applied sciences

MDPI

Article

Contributions of Temporal Modulation Cues in Temporal Amplitude Envelope of Speech to Urgency Perception

Masashi Unoki [1,*], Miho Kawamura [1], Maori Kobayashi [1], Shunsuke Kidani [1], Junfeng Li [2] and Masato Akagi [1]

[1] School of Information Science, Japan Advanced Institute of Science and Technology, 1-1 Asahidai, Nomi 923-1292, Japan
[2] Institute of Acoustics, Chinese Academy of China, No. 21, Beisihuan Xilu, Haidian, Beijing 100190, China
* Correspondence: unoki@jaist.ac.jp; Tel.: +81-761-51-1237

Featured Application: The findings of this study can be applied to speech processing systems that communicate urgency by speech to cochlear implant users.

Abstract: We previously investigated the perception of noise-vocoded speech to determine whether the temporal amplitude envelope (TAE) of speech plays an important role in the perception of linguistic information as well as non-linguistic information. However, it remains unclear if these TAEs also play a role in the urgency perception of non-linguistic information. In this paper, we comprehensively investigated whether the TAE of speech contributes to urgency perception. To this end, we compared noise-vocoded stimuli containing TAEs identical to those of original speech with those containing TAEs controlled by low-pass or high-pass filtering. We derived degrees of urgency from a paired comparison of the results and then used them as a basis to clarify the relationship between the temporal modulation components in TAEs of speech and urgency perception. Our findings revealed that (1) the perceived degrees of urgency of noise-vocoded stimuli are similar to those of the original, (2) significant cues for urgency perception are temporal modulation components of the noise-vocoded stimuli higher than the modulation frequency of 6 Hz, (3) additional significant cues for urgency perception are temporal modulation components of the noise-vocoded stimuli lower than the modulation frequency of 8 Hz, and (4) the TAE of the time-reversal speech is not likely to contain important cues for the perception of urgency. We therefore conclude that temporal modulation cues in the TAE of speech are a significant component in the perception of urgency.

Keywords: noise-vocoded speech; temporal modulation cue; urgency perception; temporal amplitude envelope

check for updates

Citation: Unoki, M.; Kawamura, M.; Kobayashi, M.; Kidani, S.; Li, J.; Akagi, M. Contributions of Temporal Modulation Cues in Temporal Amplitude Envelope of Speech to Urgency Perception. *Appl. Sci.* **2023**, *13*, 6239. https://doi.org/10.3390/app13106239

Academic Editors: Douglas O'Shaughnessy and Javier Hernando

Received: 3 March 2023
Revised: 26 April 2023
Accepted: 13 May 2023
Published: 19 May 2023

Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Speech is a natural and vital means of human communication in terms of expressing linguistic as well as non-linguistic and para-linguistic information. Humans typically utilize non-linguistic information such as vocal emotion and speaker individuality to enhance their speech and use linguistic information (speech intelligibility) to convey messages. They also utilize para-linguistic information such as emphasis and intention to convey greater nuance. Speech communication is also directly related to the behavior of the listeners. When a disaster occurs, for example, the sense of urgency can be conveyed through speech to properly facilitate the evacuation procedures. However, it is not yet clear which specific components of speech promote the perception of para-linguistic information as urgent.

Speech signals often redundantly contain key acoustical features pertaining to linguistic, non-linguistic, and para-linguistic information, and it is easy for people to recognize this information even if some of the features (fundamental frequency, vocal-intensity, spectral characteristics, etc.) are difficult to distinguish as a result of noise or reverberation. Important acoustical features have been studied in both time and frequency domains based on

the source filter model related to the speech production system. These features are spectral and temporal fine structures (harmonicity and periodicity) related to the fundamental frequency, formants, spectral tilt, and temporal power fluctuation, and are essential for speech perception.

On the other hand, the temporal amplitude envelope (TAE) and temporal fine structure (TFS) play important roles in auditory perception [1]. In particular, the role of TAE and TFS information was reviewed from aspects of auditory perception such as modulation perception and pitch perception. Drullman et al. reported that the cue of the TAE is more important for speech perception than that of the TFS [2]. A recent study by Atlas et al. also reported that the TAE or its modulation spectrum conveys linguistic information of speech [3]. We are, therefore, interested in understanding how the TAE of speech contributes to perception of non-linguistic information.

Recent psychoacoustical studies based on a noise-vocoded speech (NVS) scheme have demonstrated that the TAE of speech is an important cue for speech recognition (speech perception regarding linguistic information, i.e., speech intelligibility) [4–7]. NVS is generated by replacing the TFS of speech with band-limited noise while preserving the temporal amplitude envelope of speech, which enables the temporal characteristics to be preserved while drastically reducing the number of spectral characteristics.

Several prior works have investigated the NVS scheme in the context of speech recognition. A work by Shannon et al. demonstrated that NVS with just four bands is sufficient to achieve a good speech recognition (vowel, consonant, and sentence recognition) [3]. Drullman et al. investigated the importance of modulation-frequency bands for speech recognition by applying low- and high-pass filtering on the TAE [8,9], and Atlas et al. also reported that the TAE (or its modulation spectrum of speech) conveys the linguistic information of speech [3]. Xu et al. attempted to elucidate the importance of temporal cues for phoneme recognition using NVS [10]. These studies have indicated that modulation frequency components ranging from 4 to 16 Hz are key in speech recognition [8,11], which suggests that people can utilize the TAE of speech signals as a primary temporal modulation cue for the successful perception of linguistic information.

Our previous experiments in which we systematically varied the number of channels in the NVS scheme and the upper limitation of the modulation frequency components [11,12] demonstrated that temporal modulation cues play a key role in the recognition of both vocal emotion and speaker individuality, particularly at modulation frequencies lower than 8 Hz. Our research has also demonstrated that these features are robust to noise and reverberation [13,14]. However, we have still not clarified the influence of temporal modulation cues in non-linguistic information such as that used in emergency situations.

We have performed preliminary and experimental investigations to determine whether or not the TAE of speech affects the perception of urgency [15]. To ensure the results are exhaustive, in the current work, we performed four psychoacoustical experiments in which we investigated (1) whether the TAE of speech contributes to the perception of urgency when original and noise-vocoded stimuli are used individually, (2) whether the TAE of speech contributes to urgency perception when these stimuli are used simultaneously, (3) which temporal modulation component in a TEA of speech is a temporal modulation cue of urgency perception, where we restricted the lower or upper modulation frequency of the TAE, and (4) whether a temporal modulation cue in the TAE of time-reversal speech also contributes to the perception of urgency.

2. Stimulus Generation Based on Noise-Vocoded Speech

2.1. Speech Data in Evacuation Announcements

We utilized the recorded speech data from actual evacuation announcements in our experiments. These were recorded when a tsunami (tidal wave) was predicted after an earthquake off the coast of Fukushima, Japan on 22 November 2016 [16]. The announcements were spoken by multiple professional announcers with various speaking styles on different TV channels. We selected one real speech dataset (the same speaker, the same

sentence) by a professional male speaker (the same used by Kobayashi & Akagi [16]) as the original stimuli, which was/i/ma/su/gu/ni/ge/te/ku/da/sa/i/("Please evacuate now" in English). This announcement was made at four different levels of urgency ("A," "B," "C," and "D", corresponding to noise-vocoded stimuli labeled "a," "b," "c," and "d") to determine how the urgency was perceived for each.

2.2. Noise-Vocoded Speech

The process flow of the speech analysis/synthesis method we utilized to generate NVS as noise-vocoded stimuli is shown in Figure 1. In the NVS scheme, a speech signal can be re-synthesized by replacing the temporal fine structure of speech in the sub-bands with band-limited noise carriers while preserving the TAE. This scheme is therefore suitable for investigating the role of the TAE in urgency perception.

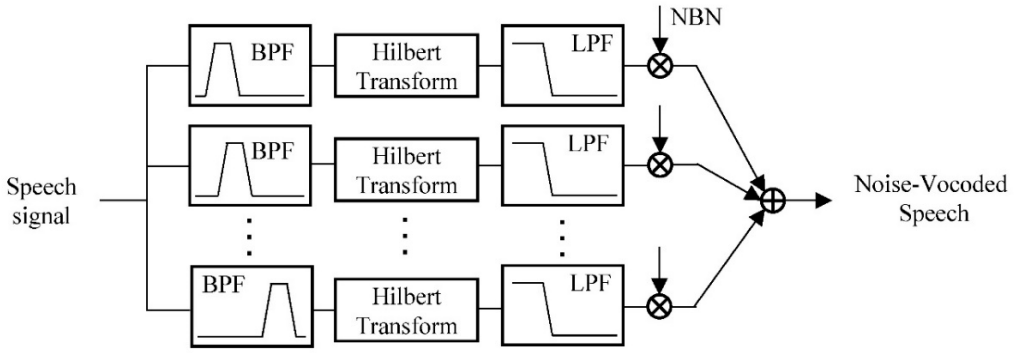

Figure 1. Process flow of speech analysis/synthesis method utilized to generate noise-vocoded stimuli (BPF: band-pass filter, LPF: low-pass filter, and NBN: narrow-band noise).

First, we reduced the effect of the average intensity by normalizing the active speech levels of all speech signals to −26 dBov with a P.56 speech voltmeter [14]. We then implemented band-pass filters (BPF) (essentially functioning as a band-pass filterbank) to divide the signals into several frequency bands. Next, we defined the bandwidth and center frequencies of the band-pass filterbank utilizing equivalent rectangular bandwidth (ERB_N) and ERB_N-number scales in Cam [17]. The ERB_N-number scale is similar to a distance scale from the basal side to the apical side of the cochlea, so it is fairly straightforward to replicate the frequency selectivity of the auditory system by simply decomposing the frequency bands on the basis of the ERB_N-number.

We respectively define the ERB_N and ERB_N-number as

$$\mathrm{ERB_N} = 24.7(4.37f/1000 + 1), \tag{1}$$

$$\mathrm{ERB_N - number} = 21.4\log_{10}(4.37f/1000 + 1), \tag{2}$$

where f refers to the center frequency in Hz. We define the center frequencies of the band-pass filterbank from 3 to 35 Cam with the bandwidth set to 2× ERB_N, resulting in a band-pass filterbank with 16 channels. The band-pass filterbank was implemented using 6th-order Butterworth infinite impulse response (IIR) filters.

We extracted the TAE of each channel signal from the output of the filterbank by utilizing the Hilbert transformation and a low-pass filter (LPF). The LPF was implemented using a 2nd-order Butterworth IIR filter and then, on the basis of its cut-off frequency, we set the upper limit of the modulation frequency to 64 Hz. This upper limit is related to the temporal resolution in that the higher the upper limit, the higher the obtained temporal resolution.

In the final step, the TAE in each channel was utilized to modulate the temporal amplitude with the narrow band-limited noise (NBN) generated by band-pass filtering white

noise at the same channel (band-frequency). The noise-vocoded stimuli were generated by summing up all amplitude-modulated NBNs.

3. Experiment I: Urgency Perception of TAE of Speech

The objective of experiment I was to determine whether the TAE of speech contributes to the perception of speech urgency.

3.1. Stimuli

We utilized the four original stimuli of the evacuation announcement and the corresponding noise-vocoded stimuli. Two randomly paired stimuli were used for paired comparison and 0.5 s of silence was added between the first and second stimuli. The total number of paired stimuli was 12 and the total execution time was approximately 10 min.

3.2. Participants

The participants were ten native Japanese speakers (three women and seven men) ranging in age from 22 to 25 years old. The absolute threshold of all participants measured through a standard audiometric tone test with a RION AA-72B audiometer was the hearing level of 12 dB or less for both ears at octave frequencies between 125 and 8000 Hz (i.e., all had normal hearing).

3.3. Procedure

The participants sat in a soundproof room and stimuli were simultaneously presented to both ears through a PC, an audio interface (RME Fireface UCX), and a set of headphones (Sennheiser HDA 200). We utilized a head and torso simulator (B&K type 4128) and sound level meter (B&K type 2231) to calibrate the sound pressure to the same level for all participants.

Scheffé's paired comparison method [18] was used to evaluate the degree of urgency of the stimuli. We instructed the participants to indicate whether the first or second stimulus was more urgent on a five-point scale (-2: second stimulus is significantly more urgent, -1: second stimulus is somewhat more urgent, 0: both are the same, $+1$: first stimulus is somewhat more urgent, and $+2$: first stimulus is significantly more urgent).

3.4. Results

Figures 2 and 3 depict the degrees of urgency for the original stimuli and noise-vocoded stimuli, respectively, as based on Scheffé's paired comparison method, where the horizontal axis shows the degree of urgency (positive values mean higher urgency). As we can see, the degrees of urgency from lowest to highest were A, B, D, and C for the original stimuli and a, b, d, and c for the noise-vocoded stimuli.

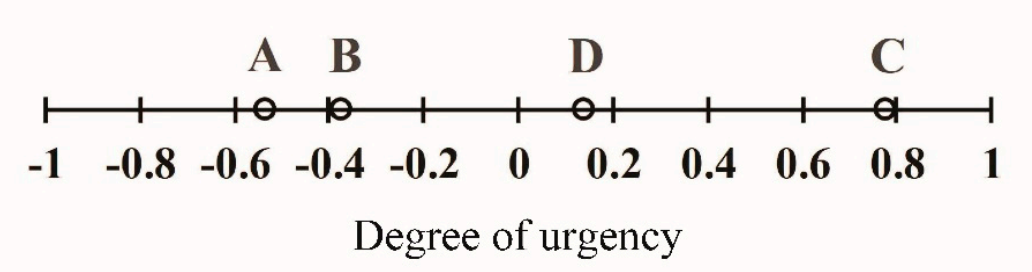

Figure 2. Results of experiment I: Degrees of urgency perception of original stimuli.

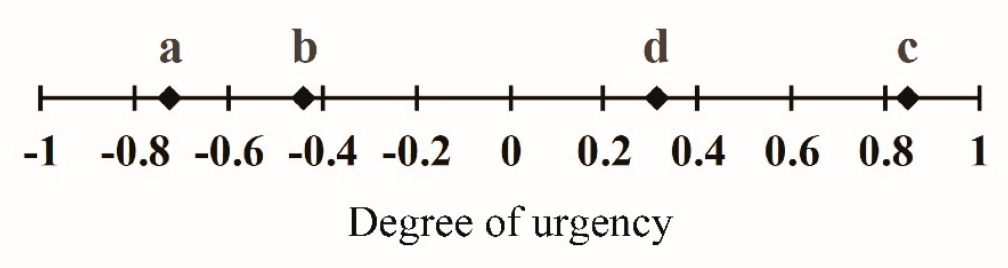

Figure 3. Results of experiment I: Degrees of urgency perception of noise-vocoded stimuli.

Analysis of variance (ANOVA) indicated a significant main effect ($F(3,77) = 34.42$, $p < 0.01$) of the urgency perception of the original stimuli and significant differences between all pairs except A and B ($p < 0.01$). We also found a significant main effect ($F(3,77) = 125.8,\ p < 0.01$) of the urgency perception of the noise-vocoded stimuli and significant differences between all pairs ($p < 0.01$).

3.5. Consideration

The order of the degrees of urgency from highest to lowest matched up perfectly for the noise-vocoded stimuli and the original stimuli (C (c), D (d), B (b), and A (a)), which confirms that the TAE contributes to the perception of urgency. We also clarified that the degrees of urgency for both types of stimuli were different.

4. Experiment II: Effect of Noise-Vocoded Stimuli on Urgency Perception

The objectives of experiment II were to corroborate the results of experiment I and to determine whether the TAE of speech contributes to the perception of urgency.

4.1. Stimuli

We utilized eight stimuli: the four original stimuli and the four corresponding noise-vocoded stimuli. As in experiment I, two randomly paired stimuli were used for comparison and 0.5 s of silence was added between the first and second stimuli. The total number of paired stimuli was 56 and the total execution time was approximately 20 min.

4.2. Participants

The participants were ten native Japanese speakers (two women and eight men) ranging in age from 22 to 25 years old. All had normal hearing as determined through the same testing as experiment I.

4.3. Procedure

The participants sat in a soundproof room and were presented with the same stimuli as experiment I. We again carried out Scheffé's paired comparison method to evaluate the degrees of urgency of the original and noise-vocoded stimuli.

4.4. Results

The results are shown in Figure 4, where the horizontal axis indicates the degree of urgency. As we can see, the degrees of urgency from lowest to highest are ordered as A, a, B, b, D, d, C, and c. ANOVA analysis indicated a significant main effect of the urgency perception of these stimuli ($F(3459) = 100.8,\ p < 0.01$) and significant differences between all pairs except a & B, a & b, D & d, D & C, d & C, d & c, and C & c ($p < 0.01$).

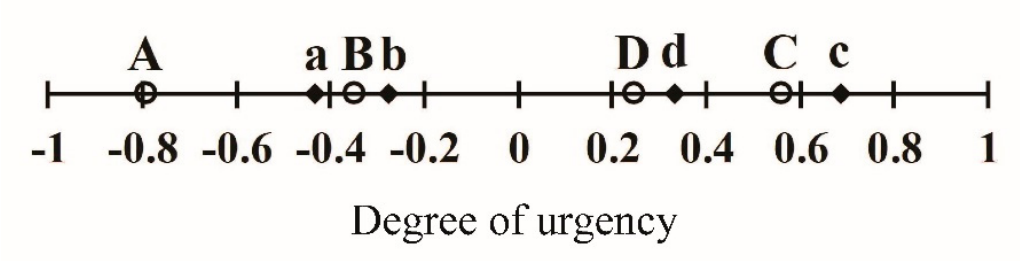

Figure 4. Results of experiment II: Degrees of urgency perception of all original and noise-vocoded stimuli presented in experiment I.

4.5. Consideration

The order of the degrees of urgency from highest to lowest was c, C, d, D, b, B, a, and A. We found that the order here is consistent with that derived in experiment I.

We found that the degrees of urgency of the original stimuli were lower than those of the noise-vocoded stimuli when the same stimulus pair of the same urgency level (e.g., A and a) was compared. In the process of generating NVS, the TAE of the original stimuli was preserved and driven by band-limited random noise as a carrier signal, which is presumably why the spectral centroid differed slightly from that of the original stimuli.

Our findings suggest that the shape of the frequency spectrum in the middle and higher ranges is related to the sense of urgency perception [19]. Therefore, we investigated why there is a difference in urgency perception between the original and noise-vocoded stimuli by utilizing sharpness [20], which is a common sound-quality metric.

The results of our analysis are listed in Table 1, where we can see that the sharpness of the noise-vocoded stimuli was somewhat higher than that of the original stimuli when the sharpness differed between stimuli of the same level of urgency (see difference in the fourth row). It is also clear that the difference was almost constant regardless of the level of urgency. These results suggest that the difference in sharpness may have affected the difference in the perception of urgency between the original and noise-vocoded stimuli observed in experiments I and II.

Table 1. Sharpness differences between original and noise-vocoded stimuli. The unit of sharpness is [acum].

Stimulus Label	A (a)	B (b)	C (c)	D (d)
Original	1.45	1.55	1.55	1.51
Noise-vocoded	1.73	1.84	1.82	1.77
Difference	0.28	0.29	0.27	0.26

5. Experiment III: Effect on Restricting Temporal Modulation Components of TAE

The objective of experiment III was to determine which temporal modulation component in the TAE of speech acts as a temporal modulation cue of urgency perception. We compared noise-vocoded stimuli consisting of TAEs that were identical to those of the original stimuli and that were restricted using LPF or a high-pass filter (HPF).

5.1. Stimuli

We utilized four corresponding noise-vocoded stimuli and set seven cut-off frequencies (2, 4, 6, 8, 12, 16, and 32 Hz) for the modulation frequency components of the restricted TAEs. Scheffé's paired comparison method was utilized to randomly generate two stimulus pairs with each speech stimulus, for each of the LPFs and HPFs. We utilized a total of 756 paired stimuli and set 0.5 s of silence between the first and second stimuli. All paired stimuli were divided into 14 sessions containing 54 judgements each. We scheduled a

90-min break after the first seven sessions to give the participants a chance to rest, which brought the total execution time to approximately 180 min.

5.2. Participants

The participants were ten native Japanese speakers (three women and seven men) ranging in age from 22 to 25 years old. All had normal hearing as determined through the same testing as experiment I.

5.3. Procedure

The participants sat in a soundproof room and were presented with the same stimuli as experiment I. We again carried out Scheffé's paired comparison method to evaluate the degrees of urgency of the stimuli.

5.4. Results

The results under the LPF condition are shown in Figure 5, where the horizontal axis indicates the cut-off frequency of the LPF, and the vertical axis shows the urgency scale (positive values indicate a higher degree of urgency). As we can see, the degrees of urgency from lowest to highest are ordered as the same as the results of experiments I and II. The degrees of urgency for stimuli c and d decreased as the cut-off frequency decreased. ANOVA analysis indicated a significant main effect of urgency perception ($F(24, \infty) = 200.1,\ p < 0.01$) and significant differences among all stimuli in c and d, where the cut-off frequencies ranged from 4 to 32 Hz ($p < 0.01$). With the cut-off frequency of 2 Hz, significant differences were observed among all stimulus pairs except b & d and d & c ($p < 0.01$).

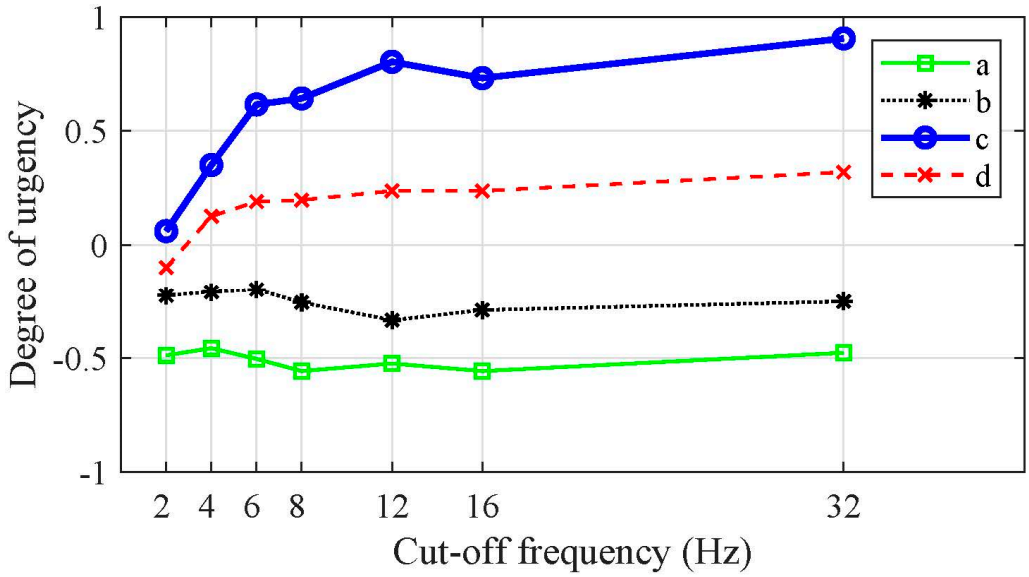

Figure 5. Results of experiment III: Urgency perception of noise-vocoded stimuli with the upper temporal modulation component of TAE restricted using LPF.

The results under the HPF condition are shown in Figure 6, where the horizontal and vertical axes are the same as in Figure 5. We can see here that the degrees of urgency from lowest to highest are ordered and the degrees of urgency for stimuli c and d decreased as the cut-off frequency increased. ANOVA analysis indicated a significant main effect of urgency perception ($F(24, \infty) = 1.517,\ p < 0.01$). 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 were from 2

to 8 Hz ($p < 0.01$). With the cut-off frequencies of 8 and 12 Hz, significant differences were observed among all stimulus pairs except c & d ($p < 0.01$).

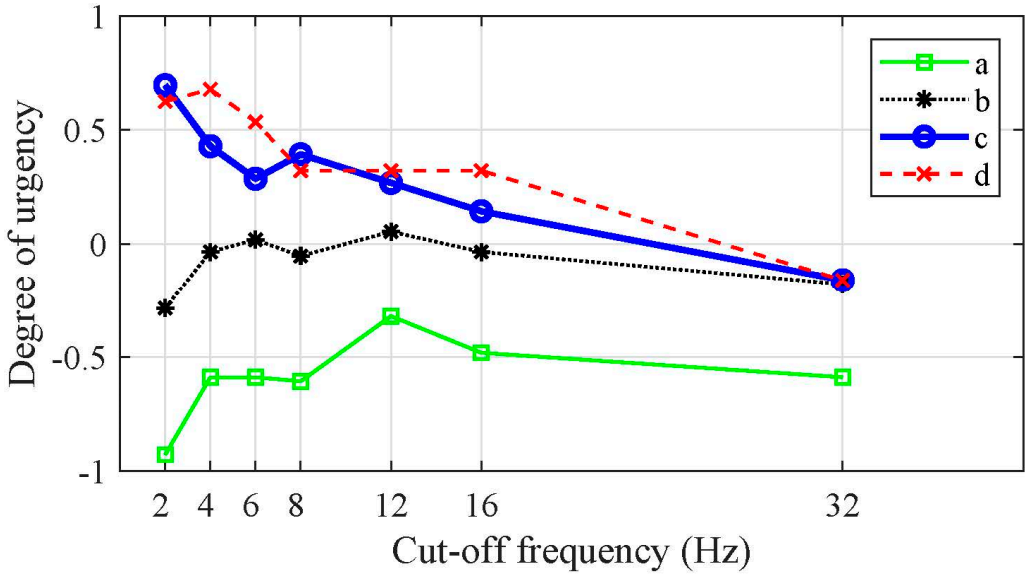

Figure 6. Results of experiment III: Urgency perception of noise-vocoded stimuli with the lower temporal modulation component of TAE restricted using HPF.

5.5. Consideration

The orders of urgency perception among the four stimuli regarding the LPF cut-off frequency (Figure 5) and HPF cut-off frequency (Figure 6) are in good agreement. The significant difference between the degrees of urgency at the cut-off frequencies of 4 and 6 Hz in Figure 5 suggests that the temporal modulation components of the noise-vocoded stimuli upwards of 6 Hz are significant cues for urgency perception. Similarly, the significant difference between the cut-off frequencies of 8 and 12 Hz in Figure 6 indicates that those downwards of 8 Hz are significant cues. These findings demonstrate that temporal modulation cues in a TAE from 6 to 8 Hz have a key role to play in the perception of urgency.

6. Experiment IV: Effect on Temporal Amplitude Envelope of Time-Reversed Speech

The results of experiment III indicate that the modulation frequency components between 6 and 8 Hz in the TAE of speech are key elements in the perception of urgency as para-linguistic information. Moreover, since para-linguistic information is independent of linguistic information, the modulation frequency components between 6 and 8 Hz are also presumably important. To confirm this, we conducted an urgency-perception experiment on time-reversed noise-vocoded stimuli utilizing LPFs and HPFs, the same as in experiment III.

6.1. Stimuli

We utilized time-reversed versions of the four noise-vocoded stimuli from experiment III. The modulation frequency components of the TAE were restricted using LPFs and HPFs at seven different cut-off frequencies (2, 4, 6, 8, 12, 16, and 32 Hz). Two stimulus pairs were randomly generated using Scheffé's paired comparison method with each speech stimulus for both the LPFs and HPFs. We utilized a total of 756 paired stimuli and set 0.5 s of silence between the first and second stimuli. All paired stimuli were divided into 14 sessions containing 54 stimulus pairs each. We scheduled a 90-min break after the first seven sessions, so the duration of the experiment was approximately 180 min.

6.2. Participants

The participants were 12 native Japanese (two women and ten men) ranging in age from 22 to 25 years old. All had normal hearing as determined through the same testing as experiment 1.

6.3. Procedure

We conducted the experiment in a soundproof room using the same stimulus presentation, experimental setup, and procedures as experiment I.

6.4. Results

In this experiment, we set the upper and lower limits of the modulation frequency component of the TAE of time-reversed speech by using an LPF and HPF, respectively. Figure 7 shows the results for the LPF condition and Figure 8 for the HPF condition, where the formatting is the same as in experiment III.

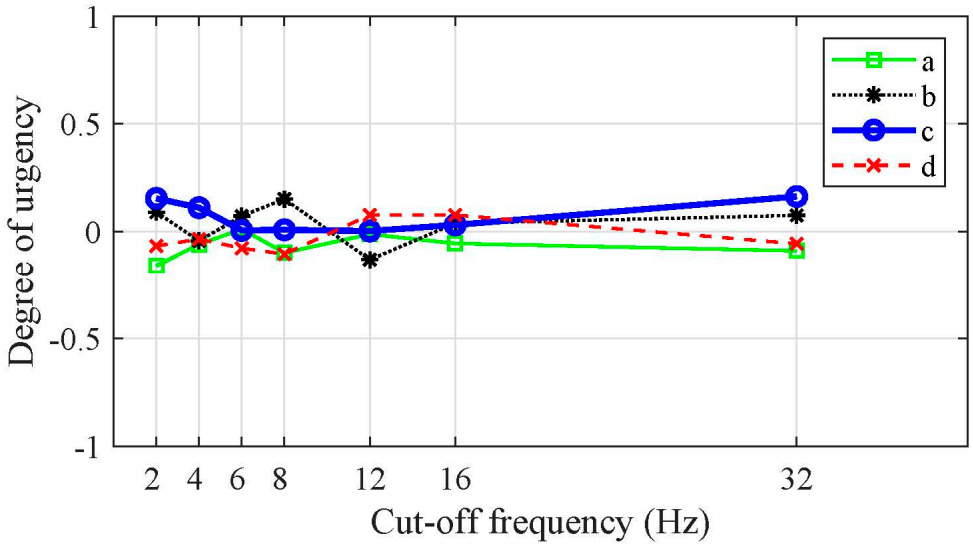

Figure 7. Results of experiment IV: Urgency perception of time-reversed noise-vocoded stimuli with the upper modulation component of TAE restricted using LPF.

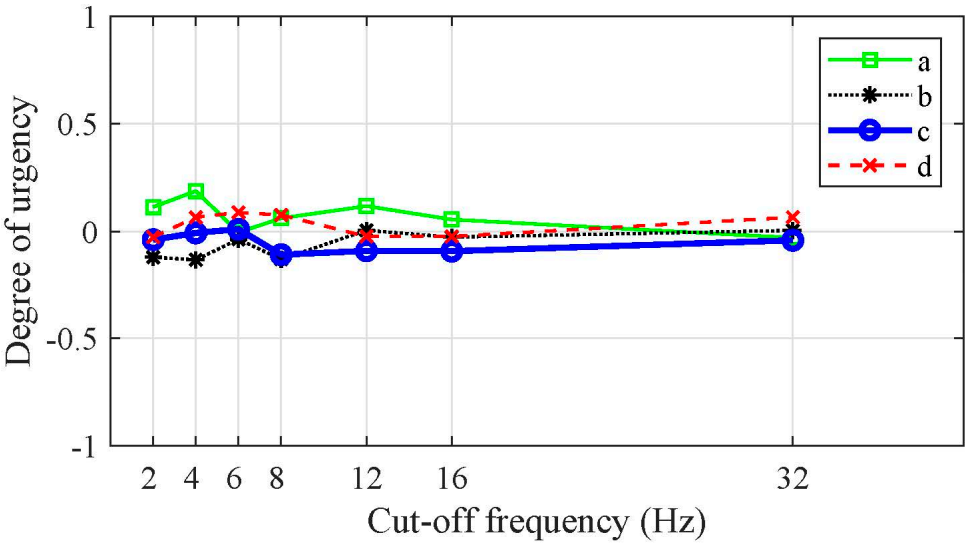

Figure 8. Results of experiment IV: Urgency perception of time-reversed noise-vocoded stimuli with the lower modulation component of TAE restricted using HPF.

As we can see here, in contrast to the results of experiment III, the degree of urgency was almost the same for all four stimulus conditions, and the degree of urgency was near zero regardless of the LPF and HPF cut-off frequencies.

6.5. Consideration

The LPF cut-off frequency of 32 Hz in Figure 7 and the HPF cut-off frequency of 2 Hz in Figure 8 correspond roughly to the perceived urgency in the time-reversed noise-vocoded stimuli demonstrated in experiments I and II. Both the different order in the degrees of urgency for each stimulus condition and the fact that the degree of urgency was almost zero indicate that the temporal ordering and changes in the TAE of speech information are key factors influencing the perception of urgency. Of course, the orderliness of the temporal direction of the TAE and its changes are also necessary for the perception of the linguistic information of speech.

7. General Discussion

The NVS scheme has been used as a cochlear implant (CI) simulator to simulate the poor spectral cue and enough temporal cues provided by a CI device [21,22]. Since CIs are known to lack the information necessary for recognizing non-linguistic and para-linguistic information [23], the results of our study could be helpful for improving the non-linguistic and para-linguistic information recognition of CI users by increasing or decreasing specific modulation frequency components of a TAE.

Our previous studies in which we systematically varied the number of channels in the NVS scheme and the upper limitation of the modulation frequency components [11,12] demonstrated that the TAE plays an important role in the recognition of both vocal emotion and speaker individuality as non-linguistic information. These studies found that significant cues for perception of non-linguistic information are temporal modulation components of the noise-vocoded stimuli lower than the modulation frequency of 8 Hz. These studies also suggested that recognition of vocal-emotion and speaker individuality using the NVS scheme with NH listeners can be used to predict the response from CI listeners as CI simulators [24–26].

In this paper, we found with experiments I and II that NVS contains features that contribute to a sense of urgency as para-linguistic information. The perception of urgency might also involve human perceptual mechanisms of amplitude modulation. Additionally, according to the results of experiment III, the modulation frequency component of a TAE that contributes to the perception of urgency is from 6 to 8 Hz. Din et al. previously showed that the modulation frequency of speech peaks near 4 Hz regardless of language [21], which suggests that the modulation frequency of linguistic information is near 4 Hz. Our own findings thus indicate that the modulation frequency component of contributing to urgency perception is higher than that of linguistic information.

Our experimental findings indicate that even para-linguistic information might not be completely unrelated to linguistic information. For example, the results of experiment IV suggest that the sense of urgency may also be lost due to the loss of the linguistic information of speech stemming from the use of reversed speech. However, the modulation frequency components at the core of linguistic information and those important for the perception of urgency differ, as noted above [23]. Therefore, it is unlikely that the loss of a sense of urgency observed in experiment IV is related to the loss of linguistic information of speech.

We should mention that the modulation frequency observed in experiment III was calculated in the sense of a long-time average. Therefore, if the modulation frequency component in a TAE of speech plays an important role in the perception of urgency, it should be possible to perceive urgency even in a speech where the temporal direction is reversed. The results of experiment IV suggest that this was not the case, however, which leads us to suspect that the instantaneous modulation frequency (e.g., temporal variation of modulation frequency, asymmetry of its variation pattern, the difference between rising

and falling edges, etc.), as well as the modulation frequency component calculated in the sense of a long-time average, are important cues for the perception of urgency. This is our future work.

8. Conclusions

In this work, we comprehensively examined whether temporal modulation cues in the TAE of speech contribute to the perception of urgency. To this end, we performed experiments in which noise-vocoded stimuli containing TAEs identical to those of the original stimuli were compared with those in which the components were changed by low-pass or high-pass filtering. We derived the degrees of urgency from a paired comparison of the results and then used them as a basis for clarifying the relationship between the temporal modulation components and urgency perception. Our key findings are as follows.

(i) The perceived degrees of urgency of noise-vocoded stimuli are similar to those of the original stimuli.
(ii) Significant cues for urgency perception are temporal modulation components of the noise-vocoded stimuli higher than the modulation frequency of Hz.
(iii) Additional significant cues for urgency perception are temporal modulation components of the noise-vocoded stimuli lower than the modulation frequency of Hz.
(iv) A TAE of time-reversal speech is not likely to contain important cues for the perception of urgency.

In other words, the instantaneous modulation frequency (or asymmetry of the time-varying pattern of modulation frequency, etc.) is presumably more important than the modulation frequency in the long-time average sense.

Overall, our findings demonstrate that temporal modulation cues in the TAE significantly contribute to the perception of urgency.

Author Contributions: Conceptualization: M.U.; methodology: M.K. (Miho Kawamura), M.K. (Maori Kobayashi) and S.K.; experiments: M.K. (Miho Kawamura); data analysis: M.K. (Miho Kawamura), M.K. (Maori Kobayashi) and S.K.; data curation: all authors; general discussion: all authors; writing—original draft preparation: M.U. and S.K.; writing—review and editing: M.U., S.K. and J.L.; visualization: M.K. (Miho Kawamura); supervision: M. U. and M. A.; funding acquisition: M.U. and M.A. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the SCOPE Program of the Ministry of Internal Affairs and Communications (Grant No. 201605002) and JSPS-NSFC Bilateral Programs (Grant number: JSJSBP120197416). This research was also supported by a Grant-in-Aid for Innovative Areas (Grant Nos. 16H01669, 18H05004, and 21H03463), a Grant-in-Aid for Scientific Research (B) (21H03463), and a Fund for the Promotion of Joint International Research (Fostering Joint International Research (B)) (20KK0233), from MEXT.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Informed consent was obtained from all participants involved in the study.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Moore, B.C.J. The roles of temporal envelope and fine structure information in auditory perception. *Acoust. Sci. Technol.* **2019**, *40*, 61–83. [CrossRef]
2. Drullman, R. Temporal envelope and fine structure cues for speech intelligibility. *J. Acoust. Soc. Am.* **1995**, *97*, 585–592. [CrossRef] [PubMed]
3. Atlas, L.; Greenberg, S.; Hermansky, H. The Modulation Spectrum and Its Application to Speech Science and Technology. In Proceedings of the Interspeech2007, Tutorial, Antwerp, Belgium, 27 August 2007.
4. Shannon, R.V.; Zeng, F.G.; Kamath, V.; Wygonski, J.; Ekelid, M. Speech recognition with primarily temporal cues. *Science* **1995**, *270*, 303–304. [CrossRef] [PubMed]

5. Tachibana, R.O.; Sasaki, Y.; Riquimaroux, H. Relative contributions of spectral and temporal resolutions to the perception of syllables, words, and sentences in noise-vocoded speech. *Acoust. Sci. Technol.* **2013**, *34*, 263–270. [CrossRef]
6. Loizou, P.C.; Dorman, M.; Tu, Z. On the number of channels needed to understand speech. *J. Acoust. Soc. Am.* **1999**, *106*, 2097–2103. [CrossRef] [PubMed]
7. Xu, L.; Pfingst, B.E. Spectral and temporal cues for speech recognition: Implications for auditory prostheses. *Hear. Res.* **2008**, *242*, 132–140. [CrossRef] [PubMed]
8. Drullman, R.; Festen, J.M.; Plomp, R. Effect of temporal envelope smearing on speech reception. *J. Acoust. Soc. Am.* **1994**, *95*, 1053–1064. [CrossRef] [PubMed]
9. Drullman, R.; Festen, J.M.; Plomp, R. Effect of reducing slow temporal modulations on speech reception. *J. Acoust. Soc. Am.* **1994**, *95*, 2670–2680. [CrossRef]
10. Xu, L.; Thompson, C.S.; Pngst, B.E. Relative contributions of spectral and temporal cues for phoneme recognition. *J. Acoust. Soc. Am.* **2005**, *117*, 3255–3267. [CrossRef] [PubMed]
11. Zhu, Z.; Nishino, Y.; Miyauchi, R.; Unoki, M. Study on linguistic information and speaker individuality contained in temporal envelope of speech. *Acoust. Sci. Technol.* **2016**, *37*, 258–261. [CrossRef]
12. Zhu, Z.; Miyauchi, R.; Araki, Y.; Unoki, M. Contributions of temporal cue on the perception of speaker individuality and vocal emotion for noise-vocoded speech. *Acoust. Sci. Technol.* **2018**, *39*, 234–242. [CrossRef]
13. Zhu, Z.; Kawamura, M.; Unoki, M. Study on the perception of nonlinguistic information of noise-vocoded speech under noise and/or reverberation conditions. *Acoust. Sci. Technol.* **2022**, *43*, 306–315. [CrossRef]
14. Guo, T.; Zhu, Z.; Kidani, S.; Unoki, M. Contribution of common modulation spectral features to vocal-emotion recognition of noise-vocoded speech in noisy reverberant environments. *Appl. Sci.* **2022**, *12*, 9979. [CrossRef]
15. Unoki, M.; Kawamura, M.; Kobayashi, M.; Kidani, S.; Akagi, M. How the temporal amplitude envelope of speech contributes to urgency perception. In Proceedings of the 23rd International Congress on Acoustics, ICA 2019, Aachen, Germany, 9–13 September 2019; pp. 1739–1744.
16. Kobayashi, M.; Akagi, M. Psychological evaluation of evacuation announcements. *J. Acoust. Soc. Jpn.* **2018**, *74*, 633–640. (In Japanese with English Abstract)
17. Moore, B.C.J. *An Introduction to the Psychology of Hearing*, 6th ed.; Brill Academic Pub.: Leiden, The Netherlands, 2013.
18. Scheffé, H. An Analysis of Variance for Paired Comparisons. *J. Am. Stat. Assoc.* **1952**, *47*, 381–400.
19. Kobayashi, M.; Hamada, Y.; Akagi, M. Acoustic features correlated to perceived urgency in evacuation announcements. *Speech Commun.* **2022**, *139*, 22–34. [CrossRef]
20. Fastl, H.; Zwicker, E. *Psycho-Acoustics Facts and Models*; Springer: La Vergne, TN, USA, 2010.
21. Whitmal, A.N.; Poissant, F.S.; Freyman, L.R.; Helfer, S.K. Speech intelligibility in cochlear implant simulations: Effects of carrier type, interfering noise, and subject experience. *J. Acoust. Soc. Am.* **2007**, *122*, 2376–2388. [CrossRef]
22. Everhardt, M.K.; Sarampalis, A.; Coler, M.; Başkent, D.; Lowie, W. Meta-analysis on the identification of linguistic and emotional prosody in cochlear implant users and vocoder simulations. *Ear Hear.* **2020**, *41*, 1092–1102. [CrossRef]
23. Ding, N.; Patel, A.D.; Chen, L.; Butler, H.; Luo, C.; Poeppel, D. Temporal modulations in speech and music. *Neurosci. Biobehav. Rev.* **2017**, *81*, 181–187. [CrossRef]
24. Zhu, Z.; Miyauchi, R.; Araki, R.; Unoki, M. Modulation spectral features for predicting vocal emotion recognition by simulated cochlear implants. In Proceedings of the Interspeech 2016, San Francisco, CA, USA, 8–12 September 2016; pp. 262–266.
25. Zhu, Z.; Miyauchi, R.; Araki, R.; Unoki, M. Recognition of vocal emotion in noise-vocoded speech by normal hearing and cochlear implant listeners. *J. Acoust. Soc. Am.* **2016**, *140 Pt 2*, 3271. [CrossRef]
26. Zhu, Z.; Miyauchi, R.; Araki, R.; Unoki, M. Important role of temporal cues in speaker identification for simulated cochlear implants. In Proceedings of the 1st International Workshop on Challenges in Hearing Assistive Technology (CHAT-2017), Stockholm, Sweden, 19 August 2017; pp. 51–55.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.