Relationship between contributions of temporal amplitude envelope of speech and modulation transfer function in room acoustics to perception of noise-vocoded speech

Masashi Unoki* and Zhi Zhu†

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

Abstract: Speech signals can be represented as a sum of amplitude-modulated frequency bands. This sum can also be regarded as a temporal amplitude envelope (TAE) with temporal fine structure. Our previous studies using noise-vocoded speech (NVS) showed that the TAE of speech plays an important role in the perception of linguistic information (speech intelligibility) as well as non-linguistic information (e.g., vocal-emotion recognition). It was found that the upper limit of the modulation frequency from 4 to 8 Hz on the TAE is important for speech intelligibility, while that from 8 to 16 Hz is important for vocal-emotion recognition. However, speech intelligibility generally dramatically degrades due to reverberation. The concept of the modulation transfer function (MTF) takes into account the relationship between the transfer function in an enclosure in terms of input and output TAEs and characteristics of the enclosure under reverberant conditions. This concept was introduced as a measure in room acoustics for assessing the effect of an enclosure on speech intelligibility. For this study, we conducted two experiments involving word intelligibility tests and vocal-emotion recognition with NVS under reverberant conditions to investigate the relationship between the contributions of the TAE of speech and MTF of reverberation to modulation perception of NVS. We also pointed out that the straightforward scheme, i.e., the relationship between the contributions of the static features (peak/slope) in the modulation spectrum (MS) of speech and MTF of reverberation, cannot consistently account for the auditory perception of both linguistic and non-linguistic information obtained from these perceptual data of NVS under reverberant conditions. We then developed a scheme in which the relationship between the contributions of the temporal MS features and MTF of reverberation to modulation perception can consistently account for these perceptual data of NVS.

Keywords: Temporal amplitude envelope, Modulation transfer function, Noise-vocoded speech, Vocal-emotion recognition, Speech intelligibility, Temporal modulation-spectral feature

PACS number: 43.64.Bt, 43.66.Ba, 43.71.Bp, 43.71.Gv [doi:10.1250/ast.41.233]

1. INTRODUCTION

Speech is, at any time, used as an important and natural means of communication for humans to express linguistic as well as non-linguistic information. In particular, non-linguistic information, such as emotion, gender, age, and speaker individuality, is used for rich speech communications among humans, while linguistic information is used to convey a message in communications. Important features related to linguistic and non-linguistic information are redundantly contained in the speech signals. Therefore, humans can easily and correctly recognize such information even if some of the features are smeared due to noise and reverberation.

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

Important acoustical features, such as spectral and temporal fine structures (harmonicity and periodicity), formants, spectral tilt, and temporal power fluctuation, have been studied in both time and frequency domains based on the source filter model and are essential for speech perception. The temporal amplitude envelope (TAE) and temporal fine structure (TFS) also play important roles in auditory perception [1]. In particular, the role of TAE and TFS information was reviewed from aspects of auditory perception, i.e., pitch perception, sound localization, and perception of speech in background sounds. The important role of monaural and binaural TAE and TFS information was also discussed with regard to the effects of hearing loss and age.

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 (MS) conveys linguistic information of speech [3].

Recent psychoacoustical studies based on the noise-vocoded speech (NVS) scheme have revealed that the TAE of speech plays an important role in speech perception [4–7]. NVS is generated by replacing the TFS with band-limited noise, so the spectral cue decreases dramatically and the temporal cue is preserved. Shannon et al. reported that the presentation of modulation information of only a few acoustic bands, such as NVS, is sufficient for speech recognition [4]. Tachibana et al. also reported that spectral and temporal resolutions relatively contribute to the perception of syllables, words, and sentences in NVS [5]. Therefore, humans can successfully perceive linguistic information (speech intelligibility) using the TAE of speech.

The quality of speech transmission, however, must be evaluated to design the required room acoustics [8]. This would require many subjective experiments, resulting in high costs. Therefore, from the view point of auditory perception in room acoustics, objective indices and measurements in room acoustics are needed to inexpensively assess the quality and intelligibility of speech [8,9].

To this end, the articulation index (AI), degree of contribution of early reflections (or early decay time (EDT)), Deutlichkeit (early to total sound energy ratio: \(D_{30}\)), clarity (early to late arriving sound energy ratio: \(C_{50}\)), and other acoustic parameters (e.g., reverberation time: \(T_{30}\) and \(T_{60}\)) have been used to assess the quality of speech transmissions [8,9].

The concept of the modulation transfer function (MTF) was introduced as a measure in room acoustics for assessing the effect of the enclosure on speech intelligibility [10,11]. The speech transmission index (STI) is a particularly important objective measurement that can be used to assess the quality of speech transmission in room acoustics [12]. It has also recently become known that the correlation between listening difficulty ratings and STI is the strongest of all tested objective measures [13,14]. Methods of calculating the STI are currently standardized by IEC 60268-16 [15] on the MTF concept [10,11].

The amplitude-modulation domain, however, is an important dimension in hearing, i.e., “amplitude modulation,” since there is modulation-frequency selectivity in the auditory system [16]. The modulation-detection task has been typically used to investigate modulation masking, i.e., the listener is required to discriminate a sinusoidal amplitude-modulated signal from a flat fluctuated signal. In this case, the smallest detectable modulation depth as a function of modulation frequency is also referred to as the temporal MTF (TMTF) [17,18].

Modern psychoacoustical studies of temporal amplitude-modulation processing suggest that the TAE is processed using a modulation filterbank [19–22]. A filterbank can be regarded as a bank of overlapping “modulation filters,” i.e., analogous to the auditory filterbank, to be tuned to a different modulation frequency. The estimated bandwidth of the modulation filters is roughly 20 Hz and roughly equals the center modulation frequency in modulation perception [16].

The MTF used in room acoustics indicates the transfer function of modulation depth between the source and receiver sides, while the TMTF indicates the ability of modulation perception. The lower modulation components (under 20 Hz) are highly related to speech perception, especially speech intelligibility. Our research question is what role do the modulation frequencies (TAE information) play in hearing for the temporally smeared speech signal in real environments? That is, how can humans perceive linguistic and non-linguistic information by using TAE information in real environments, from the view point of “amplitude modulation” from the source (speaker) to receiver via room acoustics (transmission system)?

In this paper, we assume that speech transmission in real environments can be regarded as auditory perception based on amplitude modulation, that is, modulation perception (Fig. 1). We then introduce TAE information and/or MS features of speech for auditory perception as an important role of speech transmission in our daily environments. Thus, we aimed to investigate the important role of temporal amplitude-modulation features for speech perception by consistently considering the MS of speech, MTF in room acoustics, and modulation perception (or TMTF).

This paper is organized as follows. In Sect. 2, we introduce speech transmission and modulation perception based on amplitude-modulation perception among speech representation, speech transmission in room acoustics, and the perceptual effect on amplitude modulation. In Sect. 3, we describe the signal-processing method for generating NVS. In Sects. 4 and 5, we describe our two experiments involving word intelligibility tests and vocal-emotion recognition under reverberant conditions using NVS. In Sect. 6, we discuss the contributions of temporal MS cues and MTF of reverberation to systematically and consistently account for the results of word intelligibility and vocal-emotion recognition with NVS although the relation-
ship between contributions of MS and MTF cannot systematically and consistently account for these results. In Sect. 7, we conclude the paper.

2. SPEECH TRANSMISSION BASED ON AMPLITUDE MODULATION

Figure 1 shows a block diagram of speech transmission based on the concept of amplitude modulation. Input (source: speaker), system (transmission system), and output (receiver: listener) in this diagram correspond to the MS of speech, MTF in room acoustics, and modulation perception on the auditory filterbank. This block diagram is used to consistently and systematically investigate the relationship between the contributions of the TAE of speech and MTF of reverberation to modulation perception. In this section, we explain the details of three key concepts related to the input/system/output in Fig. 1.

2.1. Modulation Spectrum

Recent studies by Greenburg et al. [3,23] revealed that the TAE or its MS conveys linguistic information of speech, as shown in Fig. 2 (figure was originally published in [23]). This figure shows the relationship between the distribution of syllable durations for fifteen minutes of spontaneous speech, as shown in Fig. 2(a), and the corresponded MS (for an octave band between 1 and 2 kHz), as shown in Fig. 2(b). In particular, low-frequency modulations (corresponding to one over syllable duration) of sound have been shown to be the fundamental carriers of information in speech. The implications of this figure are that the MS largely reflects syllabic modulation of spontaneous speech.

Drullman [2], for example, investigated the importance of modulation frequencies for intelligibility by applying low-pass and high-pass filters to the temporal envelopes of acoustic frequency sub-bands. They showed frequencies between 4 and 16 Hz to be important for intelligibility, with the region around 4–5 Hz (peak in the MS) being the most significant. In a similar study, Arai et al. [24] showed that applying band-pass filters between 1 and 16 Hz does not impair speech intelligibility.

The MS represents how the whole power spectrum changes as a function of time, as shown in Fig. 3 (figure was originally published in [3]). The MS of speech can be derived from the short-term sound spectrogram by spectrally analyzing the mean removed time-trajectory of the sound spectrogram at specific frequencies around 1 to 2 kHz. It is these temporal changes that convey most of the linguistic information of speech.

In the above intelligibility studies, the lower limit of 1 Hz stems from the fact that the slow vocal-tract changes do not convey much linguistic information. In addition, the lower limit helps to make speech communication more robust since the majority of noise occurring in nature varies slowly as a function of time; hence, the MS is dominated by modulation frequencies below 1 Hz. The upper limit of 16 Hz is due to the physiological limitation on how fast the vocal tract is able to change with time.

Helmansky historically reviewed this point of view for automatic speech recognition then argued that modulation features and the RelAtive SpecTrAl (RASTA) filter are important for machine recognition [25]. The MS has recently been re-recognized as an important feature.

2.2. Modulation Transfer Function

Houtgast and Steeneken proposed a prediction method for assessing the effects of an enclosure on the intelligibility of speech in both noisy and reverberant environments by using the MTF concept [5]. This concept has been used to account for the relation among the degree of modulation...
of the envelopes of input and output signals, characteristics of an enclosure, and predicting the STI, as shown in Fig. 4, which is strongly related to intelligibility. The correspondence between the STI and assessed quality of speech transmission is summarized in Table 1.

The complex MTF $M(\omega)$ is defined as

$$M(\omega) = \frac{\int_0^\infty h^2(t) e^{j\omega t}dt}{\int_0^\infty h^2(t)dt},$$

(1)

where $h(t)$ is the impulse response of room acoustics (diffused field) and $\omega$ is the radian frequency [10,11]. This equation means the complex Fourier transform of the squared impulse response is divided by its total energy. Let us consider the impulse response of a room acoustics:

$$h(t) = \exp\left(-\frac{6.9t}{T_R}\right)c(t),$$

(2)

where the response has an envelope of exponential decay and white noise carrier $c(t)$ and $T_R$ is the reverberation time, that is, the time required for the power of $h(t)$ to decay by 60 dB [26]. This is the well-known stochastic approximated impulse response in room acoustics [11]. The MTF $m(\omega)$ can be obtained as

$$m(\omega) = |m(\omega)| = \left(1 + \left(\frac{\omega T_R}{13.8}\right)^2\right)^{-\frac{1}{2}}.$$  

(3)

Figure 5 shows MTF as a function of the modulation frequency $f_m$ (the dominant frequency in the temporal envelope). These theoretical curves were calculated by substituting five $T_R$s (0.1, 0.3, 0.5, 1.0, and 2.0) and $\omega = 2\pi f_m$ into Eq. (3). The $m(\omega)$ can also be regarded as the modulation index (degree of relative fluctuation in the normalized amplitude) with respect to the modulation frequency, $f_m$. These curves show how much the modulation index of the envelope will be reduced from 1.0 to 0.0 depending on the $T_R$ at a specific $f_m$. The STIs derived from the five MTFs, $m(2\pi f_m)$s, by using [5], are shown in Table 2. From these results, we can predict that the assessed quality of speech transmission in Fig. 5 is from fair to excellent.

On the one hand, $T_R$ can also be predicted from a specific $m(2\pi f_m)$ at a specific $f_m$. This is one advantage of using the MTF. Based on this concept, we can determine how much reverberation reduces the modulation index then predict reduced speech intelligibility using the MTF concept. Most temporal-deconvolution methods are aimed at restoring the reduced MTF then enhancing speech intelligibility using it. We previously proposed a method of restoring the MTF for speech intelligibility based on this idea [27,28].

The MTF concept [5] has also been used to predict the STI under noisy reverberant conditions. Since Tachibana et al. used white Gaussian noise as background stationary noise [5], the theoretical curve of under noisy conditions was derived as $1/(1 + 10^{-\text{SNR}/10})$, where SNR is signal-to-noise ratio. The MTF under noisy conditions is independent of $f_m$ and reduced as a function of SNR. The total MTF under noisy reverberant conditions was derived as multiplication of both MTFs under reverberant conditions and noisy conditions (see Fig. 6 in [29]). This property was also used to propose a blind STI-estimation method under noisy reverberant environments based on the MTF concept [29].

### Table 1 Relation between speech-transmission quality and STI.

| Quality | Bad | Poor | Fair | Good | Excellent |
|---------|-----|------|------|------|-----------|
| STI     | 0.0–0.3 | 0.3–0.45 | 0.45–0.6 | 0.6–0.75 | 0.75–1.0 |

### Table 2 STIs derived from Modulation transfer functions (MTFs) in Fig. 5.

| $T_R$ (s) | 0.1 | 0.3 | 0.5 | 1.0 | 2.0 |
|-----------|-----|-----|-----|-----|-----|
| STI       | 0.95 | 0.88 | 0.80 | 0.7  | 0.53 |

---

![Fig. 4 MTF concept.](image)

![Fig. 5 Theoretical curves representing MTF $m(\omega)$ under various conditions with reverberation time $T_R$.](image)
2.3. Modulation Filterbank

The main function of the human auditory system is to decompose sound signals into frequency components (i.e., frequency selectivity), as shown in Fig. 6. It is well known that this frequency selectivity involves nonlinear signal processing. The masking data of various masking situations have been corrected to find nonlinear frequency selectivity [30]. Nonlinear auditory filterbank, whose function is equivalent to that of the human auditory system, has recently been proposed (e.g., [31]).

The auditory filterbank decomposed the sound into the channel signals (TAE & TFS) in the time-frequency domain. Half-wave rectification and low-pass filtering were then carried out as the mechanisms for inner hair cells and neural firing to extract TAEs from the channel signals. Finally, a band-pass filterbank was used as the “modulation filterbank” to represent the modulation spectrogram from the TAEs. In other words, the human auditory peripheral system involves processes such as band division and temporal envelope extraction, as shown in Fig. 7. Amplitude modulation is an important dimension in hearing since there is modulation-frequency selectivity in the human auditory system.

There is growing psychoacoustic and physiological evidence to support the significance of the modulation domain in the analysis of speech signals [32]. Experiments conducted by Bacon and Grantham, for example, showed that there are channels in the human auditory system that are tuned to detect modulation frequencies [33]. Yost and Sheft showed that our perception of temporal dynamics corresponds to our perceptual filtering into $f_m$ channels and that faithful representation of these modulations is critical to our perception of speech [34]. Experiments conducted by Schreiner and Urbas showed that a neural representation of amplitude modulation is preserved through all levels of the mammalian auditory system, including the highest level of audition, i.e., the auditory cortex [35]. Neurons in the auditory cortex are thought to decompose the acoustic spectrum into spectro-temporal modulation content (spectro-temporal MTF) and are best driven by sounds that combine both spectral and temporal modulations [36].

3. SIGNAL PROCESSING METHOD: NOISE-VOCODED SPEECH

Modern psychophysical models of temporal-modulation processing suggest that the TAE is processed using a modulation filterbank [3]. Therefore, in the human auditory system, modulation-frequency analysis should be conducted to extract the linguistic and non-linguistic information included in the TAE of speech.

In previous studies, we investigated the contribution of TAE cues on the perception of linguistic and non-linguistic information using NVS [37–39]. For speech recognition, the results showed that the speech-recognition rate drastically decreases as the maximum modulation frequency of TAE goes below 5 Hz [37]. For vocal-emotion recognition, the largest variances at the modulation frequencies, lower than 16 Hz, were also observed in all bands [37–39]. The results from these experiments indicate that the vocal-emotion recognition rates drastically decrease as the modulation frequency goes below 8 Hz. These results also indicate that auditory modulation filtering affects the perception of both linguistic information (speech intelligibility) and non-linguistic information (vocal emotion perception) and there are dominant modulation regions for linguistic information (below 5 Hz) and non-linguistic information (below 8 Hz).

As we have already mentioned in Sect. 1, our research question is whether our daily sound environments affect speech perception, especially in the sense of TAE perception. Room acoustics, such as strong reverberation, generally smear important cues for speech perception, dramatically decreasing speech intelligibility [40]. However, our previous studies investigated the important role of TAE for speech intelligibility and vocal-emotion recog-
nition under clean conditions. Thus, we are not sure whether strong reverberation also affect vocal-emotion recognition.

Our next study in this paper, therefore, is to investigate whether reverberation affects NVS perception for linguistic as well as non-linguistic information. Thus, in this paper, we conducted speech intelligibility tests and vocal-emotion recognition under reverberant environments to consistently and systematically investigate the relationship between the contributions of the TAE of speech and MTF of reverberation to modulation perception for NVS with regard to linguistic and non-linguistic information.

Figure 8 shows a schematic diagram of noise-vocoder method used to generate stimuli (BPF: band-pass filter; LPF: low-pass filter; and NBN: narrow-band noise).

In the NVS scheme, we decomposed a speech signal into ERB\textsubscript{N}-bands by using an auditory-motivated filterbank and modulated temporal envelopes of ERB\textsubscript{N}-bands with noise-carriers of the same bands [30]. We implemented an auditory motivated filterbank as a 6th order Butterworth infinite impulse response (IIR) filterbank using ERB\textsubscript{N} and ERB\textsubscript{N}-number scale [30]. This scale is comparable to that of distance along the basilar membrane, so the frequency resolution of the human auditory system can be faithfully replicated by dividing frequency bands in accordance with the ERB\textsubscript{N}-number. The relationship between ERB\textsubscript{N}-number and acoustic frequency is defined as

\[
\text{ERB}_N\text{-number} = 21.4 \log_{10} \left( \frac{4.37f}{1000} + 1 \right),
\]

where \( f \) is acoustic frequency in Hz. The boundary frequencies of the band-pass filters were defined from 3 to 35 ERB\textsubscript{N}-number with bandwidths of 2, 4, or 8 ERB\textsubscript{N}. Therefore, the number of channels of the band-pass filterbank was 16, 8, or 4. The number of channels determines the frequency resolution of NVS: higher determines that the higher resolution will be obtained with more channels.

We then extracted the TAE of the output signal from each band-pass filter using the Hilbert transformation and used a low-pass filter (2nd order Butterworth IIR filter). The cut-off frequency of the low-pass filter determined the upper limit of modulation frequency. This frequency relates to the temporal resolution that higher TAE will be obtained with a higher upper limit of modulation frequency. Dau \textit{et al.} used a modulation filterbank to control the envelopes of octave-bands from 2 to 64 Hz [3]. Therefore, we investigated the commonalities and individual differences in modulation aspects by analyzing the MSs in all bands, for all stimuli from different speech content and speakers.

Figure 9 shows an example of modulation spectrogram for clean speech. Red trajectory in Fig. 9 is the derived MS at the specific auditory filter from the TAE in Fig. 7. These modulation spectrograms for NVS under reverberant conditions are used to investigate the important role of TAE based on the MTF concept and modulation perception for speech intelligibility and vocal-emotion perception of NVS.

![Fig. 8](image_url) Schematic diagram of noise-vocoder method used to generate stimuli (BPF: band-pass filter; LPF: low-pass filter; and NBN: narrow-band noise).

![Fig. 9](image_url) Example of modulation spectrogram for speech.
4. SPEECH INTELLIGIBILITY

4.1. Signal Generation

The test materials in this experiment were chosen from the Familiarity-controlled Word-lists (FW07) [41]. The speech signal of the word list had a sampling frequency of 48 kHz and quantization of 16 bits. The words were composed of four morae. The familiarities of all words that we used were at the highest and lowest rates. To eliminate the effect of familiarity on recognizing linguistic information, all words were used only once under all conditions.

All word stimuli were obtained by convolving the original signal with \( h(t) \) in Eq. (2). Five \( T_R (0.1, 0.2, 0.5, 1.0, \text{and} 2.0 \text{s}) \) were used. Thus, there were six conditions, i.e., no reverberation and five reverberant. All signals were then processed using a noise-vocoder method to generate the reverberant stimuli. Speech signals were first divided into 16 acoustic frequency bands using the same band-pass filterbank described in Sect. 3. Then, the temporal envelope was extracted using Hilbert transformation and a low-pass filter (2nd-order Butterworth IIR filter). The cut-off frequency of the low-pass filter was 64 Hz. The temporal envelope of each band was used to amplitude modulate the noise limited in the same band. Finally, all the amplitude-modulated band-limited noise was summed to generate the noise-vocoded reverberant speech stimuli.

4.2. Procedure

Seven Japanese speakers (5 males and 2 females) participated in this experiment. All participants had normal hearing (hearing levels were below 12 dB in the frequency range from 125 to 8,000 Hz).

All NVS stimuli were randomly presented to both ears of a participant through a PC, audio interface (RME, Fireface UCX), and headphones (SENNHEISER HDA 200) in a sound-proof room. The sound-pressure levels of background noise were lower than 25.8 dB. The sound-pressure levels of the output from the headphones were calibrated to a comfortable level (about 65 dB) by using a head and torso simulator (B&K, type 4128) and sound-level meter (B&K type 2231). There were 20 trials in a session (2 familiarities \( \times \) 10 words \( \times \) 6 conditions). The words for all trials were different. Participants were asked to input a word of four morae as they understood it by using a keyboard. Each stimulus was presented only once.

4.3. Results

Figure 10 shows the mean value and standard deviation of the word-recognition rates from the word intelligibility tests. The average recognition rates decreased when the familiarity level decreased from highest to lowest. The results of the analysis of variance (ANOVA) indicated that there were significant main effects of \( T_R \) (at the highest familiarity level, \( F(5, 30) = 40.60, p < 0.05 \); at the lowest familiarity level, \( F(5, 30) = 9.433, p < 0.05 \)). Moreover, regarding post hoc Turkey’s honestly significant difference (HSD) tests, there were significant differences between the \( T_R s \) (0–0.5 s) and 1.0 s at both the highest and lowest familiarity levels. These results indicate that reverberation smears linguistic information on the TAE, significantly reducing the word intelligibility score, independent from the familiarity level.

5. VOCAL-EMOTION RECOGNITION

5.1. Signal Generation

The emotional speech data were selected from the Fujitsu Japanese Emotional Speech Database [37,38]. This database includes five emotions (neutral, joy, cold anger, sadness, and hot anger) expressed by one professional actress. The same sentence was spoken with these five emotions. Ten utterances of each emotion were selected. The linguistic content of each sentence was semantically emotion-neutral to minimize any biasing of context. The duration of each utterance was about 3 or 4 s. The sampling frequency and quantization bits were 22.05 kHz and 16 bits.

All emotional speech signals were obtained by convolving the original signal with room impulse response \( h(t) \) in Eq. (2). Five \( T_R s \) (0.1, 0.2, 0.5, 1.0, and 2.0 s), were used. Thus, there were the six conditions, no reverberation and five reverberant conditions. All signals were then processed in the same manner as that mentioned in Sect. 4.1.

5.2. Procedure

Ten native Japanese speakers (6 males and 4 females) participated in this experiment. All participants had normal hearing (hearing levels were below 12 dB in the frequency range from 125 to 8,000 Hz).
The NVS stimuli were presented in the same manner as that mentioned in Sect. 4.2. All NVS stimuli were randomly presented to the participants. Participants were asked to indicate which of the five emotions (neutral, joy, cold anger, sadness, and hot anger) he/she thought was associated with the stimulus. Each stimulus was presented only once. Before the experiment, we confirmed that all participants could correctly recognize the emotion of the original speech.

5.3. Results

Figure 11 shows the mean value and standard deviation of the vocal-emotion recognition rates under reverberant conditions. Figure 11(a) shows the averaged emotion recognition rates for all emotions, while Figs. 11(b)–11(f) show the vocal-emotion recognition rates for each emotion.

We conducted two-way repeated measures ANOVA on these results with $T_R$ and emotion as the factors. There was a significant effect of $T_R$ ($F(5, 45) = 6.212$, $p < 0.01$). Moreover, from post hoc Turkey’s HSD tests, there were significant differences between $T_{RS}$ (0–0.5 s) and 2.0 s ($p < 0.05$). However, there were no significant differences under the other conditions. Analyses of the simple main effects of emotion types ($F(4, 36) = 13.434$, $p < 0.01$) showed that there is no significant interaction between emotion type and $T_R$ ($p = 0.46$). This suggests that the perception of vocal emotion with NVS significantly differs depending on the emotion and that reverberation does not affect vocal emotion perception with NVS.

6. GENERAL DISCUSSION

6.1. Characteristics and Trends

From the key concepts related to “amplitude modulation” in Fig. 4 and our previous studies of NVS [37,38], the following important characteristics were summarized.

(1) The MS at the lower modulation frequencies from 1 to 16 Hz plays an important role of speech perception. The peak in the MS around 4–5 Hz is the most significant for speech intelligibility [2,3,24].

(2) Auditory modulation filtering on the TAE affects the perception of both linguistic (speech intelligibility) and non-linguistic information (vocal-emotion recognition) of NVS, and there are dominant modulation regions for linguistic information (4–8 Hz) and non-linguistic information (8–16 Hz) (see Sect. 3) [37,38].

(3) The peak in the MS around 4–5 Hz does not significantly affect vocal-emotion recognition of NVS while significantly affecting the word-recognition rate of NVS from characteristics (1) and (2).

(4) The MTF of reverberation has characteristics of low-pass filtering on the TAE as a function of the $f_m$ and $T_R$ (see Fig. 5).

The following trends were found from the results of the two experiments involving word-intelligibility tests and vocal-emotion recognition under reverberant environments.

(i) Linguistic information (speech intelligibility) of NVS can be correctly perceived under clean conditions (no reverberation). The word-recognition rate depends on familiarity.
(ii) Non-linguistic information (vocal-emotion recognition) of NVS can also be correctly perceived under clean conditions. The recognition rate of cold anger was the worst among the five emotions.

(iii) Word-recognition rates significantly decreases as $T_R$ increases while vocal-emotion recognition rates do not significantly decrease as $T_R$ increases.

From these characteristics and trends, we predicted that the perception of NVS can account for smearing of the MS features of reverberant NVS derived by multiplying the MS of clean NVS and the MTF of reverberation as a function of $T_R$ in Eq. (3). Thus, we discuss whether the relationship between the contributions of the MS of speech and MTF of reverberation to modulation perception of NVS for speech intelligibility and vocal-emotion recognition by using a straightforward scheme based on the above idea.

6.2. Straightforward Scheme

This straightforward scheme is shown in Fig. 12. For example, under clean conditions (no reverberation), $m(\omega) = 1.0 \ (0 \text{ dB})$ did not affect the MS of the original NVS so that the MS feature, i.e., peak around 4–5 Hz on the MS, can be perceived in NVS. Under the long reverberant condition ($T_R = 1.0$), however, $m(\omega) \approx 0.5 \ (-3 \text{ dB})$ at $f_m = 4 \text{ Hz}$ significantly affected the MS of the original NVS. As a result, the MS of reverberant NVS decreased due to the MTF, so that the peak around 4–5 Hz on the MS decreased and tuning of the peak broadened, as shown in Fig. 12. This straightforward scheme can consistently account for the reduction in the word-recognition rate under reverberant environments by focusing on the relationship between the effect of the TAE of speech (peak around 4–5 Hz) and that of MTF in reverberant environments from the results presented in Sect. 4.

However, this straightforward scheme cannot consistently account for the shallow reduction in the vocal-emotion recognition rate under reverberant conditions by focusing on the relationship between the effects of TAE of speech and those of reverberation on NVS. The right slope in the MS (around 8–16 Hz) of reverberant NVS dramatically decreased as $T_R$ increased due to the MTF in reverberation environments; however, the vocal-emotion recognition rate did not dramatically decrease as $T_R$ increased.

From our previous findings (characteristics (2) and (3)), we considered that the peak in the MS does not affect decreasing or keeping the emotion-recognition rate under reverberant conditions. Therefore, this straightforward scheme should be improved by carefully reconsidering the modulation perception to enable taking account of both word-recognition and emotion-recognition rates under reverberation environments.

6.3. Temporal Modulation-spectral Features

We previously reported that temporal modulation-spectral features derived from the temporal-modulation spectrogram (time, acoustical frequency, and modulation frequency domains), as shown in Figs. 6 and 7, can be clarified to categorize emotion types according to the discriminability index ($d'$) in vocal-emotion recognition [42]. In particular, higher-order moments, such as modulation spectral centroid (MSCR), modulation spectral spread (MSSP), modulation spectral skewness (MSSK), modulation spectral kurtosis (MSKT), and modulation spectral tilt (MSTL), are useful modulation-spectral features. These features correspond to the balance, spread, asymmetry, sharpness, and slope of TAE-distribution functions on the temporal modulation spectrogram between acoustic frequency (number of auditory filters) and modulation frequency (number of modulation filters).

Figure 13 shows an example of calculating the modulation spectral centroid MSCR in the acoustic frequency domain (MSCR$_m$, $m = 1,2,\ldots,6$) and the modulation frequency domain (MSCR$_k$, $k = 1,2,3,4$). Here, $k$ is an index of the acoustic frequency band (4 bands from 3 to 35 ERB$_N$-number with bandwidths of 8 ERB$_N$) and $m$ is an...
index of the modulation frequency band (6 bands from 2 to 64 Hz with octave band). Figure 14 shows, for example, the estimated probability-distribution function of modulation spectral centroid $\text{MSCR}_m$. This indicates that the $\text{MSCR}_m$ of hot anger was highest and that of sadness was lowest in the 4th modulation frequency channel (8–16 Hz). In this case, we found that the $d_0$'s of these temporal modulation-spectral features in vocal emotion types are highly correlated with those derived from the results of vocal-emotion recognition, as shown in Fig. 15 [42]. Thus, $d_0$'s were obtained from the results of vocal-emotion recognition, as mentioned in Sect. 5, and based on the hit rates and false-alarm rates derived from the confusion matrix. Table 3 shows the $d_0$'s of the five emotions on NVS under reverberant conditions as a function of $T_R$. Variations in the $d_0$ of each emotion with various $T_R$'s were very small. This means that the contributions of the TAE with reverberant NVS to modulation perception for vocal-emotion recognition do not vary with $T_R$.

Since the highest similarity between $d_0$'s derived from the results of vocal-emotion recognition and those derived from the modulation-spectral features was obvious under clean conditions, as shown in Fig. 15 [42], we can predict that the temporal modulation-spectral features (MSCR, MSSSP, MSSK, MSKT, and MSTL) of the five emotions obtained from reverberant NVS do not vary with $T_R$ from the same analogy. Therefore, this finding can be regarded as another trend that consistently accounts for modulation perception of linguistic and non-linguistic information with NVS under reverberant conditions.

### 6.4. Proposed Scheme

Based on this trend, we propose a scheme to resolve the issue that appeared in the straightforward scheme by using temporal modulation-spectral features. Figure 16 shows the proposed scheme, which accounts for the modulation perception of linguistic and non-linguistic information on the MS obtained from a modulation filterbank (right). In the proposed scheme, the temporal modulation-spectral features on the temporal-modulation spectrogram (especially around $m = 4$, i.e., modulation frequencies of 8–16 Hz) obtained from the modulation filterbank in Fig. 6 can be used to address the issue instead of the temporal-MS.

The proposed scheme can consistently account for not only the relationship between the reduction in the word-

![Fig. 14 Estimated probability-distribution function of MSCR (MSCR$_m$, $m = 4$) under clean conditions [42].](image)

![Fig. 15 Similarity of each modulation-spectral feature (taken across all acoustic or modulation frequency channels) [42].](image)

![Fig. 16 Proposed scheme for modulation perception.](image)

| $T_R$ (s) | Neutral | Joy | Cold Anger | Sadness | Hot Anger | Ave. |
|----------|---------|-----|------------|---------|-----------|------|
| 0.0      | 2.17    | ∞   | 2.13       | 3.24    | 3.16      | —    |
| 0.1      | 2.25    | 3.10| 1.80       | 2.17    | 2.48      | 2.36 |
| 0.2      | 1.97    | 3.49| 1.59       | 2.28    | 2.68      | 2.40 |
| 0.5      | 2.07    | 2.28| 2.20       | 3.01    | 2.85      | 2.48 |
| 1.0      | 1.97    | 3.16| 1.75       | 2.11    | 2.79      | 2.36 |
| 2.0      | 1.46    | 2.33| 1.23       | 2.14    | 2.33      | 1.90 |

Table 3 Discriminability index in vocal-emotion recognition.
recognition rate and the effects of reverberation in NVS, from the results in Sect. 4, but also the relationship between the shallow reduction in the vocal-emotion recognition rate and the effects of MTF of reverberation in NVS, from the results in Sect. 5. In these cases, the temporal modulation-spectrogram of reverberant NVS is an important representation in the proposed scheme so that the temporal modulation-spectral features (blue fluctuation) shown in Fig. 16 can be regarded as important information on the TAEs for vocal-emotion recognition of NVS in reverberation environments.

7. SUMMARY

We introduced literature reviews of the modulation spectrum of speech, modulation transfer function (MTF) in room acoustics (reverberation), and modulation perception by using a modulation filterbank to investigate the temporal amplitude envelope (TAE) of speech for auditory perception as an important role of speech transmission in our daily sound environments. We conducted two experiments involving word-intelligibility tests and vocal-emotion recognition under reverberant conditions via noise-vocoded speech (NVS) to investigate whether the TAE plays an important role in the perception of linguistic as well as non-linguistic information and whether the effects of reverberation affect our perception of this information with NVS. We found that the perception of linguistic information is significantly affected by reverberation while the perception of non-linguistic information is not significantly affected by reverberation. We also showed that a straightforward scheme unfortunately cannot account for these results. We then developed a scheme in which the relationship between the contributions of the temporal modulation spectral features of speech and MTF of reverberation to modulation perception can consistently account for the auditory perception of linguistic information as well as non-linguistic information via NVS from the viewpoint of the important role of the TAE.

ACKNOWLEDGMENTS

This work was supported by a Grant in Aid for Scientific Research (A) (No. 25240026), Innovative Areas (No. 16H01669, No. 18H05004) from MEXT, Japan, the Mitsubishi Research Foundation, and JST-Mirai Program (Grant Number: JP15MJ18D1). This work was also supported by JSPS KAKENHI Grant Number JP. 17J08312.

REFERENCES

[1] B. C. J. Moore, “The roles of temporal envelope and fine structure information in auditory perception,” Acoust. Sci. & Tech., 40, 61–83 (2019).
[2] R. Drullman, “Temporal envelope and fine structure cues for speech intelligibility,” J. Acoust. Soc. Am., 97, 585–592 (1995).
[3] L. Atlas, S. Greenberg and H. Hermansky, “The modulation spectrum and its application to speech science and technology,” Interspeech 2007, Tutorial (2007).
[4] R. V. Shannon, F. G. Zeng, V. Kamath, J. Wygonski and M. Ekelid, “Speech recognition with primarily temporal cues,” Science, 270, 303–304 (1995).
[5] 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).
[6] 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).
[7] L. Xu and B. E. Pfingst, “Spectral and temporal cues for speech recognition: Implications for auditory prostheses,” Hear. Res., 242, 132–140 (2008).
[8] H. Kuttruff, Room Acoustics, 3rd ed. (Elsevier Science Publishers Ltd., London, 1991).
[9] ISO 3382, Acoustics—Measurement of the Reverberation Time of Rooms with Reference to Other Acoustical Parameters, 2nd ed., Genève (1997).
[10] T. Houtgast and H. J. M. Steeneken, “The modulation transfer function in room acoustics as a predictor of speech intelligibility,” Acustica, 28, 66–73 (1973).
[11] T. Houtgast, H. J. M. Steeneken and R. Plomp, “Predicting speech intelligibility in rooms from the modulation transfer function. I. General room acoustics,” Acustica, 46, 60–72 (1980).
[12] Y. Toida, “Speech intelligibility in sound fields,” J. Acoust. Soc. Jpn. (J), 51, 312–316 (1995).
[13] H. Sato, M. Morimoto, H. Sato and M. Wada, “Relationship between listening difficulty and acoustical objective measures in reverberation fields,” J. Acoust. Soc. Am., 123, 2087–2093 (2008).
[14] H. Sato, M. Morimoto, H. Sato and M. Wada, “Relationship between listening difficulty and objective measures in reverberant and noisy fields for young adults and elderly persons,” J. Acoust. Soc. Am., 131, 4596–4605 (2012).
[15] IEC 60268-16:2011. Sound system equipment—Part 16: Objective rating of speech intelligibility by speech transmission index (2011).
[16] S. D. Ewert and T. Dau, “Characterizing frequency selectivity for envelope fluctuations,” J. Acoust. Soc. Am., 108, 1181–1196 (2000).
[17] S. P. Bacon and N. F. Vienmeister, “Temporal modulation transfer functions in normal-hearing and hearing-impaired subjects,” Audiology, 24, 117–134 (1985).
[18] C. J. Plack, The Sense of Hearing (Lawrence Erlbaum Associates, New York, 2005).
[19] T. Dau, D. Puschel and A. Kohlrausch, “A quantitative model of the effective signal processing in the auditory system. I. Model structure,” J. Acoust. Soc. Am., 99, 3615–3622 (1996).
[20] T. Dau, D. Puschel and A. Kohlrausch, “A quantitative model of the effective signal processing in the auditory system. II. Simulations and measurements,” J. Acoust. Soc. Am., 99, 3623–3631 (1996).
[21] T. Dau, B. Kollmeier and A. Kohlrausch, “Modeling auditory processing of amplitude modulation. I. Detection and masking with narrow-band carriers,” J. Acoust. Soc. Am., 102, 2892–2905 (1997).
[22] T. Dau, B. Kollmeier and A. Kohlrausch, “Modeling auditory processing of amplitude modulation. II. Spectral and temporal integration,” J. Acoust. Soc. Am., 102, 2906–2919 (1997).
[23] S. Greenberg, H. Carvey, L. Hitchcock and S. Chang, “Temporal properties of spontaneous speech—A syllable-
centric perspective,” *J. Phon.*, **31**, 465–485 (2003).

[24] T. Arai, M. Pavel, H. Hermansky and C. Avendano, “Syllable intelligibility for temporally filtered LPC cepstral trajectories,” *J. Acoust. Soc. Am.*, **105**, 2783–2791 (1999).

[25] H. Hermansky, “History of modulation spectrum in ASR,” *Proc. ICASSP 2010*, pp. 5458–5461 (2010).

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

[27] M. Unoki, M. Furukawa, K. Sakata and M. Akagi, “An improved method based on the MTF concept for restoring the power envelope from a reverberant signal,” *Acoust. Sci. & Tech.*, **25**, 232–242 (2004).

[28] M. Unoki, K. Sakata, M. Furukawa and M. Akagi, “A speech dereverberation method based on the MTF concept in power envelope,” *Acoust. Sci. & Tech.*, **25**, 243–254 (2004).

[29] M. Unoki, A. Miyazaki, S. Morita and M. Akagi, “Method of blindly estimating speech transmission index in noisy reverberant environments,” *J. Inf. Hiding Multimed. Signal Process.*, **8**, 1430–1445 (2017).

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

[31] M. Unoki, T. Irino, B. Glasberg, B. C. J. Moore and R. D. Patterson, “Comparison of the roex and gammachirp filters as representations of the auditory filter,” *J. Acoust. Soc. Am.*, **120**, 1474–1492 (2006).

[32] J. Xiang, D. Poeppel and J. Z. Simon, “Physiological evidence for auditory modulation filterbanks: Cortical responses to concurrent modulations,” *J. Acoust. Soc. Am.*, **133**, EL7–12 (2013).

[33] S. P. Bacon and D. W. Grantham, “Modulation masking: Effects of modulation frequency, depth, and phase,” *J. Acoust. Soc. Am.*, **85**, 2575–2580 (1989).

[34] W. A. Yost and S. Sheft, “Modulation detection interference: Across-frequency processing and auditory grouping,” *Hear. Res.*, **79**, 45–58 (1994).

[35] C. E. Schreiner and J. V. Urbas, “Representation of amplitude modulation in the auditory cortex of the cat. I. The anterior auditory field (AAF),” *Hear. Res.*, **21**, 227–241 (1986).

[36] T. Chi, Y. Gao, M. C. Guyton, P. Ru and S. Shamma, “Spectro-temporal modulation transfer functions and speech intelligibility,” *J. Acoust. Soc. Am.*, **106**, 2719–2732 (1999).

[37] 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).

[38] 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).

[39] Z. Zhu, R. Miyauchi, Y. Araki and M. Unoki, “Modulation spectral features for predicting vocal emotion recognition by simulated cochlear implants,” *Proc. Interspeech 2016*, pp. 262–266 (2016).

[40] 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).

[41] T. Kondo, S. Amano, S. Sakamoto and Y. Suzuki, “Development of familiarity-controlled word-lists (FW07),” *IEICE Tech. Rep.*, **107**, 43–48 (2008).

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