PROCEEDINGS of the
23rd International Congress on Acoustics

9 to 13 September 2019 in Aachen, Germany

How the temporal amplitude envelope of speech contributes to urgency perception

Masashi UNOKI; Miho KAWAMURA; Maori KOBAYASHI; Shunsuke KIDANI; Masato AKAGI

School of Information Science, Japan Advanced Institute of Science and Technology, Japan

ABSTRACT

Speech communicates non-linguistic and para-linguistic information as well as linguistic information. Our previous studies on noise-vocoded speech (NVS) showed that temporal modulation cues provided by the temporal amplitude envelope (TAE) affect how vocal emotion and speaker individuality are perceived. However, it is still unclear if temporal modulation cues affect the perception of urgency. Here, we experimentally investigated whether the TAE of speech affects the perception of para-linguistic information, particularly urgency. We compared NVS in which the TAEs were identical to those of the original speech and NVS in which the TAEs had undergone low-pass or high-pass filtering. Urgency scales were derived from a paired comparison of the results and used to investigate the relationship between the temporal modulation components and urgency perception. Our findings were (1) the degree of urgency of the NVS stimuli was perceived as being similar to that of the original; (2) temporal modulation components of NVS upwards of 6 Hz were significant cues for urgency perception; and (3) temporal modulation components of NVS downwards of 8 Hz were significant cues for urgency perception. The results suggest that temporal modulation cues in the TAE play an important role in urgency perception.

Keywords: Urgency perception, noise-vocoded speech, temporal amplitude envelope, modulation perception

1. INTRODUCTION

Speech is the most important and natural way for people to express linguistic, nonlinguistic, and para-linguistic information. In particular, in addition to using linguistic information to convey messages, people use nonlinguistic information such as emotion and speaker individuality to enrich their speech. In addition, para-linguistic information such as emphasis and intention by people's actions is also used for richer speech. Important features related to linguistic, nonlinguistic, and para-linguistic information are redundantly contained in the speech signals. Therefore, people can easily and correctly recognize this information in daily life even if some features are obscured due to noise and reverberation.

The temporal amplitude envelope (TAE) of speech has been proved to be an important cue for speech perception from the studies using noise-vocoded speech (NVS) [1-4]. NVS is generated by replacing the carriers with band-limited noise, so the spectral cue is reduced dramatically and the temporal cue is preserved. Shannon et al. showed that the presentation of a dynamic temporal pattern in only a few broad spectral regions is sufficient for listeners to recognize linguistic information [1]. The modulation frequency bands from 4 to 16 Hz have been shown to be important regions in speech recognition [5]. Therefore, people can successfully perceive linguistic information using the TAE of speech signals as a primary cue.

In our previous study, the relative contributions of spectral and temporal cues in vocal emotion recognition and speaker individuality for NVS were clarified by systematically varying the number of channels and upper limitation of envelope frequency [6, 7]. The results demonstrated that temporal resolution contributes to recognizing both speaker and vocal emotion. Therefore, temporal cues lower than the modulation frequency of 8 Hz are found to be important for perceiving not only linguistic information but also speaker individuality and vocal emotion. However, the role of temporal cues in para-linguistic information is still unknown.

This paper aims to investigate the effect of TAE of speech on urgency perception as non-linguistic

information. Thus, three experiments are conducted (1) to investigate whether or not the TAE of speech is a cue of urgency perception by using original and NVS stimuli individually; (2) to investigate whether or not the TAE of speech is a cue of urgency perception by using these stimuli simultaneously; and (3) to investigate which temporal modulation component in the TEA of speech is a cue of urgency perception by restricting the upper or lower modulation frequency of TAE.

2. NOISE-VOCODED SPEECH

2.1 Speech data

In this paper, real speech data of evacuation announcements by a professional male speaker that Kobayashi and Akagi [8] used are used as original stimuli. Content of these stimuli was a Japanese evacuation announcement: "*Ima sugu nigetekudasai*" ("Please evacuate now"). In this study, the evacuation announcement was made in four different styles to investigate perception of urgency for these stimuli. These were labeled "A," "B," "C," and "D." The corresponding NVS stimuli were also labeled "a," "b," "c," and "d."

2.2 Speech synthesis based on NVS

Figure 1 schematically illustrates 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 a P.56 speech voltmeter [7]. Speech signals were then divided into several frequency bands by using a band-pass filterbank. The bandwidth and boundary frequencies of the band-pass filter (BPF) (6th-order Butterworth infinite impulse response (IIR) filter) were defined using Equivalent Rectangular Bandwidth (ERB_N) and ERB_N-number scale [9]. The ERB_N-number scale is comparable to a scale of distance along the basilar membrane, so the frequency resolution of the auditory system can be faithfully replicated by dividing frequency bands in accordance with the ERB_N-number. The relationship between ERB_N-number and acoustic frequency is defined as follows:

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

where f is acoustic frequency in Hz. The boundary frequencies of the BPFs were defined from 3 to 35 ERB_N-number with bandwidth as 2 ERB_N. Therefore, the band-pass filterbank had 16 channels.

Then, the TAE of the output signal from each BPF was extracted by using the Hilbert transformation and performing a low-pass filter (LPF) (2nd-order Butterworth IIR filter). The cut-off frequency of the LPF determined the upper limit of modulation frequency of 64 Hz. The upper limit of modulation frequency relates to the temporal resolution that higher temporal resolution will be obtained with a higher upper limit of the modulation frequency.

Finally, the TAE in each channel served to modulate amplitude with the narrow band-limited noise (NBN) that was generated by band-pass filtering white noise at the same boundary frequency. All amplitude-modulated NBNs were summed to generate the NVS stimulus.

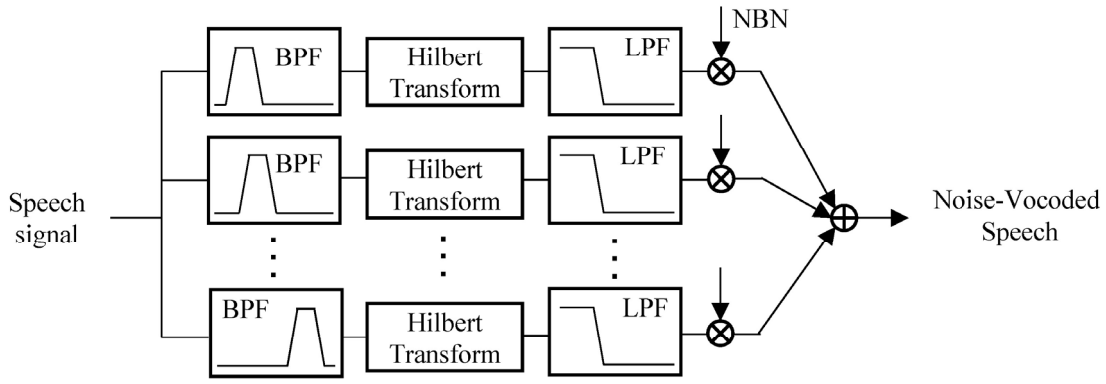

Figure 1 – Schematic diagram of noise-vocoder method used to generate stimuli.

3. EXPERIMENT I: Urgency perception on temporal envelope of speech

Psychoacoustical Experiment I was conducted to investigate whether temporal modulation cues provided by the TAE affect how urgency of speech is perceived by using NVS.

3.1 Stimuli, Participants, and Procedure

Each four stimuli of original evacuation announcements and the corresponding NVSs were used. Two stimuli were paired randomly to make a paired comparison experiment. The total number of paired stimuli was 12. Silence between the first and second stimuli was 0.5 s. Total execution time was about 10 min.

Ten native Japanese speakers (three females and seven males, aged 22 to 25 years old) participated in this experiment. All participants had normal hearing (hearing losses of the participants were below the hearing level of 12 dB in the frequency range from 125 to 8000 Hz).

The experiment was conducted while the participants were in a soundproof room. 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 for all participants by using a head and torso simulator (B&K, type 4128) and sound level meter (B&K type 2231).

This experiment was carried out using Scheffe's method of paired comparison to evaluate the degree of urgency of stimuli. Participants were asked to evaluate whether or not the first stimulus was more urgent than the second one by using a five-grade evaluation measure ($+2$: the first stimulus is rather more urgent, $+1$: the first stimulus is somewhat more urgent, 0: even, -1: the second stimulus is somewhat more urgent, and -2: the second stimulus is rather more urgent).

3.2 Results

Figure 2 shows the degrees of urgency for original stimuli derived from the results of Experiment 1 by using Sheffe's method of paired comparison. Figure 3 also shows the degrees of urgency for NVS stimuli. The horizontal axis in both figures shows the urgency scale, where the positive scale indicates a higher degree of urgency. These results reveal the order of degrees of urgency from highest to lowest to be C, D, B, and A for the original speech and c, d, b, and a for the NVS stimuli. In addition, the results of ANOVA revealed a significant main effect ($F(3,77) = 34.42, p < 0.01$) of urgency perception of original stimuli. There were significant differences between all pairs except A and B ($p < 0.01$). The results of ANOVA revealed a significant main effect ($F(3,77) = 125.8, p < 0.01$) of urgency perception of NVS stimuli. There were significant differences between all pairs ($p < 0.01$).

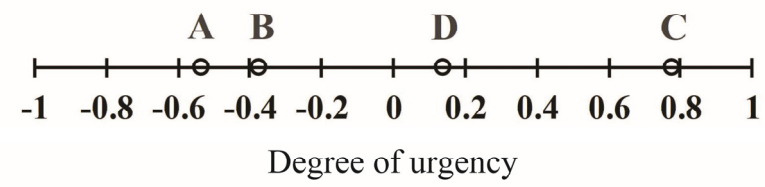

Figure 2 – Urgency perception of original stimuli

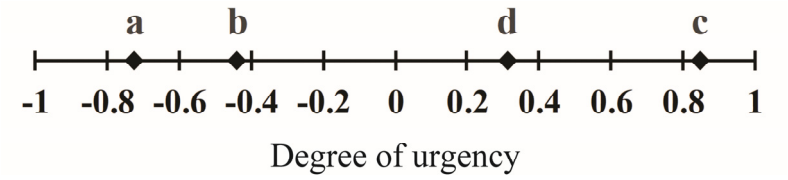

Figure 3 – Urgency perception of NVS stimuli

3.3 Considerations

In these results, degrees of urgency were ordered from C (c), D (d), B (b), to A (a) so that the order of urgency perception of the NVS stimuli is the same as that of the original stimuli. Therefore, temporal modulation cues provided by the TAE were found to affect perception of urgency. By comparing the results in Figs. 2 and 3, degrees of urgency for both stimuli were found to be different.

4. EXPERIMENT II: Effect on urgency perception of NVS stimuli

Psychoacoustical Experiment II was conducted to investigate whether or not urgency perception of NVS stimuli is consistent with that of original evacuation announcements.

4.1 Stimuli, Participants, and Procedure

Eight stimuli were used: four original evacuation announcements and the four corresponding NVSs. Two stimuli were paired randomly to make a paired comparison psychoacoustical experiment. The total number of paired stimuli was 56. Silence between the first and second stimuli was 0.5 s. Total execution time was about 20 min.

Ten native Japanese speakers (two females and eight males, aged 22 to 25 years old) participated in this experiment. All participants had normal hearing (hearing losses of the participants were below the hearing level of 12 dB in the frequency range from 125 to 8000 Hz).

The experiment was conducted while the participants were in a soundproof room. The presented stimuli were the same as in Experiment I. This experiment was carried out using Scheffe's method of paired comparison to evaluate the degree of urgency of stimuli. The procedure was the same as in Experiment I.

4.2 Results

Figure 4 shows the degrees of urgency for the mixture of original and NVS stimuli derived from the results of Experiment II by using Sheffe's method of paired comparison. The horizontal axis is the same as in Figures 2 and 3. These results reveal the order of degrees of urgency from highest to lowest to be c, C, d, D, b, B, a, and A. In addition, the results of ANOVA revealed a significant main effect of urgency perception of these stimuli ($F(3, 459) = 100.8, p < 0.01$). There were significant differences between all pairs except a & B, a & b, D & d, D & C, d & C, d & c, and C & c ($p<0.01$).

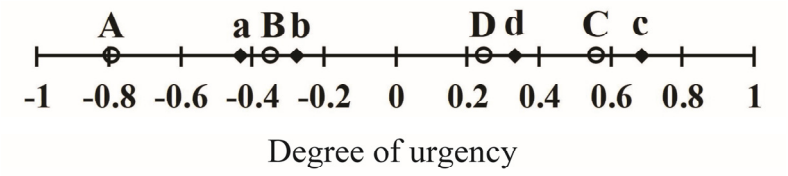

Figure 4 – Urgency perception of all stimuli

4.3 Considerations

From the results, degrees of urgency were ordered from c, C, d, D, b, B, a, and A. By comparing pairs of the same conditions such as a and A, the degree of urgency of NVS was found to be higher than that of original evacuation announcement. The order of urgency perception was also found to be consistent with the order of urgency perception derived from the results of Experiment I.

5. EXPERIMENT III: Effect on restricting modulation frequency

Psychoacoustical Experiment III was conducted to investigate which temporal modulation component in the TAE of speech is a cue of urgency perception. We compared NVS in which the TAEs were identical to those of the original speech and NVS in which the TAEs had undergone LPF or HPF.

5.1 Stimuli, Participants, and Procedure

Stimuli of the four corresponding NVSs were used. Seven conditions of the cut-off frequency (2, 4, 6, 8, 12, 16, and 32 Hz) were used for the four NVS stimuli in cases of the use of LPF and HPF. Two stimuli in each experiment (the use of LPF or HPF) were paired randomly to make a paired comparison experiment. The total number of paired stimuli in each experiment was 756. Silence between the first and second stimuli was 0.5 s. All paired stimuli of 756 were split into 14 sections in which each section has 54 judgements. Rest time between the first and last 7 sections was over 90 min to give sufficient rest time to all participants. Thus, total execution time including rest time in each experiment was about 180 min.

Ten native Japanese speakers (three females and seven males, aged 22 to 25 years old) participated in the two experiments. All participants had normal hearing (hearing losses of the participants were below the hearing level of 12 dB in the frequency range from 125 to 8000 Hz).

The experiment was conducted while the participants were in a soundproof room. The presented stimuli were the same as in Experiment I. This experiment was carried out using Scheffe's method of paired comparison to evaluate the degree of urgency of stimuli. The procedure was the same as in Experiment I.

5.2 Results

Figure 5 shows the degree of urgency for NVS stimuli under LPF conditions that are derived from the results of Experiment III by using Sheffe's method of paired comparison. The horizontal axis shows the cut-off frequency of the LPF, and the vertical axis shows the urgency scale, in which the positive scale indicates a higher degree of urgency. These results reveal that the orders of degree of urgency of stimuli c and d decrease as the cut-off frequency decreases. The results of ANOVA revealed a significant main effect of urgency perception of these stimuli ($F(24, \infty) = 200.1,\ p < 0.01$). There were significant differences between all stimuli in c and d, in which the cut-off frequencies are from 4 to 32 Hz. ($p < 0.01$). In the case of the cut-off frequency of 2 Hz, there were significant differences among stimulus conditions except b & d and d & c ($p < 0.01$).

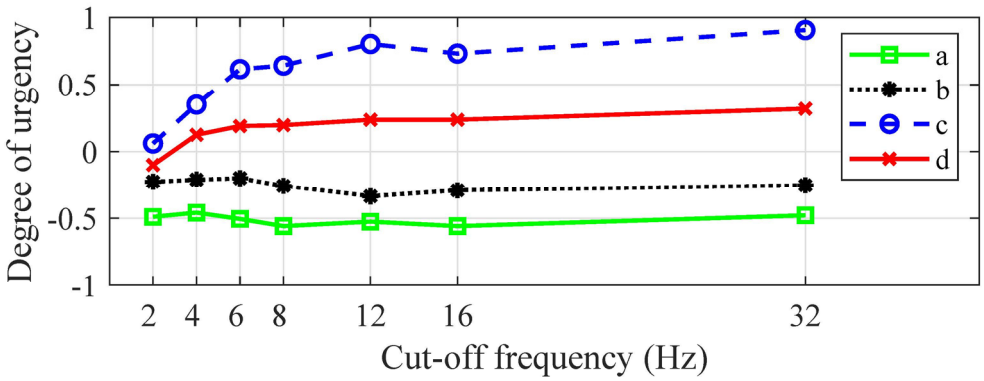

Figure 5 – Urgency perception of NVS stimuli in which upper modulation frequency of the temporal amplitude envelope was restricted by high-pass filtering.

Figure 6 shows the degree of urgency for NVS stimuli under LPF conditions derived from the results of Experiment III by using Sheffe's method of paired comparison. The horizontal and vertical axes are the same as in Figure 5. These results reveal that the orders of degree of urgency of stimuli c and d decrease as the cut-off frequency increases. The results of ANOVA revealed a significant main effect of urgency perception of these stimuli ($F(24, \infty) = 1.517,\ p < 0.01$). There were no significant differences between all stimuli in c and d with regard to the cut-off frequency, although there were significant differences between all stimuli in b, in which the cut-off frequencies are from 2 to 8 Hz. ($p < 0.01$). In the case of the cut-off frequencies of 8 and 12 Hz, there were significant differences among stimulus conditions c & d ($p < 0.01$).

5.3 Considerations

The order of urgency perception among the four stimuli with regard to the cut-off frequency of the LPF in Fig. 5 is consistent with that of HPF in Fig. 6. Fig. 5 shows a significant difference between the cut-off frequencies of 4 and 6 Hz. Thus, the temporal modulation components of NVS upwards of 6 Hz were significant cues for urgency perception from the results of Experiment III under LPF conditions. Fig. 6 shows a significant difference between the cut-off frequencies of 8 and 12 Hz. Thus, the temporal modulation components of NVS downwards of 8 Hz were significant cues for urgency perception. In summary, these results suggest that that temporal modulation cues in the TAE from 6 to 8 Hz play an important role in urgency perception.

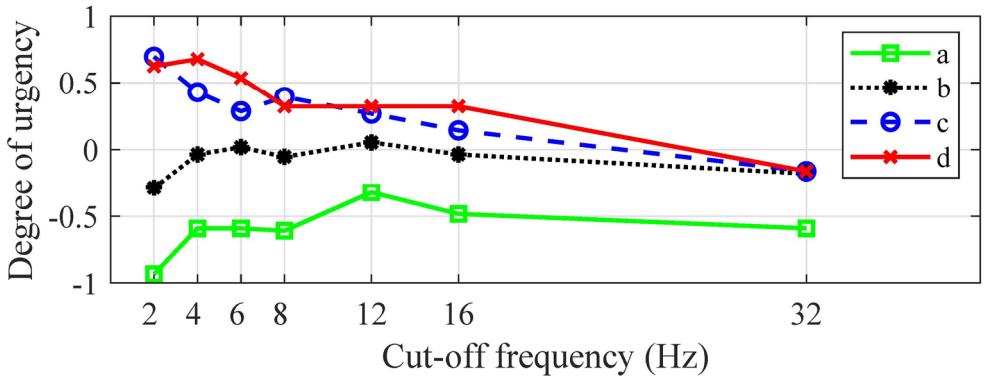

Figure 6 – Urgency perception of NVS stimuli in which lower modulation frequency of the temporal amplitude envelope was restricted by high-pass filtering.

6. CONCLUSIONS

In this paper, we experimentally investigated whether the temporal amplitude envelope (TAE) of speech affects the perception of urgency. We compared noise vocoded-speech (NVS) in which the TAEs were identical to those of the original speech and NVS in which the TAEs had undergone low-pass or high-pass filtering. Urgency scales were derived from a paired comparison of the results and used to investigate the relationship between the temporal modulation components and urgency perception. Urgency scales were derived from a paired comparison of the results and used to investigate the relationship between the temporal modulation components and urgency perception. Our findings were (1) the degree of urgency of the NVS stimuli was perceived as being similar to that of the original; (2) temporal modulation components of NVS upwards of 6 Hz were significant cues for urgency perception; and (3) temporal modulation components of NVS downwards of 8 Hz were significant cues for urgency perception. The results suggest that temporal modulation cues in the TAE play an important role in urgency perception.

ACKNOWLEDGEMENTS

This work was supported by a Grant in Aid for Innovative Areas (No. 16H01669, No. 18H05004) from MEXT, Japan. This work was also supported by SECOM Science and Technology foundation and JST-Mirai Program (Grant Number: JPMJMI18D1).

REFERENCES

1. 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.
2. 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., vol. 34, pp. 263-270, 2013.
3. P. C. Loizou, M. Dorman, and Z. Tu, On the number of channels needed to understand speech, J. Acoust. Soc. Am., vol. 106, pp. 2097-2103, 1999.
4. L. Xu, and B. E. Pfingst, Spectral and temporal cues for speech recognition: Implications for auditory prostheses, Hear. Res., vol. 242, pp. 132-140, 2008.
5. R. Drullman, J. Festen and R. Plomp, Effect of reducing slow temporal modulations on speech reception, J. Acoust. Soc. Am., vol. 95, no. 5, pp. 2670–2680, 1994.
6. Z. Zhu, Y. Nishino, R. Miyauchi, 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.
7. Z. Zhu, R. Miyauchi, Y. Araki, and M. Unoki, Contributions of temporal cue on the perception of speaker individuality and vocal emotion for noise-vocoded speech, Acoustical Science and Technology, vol. 39, no. 3, pp. 234–242, 2018.
8. M. Kobayashi and M. Akagi, Psychological evaluation of evacuation announcements, J. Acoust. Soc. Jpn. vol. 74, no. 12, pp. 633-640, 2018 (in Japanese).
9. B. C. J. Moore, An Introduction to the Psychology of Hearing, sixth edition, Brill Academic Pub., 2013.