prosodic structure of Japanese is based on mora. Mora is repeated with a steady rhythm. This means that the cycle of morae is important for Japanese. The total length of the speech signals of four morae used in the experiment was about 1,000 ms; thus, the duration of one mora was about 250 ms. If this duration is one cycle of morae repetition, the outline of the primary shape of the amplitude modulation based on the moraic syllable structure should be represented by 4 Hz modulation. Therefore, the moraic syllable structure will be destroyed when the cut-off frequency of the low-pass is lower than 5 Hz. The results of experiment 1 showed that the linguistic information decreased when the maximum modulation frequency was less than 5 Hz. This suggests that the shape of the temporal envelope of the moraic syllable structure is an important factor in recognizing linguistic information.

The results of the speaker identification experiment in Fig. 4 showed that it is difficult to recognize the speaker when the cut-off frequency of the low-pass filter is 1 Hz. It is confirmed that the modulation components at least beyond 1 Hz does contribute to the perception of speaker individuality information. Previous work showed that the modulation frequency band ranging from 3 to 15 Hz can be used by a machine to identify the speaker [5]. That means there is speaker individuality information in that modulation frequency band. The results in the experiment 2 suggest that this speaker individuality information may also be used by humans in speaker identification. Different from the perception of linguistic information, higher variations of temporal envelope are important for speaker identification.

5. Conclusions

In this study, the effect of controlling the highest modulation frequency of noise-vocoded speech on the recognition of words and speaker were investigated to clarity the modulation frequency bands related to the perception of linguistic and speaker individuality information. The highest modulation frequency of the speech signal was controlled by low-pass filtering the temporal amplitude envelope of the speech. The results of word intelligibility tests showed that the average correct number of morae decreased when the highest modulation frequency was less than 5 Hz. This suggests that the shape of the temporal envelope of the moraic syllable structure is an important factor in recognizing linguistic information. The results of the speaker identification experi-

ment showed that the modulation components less than about 8 or 16 Hz should contribute to the perception of speaker individuality information. Different from the perception of linguistic information, higher variations of the temporal amplitude envelope are important for speaker identification.

Acknowledgments

This work was supported by a Grant-in-Aid for Scientific Research (A) (No. 25240026), Young Scientists (A) (No. 24683026) and the Fostering Information Communications Technology (ICT) Global Leader Program of the Japan Advanced Institute of Science and Technology (JAIST). This work was also supported by an A3 fore-sight program made available by the Japan Society for the Promotion of Science (JSPS) and Mitani Foundation for Research and Development.

References

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

[2] M. Kazama, M. Tohyama and Y. Yamasaki, "Speaker characteristics represented by narrow-band temporal-envelope correlation matrices," *IEICE Trans.*, **J92-A**, 205–215 (2009) (in Japanese).

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

[4] R. O. 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," *Acoust. Sci. & Tech.*, **34**, 263–270 (2013).

[5] T. H. Falk and W. Y. Chan, "Modulation spectral features for robust far-field speaker identification," *IEEE Trans. Audio Speech Lang. Process.*, **18**, 90–100 (2010).

[6] S. Amano, K. Kondo, Y. Suzuki and S. Sakamoto, "Speech data set for word intelligibility test based on word familiarity (FW03)," *NII Speech Resources Consortium* (2006).

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

[8] T. Kitamura and M. Akagi, "Speaker individualities in speech spectral envelopes," *J. Acoust. Soc. Jpn.* (*E*), **16**, 283–289 (1995).