applied sciences

MDPI

Article

Contribution of Common Modulation Spectral Features to Vocal-Emotion Recognition of Noise-Vocoded Speech in Noisy Reverberant Environments

Taiyang Guo [1], Zhi Zhu [2], Shunsuke Kidani [1] and Masashi Unoki [1,*]

[1] Japan Advanced Institute of Science and Technology, 1-1 Asahidai, Nomi 923-1292, Japan
[2] Fairy Devices Inc., 7F Yushima Urban Bldg., 2-31-22 Bunkyo-ku, Tokyo 113-0034, Japan
* Correspondence: unoki@jaist.ac.jp

Abstract: In one study on vocal emotion recognition using noise-vocoded speech (NVS), the high similarities between modulation spectral features (MSFs) and the results of vocal-emotion-recognition experiments indicated that MSFs contribute to vocal emotion recognition in a clean environment (with no noise and no reverberation). Other studies also clarified that vocal emotion recognition using NVS is not affected by noisy reverberant environments (signal-to-noise ratio is greater than 10 dB and reverberation time is less than 1.0 s). However, the contribution of MSFs to vocal emotion recognition in noisy reverberant environments is still unclear. We aimed to clarify whether MSFs can be used to explain the vocal-emotion-recognition results in noisy reverberant environments. We analyzed the results of vocal-emotion-recognition experiments and used an auditory-based modulation filterbank to calculate the modulation spectrograms of NVS. We then extracted ten MSFs as higher-order statistics of modulation spectrograms. As shown from the relationship between MSFs and vocal-emotion-recognition results, except for extremely high noisy reverberant environments, there were high similarities between MSFs and the vocal emotion recognition results in noisy reverberant environments, which indicates that MSFs can be used to explain such results in noisy reverberant environments. We also found that there are two common MSFs ($MSKT_k$ (modulation spectral kurtosis) and $MSTL_k$ (modulation spectral tilt)) that contribute to vocal emotion recognition in all daily environments.

Keywords: modulation spectral feature; vocal emotion recognition; noise-vocoded speech; noisy reverberant environment

Citation: Guo, T.; Zhu, Z.; Kidani, S.; Unoki, M. Contribution of Common Modulation Spectral Features to Vocal-Emotion Recognition of Noise-Vocoded Speech in Noisy Reverberant Environment. *Appl. Sci.* **2022**, *12*, 9979. https://doi.org/10.3390/app12199979

Academic Editor: Douglas O'Shaughnessy

Received: 1 September 2022
Accepted: 24 September 2022
Published: 4 October 2022

1. Introduction

The ability to correctly identify the emotion of a speaker is an indispensable aspect of our daily lives, especially in situations where the speaker's facial expressions and gestures are not visible (when only the voice is observed). Thus, to obtain high-quality communication, it is necessary to clarify the mechanism of vocal emotion recognition.

Previous studies on vocal emotion recognition were based on acoustic features and sound patterns. For example, one study investigated the acoustic features on the basis of the source-filter model (fundamental frequency (F0), speaking rate, intensity, duration, etc.) to illustrate that the confusion patterns were close to those found for listener-judges [1]. To model the expressive perceptions of speech, Huang and Akagi proposed a three-layered model with acoustic features in the bottom layer [2]. The results confirmed significant relationships between expressive speech, semantic primitives, and acoustic features.

The acoustic features required for the vocal emotion recognition of some listeners cannot be satisfied, such as for listeners who use cochlear implants (CIs). CIs have proved to be very useful artificial organs for listeners with hearing loss because CIs enable users to achieve speech intelligibility at a high level in a clean auditory environment. In a study on vocal emotion perception involving listeners with CIs, acoustic features had difficulty

accounting for the human response from the CI listeners [3]. The probable reason is that CIs provide the temporal amplitude envelope (TAE) information as a primary cue but with the loss of spectro-temporal detail, which is needed to support the perception of harmonic pitch. As a result, pitch-dominant aspects of speech, such as vocal emotion and lexical tone recognition, are weakly transmitted by CIs [4]. For the normal cochlea, speech is decomposed into narrowband signals, each with a relatively slowly varying TAE and temporal fine structure (TFS). As illustrated in previous research, certain hearing problems, such as hearing loss with age, affect the processing of TFS information but do not affect the processing of TAE information [5]. The above insights have revealed the insufficiency of acoustic features, as they leave out the important contributions of TAE to the vocal emotion recognition of listeners. They also highlight the need to clarify the important role of TAE information and its cues in vocal emotion recognition. It is beneficial to improve the vocal-emotion-recognition quality of CI listeners and to elucidate the vocal-emotion-recognition mechanism of normal hearing.

To clarify how TAE information works in the perception of speech information, both modern physiological [6] and psychological [7] evidence suggest the existence of a modulation filterbank in the auditory system, which is used to analyze the modulation-frequency components of the TAE. In the peripheral hearing system, the cochlea converts speech signals from the middle ear into the corresponding neural signals [5]. The signal processes in the peripheral auditory system can be computationally modeled as a band-pass filterbank, envelope extraction, and amplitude compression [8,9]. On the basis of this peripheral auditory system, previous studies have proved that TAE information of speech signals plays an important role in vocal emotion recognition [10–13].

Zhu et al. used auditory-based modulation analysis to extract the modulation spectral features (MSFs) from emotional speech and investigated the contribution of MSFs to vocal emotion recognition using noise-vocoded speech (NVS) in a clean environment with no noise and no reverberation [14,15]. The results indicated that there were high similarities between MSFs and the results from vocal-emotion-recognition experiments, suggesting that the MSFs of TAE are useful in accounting for the perceptual processing of vocal-emotion with NVS. Unfortunately, daily sound environments contain noise and reverberation, and it is well known that noise and reverberation have significant effects on the perception of speech. For CI listeners, a previous study found that the comprehension of simulated speech from CIs was significantly lower than that of the normal speech (original speech) due to the addition of noise and reverberation [16]. Previous research related to the perception of nonlinguistic information of NVS suggested that, except for extremely poor sound conditions, in daily environments (signal-to-noise ratio (SNR) is greater than 10 dB and reverberation time (T_R) is less than 1.0 s), there were no significant effects of noise and reverberation in nonlinguistic information recognition [17].

However, how the noise and reverberation affect the contribution of MSFs to the perception of vocal-emotion is still unclear. We used qualitative and quantitative analyses, focusing on the relationship between the results of vocal-emotion-recognition experiments and MSFs to investigate the following two research questions:

1. Can MSFs be used to explain the vocal-emotion-recognition results in noisy reverberant environments?
2. Are there common MSFs that contribute to vocal emotion recognition in all daily environments?

This paper is organized as follows: Section 2 introduces the results of vocal-emotion-recognition experiments obtained from previous research. Section 3 presents the analysis of these vocal emotion recognition experiments. Section 4 presents the analysis of MSFs. Section 5 describes the qualitative and quantitative analyses of the relationship between MSFs and the vocal-emotion-recognition results. Section 6 discusses the contribution of common MSFs to vocal emotion recognition. Section 7 summarizes the results.

2. Results of Vocal-Emotion-Recognition Experiments

2.1. Speech Data

The results of vocal-emotion-recognition experiments were collected from a previous study [17]. Zhu et al. used the Fujitsu Japanese Emotional Speech Database as the original speech data. In this database, a professional actress's sentences are expressed, and each sentence contains one of five emotions (neutral, joy, cold anger, sadness, and hot anger). In these experiments, the experimental stimuli of NVS were in three environments: noisy, reverberant, and noisy reverberant. The experimental stimuli were created with the following procedure.

To produce noisy speech, Zhu et al. used stationary noise (white Gaussian noise), and the adjusted noise was added to the speech so that the SNRs of the original speech and noise would differ. As the noise conditions, SNRs of ∞, 20, 15, 10, 5, 0, and -5 dB were selected as the noise conditions, and SNR = ∞ means a clean environment without noise. Therefore, there were a total of seven noise conditions.

To produce reverberant speech, they used a statistical room-impulse response (Schroeder model) [18] and convoluted five types of room-impulse responses with T_R of 0.1, 0.2, 0.5, 1.0, and 2.0 s into the original speech to make experimental stimuli in the reverberant environment. Since a no-reverberation environment, i.e., T_R = 0 s, was added, there were a total of six reverberation conditions.

For the noisy reverberant environment, a reverberant speech was created by convolving three types of room-impulse responses with T_R of 0.5, 1.0, and 2.0 s into the original speech. Five types of constant noise (white Gaussian noise) with SNRs of 20, 10, 5, 0, and -5 dB were then added to the reverberant speech. From the combination of the above cases, there were a total of 15 reverberation conditions.

After adding noise and reverberation, the experimental stimuli of NVS were created. NVS is speech obtained by driving TAE information using band-limited random noise as a carrier signal [15,17,19]. NVS is primarily used to simulate the perception of speech as heard from CIs for normal-hearing listeners. Even in daily noise reverberation environments (SNR is greater than 10 dB and T_R is less than 1.0 s), the impact of these types of disturbances on experimental emotional perception using NVS was not significant [17]. Figure 1 shows the generation of the NVS stimuli. As the input signal, noisy reverberant emotional speech was divided into several frequency bands by using an auditory filterbank that simulates human frequency selectivity. The 6th-order Butterworth infinite impulse response (IIR) band-pass filters (BPFs) were used as the auditory filterbank. The bandwidth of each filter was that of the human auditory filter, and the order of the filters was determined in accordance with the equivalent rectangular bandwidth (ERB_N) and ERB_N-number scale [20]. The unit of ERB_N-number is Cam. The relationship between ERB_N-number and acoustic frequency is defined as

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

where f is the frequency in Hz, and subscript N indicates the characteristics of normal hearing. The signal was then constructed at the boundary frequencies of the BPFs, which were defined as an ERB_N-number from 3 to 35 Cam with bandwidths of 2 ERB_N, and the number of channels was 16.

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

The TAE in each channel was then 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.

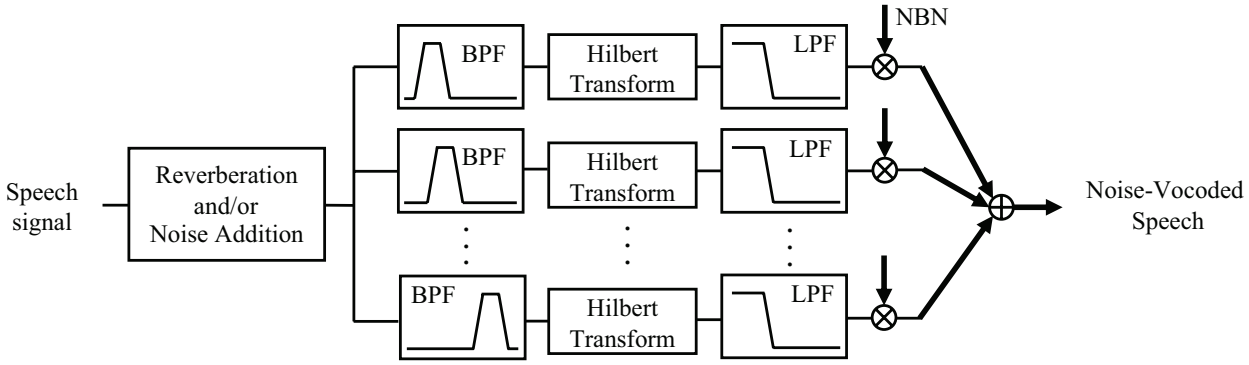

Figure 1. Schematic diagram of noise-vocoded method used for generating experiment stimuli (NBN: narrow band noise). BPFs were defined as ERB_N-number from 3 to 35 Cam with bandwidths of $2ERB_N$. Number of channels was 16 [17].

2.2. Participants and Procedure

In Zhu et al.'s experiment, all experimental stimuli were presented to the participants randomly, and they were instructed to choose one emotion from five categories. The stimuli were presented only once, and no repetition was allowed. Five sentences were prepared for each emotion.

Ten native Japanese speakers with normal hearing (seven men and three women, all in their 20 s) were recruited for the experiments.

2.3. Results of Vocal Emotion Recognition Experiments in Noisy and/or Reverberant Environments

Figure 2 shows the results of the vocal-emotion-recognition experiment in the noisy environment conducted by Zhu et al. [17]. The horizontal axis shows the SNR, and the vertical axis shows the vocal-emotion-recognition rate. In Figures 2–4, the circles indicate the mean of the vocal-emotion-recognition rate, and the error bars indicate the standard error. Except in extremely high noisy environments, there were no significant effects of noise that could affect vocal emotion recognition under daily noisy conditions (SNR is greater than 0 dB).

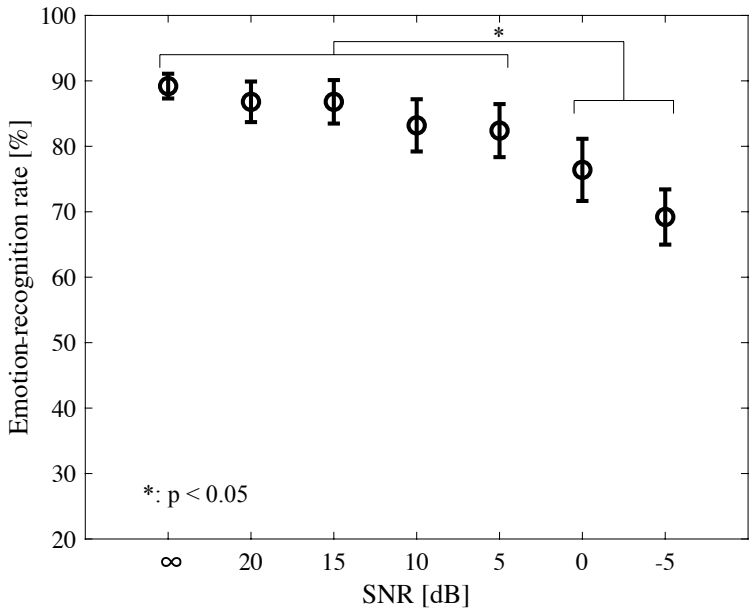

Figure 2. Vocal−emotion−recognition results in noisy environments [17].

Figure 3 shows the results from the vocal-emotion-recognition experiment in the reverberant environment. The horizontal axis shows the T_R conditions, and the vertical axis shows the vocal-emotion-recognition rate. Except in extremely high noisy reverberant environments, there were no significant effects of reverberation that could affect vocal emotion recognition under daily reverberant conditions (T_R is less than 2.0 s).

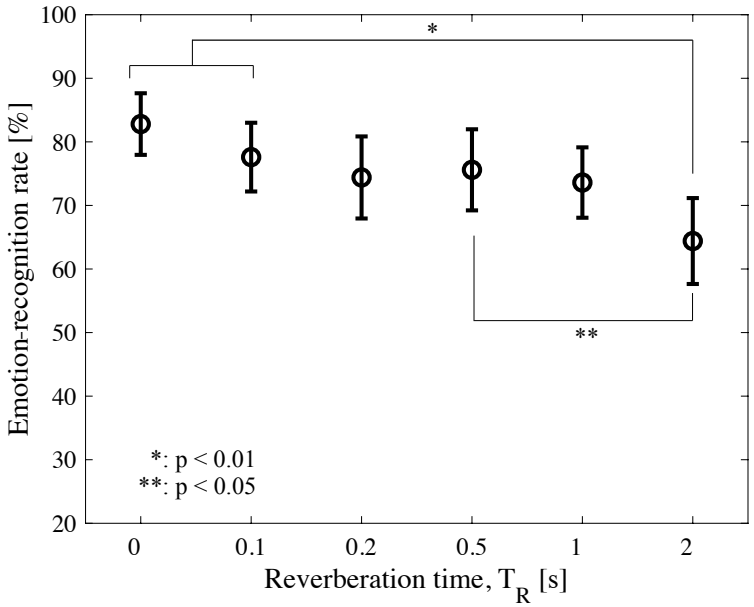

Figure 3. Vocal−emotion−recognition results in reverberant environments [17].

Figure 4 shows the results of the vocal-emotion-recognition experiments in the noisy reverberant environment. Except in extremely high noisy reverberant environments, there were no significant effects of noise and reverberation that could affect vocal emotion recognition under daily noise and reverberation conditions (SNR is greater than 10 dB and T_R is less than 1.0 s).

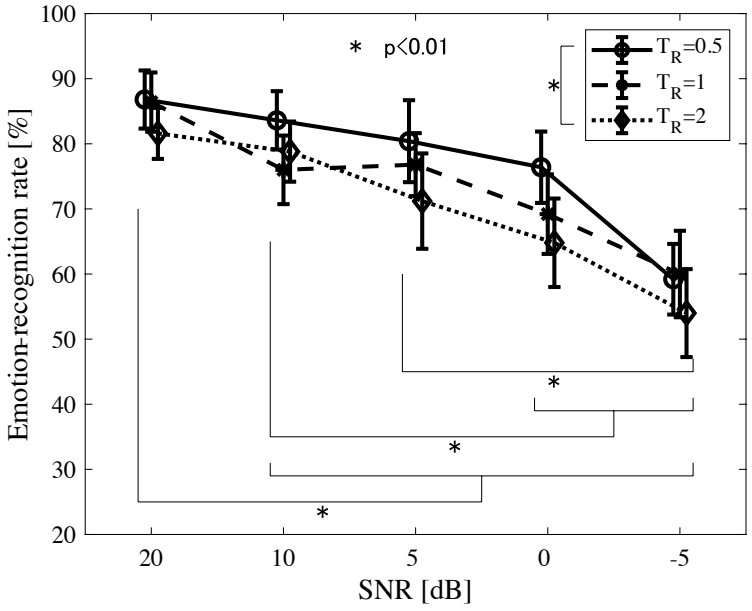

Figure 4. Vocal−emotion−recognition results in noisy reverberant environments [17].

From above vocal-emotion-recognition experiments in noisy reverberant environments, it was found that except under extremely high noisy reverberant conditions, there were no significant effects of noise and reverberation in vocal emotion recognition under daily noise and reverberant conditions.

3. Analysis of Vocal-Emotion-Recognition Results

The calculation and analysis approach used in this study were the same in a previous study [15]. As shown in Figure 5, to better understand the above vocal-emotion-recognition results, we calculated the discriminability index (d'_{P}) for emotion recognition from the confusion matrix of the results. The d'_{P} were based on the hit rates and false-alarm rates derived from the confusion matrix as follows:

$$d'_{\mathrm{P}} = \mathcal{Z}(H) - \mathcal{Z}(F), \tag{2}$$

where H and F are the hit rate and false-alarm rate, respectively, and $\mathcal{Z}(\cdot)$ is the inverse of the normal distribution function. Generally, high d'_{P} values are derived from high hit rates and low false-alarm rates.

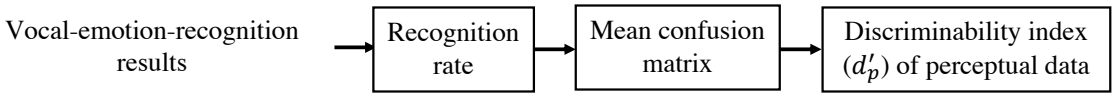

Figure 5. Process of calculating the discrimination index of experimental results.

4. Analysis of Modulation Spectral Features

To extract MSFs from emotional speech signals, it is first necessary to calculate the modulation spectrogram using a modulation filterbank. Figure 6 shows the modulation process we used. Emotional speech signals s were divided into several frequency bands by using an auditory-based band-pass filterbank:

$$s_k(n) = s(n) * h_k(n), \tag{3}$$

where * denotes the convolution operator, $h_k(n)$ is the impulse response of the kth channel, and n is the sample number in the time domain. The bandwidth and boundary frequencies of the 6th-order Butterworth IIR BPFs were the same as those of the auditory filterbank mentioned in Section 2.1. The boundary frequencies of the BPFs were defined as the $\mathrm{ERB_N}$-number from 3 to 35 Cam with an 8 $\mathrm{ERB_N}$ bandwidth, and the number of channels was four.

The temporal envelope of the output signal from each BPF $s_k(n)$ was extracted using the Hilbert transform, and a 2nd-order Butterworth IIR LPF (cut-off frequency 64 Hz) was used as follows:

$$e_k(n) = \mathrm{LPF}[|s_k(n) + j\mathcal{H}[s_k(n)]|], \tag{4}$$

where $\mathcal{H}$ denotes the Hilbert transform.

The next step involved decomposing the temporal envelope into several modulation-frequency bands by using a modulation filterbank:

$$E_{k,m}(n) = g_m(n) * (e_k(n) - \overline{e_k(n)}), \tag{5}$$

where m is the channel number of the modulation filter, $g_m(n)$ is the impulse response of the modulation filterbank, and $\overline{e_k(n)}$ is the time-averaged amplitude of $e_k(n)$. The modulation filterbank consisted of six filters (one LPF and five BPFs). The boundary frequencies of the filters were spaced on an octave frequency band from 2 to 64 Hz.

The root-mean square (RMS) of $E_{k,m}(n)$ was calculated as the modulation spectrogram,

$$\bar{E}_{k,m} = \sqrt{\frac{1}{N}\sum_{n=1}^{N} E_{k,m}^2(n)}, \tag{6}$$

where N is the length of the speech signal $s(n)$. The $\bar{E}_{k,m}$ was then used to calculate modulation spectral features later.

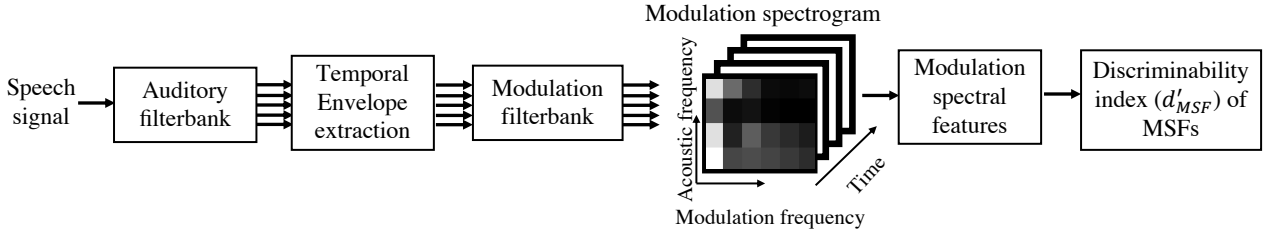

Figure 6. Process of calculating the discrimination index of MSFs [15].

In the above analysis, the modulation spectrogram of modulation speech signals and the confusion matrix of the vocal-emotion-recognition results could provide qualitative evidence to clarify the contribution of modulation information of TAE to vocal emotion recognition. In the subsequent analysis, the similarity between the discriminability index of the vocal-emotion-recognition results (d'_{P}) and MSFs (d'_{MSF}) was then calculated to provide quantitative evidence.

Ten MSFs should be extracted from the modulation spectrograms. Ten are those in the acoustic-frequency domain (the subscript is m) and in the modulation-frequency domain (the subscript is k): the modulation spectral centroid ($\mathrm{MSCR}_{m/k}$), modulation spectral spread ($\mathrm{MSSP}_{m/k}$), modulation spectral skewness ($\mathrm{MSSK}_{m/k}$), and modulation spectral kurtosis ($\mathrm{MSKT}_{m/k}$), which are defined as follows:

$$\mathrm{MSCR_m} = \frac{\sum_{k=1}^{K} k\bar{E}_{k,m}}{\sum_{k=1}^{K} \bar{E}_{k,m}} \tag{7}$$

$$\mathrm{MSSP_m} = \frac{\sum_{k=1}^{K} [k - \mathrm{MSCR_m}]^2 \bar{E}_{k,m}}{\sum_{k=1}^{K} \bar{E}_{k,m}} \tag{8}$$

$$\mathrm{MSSK_m} = \frac{\sum_{k=1}^{K} [k - \mathrm{MSCR_m}]^3 \bar{E}_{k,m}}{\sum_{k=1}^{K} \bar{E}_{k,m}} \tag{9}$$

$$\mathrm{MSKT_m} = \frac{\sum_{k=1}^{K} [k - \mathrm{MSCR_m}]^4 \bar{E}_{k,m}}{\sum_{k=1}^{K} \bar{E}_{k,m}} \tag{10}$$

$$\mathrm{MSCR_k} = \frac{\sum_{m=1}^{M} m\bar{E}_{k,m}}{\sum_{m=1}^{M} \bar{E}_{k,m}} \tag{11}$$

$$\mathrm{MSSP_k} = \frac{\sum_{m=1}^{M} [m - \mathrm{MSCR_k}]^2 \bar{E}_{k,m}}{\sum_{m=1}^{M} \bar{E}_{k,m}} \tag{12}$$

$$\mathrm{MSSK_k} = \frac{\sum_{m=1}^{M} [m - \mathrm{MSCR_k}]^3 \bar{E}_{k,m}}{\sum_{m=1}^{M} \bar{E}_{k,m}} \tag{13}$$

$$\mathrm{MSKT_k} = \frac{\sum_{m=1}^{M} [m - \mathrm{MSCR_k}]^4 \bar{E}_{k,m}}{\sum_{m=1}^{M} \bar{E}_{k,m}}, \tag{14}$$

The final two MSFs in the acoustic-frequency and modulation-frequency domains were modulation spectral tilts ($MSTL_m$ and $MSTL_k$), which are the linear regression coefficients obtained by fitting the first-degree polynomial to the modulation spectrograms. Figure 7 shows an example of calculating $MSCR_m$ and $MSCR_k$.

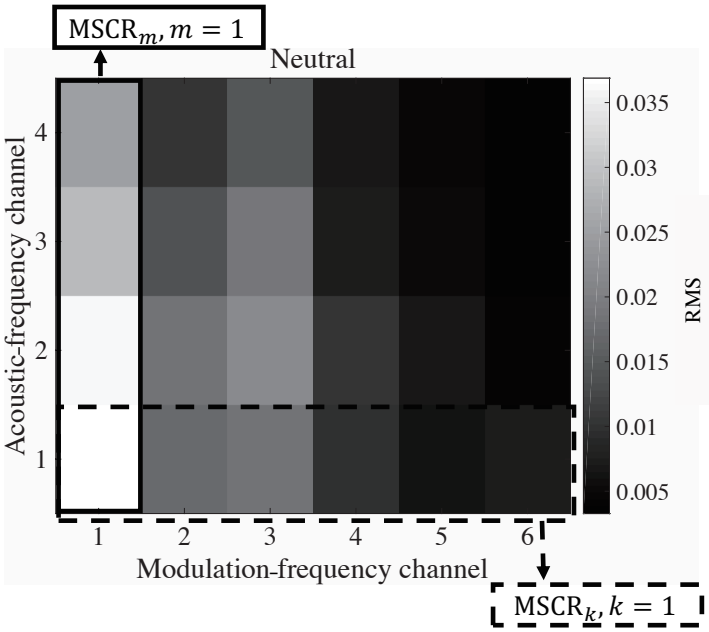

Figure 7. Calculation example of $MSCR_m$ and $MSCR_k$ [15].

The discriminability indexes of the MSFs (d'_{MSF}) were then calculated using the following:

$$d'_{\mathrm{MSF}}(em1, em2) = \frac{|\mu_{em1} - \mu_{em2}|}{\sqrt{\frac{1}{2}(\sigma^2_{em1} + \sigma^2_{em2})}}, \tag{15}$$

where μ and σ^2 are the mean and variance of one MSF (taken across the 10 utterances of each emotion), $em1$ and $em2$ are two different emotions. The $d'_{\mathrm{MSF}}(em1, em2)$ is the discriminability index of one MSF between two different emotions. The $\bar{d}'_{\mathrm{MSF}}(em)$ is the mean of all the d'_{MSF} between one emotion and other four emotions for each emotion, which was computed as the approximate measure of the net discriminability of the MSFs. The $\bar{d}'_{\mathrm{MSF}}(em)$ represents the mean distance of the MSFs between different emotions.

We calculated the similarity value to quantitatively evaluate the relationship between the vocal emotion recognition results and MSFs:

$$\text{Similarity} = \frac{\sum_{em=1}^{5} A(em)B(em)}{\sqrt{\sum_{em=1}^{5} A(em)^2}\sqrt{\sum_{em=1}^{5} B(em)^2}} \tag{16}$$

$$A(em) = d'_p(em) - \frac{1}{5}\sum_{em=1}^{5} d'_p(em) \tag{17}$$

$$B(em) = \bar{d}'_{\mathrm{MSF}}(em) - \frac{1}{5}\sum_{em=1}^{5} \bar{d}'_{\mathrm{MSF}}(em), \tag{18}$$

where $d'_p(em)$ is the discriminability index (d'_p) for different emotions, as shown in Equation (2). In the above equations, em is the emotion, which could be 1 (neutral), 2 (joy), 3 (cold anger), 4 (sadness), or 5 (hot anger).

5. Results

5.1. Qualitative Analysis: Modulation Spectrogram and Discrimination Index of Vocal-Emotion-Recognition Results

From the analysis of the modulation spectrogram and vocal emotion recognition results, we can initially clarify the relationship between these results and the modulation spectrum. We have to turn the time-averaged modulation spectrogram mentioned in Section 4 into a three-dimensional modulation representation to analyze the energy change over the modulation- and acoustic-frequency channels. According to the conclusions of Zhu et al.'s study that daily noise and reverberation cannot affect vocal emotion recognition [17], we selected SNR = 5 dB, T_R = 1.0 s, SNR = 10 dB, and T_R = 0.5 s as the representative daily conditions of the noisy, reverberant, and noisy reverberant environments. The results are shown in Figure 8. Figure 8a–d represents the four environments (including that with no noise or reverberation, i.e., clean), and in each environment, the results of the time-averaged modulation representations are displayed from left to right in the order of neutral, joy, cold anger, sadness, and hot anger. The x-axis denotes six modulation-frequency channels, the y-axis denotes four acoustic-frequency channels, and the values of the z-axis denote the RMSs of modulation spectrograms, where each one is the average over all the time frames for an emotion.

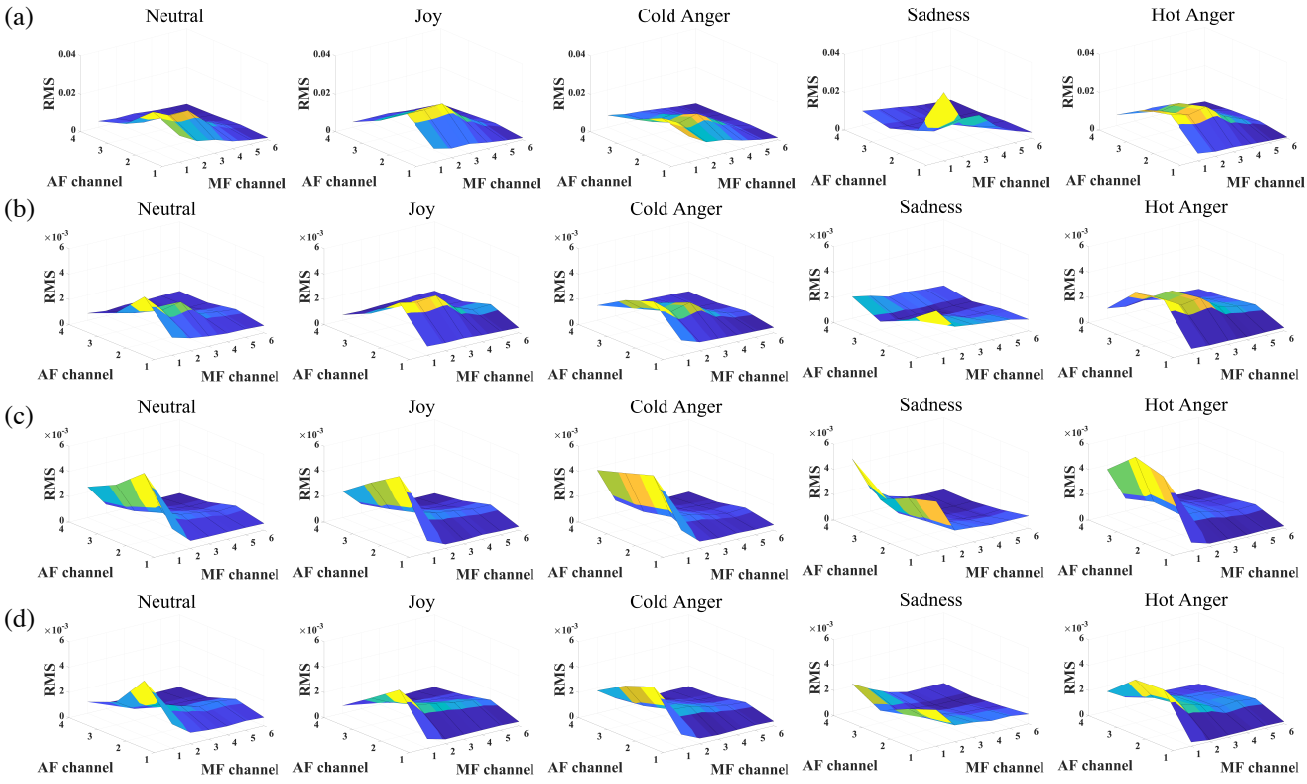

Figure 8. Time-average modulation representation in (**a**) a clean environment, (**b**) noisy environment (SNR = 5 dB), (**c**) reverberant environment (T_R = 1.0 s), and (**d**) noisy reverberant environment (SNR = 10 dB, T_R = 0.5 s). MF channel denotes six modulation-frequency channels, AF channel denotes four acoustic-frequency channels.

As Figure 8 illustrates, in the clean environment, neutral speech produced a peak of power around the first and second modulation-frequency channels (frequency range: 0 to 4 Hz), which is consistent with previous research; the power is mostly concentrated at the lower modulation-frequency channel with a peak at 4 Hz for neutral emotion [21–23], and the peak shifted to a higher modulation-frequency channel for joy and hot anger, indicating that joy and hot anger have faster speaking rates and higher RMSs. In all environments, neutral, joy, cold anger, and hot anger speech had the same power pattern over acoustic-frequency channels, which is shaped like a mountain. In the second and third acoustic-frequency channels, these four categories of emotional speech had higher power, but lower power in the first and four acoustic-frequency channels. Due to the same power patterns, these emotional speeches may not be easily distinguishable in subsequent analysis. However, neutral, joy, cold anger, and hot anger speech also had the same power patterns over all modulation-frequency channels, which are lower and stable. For sad speech in all environments, RMSs in the second and third acoustic-frequency channels were lower than in the first and fourth acoustic-frequency channels. The RMS patterns for sadness are shaped like a valley. In other words, sad speech has a different power pattern from the other four categories of emotional speech, which indicates that sad speech might also be easily recognized in subsequent analysis.

To explain how challenging it was people to distinguish different emotional speeches in the vocal-emotion-recognition experiments, as shown in Tables 1 and 2, we calculated the confusion matrix and discrimination index of the vocal emotion recognition results in the noisy reverberant environment (under the conditions mentioned in Section 2, which do not affect vocal emotion recognition). In Table 1, there are four sub-tables for the four environments. In each sub-table, the rows represent the speech stimuli of the five categories of emotional speech, the columns represent the vocal-emotion-recognition results of the participants, and the values are the ratio of the number of times the subjects chose this emotion as the answer to the total answers (in percent). The diagonal values are the correct recognition rates for the emotions, and the other values are the percentages of these emotions misrecognized as other emotions. The higher the correct recognition rate, the easier it is to recognize this emotion. In Table 2, the rows represent the same four environments as in Table 1, and the columns represent the five categories of emotional speech, and each value represents the discrimination index of that emotion in that environment. From the equation of discrimination index in Section 3, the higher the discrimination index value, the easier to distinguish this emotion from all emotion categories.

As Table 1 shows, in almost all environments, sad speech had higher correct recognition rates than other emotional speech, which suggests that sadness may be easier to distinguish from other emotions. The same trend is shown in Table 2: in almost all environments, sad speech had a higher discriminability index than other emotional speech. The misrecognition rates of joy and hot anger speech are higher in Table 1, which suggests that these two emotion categories are easily confused by the listener. Neutral speech had lower correct recognition rates, and the high misrecognition rates might be due to the experimental procedure. When participants were confused about the emotion of the experiment stimuli, most chose neutral emotion as the answer.

Table 1. Mean confusion matrix of vocal-emotion-recognition results (in percent). The confusion matrix is presented as a percentage with stimuli organized vertically and response categories organized horizontally.

Clean	**Neutral**	**Joy**	**Cold Anger**	**Sadness**	**Hot Anger**
Neutral	62.27	3.65	20.00	6.63	2.73
Joy	22.73	21.82	18.18	1.18	35.45
Cold anger	33.64	1.82	40.00	20.00	4.55
Sadness	4.55	0.00	5.46	90.00	0.00
Hot anger	16.36	2.73	5.46	0.91	74.55
SNR = 5 dB	**Neutral**	**Joy**	**Cold Anger**	**Sadness**	**Hot Anger**
Neutral	96.00	0.00	2.00	2.00	0.00
Joy	4.00	88.00	0.00	2.00	6.00
Cold anger	24.00	0.00	58.00	18.00	0.00
Sadness	0.00	0.00	2.00	98.00	0.00
Hot anger	6.00	10.00	12.00	0.00	72.00
T_R = 1.0 s	**Neutral**	**Joy**	**Cold Anger**	**Sadness**	**Hot Anger**
Neutral	86.96	0	4.35	8.70	0.00
Joy	0.00	93.33	0.00	2.22	4.44
Cold anger	30.23	0.00	44.19	25.58	0.00
Sadness	11.11	0.00	4.44	84.44	0.00
Hot anger	17.78	6.67	0.00	0.00	75.56
SNR = 10 dB T_R = 0.5 s	**Neutral**	**Joy**	**Cold Anger**	**Sadness**	**Hot Anger**
Neutral	90.00	0.00	6.00	4.00	0.00
Joy	8.00	80.00	4.00	2.00	6.00
Cold anger	26.00	0.00	58.00	16 .00	0.00
Sadness	4.00	0.00	4.00	92.00	0.00
Hot anger	14.00	14.00	10.00	0 .00	62.00

Table 2. Discrimination index of vocal-emotion-recognition results in clean/noisy/reverberant/noisy reverberant environments.

d'_p of Experimental Result	**Neutral**	**Joy**	**Cold Anger**	**Sadness**	**Hot Anger**	**Average**
Clean	1.31	1.27	0.91	2.74	1.90	1.63
SNR = 5 dB	3.12	3.14	1.95	3.65	2.75	2.92
T_R = 1.0 s	1.97	3.16	1.75	2.11	2.79	2.36
SNR = 10 dB, T_R = 0.5 s	2.73	2.97	2.52	3.44	3.22	2.98

The analysis of the vocal-emotion-recognition results and modulation spectrogram show a consistent trend. Both evinced that the vocal emotion recognition of sadness is easier than that of other emotion categories. However, we need more quantitative evidence to rigorously evaluate the above results. Therefore, to deeply investigate how MSFs contribute to the perception of vocal emotion recognition, we extracted the MSFs from the modulation spectrograms then calculated the similarity value between MSFs and the vocal-emotion-recognition results.

5.2. Quantitative Analysis: Modulation Spectral Features and Vocal-Emotion-Recognition Results

The results of the highest similarity (over all modulation-frequency channels or over all acoustic-frequency channels) between each MSF and the vocal-emotion-recognition results in clean/noisy/reverberant/noisy reverberant environments are shown in Figure 9. We also selected SNR = 5 dB, T_R = 1.0 s, SNR = 10 dB, and T_R = 0.5 s as the representative conditions for each environment. The horizontal axis denotes the ten types of MSFs extracted from the modulation spectrograms. The value of each bar denotes the similarity between the vocal-emotion-recognition results and MSFs. The higher the value, the higher the similarity between these results and MSFs.

If the similarity exceeds 0.4, it is generally considered that the similarity is high. As illustrated in Figure 9b–d, there are 8, 7, and 10 MSFs that have similarity with the vocal-emotion-recognition results of over 0.4, respectively. This suggests that the contribution of the MSFs to NVS vocal emotion recognition cannot be affected by daily noise and reverberation. It also means that MSFs can be used to explain the vocal-emotion-recognition results in daily noisy reverberant environments.

The bar graphs for the highest similarity in all daily noisy reverberant environments are not shown due to the page limitation, but we still want to clarify the following two issues: (1) whether the conclusion of high similarity between vocal-emotion-recognition results and MSFs still exists in all daily environments, and (2) whether there exist common MSFs in all daily environments.

To investigate the similarity between MSFs and vocal-emotion-recognition results in all daily environments and further determine common features in these environments, we plotted radar charts, as shown in Figure 10. The different axes represent 10 MSFs, and the values on each axis represent the similarity values between the vocal-emotion-recognition results and MSFs. In Figure 10a–f, different colored lines represent different environments; the results for daily environments are shown with solid lines, and extremely high noisy and/or reverberant environments are shown with dashed lines. In Figure 10g,h, all conditions were for daily environments, so that all of the results are represented as solid lines. This is the same for the bar graphs; the higher the value, the higher the similarity between the vocal-emotion-recognition results and MSFs. The reason for the existence of negatively similarity values is shown in Equations (16)–(18). When a certain emotion's discriminability index of the vocal-emotion-recognition results or MSFs is lower than the average discriminability index of all five emotion categories, the similarity calculated using Equation (16) is negative. The thick black line denotes the similarity of 0.4. The MSFs in red circles mean that they have similarities with the vocal-emotion-recognition results over 0.4 in all daily environments. These MSFs are considered common MSFs.

As shown in Figure 10a,b, MSSP_m, MSKT_m, MSKT_k, and MSTL_k were common MSFs in all daily noisy environments. As shown in Figure 10c,d, MSSK_m, MSKT_k, MSTL_k, and MSSK_k were common MSFs in all daily reverberant environments. As shown in Figure 10e,f, MSCR_m, MSTL_m, MSSK_m, MSKT_k, MSTL_k, MSCR_k, and MSSK_k were common MSFs in all daily noisy reverberant environments. As shown in Figure 10g,h, MSKT_k and MSTL_k were common MSFs in all daily environments.

Combining all the above radar chart results, the conclusion of high similarity between the vocal-emotion-recognition results and MSFs exist in all noisy reverberant environments. We also clarified that there are two common MSFs: MSKT_k and MSTL_k. In all daily environments, these two MSFs have a strong relation with the vocal-emotion-recognition results, which indicates that the important role MSKT_k and MSTL_k play in the vocal emotion recognition of NVS is not affected by noise and reverberation.

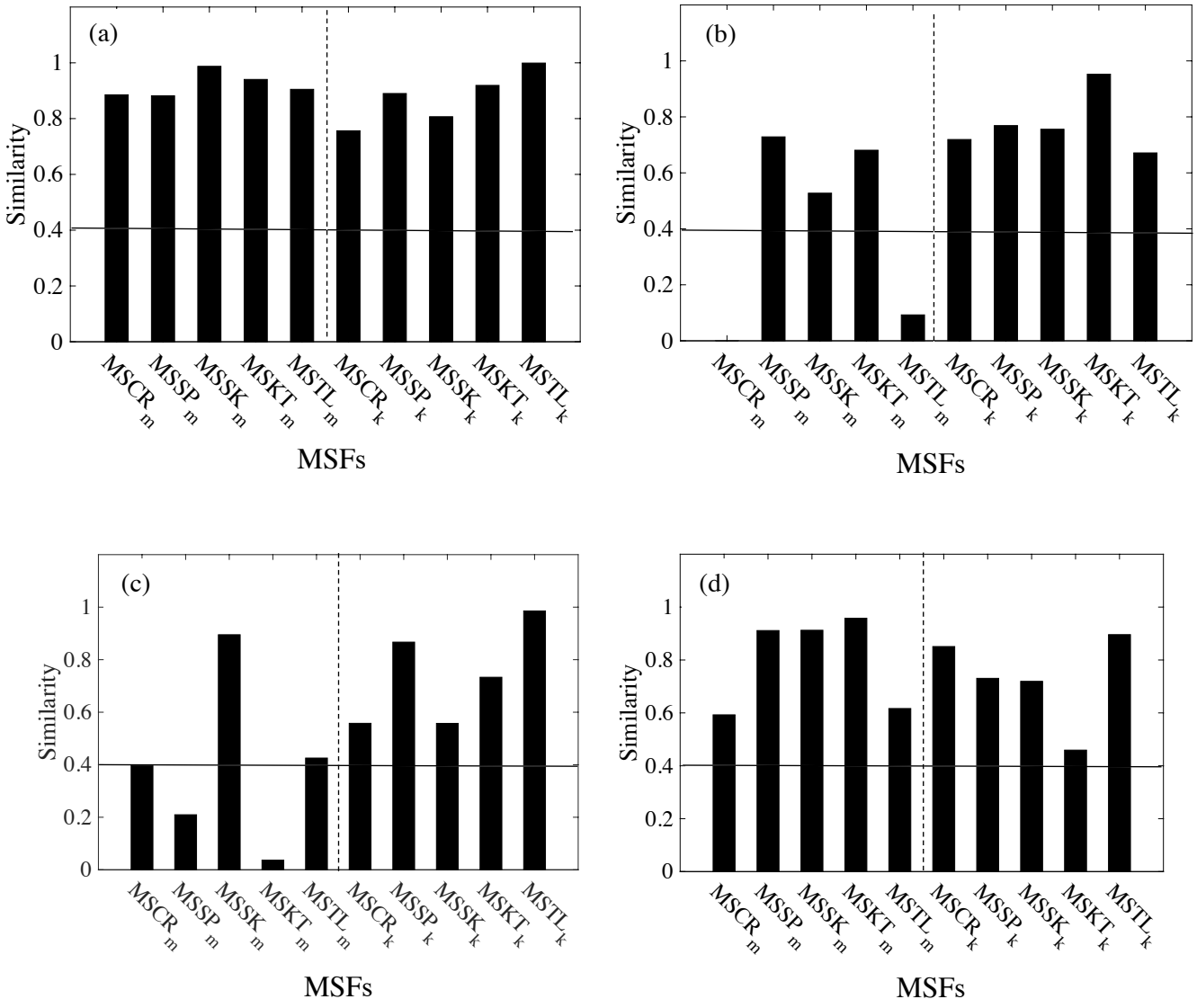

Figure 9. Highest similarity of each MSF in clean/noisy/reverberant/noisy reverberant environments: (**a**) clean environment [15], (**b**) noisy environment (SNR = 5 dB), (**c**) reverberant environment (T_R = 1.0 s), and (**d**) noisy reverberant environment (SNR = 10 dB, T_R = 0.5 s).

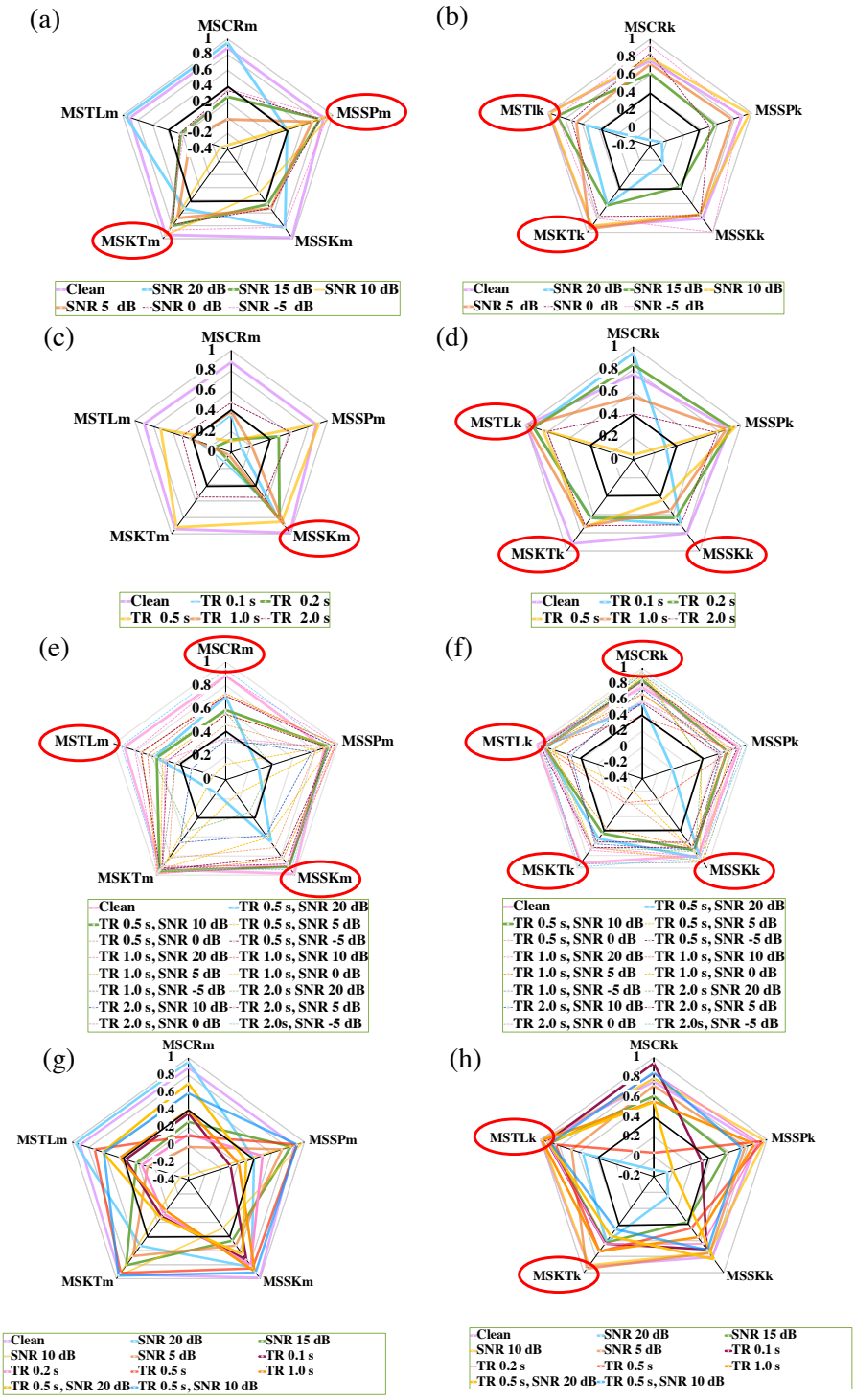

Figure 10. Radar charts of highest similarity of each MSF in clean/noisy/reverberant/noisy−and−reverberant environments: (**a**) MSFs in acoustic−frequency domain in all noisy environments, (**b**) MSFs in modulation−frequency domain in all noisy environments, (**c**) MSFs in acoustic−frequency domain in all reverberant environments, (**d**) MSFs in modulation−frequency domain in all reverberant environments, (**e**) MSFs in acoustic−frequency domain in all noisy reverberant environments, (**f**) MSFs in modulation−frequency domain in all noisy reverberant environments, (**g**) MSFs in acoustic−frequency domain in all daily environments, and (**h**) MSFs in modulation−frequency domain in all daily environments).

6. Discussion

In the above analyses, we considered ten MSFs and found that there were high similarities between MSFs and the vocal-emotion-recognition results in the noisy reverberant environments. We also clarified that there are two common MSFs ($MSKT_k$ and $MSTL_k$) in all daily environments.

We focused on specific MSFs and discussed the contributions of two common MSFs to vocal emotion recognition from the following two perspectives: the emotion perception characteristics of different emotions, and the different properties of acoustic- and modulation-frequency channels. The distribution of each MSF in each emotion is a real-valued random variable, and it has a finite mean and variance. Therefore, this study assumed that the distribution of each MSF in each emotion conforms to a normal distribution. As shown in Figure 11, we estimated its distribution using the probability density functions (PDFs) of the normal distribution; μ and σ^2 are the mean and variance of each MSF taken across the 10 utterances of each emotion, which are the same as shown in Equation (15). Figure 11a–d represent the four environments, and the PDF results of $MSKT_k$ and $MSTL_k$ are shown in order from top to bottom in each environment; the five colors represent five different emotions, where K means the Kth acoustic frequency channel. Due to the page limitation, only the highest similarity (from the first to fourth acoustic-frequency channel) of each MSF is shown in Figure 11.

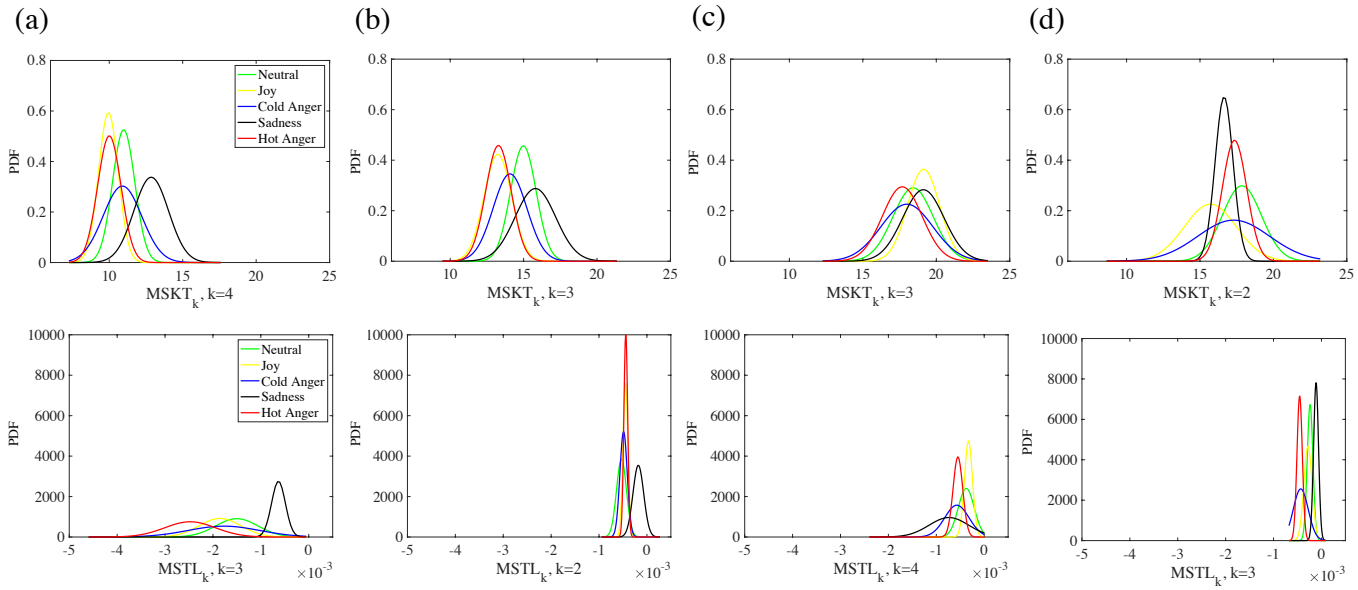

Figure 11. PDFs of $MSKT_k$ and $MSTL_k$ for each emotion in noisy/reverberant/noisy reverberant environments. Only MSFs with highest similarity are shown: (**a**) clean environment [15], (**b**) noisy environment (SNR = 5 dB), (**c**) reverberant environment (T_R = 1.0 s), and (**d**) noisy reverberant environment (SNR = 10 dB, T_R = 0.5 s).

In Figure 11, $MSTL_k$'s PDFs of sadness have less overlap with other emotions, as explained in Section 4, and the $\bar{d}'_{MSF}$ represents the physical distance of the MSFs between different emotions. Therefore, the large distance between sadness and other emotions indicates that using $MSTL_k$ makes it easier to distinguish sadness from other emotions. The reason can be explained by the emotion-perception characteristic of sadness. Sad speech is a type of breathy speech produced when the vocal fold motion is not broad enough to close the glottis completely during the vibration cycle. This phenomenon perceptually decreases the loudness of the speech unit produced, so that a breathy voice has a spectrum with a strong slope. Conversely, for loud speech, such as hot-anger speech and joyous speech, which see a rapid closure of the vocal folds and a short open phase of the glottis, the spectral envelope is flatter (the spectral tilt is less) [24–26]. Combining the analysis of sad speech, as the results show in Figure 8, we can also see that sadness has a different power

pattern in time-average modulation representation. In all environments, only sadness had a lower RMS in the second and third acoustic-frequency channels than in the first and fourth. This might also be due to sad speech being more breathy than other emotional speech. It is known that breathy voices are characterized by a steep decrease in spectral level from the 0–2 to the 2–5 kHz band. Conversely, "vocal fry/creaky" voices, such as those for hot anger and joy, were correlated with a steep decrease in level from the 2–5 to 5–8 kHz band [24,27]. As we can see in the time-average modulation representation in Figure 8, sadness has an RMS that steeply decreased in the lower acoustic-frequency range. Hot anger and joy had RMSs that steeply decreased in the higher acoustic-frequency range. Consequently, $MSTL_k$ is important in sadness recognition.

For both $MSKT_k$ and $MSTL_k$, in most environments there was less overlap between sadness and other emotions. As shown in Figure 11a,d, for the PDF of $MSTL_k$, sadness and hot anger hardly overlapped, which indicates that by using these common features, it is easy to distinguish sadness from hot anger. Both common MSFs are in the modulation-frequency domain (calculated by six modulation frequency channels); no MSFs in the acoustic-frequency domain (calculated by four acoustic frequency channels) is regarded as a common feature contributing to vocal emotion recognition in all daily environments. The reason might be the different properties between acoustic- and modulation-frequency channels shown in Figure 8. Different power patterns of five emotions were only obvious in the modulation-frequency domain, suggesting no significant difference in the acoustic-frequency domain. Therefore, the MSFs that reflect the characteristics of six modulation-frequency channels are more important in vocal emotion recognition.

Common MSFs contribute to vocal emotion recognition, which might be due to their emotion-perception characteristics. For example, the relationship between the breathy voice and spectral tilt makes the spectral tilt possibly account for the perception of vocal emotion recognition. Moreover, the differentiating properties of acoustic- and modulation-frequency channels are useful in the explanation of different performances of MSFs when distinguishing different emotions. We used the vocal-emotion-recognition results in noisy reverberant environments [17], which involved NVS simulations with normal-hearing listeners to predict the responses from CI listeners. One study on the important role of temporal cues in speaker identification for simulated CIs also suggested that temporal modulation cues contribute to speaker identification and have the potential to improve speaker identification if enhanced [28]. Therefore, the results of this study also indicate that MSFs might have the potential to be used in the explanation of nonlinguistic information perception when using CIs, such as vocal emotion recognition and speaker identification in noisy reverberant environments.

7. Conclusions

We investigated the contribution of MSFs to vocal emotion recognition of NVS in noisy and/or reverberant environments. To clarify the relationship between MSFs and previous results from vocal-emotion-recognition experiments in noisy and/or reverberant environments, we analyzed the results of those experiments regarding NVS, and then obtained modulation spectrograms of NVS by using an auditory-based modulation filterbank and extracted MSFs from modulation spectrograms. We then calculated the similarity between MSFs and the vocal-emotion-recognition results. The qualitative analysis between modulation spectrograms and discrimination index of the vocal-emotion-recognition results and the quantitative analysis between MSFs and these results indicate that

1. MSFs can be used to explain the results of vocal emotion recognition in noisy and/or reverberant environments. Except for extremely high noisy reverberant environments, in daily environments, the contribution of MSFs in vocal emotion recognition is not affected by noise and reverberation.
2. There are two common MSFs ($MSKT_k$ (modulation spectral kurtosis) and $MSTL_k$ (modulation spectral tilt)) that have high similarity with the vocal-emotion-recognition results in all daily environments.

MSFs, especially the two common MSFs, are considered to play an important role in vocal emotion recognition, and are considered to have the potential to be used in the explanation of nonlinguistic information perception by using CIs in noisy and/or reverberant environments, which still requires further research. For future work, due to the Fujitsu Japanese Emotional Speech Database used in this study being created using a professional actress (containing five clear acted emotions), these emotions can be considered the typical representations of diverse emotions, but they may differ from the emotional speech communication in our daily lives (containing ambiguous natural emotions). Therefore, it is necessary to use more natural emotion databases containing more emotions to examine the contribution of MSFs to vocal emotion recognition in daily communication.

Author Contributions: Conceptualization, T.G.; resources, Z.Z.; project administration, S.K. and M.U.; supervision, M.U.; writing—original draft, T.G.; writing—review and editing, S.K. and M.U. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by JST SPRING (JPMJSP2102), a Grant-in-Aid for Scientific Research (B) (21H03463), a Fund for the Promotion of Joint International Research (Fostering Joint International Research (B)) (20KK0233), and the SCOPE Program of Ministry of Internal Affairs and Communications (201605002).

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

References

1. Banse, R.; Scherer, K.R. Acoustic profiles in vocal-emotion expression. *J. Personal. Soc. Psychol.* **1996**, *70*, 614. [CrossRef]
2. Huang, C.F.; Akagi, M. A three-layered model for expressive speech perception. *Speech Commun.* **2008**, *50*, 810–828. [CrossRef]
3. Chatterjee, M.; Zion, D.J.; Deroche, M.L.; Burianek, B.A.; Limb, C.J.; Goren, A.P.; Kulkarni, A.M.; Christensen, J.A. Voice emotion recognition by cochlear-implanted children and their normally-hearing peers. *Speech Commun.* **2015**, *322*, 151–162. [CrossRef] [PubMed]
4. Chatterjee, M.; Peng, S.C. Processing F0 with cochlear implants: Modulation frequency discrimination and speech in-tonation recognition. *Hear. Res.* **2008**, *235*, 143–156. [CrossRef] [PubMed]
5. Moore, B.C.J. The roles of temporal envelope and Fine Structure Information in auditory perception. *Acoust. Soc. Technol.* **2019**, *40*, 61–83. [CrossRef]
6. Xiang, J.; Poeppel, D.; Simon, J.Z. Physiological evidence for auditory modulation filterbanks: Cortical responses to concurrent modulations. *J. Acoust. Soc. Am.* **2013**, *133*, EL7–EL12. [CrossRef]
7. Ewert, S.D.; Dau, T. Characterizing frequency selectivity for envelope fluctuations. *J. Acoust. Soc. Am.* **2000**, *108*, 1181–1196. [CrossRef] [PubMed]
8. Dau, T.; Puschel, D.; Kohlrausch, A. A quantitative model of the "effective" signal processing in the auditory system. I. Model structure. *J. Acoust. Soc. Am.* **1996**, *99*, 3615–3622. [CrossRef]
9. Dau, T.; Puschel, D.; Kohlrausch, A. A quantitative model of the "effective" signal processing in the auditory system. II. Simulations and measurements. *J. Acoust. Soc. Am.* **1996**, *99*, 623–3631. [CrossRef]
10. Zhu, Z.; Miyauchi, R.; Araki, Y.; Unoki, M. Contributions of temporal cue on the perception of speaker individuality and vocal emotion for noise-vocoded speech. *Acoust. Soc. Technol.* **2018**, *39*, 234–242. [CrossRef]
11. Tachibana, R.O.; Yasunari, S.; Hiroshi, R. Relative contributions of spectral and temporal resolutions to the perception of syllables, words, and sentences in noise-vocoded speech. *Acoust. Sci. Technol.* **2013**, *34*, 263–270. [CrossRef]
12. Xu, L.; Bryan, E.P. Spectral and temporal cues for speech recognition: Implications for auditory prosthe-ses. *Hear. Res.* **2008**, *242*, 132–140. [CrossRef] [PubMed]
13. Unoki, M.; Kawamura, M.; Kobayashi, M.; Kidani, S.; Akagi, M. *How the Temporal Amplitude Envelope of Speech Contributes to Urgency Perception*; Universitätsbib-liothek der RWTH Aachen: Aachen, Germany, 2019.
14. Zhu, Z.; Miyauchi, R.; Araki, Y.; Unoki, M. Modulation Spectral Features for Predicting Vocal-Emotion Recognition by Simulated Cochlear Implants. In Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, San Francisco, CA, USA, 8–12 September 2016; pp. 262–266.
15. Zhu, Z.; Miyauchi, R.; Araki, Y.; Unoki, M. Contributions of modulation spectral features on the perception of vocal-emotion using noise-vocoded speech. *Acoust. Soc. Technol.* **2018**, *39*, 379–386. [CrossRef]
16. Helms, T.K.; Brown, C.A.; Bacon, S.P. Comparing the effects of reverberation and of noise on speech recognition in simulated electric-acoustic listening. *J. Acoust. Soc. Am.* **2012**, *131*, 416–423. [CrossRef] [PubMed]
17. Zhu, Z.; Kawamura, M.; Unoki, M. Study on the perception of nonlinguistic information of noise-vocoded speech under noise and/or reverberation conditions. *Acoust. Soc. Technol.* 2022, *in press*.
18. Schroeder, M.R. Modulation transfer functions:10 definition and measurement. *Acta Acust. United Acust.* **1981**, *49*, 179–182.
19. Newman, R.; Chatterjee, M. Toddlers' recognition of noise-vocoded speech. *J. Acoust. Soc. Am.* **2013**, *133*, 483–494. [CrossRef]

20. Moore, B.C.J. *An Introduction to the Psychology of Hearing*, 6th ed.; Brill: Leiden, The Netherlands, 2013.
21. Wu, S.; Falk, T.H.; Chan, W.Y. Automatic speech emotion recognition using modulation spectral features. *Speech Commun.* **2011**, *53*, 768–785. [CrossRef]
22. Peng, Z.; Zhu, Z.; Unoki, M.; Dang, J.; Akagi, M. Auditory-inspired end-to-end speech emotion recognition using 3D convolutional re-current neural networks based on spectral-temporal representation. In Proceedings of the 2018 IEEE International Conference on Multimedia and Expo, San Diego, CA, USA, 23–27 July 2018; pp. 1–6.
23. Kanedera, N.; Arai, T.; Hermansky, H.; Pavel, M. On the relative importance of various components of the modulation spectrum for automatic speech recognition. *Speech Commun.* **1999**, *28*, 43–55. [CrossRef]
24. Ishi, C.T.; Ishiguro, H.; Hagita, N. Analysis of the roles and the dynamics of breathy and whispery voice qualities in dialogue speech. *EURASIP J. Audio Speech Music. Process.* **2010**, *2010*, 1–12. [CrossRef]
25. Koolagudi, S.G.; Ray, S.; Sreenivasa, R.K. Emotion classification based on speaking rate. In Proceedings of the International Conference on Contemporary Computing, Noida, India, 9–11 August 2010; Springer: Berlin/Heidelberg, Germany, 2010.
26. Childers, D.G.; Lee, C.K. Vocal quality factors: Analysis, synthesis, and perception. *J. Acoust. Soc. Am.* **1991**, *90*, 2394–2410. [CrossRef] [PubMed]
27. Monson, B.B.; Hunter, E.J.; Lotto, A.J.; Story, B.H. The perceptual significance of high-frequency energy in the human voice. *Front. Psychol.* **2014**, *5*, 587. [CrossRef] [PubMed]
28. Zhu, Z.; Miyauchi, R.; Araki, Y.; Unoki, M. Important role of temporal cues in speaker identification for simulated cochlear implants. In Proceedings of the 1st International Workshop on Challenges in Hearing Assistive Technology, Stockholm University, Collocated with Interspeech, Stockholm, Sweden, 19 August 2017.