Acoust. Sci. & Tech. **43**, 6 (2022)

TRANSLATED PAPER

Study on the perception of nonlinguistic information of noise-vocoded speech under noise and/or reverberation conditions

Zhi Zhu, Miho Kawamura and Masashi Unoki*

School of Information Science, Japan Advanced Institute of Science and Technology, 1–1 Asahidai, Nomi, Ishikawa, 923–8178 Japan

(*Received 17 May 2022, Accepted for publication 4 July 2022*)

Abstract: It has been known that noise and reverberation greatly affect the perception of linguistic information, in particular speech intelligibility. However, the effect of noise and reverberation on the perception of non-linguistic information has not been clarified. We investigated how these types of disturbances affect the perception of non-linguistic information (speaker individuality and vocal emotion) of noise-vocoded speech. We conducted speaker-distinction and vocal-emotion-recognition experiments using noise-vocoded speech created from the speech in noisy, reverberation, and noisy reverberant environments as stimuli. We used seven noise conditions (signal-to-noise ratio (SNR) $= \infty, 20, 15, 10, 5, 0, -5$ dB) and six reverberation conditions (reverberation time (T_R) $= 0.0, 0.1, 0.3, 0.5, 1.0, 2.0$ s). In both speaker-distinction and vocal-emotion-recognition experiments, the main effects of noise and reverberation were significant, but the interaction was not significant. From these results, except for extremely poor sound conditions, under daily noise and reverberation conditions (an SNR of more than 10 dB and T_R less than 1.0 s), there were no significant effects of noise and reverberation.

Keywords: Modulation perception, Noise-vocoded speech, Speaker distinction, Vocal-emotion recognition, Noisy reverberant environments

1. INTRODUCTION

The temporal amplitude envelope (TAE) of speech plays an important role in speech perception. It was found from Shannon *et al.*'s study on noise-vocoded speech (NVS) [1]. NVS is generated by modulating band-limited noise with a TAE in sub-bands, thus it can be used for cochlear-implant simulation. It was reported that NVS with dynamic temporal patterns in only four broad spectral regions is sufficient for listeners to recognize linguistic information [1–5].

Speech is used as an important means of communication for humans to express linguistic information as well as non-linguistic information such as speaker individuality and vocal-emotion [6]. Studies revealed that TAEs of speech and their modulation spectral components contribute to the recognition of linguistic information as well as perception of non-linguistic information. We previously investigated the effect of controlling the upper limit of the modulation frequency of a TAE on the perception of non-linguistic information, i.e., speaker individuality and vocal-emotion, using NVS [7–9]. It was found that the speaker-distinction and vocal-emotion-recognition rates decrease as the upper limit of the modulation frequency becomes lower [7–9].

These studies focused on the role of the modulation component of a TAE in the perception of linguistic and non-linguistic information, so we conducted listening experiments in a quiet environment without noise or reverberation. However, in an actual listening environment, noise and reverberation disturb our listening to target speech. Understanding how the perception of linguistic and non-linguistic information is affected in such a speech-listening environment is not only important for modulation perception of speech but also for exploring the essence of speech communication.

It is known that noise and reverberation greatly affect the perception of linguistic information, in particular speech intelligibility. Tillery *et al.* conducted speech intelligibility tests with cochlear-implant-simulated speech in noisy reverberant environments [10]. They found that the speech intelligibility of the cochlear-implant-simulated speech was significantly lower than that of normal speech (original speech) due to the addition of noise and

*e-mail: unoki@jaist.ac.jp

The original paper (in Japanese) is published in *the Journal of the Acoustical Society of Japan*, **76**(6), 317–326 (2020).
[doi:10.1250/ast.43.306]

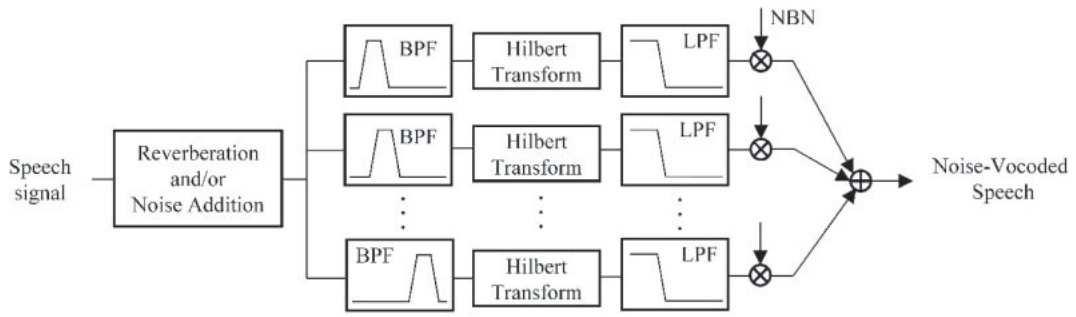

Fig. 1 Schematic diagram of noise-vocoder method used to generate stimuli (BPF: band-pass filter, LPF: low-pass filter, and NBN: narrow band noise). BPFs were defined from 3 to 35 ERB_N-number with bandwidth as $2ERB_N$. Number of channels was 16.

reverberation. However, it is still unclear how these types of disturbances affect the perception of non-linguistic information.

For this study, we investigated the perception of non-linguistic information of NVS, assuming a speech-listening environment in which background noise and room reverberation simultaneously exist. Therefore, we conducted experiments on speaker distinction and vocal-emotion recognition of NVS created from a speech signal with noise and reverberation added. For background noise, stationary noise (white Gaussian noise) was assumed. As the noise condition, the adjusted noise was added to the speech so that the signal-to-noise ratio (SNR) of the original signal and noise would have different values. A statistical room impulse response (Schroeder model) [11] was assumed for reverberation, and a reverberation impulse response with a different reverberation time convoluted into speech was assumed for the reverberation condition. The NVS of noise and reverberation was obtained by first convolving the reverberation into the original voice, adding noise to it, then driving it with noise.

This paper is organized as follows. Section 2 introduces the method for creating NVS, which was used as the experimental stimuli. Section 3 gives an overview of our speaker-distinction and vocal-emotion-recognition experiments using NVS in noisy environments. Section 4 describes these experiments using NVS in a reverberant environment. Section 5 describes these experiments using NVS in a noisy reverberant environment. Section 6 discusses the perception of linguistic and non-linguistic information of NVS in noisy and reverberant environments. Finally, Section 7 provides a discussion and summarizes the results.

2. STIMULUS CONFIGURATION

We investigated the perception of non-linguistic information of NVS in three environments: noisy, reverberant, and noisy reverberant. Figure 1 shows the schematic diagram of the noise-vocoder method for generating the NVS stimuli used in this study [12]. This method was used to equalize the root mean squared (RMS) level of all speech data to -26 dB.[1] Next, noise and reverberation were added to the speech in the three environments. When noise and reverberation were added simultaneously (noisy reverberation environment), reverberation was added first then noise. Finally, NVS stimuli were created on the basis of the speech with noise and reverberation added.

2.1. Noise Addition

For the noise condition, the powers of the original speech and noise were obtained and background noise (additive stationary noise) with an arbitrarily determined SNR was added to the original speech signal $x(t)$. The SNR is defined as follows.

$$\mathrm{SNR} = 10\log_{10}\frac{\int_0^T x^2(t)dt}{\int_0^T a_N^2 n_N^2(t)dt}, \tag{1}$$

where $n_N(t)$ is white Gaussian noise, a_N is the amplitude of the noise, and T is the signal length. Here, a_N is adjusted to obtain an arbitrary SNR, and the noisy speech $y(t)$ is represented as follows:

$$y(t) = x(t) + a_N n_N(t). \tag{2}$$

2.2. Reverberation Addition

For reverberation, Schroeder's statistical room impulse response $h(t)$ [11] defined by Eq. (4) was used to add reverberation to the original speech to create a reverberant speech. The reverberation time T_R of the statistical room

[1]When the reference signal was recorded as the digital signal, the signal level in the speech section must be set to be -26 dBov by following ITU-T Rec. P56 to avoid the clipping issue. In this case, A-weighted noise level is less than -80 dBov. Decibel overload is referred to as dBov. In this paper, dBov is represented in dB according to "guide for contributor" in Acoustical Science and Technology.

impulse response was set to an arbitrary value, and the reverberation speech $y(t)$ was created by convoluted $h(t)$ with the original speech $x(t)$.

$$y(t) = h(t) * x(t), \tag{3}$$

$$h(t) = a_{\mathrm{R}} \exp\left(-\frac{6.9t}{T_{\mathrm{R}}}\right) n_{\mathrm{R}}(t), \tag{4}$$

where "$*$" is the convolution operator, a_{R} is the amplitude, and $n_{\mathrm{R}}(t)$ is the white Gaussian noise carrier. Since a diffused sound field was assumed, the TAE of the room impulse response was represented as exponentially decaying with time. Since the statistical room impulse response was used, a different $n_{\mathrm{R}}(t)$ was used for each stimulus speech.

2.3. Additions of Noise and Reverberation

As the noisy reverberant condition in which both background noise and room reverberation exist simultaneously, a noisy reverberant speech $y(t)$ was represented as follows: after changing the SNR and T_{R}, as defined by the following equation, the $y(t)$ was created by convoluting the original speech with the room impulse response $h(t)$ using the method described in Sect. 2.2 then adding the background noise $a_{\mathrm{N}} n_{\mathrm{N}}(t)$ using the method described in Sect. 2.1.

$$y(t) = h(t) * x(t) + a_{\mathrm{N}} n_{\mathrm{N}}(t), \tag{5}$$

with which the SNR was calculated only for the speech section of the original signal and the amplitude term a_{N} was adjusted to achieve an arbitrary SNR.

There is another formulation for creating $y(t)$ which the background noise is considered as a noise source and, after the background noise is added to the original speech, reverberation is added to the whole speech. Since we considered speech-signal processing at the listening side, as shown in Fig. 1, we adopted the formulation in which background noise is added to the reverberated speech, instead of the formulation in which both the original speech and background noise are affected by reverberation. This formulation has been used in studies of speech-recovery methods in noisy reverberant environments and in front-end studies of speech recognition (e.g., [13,14]). This is also due to the fact that the rear reverberation component, which is the main component of superimposed noise, can be regarded as additive noise.

The room impulse response used in Sect. 2.2 and the stationary noise used in Sect. 2.1 are composed of a white Gaussian noise carrier that is independent of each other. Even if it is assumed that the background noise is affected by the reverberation in accordance with another formulation, the reverberation noise can also be regarded statistically as white Gaussian noise (stationary noise) [15]. Both formulations are equivalent in terms of statistical signal processing even if the order of adding background noise and reverberation is changed. Therefore, we added noise and reverberation by using Eq. (5), and the one created with this method is referred to as the noisy reverberant speech.

2.4. How to Generate Noise-vocoded Speech

The following procedure was used to generate the stimulus of NVS of noisy speech, reverberant speech, or noisy reverberant speech.

First, the input signal (noisy speech, reverberant speech, or noisy reverberant speech) was divided into several frequency bands by using an auditory filterbank that simulates human frequency selectivity. The auditory filterbank we used is the 6th-order Butterworth infinite impulse response (IIR) filterbank. The bandwidth of each filter was the bandwidth of the human auditory filter, and the order of the filters was determined in accordance with the equivalent rectangular bandwidth ($\mathrm{ERB_N}$) and $\mathrm{ERB_N}$-number scale [16].

$$\mathrm{ERB_N\text{-number}} = 21.4 \log_{10}\left(\frac{4.37f}{1000} + 1\right), \tag{6}$$

where f is the frequency in Hz and the subscript N indicates the characteristics of normal hearing. The $\mathrm{ERB_N}$ approximates the bandwidth of the auditory filter by an equivalent rectangular bandwidth, and $\mathrm{ERB_N}$-number expresses the frequency with its width set to 1.[2]

The speech signal was analyzed, and the NVS was then constructed under the condition that the sixteen band-pass filters were defined from 3 to 35 $\mathrm{ERB_N}$-number with bandwidths of 2 $\mathrm{ERB_N}$, as our audible range. This bandwidth was determined on the basis of our previous studies [8,9].

In each frequency band, the TAE of the signal was extracted using the Hilbert transformation and a 2nd-order Butterworth IIR low-pass filter (LPF) with a cutoff frequency of 64 Hz.

Finally, the TAE in each channel served with the band-limited noise generated by band-pass filtering white Gaussian noise at the same boundary frequency. All amplitude modulated noise was summed to generate the NVS stimulus. The sampling frequency for stimulus creation was unified at 20 kHz to carry out speaker distinction and vocal-emotion recognition under common conditions in a noisy reverberant environment.

[2]ERB scale is an auditory frequency representation with regard to the bandwidth of the auditory filter, $\mathrm{ERB_N}$, such as Bark scale. Thus this is referred to as $\mathrm{ERB_N}$-number. Note that the previous representation was ERB-rate but is currently not used in this research field.

3. EXPERIMENT I: SPEAKER DISTINCTION AND VOCAL-EMOTION RECOGNITION OF NVS IN NOISY ENVIRONMENT

3.1. Conditions

In these experiments, speaker distinction and vocal-emotion recognition were carried out for NVS generated from noisy speech. We investigated the experimental conditions from that under which it is easy to listen to speech stimulus to that under which it is quite difficult to listen, SNR $= \infty$, 20, 15, 10, 5, 0, and -5 dB were set as the noise conditions. Note that SNR $= \infty$ means no noise condition. Therefore, there were seven experimental conditions.

3.2. Speaker Distinction

3.2.1. Procedure

From the results of Kitamura *et al.*'s study [17], the speech data used in this study were selected from the speeches in the ATR Japanese Speech Database c set [18]. These data consist of five pairs with different speaker similarity for male and female pairs; thus, there are 20 speakers. The five pairs we selected are typical speaker pairs which are comprehensively distributed from those with lower to higher similarity, as analyzed by Kitamura *et al.* [17].

Table 1 lists the speaker pairs we used. All the speech data were recorded at a 20-kHz sampling frequency and 16-bit quantization. Each sentence was uttered for about 4 to 5 s.

We used the XAB method for the speaker pairs listed in Table 1 in this experiment. One trial consisted of the following three different speech signals (X, A, B):

X: NVS,

A: NVS with a different sentence of X and the same speaker as X and

B: NVS with a different sentence of X and a different speaker from X,

where the sentences in A and B are different. The above stimulus was presented with a 0.5-s silence interval.

Eight native Japanese speakers (four males and four females, all in their 20s) with normal hearing participated in this experiment. Three different speech signals (X, A, B) were presented to the participants and the participants were then asked to compare the speakers of A and B with the speaker of X to select which one was more similar to speaker X. Both stimuli with XAB and XBA orders were presented to counterbalance any effects due to the order of presentation. Thus, there were four stimulus-presentation patterns for each speaker pair: AAB, ABA, BAB, and BBA. All stimulus sets were presented randomly to the participants. The stimuli were presented only once with no repetition allowed.

The experiment was conducted while the participants were in a sound-proof room. The sound-pressure level of background noise was from 23.7 to 25.8 dB. The stimuli were simultaneously presented to both ears of a participant through a PC (Windows10, matlab), 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). The experiment was conducted with sufficient breaks so that the participants would not become fatigued.

3.2.2. Results

Figure 2 shows the results of the speaker-distinction experiment. The circles indicate the mean of the speaker distinction and error bars indicate the standard error of mean. The mean of speaker distinction decreased when the SNR was low. A one-way repeated-measures analysis of

Table 1 Speaker pairs selected from ATR database and their average similarity index measured by Kitamura *et al.* [17]. Left and right halves show female and male speaker pairs, respectively.

Speaker pair		Similarity
F507	F609	1.45
F407	F702	1.97
F213	F214	2.42
F611	F614	2.93
F606	F704	3.32
M504	M601	1.61
M614	M710	1.83
M214	M519	2.36
M509	M603	2.68
M409	M705	3.38

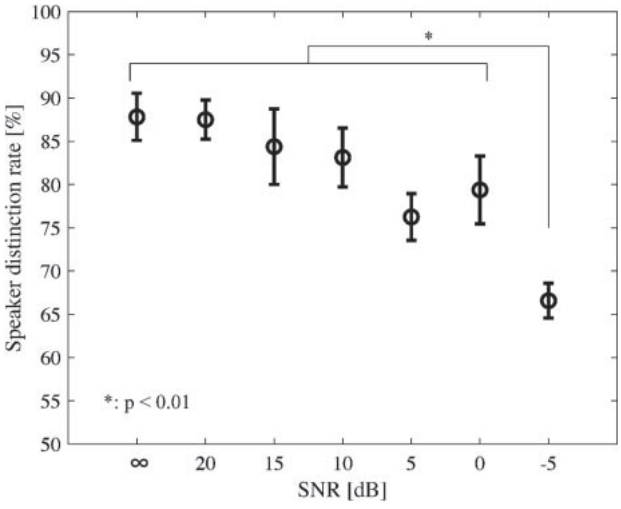

Fig. 2 Results from speaker-distinction experiment in noisy environment.

variance (ANOVA) was conducted on these results with the SNR. There was a main effect ($F(6, 42) = 16.41$, $p < 0.01$) for the noise condition. The results of multiple comparisons using Tukey's test indicate a significant difference ($p < 0.01$) between the SNR of -5 dB and the other SNRs. However, no significant difference was observed at other SNRs.

3.3. Vocal-emotion Recognition

3.3.1. Procedure

The Fujitsu Japanese Emotional Speech Database [19,20] used by Huang & Akagi was used as the original speech data and in the same manner as in the speaker-distinction experiment. In this database, emotional speech signals were recorded so that the speaking rate and voice quality would be uniform throughout the sentence. For each sentence, speech data were included for five emotions (neutral, joy, cold anger, sadness, and hot anger). The speech data were recorded with a sampling frequency of 20 kHz (the original signal was 22.05 kHz but was resampled to 20 kHz to match the conditions of other experiments) and 16-bit quantization, and the duration of each utterance was about 3 to 4 s.

Similar to the speaker-distinction experiment, NVS was generated from noisy speech after adding noise to the original speech. We also used speech signals obtained by adding noise under six noise (SNR) conditions and clean speech. Therefore, there were a total of seven experimental conditions.

Ten native Japanese speakers with normal hearing (seven males and three females, all in their 20s) participated in this experiment. All stimuli were randomly presented to the participants. The participants were asked to forcibly indicate one emotion from the above five emotions. The stimuli were presented only once with no repetition allowed. Five sentences were prepared for each emotion, and NVS stimuli were generated under the above experimental conditions. There were five sentences, five emotions, and seven noise conditions, so the total number of stimuli was 175. The experimental setting was the same as in the speaker-distinction experiment.

3.3.2. Results

Figure 3 shows the results of the emotion-recognition experiment. The horizontal axis shows the SNR and the vertical axis shows the vocal-emotion-recognition rate. The circles indicate the mean of the vocal-emotion recognition rate and error bars indicate the standard error. The vocal-emotion-recognition rate decreased when the SNR was low. The difference between the vocal-emotion-recognition rate when the SNR was -5 dB and that when the SNR was 0 dB was the largest. A two-way repeated-measures ANOVA was conducted on these results with the noise conditions and emotions. There was a main effect of the noise condition ($F(6, 54) = 11.71$, $p < 0.01$). The results of multiple comparisons using Tukey's test indicate a significant difference ($p < 0.05$) between the SNRs of -5 and 0 dB and the higher SNRs. However, there was no significant difference among the other SNRs.

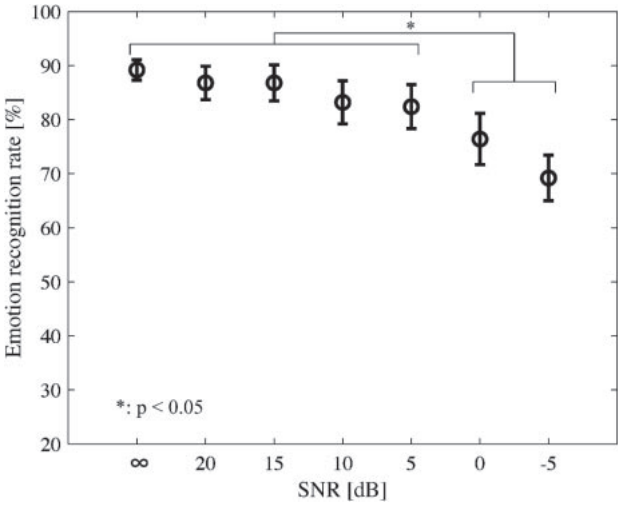

Fig. 3 Results from vocal-emotion-recognition experiment in noisy environment.

The results for each emotion are not shown due to the page limitation, but when the emotion and noise condition were examined, there was a main effect of the emotion ($F(4, 36) = 13.05$, $p < 0.01$). There was also a significant difference in the interaction between noise condition and emotion ($F(24, 216) = 2.470$, $p < 0.01$).

3.4. Discussion

As the ANOVA results for the speaker-distinction experiment indicate that there was a main effect on the noise condition. However, Tukey's test revealed a significant difference only when the SNR was -5 dB or higher. Therefore, it can be considered that the perception of speaker individuality is basically unaffected by noise unless the noise conditions are extremely poor for speech listening (conditions with extremely low SNR).

The ANOVA results for the vocal-emotion recognition experiment indicate that there was a main effect on the noise condition. However, Tukey's test showed no significant difference between these conditions when the SNR was higher than 0 dB. Therefore, it can be considered that vocal-emotion perception is basically unaffected by noise as well as speaker individuality unless the noise condition is extremely poor for speech perception. However, a significant difference was observed between when the SNR was 0 dB or less and the other SNRs, suggesting that noise affects emotion perception when the SNR is 0 dB or less.

It can be concluded that there is basically no effect of noise on the perception of speaker individuality and vocal

emotion. However, it was suggested that noise may have an effect when the SNR becomes extremely low.

4. EXPERIMENT II: SPEAKER-DISTINCTION AND VOCAL-EMOTION RECOGNITION OF NVS IN REVERBERANT ENVIRONMENT

4.1. Condition

In these experiments, speaker distinction and vocal-emotion recognition were conducted with NVS generated from reverberant speech signals. Five room impulse responses with $T_R = 0.1$, 0.2, 0.5, 1.0, and 2.0 s were used as reverberation conditions, and reverberation was added to the original speech by using the method described in Sect. 2.2. Since the condition of no reverberation, that is, $T_R = 0$ s, was added, there were a total of six reverberation conditions. Under these conditions, the stimulating speech of the NVS was generated from the reverberant speech created using the method described in Sect. 2.4.

4.2. Speaker Distinction

4.2.1. Procedure

As in Sect. 3.2.1, for each of the speaker pairs listed in Table 1, this speaker-distinction experiment was conducted using the XAB method. Nine native Japanese speakers (six males and three females, all in their 20s) with normal hearing participated in the experiment. The experimental setting was the same as in Experiment I.

4.2.2. Results

Figure 4 shows the results of the speaker-distinction experiment. The circles indicate the mean of the speaker-distinction rate, and the error bars indicate their standard error. The horizontal axis is T_R, where 0 s means no reverberation. As T_R increased, the speaker-distinction rate tended to decrease. However, the results of a one-way repeated-measures ANOVA revealed no main effect on the reverberation condition ($F(5, 40) = 1.802$, $p = 0.13$).

4.3. Vocal-emotion Recognition

4.3.1. Procedure

As in Experiment I, the Fujitsu Japanese Emotional Speech Database was used in this experiment. Ten native Japanese speakers (seven males and three females, all in their 20s) with normal hearing participated in this experiment. The experimental setting was the same as in Experiment I.

4.3.2. Results

Figure 5 shows the results of the vocal-emotion-recognition experiment. The circles indicate the mean of the emotion-recognition rate for all five emotions, and the error bars indicate the standard error. As in the speaker-distinction experiment, the emotion-recognition rate tended to decrease as T_R increased. A two-way repeated-measures ANOVA was conducted with the reverberation conditions and emotions as factors. The results indicate a main effect for the reverberation condition ($F(5, 45) = 6.201$, $p < 0.01$). A main effect on emotion was also observed ($F(4, 36) = 13.43$, $p < 0.01$), but the interaction between the reverberation condition and emotion was not observed ($F(20, 180) = 1.190$, $p = 0.27$). As a result of Tukey's test for reverberation conditions, a significant difference ($p < 0.05$) was found between a T_R of 2 s and with no reverberation, 0.1 s, and 0.5 s.

4.4. Discussion

The results of the speaker-distinction experiment revealed no main effect of reverberation condition. Unlike the perception of vocal-emotion information, the results

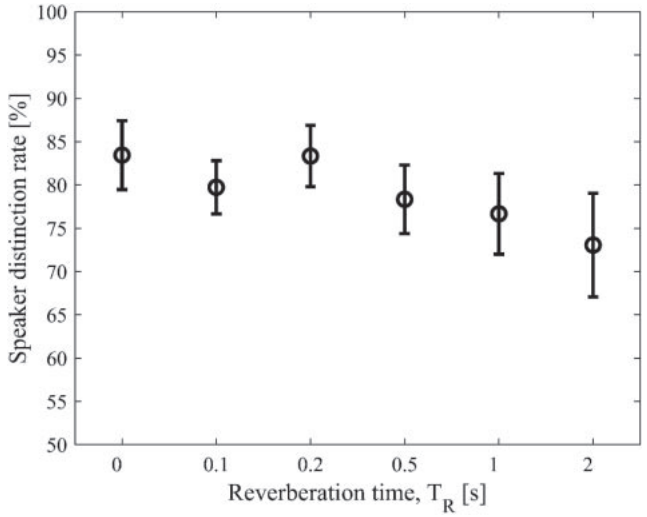

Fig. 4 Results from speaker-distinction experiment in reverberant environment.

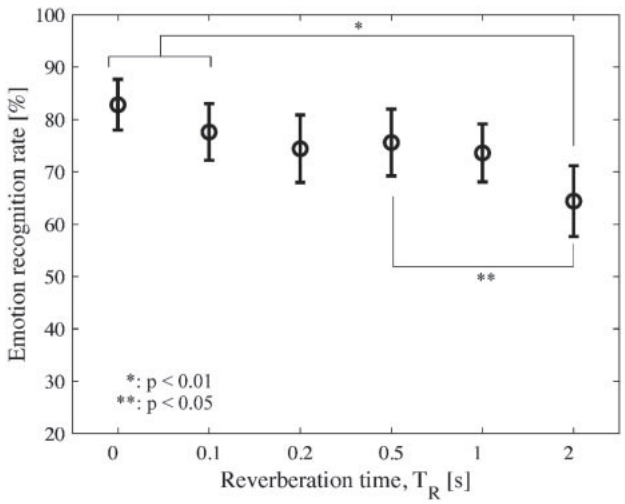

Fig. 5 Results from vocal-emotion-recognition experiment in reverberant environment.

suggest that reverberation did not significantly affect the perception of speaker individuality of NVS. However, the results of the vocal-emotion-recognition experiment indicate a main effect for the reverberation condition, but the results of the Tukey's test indicate that only the T_R of 2 s was significantly different from the other reverberation conditions. Conditions with a T_R of more than 2 s are poor conditions that are not often experienced in daily life. The results suggest that speaker distinction and vocal-emotion recognition of NVS are not affected by reverberation except under poor reverberation conditions. This is very different from previous studies in linguistic information perception, where Tillery *et al.* showed that even short reverberation durations (125 ms) significantly reduced the intelligibility of NVS [10].

From the above, the effect of reverberation on perception of non-linguistic information is smaller than that of linguistic information, and this effect is not obvious.

5. EXPERIMENT III: SPEAKER DISTINCTION AND VOCAL-EMOTION RECOGNITION OF NVS IN NOISY REVERBERANT ENVIRONMENT

5.1. Condition

In these experiments, we investigated non-linguistic information of NVS in a noisy reverberation environment where background noise and room reverberation exist simultaneously. Therefore, this experiment evaluated speaker distinction and vocal-emotion recognition of NVS created from speech signals with added noise and reverberation. Reverberation was first added to the original speech using the method described in Sect. 2.2. Three room impulse responses with T_R of 0.5, 1.0, and 2.0 s were used. Next, five types of stationary noise with SNRs of 20, 10, 5, 0, and -5 dB were added to the reverberant speech using the method described in Sect. 2.1. Finally, the stimulus of the NVS was generated on the basis of the noisy reverberant speech using the method described in Sect. 2.4. There were a total of 15 experimental conditions; three T_R and five SNR conditions. These experiments did not include a clean condition with no noise and no reverberation.

5.2. Speaker Distinction

5.2.1. Procedure

As in Sect. 3.2.1, for each of the speaker pairs listed in Table 1, this speaker-distinction experiment was conducted using the XAB method. Nine native Japanese speakers with normal hearing (six males and three females, all in their 20s) participated in the experiment. The experimental setting was the same as in Experiment I.

5.2.2. Results

Figure 6 shows the results of the speaker-distinction experiment. The circles indicate the mean of the speaker-distinction rate, and the error bars indicate the standard error. Figure 7 shows the contour of the average speaker-distinction rate. A two-way repeated-measures ANOVA was conducted under noise (SNR) and reverberation (T_R) as factors. The results revealed the main effects of SNR ($F(4, 32) = 15.79$, $p < 0.01$) and T_R ($F(2, 16) = 4.888$, $p < 0.05$). There was no interaction between the SNR and T_R ($F(8, 64) = 0.3651$, $p = 0.94$). The decrease in the speaker-distinction rate of NVS can be explained by SNR and T_R.

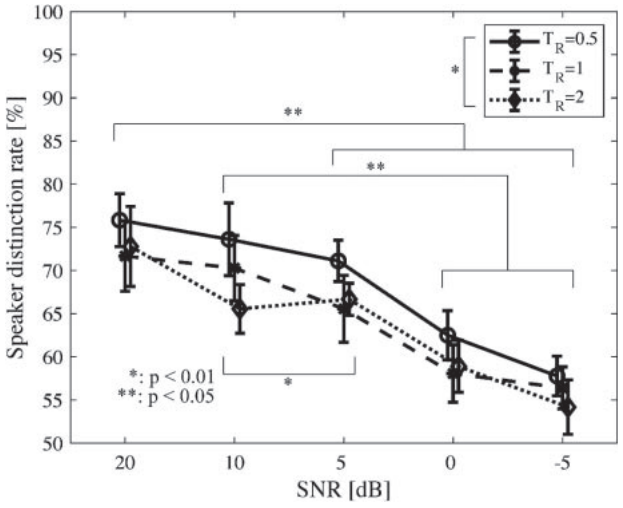

Fig. 6 Results from speaker-distinction experiment in noisy reverberant environment.

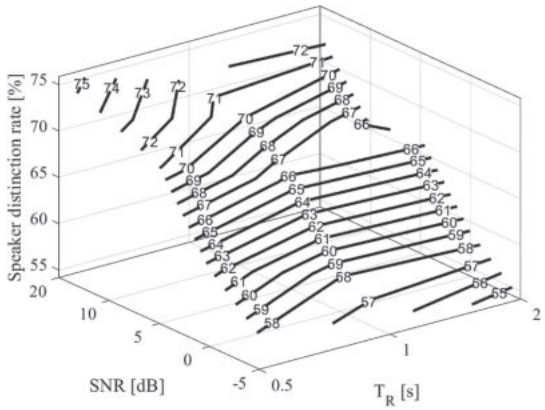

Fig. 7 Contour of results from speaker-distinction experiment in noisy reverberant environment.

5.3. Vocal-emotion Recognition

5.3.1. Procedure

The Fujitsu Japanese Emotional Speech Database was also used in this experiment. Ten native Japanese speakers (seven males and three females, all in their 20s) with normal hearing participated in this experiment. The experimental setting was the same as in Experiment I.

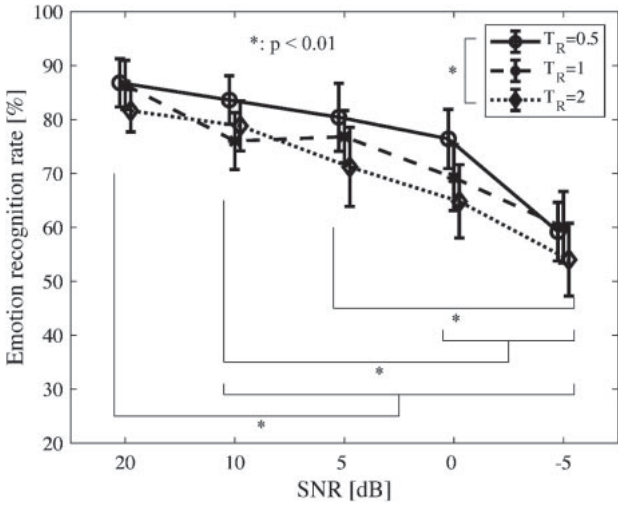

Fig. 8 Results from vocal-emotion-recognition experiment in noisy reverberant environment.

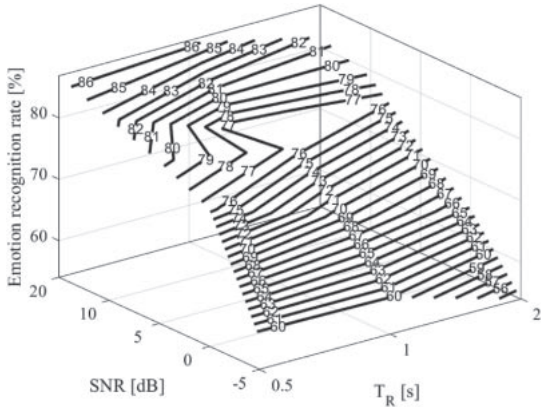

Fig. 9 Contour of results from vocal-emotion-recognition experiment in noisy reverberant environment.

5.3.2. Results

Figure 8 shows the results of the vocal-emotion-recognition experiment. The circles indicate the mean of the emotion-recognition rate, and the error bars show the standard error. Figure 9 shows a contour of the average emotion-recognition rate. A two-way repeated-measures ANOVA was conducted with the SNR and T_R as factors. The results revealed that there were main effects of the SNR ($F(4, 36) = 35.48$, $p < 0.01$) and T_R ($F(2, 18) = 10.57$, $p < 0.01$). There was no interaction ($F(8, 72) = 1.543$, $p = 0.16$) between the SNR and T_R. Therefore, the change in the vocal-emotion recognition of NVS in a noisy reverberation environment can be explained by the SNR and T_R.

5.4. Discussion

The effects of noise and reverberation in Experiment III were similar to those in Experiments I and II. It was shown that there is no effect of noise and reverberation except under extremely low noise and poor reverberation conditions for listening to target speech. In both speaker-distinction and vocal-emotion-recognition experiments, the effects of noise and reverberation were independent of each other within the significance level in this study. In other words, the addition of noise had little effect on the effect of reverberation, and that the addition of reverberation had little effect on the effect of noise. This suggests that there is no interaction between noise and reverberation for the perception of non-linguistic information of NVS.

The contour of the average speaker-distinction rate and vocal-emotion-recognition rate (as shown in Figs. 7 and 9, respectively) showed the existence of a region where the speaker-distinction and vocal-emotion-recognition rates remained almost unchanged. In the speaker-distinction experiment in the noisy environment, the speaker-distinction rate dropped significantly when the SNR was 10 dB or less. In the vocal-emotion-recognition experiment, the SNR decreased moderately between 20 and 0 dB, but the difference between 0 and −5 dB was the largest. Regarding the reverberation environment, it was found that the speaker-distinction rate and vocal-emotion-recognition rate decreased more significantly when the T_R increased from 0.5 to 1 s, but there was almost no change between 1 and 2 s.

6. DISCUSSION

It is known that the perception of linguistic information of speech (speech intelligibility) is significantly affected by noise and reverberation. For example, Plomp and Mimpen indicated a sharp increase in the voice-listening threshold of normal hearing people as the noise level rises [21], and Duquesnoy and Plomp showed that noise and reverberation become the voice-listening threshold of elderly people [22]. Kobayashi and Kondo reported that the intelligibility of Japanese speech decreases due to the effect of noise reverberation [23]. Hazrati and Loizou also reported that cochlear-implant wearers' speech intelligibility significantly decreases in noisy and reverberant environments [24]. All these studies showed the effects of noise and reverberation in different languages for normal-hearing listeners, elderly listeners, hearing-impaired listeners, and cochlear-implant users.

Tillery *et al.* conducted a cochlear-implant speech-comprehension experiment in a noisy reverberant environment using cochlear-implant-simulated speech instead of the original speech for normal-hearing listeners [10]. They found that the intelligibility of the speech was significantly lower than that of the normal speech (original speech) due to the addition of noise and reverberation. This is an important result showing that cochlear-implant user's speech perception is affected more by noise reverberation

than normal-hearing listener's speech perception. However, it is still unclear how noise and reverberation affect cochlear-implant users' perception of non-linguistic information (speaker individuality and vocal emotion).

We investigated the effects of noise and reverberation on the perception of non-linguistic information using NVS as cochlear-implant-simulated speech. Since our results were only from Japanese, they are limited to the perception of non-linguistic information in Japanese speech. However, since non-linguistic information is generally considered to be language-independent and culturally related, the results of this study may be able to explain the non-linguistic information of other languages as well.

We then investigated the relationship between TAE information and noise/reverberation. We used an amplitude modulation model that expresses the signal by decomposing its main components into TAE information (amplitude modulation component) and temporal fine structure (carrier component) [15,25]. We replaced the temporal fine structure with a white Gaussian noise carrier and investigated the effect of TAE information on speech perception using the NVS synthesis shown in Fig. 1.

As defined in Sect. 2.3, the noisy reverberant environment was used in this signal expression by convolving the room impulse response with the original signal then adding the background noise into the reverberant signal. The room impulse response was defined as Schroeder's statistical room impulse response (composed of exponentially attenuated amplitude envelope information and white Gaussian noise carrier), and the background noise was defined as white Gaussian noise. Since these signals are mutually independent, those with reverberation added to the background noise also result in white Gaussian noise. Since the original speech is also represented by an amplitude modulation model using white Gaussian noise as a carrier, as described above, the carrier signal of the original speech is also mutually independent of carrier signals of the background noise and reverberation.

When simulating a noisy reverberant environment, noisy reverberant speech in which the background noise is added to the reverberant speech (the original speech with reverberation) and the noisy reverberant speech in which the reverberation is convolved with the noisy speech (the original speech with noise) are equivalent in the statistical sense. In the representation of the amplitude modulation model, the temporal amplitude envelopes of the original, noisy, reverberant, and noisy reverberant speech signals can be explained by relationships among the temporal amplitude envelopes of the original signal, noise signal, and reverberant impulse response (addition and convolution). For a detailed explanation of these envelopes from the perspective of statistical signal processing, refer to previous studies [15,25].

We also discussed the noise and reverberation conditions related to speech perception. The original speech, background noise (white Gaussian noise), and statistical room impulse response (Schroeder's model) were used with the amplitude modulation model described above. The advantage of using this model is that the SNR for noise conditions and the T_R for reverberation conditions can be used to independently control the noise power and reverberation characteristics, respectively. Therefore, the effect of noise and reverberation on the perception of non-linguistic information in speech can be investigated by focusing on TAEs.

Regarding the speech-listening environment of this study, it is difficult to judge whether it is a daily auditory environment or a poor auditory environment under what SNR and how long the T_R is because this may be a subjective division. On the basis of our previous studies, we considered the auditory environment to be poor if the SNR is less than 0 dB or T_R is 1.0 s or longer.

From the results of classroom acoustic measurements by Sato *et al.* [26], for example, it is known that the average SNR in the classroom during class is about 11 dB, and the average T_R in the classroom with students is 0.41 s. If such an auditory environment can be judged as our daily noisy reverberation environment, perhaps based on the results of previous studies, the intelligibility of linguistic-information perception is slightly lower for normal listeners. However, this situation is considerably lower for hearing-impaired people or cochlear-implant users. From our results, if NVS is interpreted as cochlear-simulated speech, it can be considered that the perception of non-linguistic information of NVS is not affected by noise or reverberation.

There are conditions in which noise and reverberation do not affect the perception of non-linguistic information of NVS. It was suggested that when the SNR and T_R exceed a certain threshold, the effect of noise and reverberation on the perception of non-linguistic information gradually increases.

7. SUMMARY

We conducted speaker-distinction and vocal-emotion-recognition experiments of noise-vocoded speech (NVS) generated from noisy and reverberant speech to investigate the perception of non-linguistic information (speaker individuality and vocal-emotion) of NVS in either a noisy or reverberant environment. Speaker distinction and vocal-emotion recognition were also investigated in a noisy reverberant environment to investigate how noise and reverberation interactively affect the perception of non-linguistic information of NVS. From both speaker distinction and vocal-emotion recognition in noisy and reverberant environments, it was found that perception of non-

linguistic information of NVS are not affected by noisy or reverberant environments except in extremely poor sound environments for listening to target speech. The same trend has been observed in both noisy and reverberant environments except in extremely poor sound environments. It was also found that there is no interaction between noise and reverberation for the perception of non-linguistic information of NVS and that there is a region where the speaker-distinction and vocal-emotion-recognition rates remain almost unchanged. Within the significant levels of the results of these experiments, both background noise and reverberation do not affect the perception of non-linguistic information of NVS in noisy and reverberant environments (SNR is greater than 10 dB and reverberation time is less than 1.0 s) except in extremely poor sound environments.

We will investigate how TAEs play an important role in the perception of non-linguistic information of NVS and investigate the relationship between the TAEs of NVS and modulation-transfer function in noisy reverberant environments to verity the results of this study, according to our previous reports [9,25].

ACKNOWLEDGMENTS

This work was supported by a Grant in Aid for Innovative Areas (No. 16H01669, No. 18H05004), from MEXT, Japan, Grant in Aid for JSPS Fellows (No. 17J08312), and JST-Mirai Program (Grant Number JPMJMI18D1). We would like to thank Mr. Shinichi SEKIYA to contribute our listening experiments.

REFERENCES

[1] R. V. Shannon, F. G. Zeng, V. Kamath, J. Wygonski and M. Ekelid, "Speech recognition with primarily temporal cues," *Science*, **270**(5234), 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.*, **34**, 263–270 (2013).

[3] K. Ueda, T. Araki and Y. Nakajima, "Frequency specificity of amplitude envelope patterns in noise-vocoded speech," *Hear. Res.*, **367**, 169–181 (2018).

[4] P. C. Loizou, M. Dorman and Z. Tu, "On the number of channels needed to understand speech," *J. Acoust. Soc. Am.*, **106**, 2097–2103 (1999).

[5] L. Xu and B. E. Pfingst, "Spectral and temporal cues for speech recognition: Implications for auditory prostheses," *Hear. Res.*, **242**, 132–140 (2008).

[6] H. Mori, K. Maekawa and H. Kasuya, *Speech Science of Emotions, Paralinguistic Information, and Personal Information* (CORONA PUBLISHING CO., LTD., Tokyo, 2014) (in Japanese).

[7] Z. Zhu, Y. Nishino, R. Miyauchi and M. Unoki, "Study on linguistic information and speaker individuality contained in temporal envelope of speech," *Acoust. Sci. & Tech.*, **37**, 258–261 (2016).

[8] 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," *Acoust. Sci. & Tech.*, **39**, 234–242 (2018).

[9] Z. Zhu, R. Miyauchi, Y. Araki and M. Unoki, "Contributions of modulation spectral features on the perception of vocal emotion using noise-vocoded speech," *Acoust. Sci. & Tech.*, **39**, 379–386 (2018).

[10] K. H. Tillery, C. A. Brown and S. P. Bacon, "Comparing the effects of reverberation and of noise on speech recognition in simulated electric-acoustic listening," *J. Acoust. Soc. Am.*, **131**, 416–423 (2012).

[11] M. R. Schroeder, "Modulation transfer functions: Definition and measurement," *Acustica*, **49**, 179–182 (1981).

[12] International Telecommunication Union, "Objective measurement of active speech level," ITU-T, P.56, Switzerland (1993).

[13] J. B. Crespo and R. C. Hendriks, "Speech reinforcement in noisy reverberant environments using a perceptual destortion measure," *Proc. ICASSP 2014*, pp. 910–914 (2014).

[14] X. Feng, Y. Zhang and J. Glass, "Speech feature denoising and dereverberation via deep autoencoder for noisy reverberant speech recognition," *Proc. ICASSP 2014*, pp. 1778–1782 (2014).

[15] M. Unoki and X. Lu, "Unified denoising and dereverberation method used in restoration of MTF-based power envelope," *Proc. Int. Symp. Chinese Spoken Language Processing* (*ISCSLP 2012*), pp. 215–219, Hong Kong (2012).

[16] B. C. J. Moore, *An Introduction to the Psychology of Hearing*, 6th ed. (Brill Academic Publishers, Leiden, 2013).

[17] T. Kitamura, T. Nakama, H. Ohmura and H. Kawamura, "Measurement of perceptual speaker similarity for sentence speech in ATR speech database," *J. Acoust. Soc. Jpn.* (*J*), **71**, 516–525 (2015) (in Japanese).

[18] T. Takezawa, A. Nakamura and E. Sumita, "Databases for Conversation Speech Translation Research at ATR," *J. Phon. Soc. Jpn.*, **4**(2), pp. 16–23 (2000) (in Japanese).

[19] C.-F. Huang and M. Akagi, "A three-layered model for expressive speech perception," *Speech Commun.*, **50**, 810–828 (2008).

[20] M. Akagi, "Emotion recognition in speech: How do we describe an emotion space?" *J. Acoust. Soc. Jpn.* (*J*), **66**, 393–398 (2010) (in Japanese).

[21] R. Plomp and A. M. Mimpen, "Speech-reception threshold for sentences as a function of age and noise level," *J. Acoust. Soc. Am.*, **66**, 1333–1342 (1979).

[22] A. J. Duquesnoy and R. Plomp, "Effect of reverberation and noise on the intelligibility of sentences in cases of presbyacusis," *J. Acoust. Soc. Am.*, **68**, 537–544 (1980).

[23] Y. Kobayashi and K. Kondo, "Japanese speech intelligibility estimation and prediction using objective intelligibility indices under noisy and reverberant conditions," *Appl. Acoust.*, **165**, 327–335 (2019).

[24] O. Hazrati and P. C. Loizou, "The combined effects of reverberation and noise on speech intelligibility by cochlear implant listeners," *Int. J. Audiol.*, **51**, 437–443 (2012).

[25] M. Unoki and Z. Zhu, "Relationship between contributions of temporal amplitude envelope of speech and modulation transfer function in room acoustics to perception of noise-vocoded speech," *Acoust. Sci. & Tech.*, **41**, 233–244 (2020).

[26] H. Sato and J. S. Bradley, "Evaluation of acoustical conditions for speech communication in working elementary school classrooms," *J. Acoust. Soc. Am.*, **123**, 2064–2077 (2008).