Increasing speech intelligibility and naturalness in noise based on concepts of modulation spectrum and modulation transfer function

Thuanvan Ngo[a,*], Rieko Kubo[b], Masato Akagi[a]

[a]*Graduate school of advanced science and technology, Japan Advanced Institute of Science and Technology, 1-1 Asahidai, Nomi, Ishikawa, 923-1292, Japan*
[b]*Graduate school of frontier biosciences, Osaka University, 1-4 Yamadaoka, Suita City, Osaka, 565-0871, Japan*

Abstract

This study focuses on identifying effective features for controlling speech to increase speech intelligibility under adverse conditions. Previous approaches either cancel noise throughout speech presentation or preprocess speech by controlling its intensity and/or spectra. Among them, a method based on modulation transfer function theory, inverting the environmental effects to anticipate attenuation of speech modulation spectrum, shows excellent potential due to its systematic and explicit derivation of intelligibility enhancement against environmental smears. However, strictly following the inverse modulation transfer function is dangerous and inefficient as important speech features can be damaged, and it costs lots of energy to boost all smeared regions. This study takes a different approach: analyzing the relations of smeared modulation spectra by the environments for intelligibility to extract effective modifying features. First, we conduct listening tests for intelligibility in noise with different types of enhanced speech. Next, we extract acoustic and modulation frequency components in the smeared modulation spectra by noise showing high correlation with intelligibility scores. Finally, we examine the intelligibility benefits of modifying these components by performing listening tests. The results show that these components effectively increase intelligibility by at most 18%, which demonstrates that our concept is valid.

Keywords: Modulation transfer function, modulation spectrum, intelligibility.

1. Introduction

When people listen to an announcement in train stations, factories, or airports, the presence of noise often smears the speech signals, thus making the systems hard to deliver the speech content to most of them. Speech intelligibility could be maintained by canceling noise throughout the delivery. However, this is impractical for the locations with the complex architectures and inefficient for the cost of installing the necessary devices. A more practical and efficient approach is to enhance the speech before presenting it to compensate for degradation in intelligibility due to smearing. Two conventional methods to enhance the intelligibility of the presented speech are by controlling acoustic features directly or using models.

Acoustic features for speech intelligibility can be directly controlled by spectral shaping, intensity amplifiers, and equalizers. Spectral shaping modifies speech spectra by increasing spectral regions for intelligibility (e.g., 1-4 kHz (Zorila et al., 2012) or 2-6 kHz (Ngo et al., 2020b), and formants (Zorila et al., 2012, Ngo et al., 2020b)). Intensity amplifiers increase speech intensity by adjusting the gain of speech. Equalizers are audio processors that use a combination of different filters to alter the balance of frequencies in an audio signal (EQ, Westerlund et al., 2002). Thus, it can be used to boost important frequency regions for intelligibility. These important acoustic features to control were mainly extracted from clear speech (Picheny et al., 1986) and Lombard speech (Lombard, 1911). Increasing the frequency regions between 1-4 kHz or 2-6 kHz is to decrease spectral tilt, modification on formants is to increase formant frequencies and amplitudes (Parikh and Loizou, 2005, Ngo et al., 2020a,b). Increasing spectra in different frequency regions also have different effects on the perception of related factors of the naturalness of speech. For example, decreasing spectral regions below 1 kHz reduces fullness, while increasing the spectral region above 1 kHz extends brightness (Raake, 2006). Increasing speech intensity is to increase vocal strength as in Lombard speech. Generated voices can be intelligibible and extreme as Lombard speech. However, it is unable for these methods to control acoustic features to changing phenomena of environments appropriately.

Models as perceptual models, compressing models and room acoustic models can be used to estimate the equivalent amount of environment phenomena such as the signal-to-noise ratio (SNR) and noise level needed to compensate for degradation in intelligibility. These methods were also mentioned as NELE (Near-end speech enhancement), which is a technique to enhance the speech intelligibility in environmental noise by adaptively modifying the speech based on a noise estimate (Niermann et al.,

*Corresponding author
Email addresses: vanthuanngo@jaist.ac.jp (Thuanvan Ngo), rkubo@fbs.osaka-u.ac.jp (Rieko Kubo), akagi@jaist.ac.jp (Masato Akagi)

2016)). Speech intelligibility has been improved in noisy environments (Sauert and Vary, 2010, Taal and Jensen, 2013, Taal et al., 2014, Tang and Cooke, 2018) by optimizing the index of the perceptual model used for intelligibility measurement such as the Speech Intelligibility Index (SII) (ANSI, 1997), the Speech Transmission Index (STI) (CODE, 2003), and the high energy glimpse portion (HEGP) (Tang et al., 2016). Further analyses of the speech after index optimization indicated that increasing the spectrum above 1 kHz increases intelligibility (Tang and Cooke, 2018). Compressing models that uses a signal processing operation like dynamic range compression (DRC) (Zorila et al., 2012) has been used to reduce the speech amplitude on the basis of an input-output energy curve. Different configurations in the curve yield different effects on modified speech, which implicitly responds to the environmental factors of SNR and/or noise levels. The DRC emphasizes loudness in the voice onsets and offsets and in the stops and nasals, thereby increasing intelligibility. Xu et al. (2019) showed that an intensity range around peak amplitude yielded better intelligibility performance under noisy conditions than others. This finding indicates that compressing the speech amplitude into the regions of the peak amplitude might be useful in increasing speech intelligibility, which is in line with the effects of the DRC. Besides, the DRC makes speech signals degraded, especially in naturalness.

A method based on a room acoustic model uses the modulation transfer function (MTF) to control the speech modulation spectrum (MS) and has demonstrated a more systematic and explicit derivation to enhance speech intelligibility against environmental smears. In the MTF concept, which was proposed by Houtgast and Steeneken (Houtgast and Steeneken, 1973, 1985), the reduction of the fluctuations in the envelope of an output signal relative to the envelope of the input signal during transmission in a room is described as MTF. MTF was used in the calculation of STI (Houtgast and Steeneken, 1973), which is an important objective index for speech intelligibility under noisy reverberant environments. In the MS concept, the speech MS is produced by spectral analysis of the temporal amplitude envelope of the frequency spectra. The dominant MS component of continuous speech lies between modulation frequencies of 1 and 16 Hz, with a peak around 4 Hz (Houtgast and Steeneken, 1973, Hermansky, 1998, Kusumoto et al., 2005). Some recent studies (Bosker and Cooke, 2020, Hansen et al., 2020) reported that better intelligibility is obtained when increasing the MS indexes as high as in Lombard speech. In the other words, the higher the MS index in these modulation frequencies, the better the intelligibility.

As a result, the room acoustic model based on the MS and MTF concepts obtain archivements in different fields such as speech restoration (Unoki et al., 2004, Liu et al., 2016), speech perception (Unoki and Zhu, 2020a), especially speech enhancement (Kusumoto et al., 2005, Koutsogiannaki and Stylianou, 2016). In short, speech signals and environmental effects seem to be approriately represented by the MTF and MS concepts considering both acoustic and modulation frequency components. The concept models can have a strong correlation with speech intelligibility. The systematic estimations of the concept models, which can highly include controllable parameters, are also convenient for more insightfull and flexible analyses. With these potentials, this room acoustic model was therefore chosen as a basis of our study for enhancements of speech intelligbility and naturalness under adverse conditions.

The basic concept is that speech is intelligibly presented if its MS resists smearing of the MTF by the environments. On this concept, our defined MS is calculated for multiple narrow acoustic frequence ranges, MS is thus a 2-dimensional spetrum of two axes of acoustic frequencies (AF) and modulation frequencies (MF). For example, speech is analyzed with a BP filterbank (band-pass filterbank) on AF, each output wave of the BP filterbank is transformed into amplitude envelope and carrier, and the amplitude envelope is transformed into MS on MF at a certain AF. Thus, MS for the outputs of the BP filterbank is a 2-dimesional spectrum.

Corresspondingly, if a "smeared MS" ($MS_{smeared}$) is given by

$$MS_{smeared} = MS \times MTF \tag{1}$$

where MS is the MS of the original speech, then an "optimally resistant MS" (MS_{res}^{opt}, if only "resistant MS", i.e., MS_{res}) can be calculated using

$$MS_{res}^{opt} = MS \times MTF^{-1}. \tag{2}$$

If MS_{res}^{opt} is presented in an adverse environment with such an MTF, the MS of the speech reaching the listeners should be MS as $MS = MS_{res}^{opt} \times MTF$, which has the original intelligible MS.

Directly obtaining MTF^{-1}, which requires estimating MTF, is complicated, we thus assume that this estimation is done with the provided noise and room impulse response. Though MTF^{-1} is obtained by the assumption, using them efficiently has become a critical problem. Because if strictly following the inverse MTF to modify speech MS (Kusumoto et al., 2005, Koutsogiannaki and Stylianou, 2016), it is sometimes dangerous and ineffective because the modification might blindly destroy important speech features and it costs lots of energy to boost all the smeared regions.

In this study, our perspective is still basing on the basic concept of MS and MTF in order to propose an efficient way to modify speech MS. To this end, we concentrated on significant acoustic and modulation frequency regions and their appropriate tuning amplitude levels for improving speech intelligibility and naturalness in adverse conditions such as noise. Furthermore, the proposed concept is not fixed for noise level and SNRs. Conversion controls are based on the anytime estimated MTF. This means that we can convert MS according to the estimated MTF almost in real time, which was implied by the almost real-time estimations of STI, SNRs, and reverberation time with sub-band analyses (Duangpummet et al., 2019, Unoki et al., 2017, Morita et al., 2017, Unoki and Hiramatsu, 2008). So, when using our method, firstly we estimated MTF and secondly we determined the parameter values to convert announced voices. The detail of our proposed concept and

methodology and their implementation is described in the following sections.

2. Proposed concept and methodology

2.1. Concept

Figure 1 illustrated the detailed steps aiming to convert plain speech into intelligible and natural mondified speech by modifying plain speech MS. As was indicated in the MTF concept, the MTF by environments affects speech MS for both AF and MF regions (as AF and MF axes in the figure). Therefore, two-dimensional filters (2-D filters) were applied to modify plain speech MS efficiently in terms of both AF and MF filtering. Two problems of identifying acoustic and modulation frequency regions and their tuning amplitudes were tackled separately to construct the 2-D filters. We consulted the relations of $MS_{smeared}$ for speech intelligibility and naturalness for the identification of acoustic and modulation frequency regions. Meanwhile, we based on the 2-D filters to suggest tuning amplitude levels for these identified acoustic and modulation frequencies.

The essentials of this concept were to activate acoustic and modulation frequencies regions on $MS_{smeared}$ under the affection of environments that related to speech intelligibility and naturalness. It was then to design 2-D filters with speech MS modification to obtain an efficient MS_{res} to be presented under the final near-end conditions.

Furthermore, it can be seen that this concept aims to apply for real-time situations, however, it still needs the information from background noise and reverberation in advance but in a short period of time. For example, we can apply a method by Duangpummet et al. (2019) to estimate MTF by noise and reverberation in almost real-time situations. Regarding spectral and temporal characteristics of the background noises, it can accept any types of them. Because we based on modulation transfer function and modulation spectrum concepts, which can mostly analyze any types of noise and reverberation. For example, speech-like noise conditions are some kinds of noise with spectral information like speech. Modulation spectrum and modulation transfer function can also be applied for this noise; therefore, we can apply our method for them.

2.2. Methodology

The steps and methods for obtaining a solution of an efficient MS_{res} to modify "plain" speech (this term was inherited from Bradlow and Alexander (2007), which indicates speech uttered in quiet conditions) by the proposed concept were presented in Fig. 2. In general, there were three phases: analysis, verification, and efficient MS_{res}.

In the analysis phase, we aimed to obtain correlated acoustic and modulation frequency regions, which were called MS features by analyzing relationships (correlation, also combining existing knowledge about important acoustic and modulation frequencies) between estimated $MS_{smeared}$ and the intelligibility and naturalness scores by listening tests of enhanced and plain speech. Our method used enhancement algorithms with specific diversity rather than implementing a random walk to activate different acoustic and modulation frequency regions of $MS_{smeared}$ in response to different environments. As a result, we could save a lot of resources in searching for MS features and increase the possibility of obtaining potential features. Listening tests rather than objective measures were used to evaluate intelligibility and naturalness scores due to their reliability and to avoid ignoring unknown essential features when using objective measures.

After obtaining the perspective correlated MS features, it was necessary to investigate the intelligibility and naturalness benefits of any single and/or jointly combinations of these MS features to identify which ones were significant for improving speech intelligibility and naturalness. In this phase, we suggested an extra validation with cross features from other models to revise the completeness of the feature set extracted from the previous analysis phase. From the features, 2-D filters consulting MTF^{-1} were designed to modify MS of plain speech, i.e., MS modification methods were derived for applying these 2-D filters. The tuning amplitude levels or the gain of the 2-D filters for each feature were estimated from MTF^{-1} and limited within ranges with small gaps (almost equal gains) to obtain a fair comparison among MS features. Listening tests were still used to evaluate the intelligibility and naturalness of these modified feature combinations. At the end of this phase, the significant MS features were identified.

In the final step, the gains of 2-D filters at the regions of the identified significant MS features were tunned with trials of different limited ranges to identify the optimal filter design for modifying speech MS. This modification finally resulted in efficient MS_{res}. The implementation of the concept and methodology was presented in the next section.

3. Implementation

3.1. Phase 1: Analysis

This section aims to extract all possible correlated MS features to speech intelligibility and naturalness from typical-advanced speech enhancement methods based on MS and MTF concepts. The details are described in the following sub-sections.

3.1.1. Analysis data: Plain speech, enhancement algorithms, and environments

Plain speech was the male speech in A-set of ATR dataset (Kurematsu et al., 1990) with 600 three-mora words of an average duration about 350-450 ms, which was shown as a suited length for the MS analysis (Taal et al., 2010).

Three basic spectral shaping of the enhancement algorithms derived from different hypotheses used to enhance plain speech was illustrated in Fig. 3).

- **C2**: it was increasing spectral regions from 2-6 kHz about 13 dB and decreasing spectra at other frequencies by 15 dB. The 13 and 15 dB were tunning values to best increase speech intelligibility. This spectral shaping reflected the spectral compensation in Lombard speech by deceases in the 2^{nd} order cepstral coefficient under the speech sample frequency of 16 kHz (Ngo et al., 2020b), so-called C2.

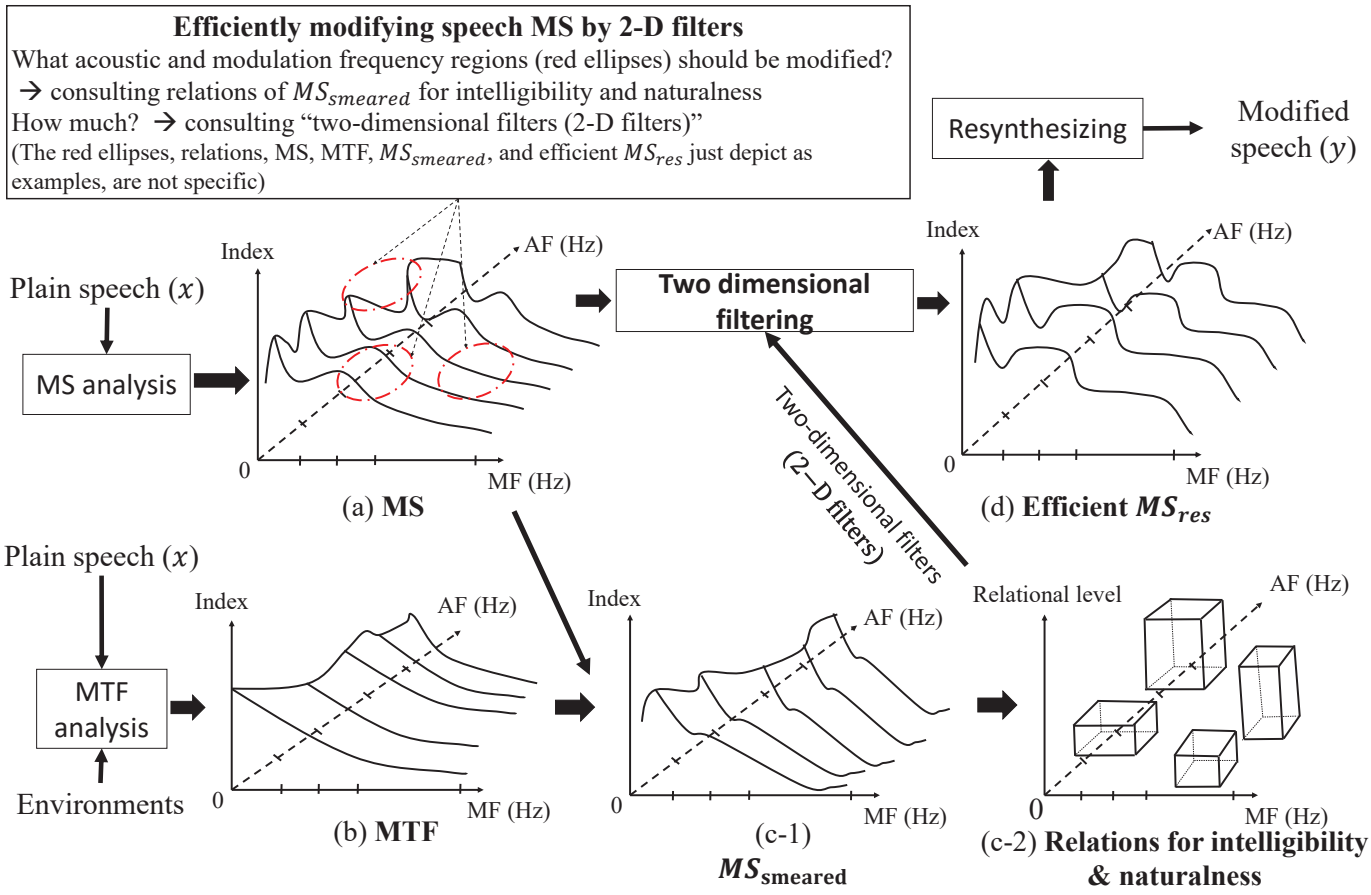


Figure 1: Proposed concept of the present study. AF and MF stand for acoustic frequency and modulation frequency respectively. Environments might contain both noise and reverberation.

- **HEGP**: it simulated the optimal spectral shaping in Tang and Cooke (2018) by increasing spectral regions above 1 kHz by 45 dB, decreasing regions below 1 kHz by 45 dB. The spectral shaping was the result of maximizing HEGP metric for speech spectra, which was later verified by listening tests.

- **SS**: it was static spectral shaping in SSDRC (spectral shaping and dynamic range compression) (Zorila et al., 2012), so-called SS. The SS included: increases in spectra from 1-4 kHz by 12 dB, decreases in spectra below 0.5 kHz by 6 dB/oct, and pre-emphasis. In addition to SS, formant sharpening (FS) and the DRC in SSDRC might well contribute to intelligibility. The DRC also showed the temporal modification. Thus, we accumulated FS and DRC with SS and investigated them as **SSFS** and **SSDRC**.

As a result, the enhanced speech was diverse with different modifications in both acoustical and modulation domains.

The environments consists of five types of noise at two SNR levels (low and high SNR levels, as shown in Fig. 4): Pink noise (Pink-Noise, 1984) (−9.5 dB of low SNR and −12 dB of high SNR), babble noise (Babble-Noise, 1990) (−12 and −9 dB), SM i.e., speech modulated noise created by multiplying the pink noise with the envelope of the babble noise (−10.5 and −7.5 dB), HP noise i.e., high-pass noise created by high-pass filtering the pink noise with cutoff frequency 0.5 kHz (−12 and −9 dB), and LP noise i.e., low-pass noise created by low-pass filtering the pink noise with cutoff frequency 4 kHz (−13 and −10 dB). The SNR levels were estimated to obtain at least 33% of correct-mora recognition for the plain speech in the present noise by pilot tests to avoid full randomness when answering keywords during listening tests. We consulted the study of Tang and Cooke (2018) for the selections of these types of noise and SNR levels.

3.1.2. Listening tests

As was described in the previous section, there were six types of speech: plain, C2, HEGP, SS, SSFS, and SSDRC, i.e., plain speech and five types of the enhanced speech by enhancing the plain speech. The synthesizing method in SSDRC was inherited to apply all those enhancements. We conducted an experiment of subjective listening tests for intelligibility and naturalness with these speech in the presence of the aforementioned five noise maskers at SNR levels. The experimental design was adopted from the study of Tang and Cooke (2018) by the following procedure.

Thirty-six thousand stimuli (600 words × 6 speech types × 5 noises × 2 SNRs) re-sampled at 44,100 Hz and normalized to their average root mean square were involved in the experiment.

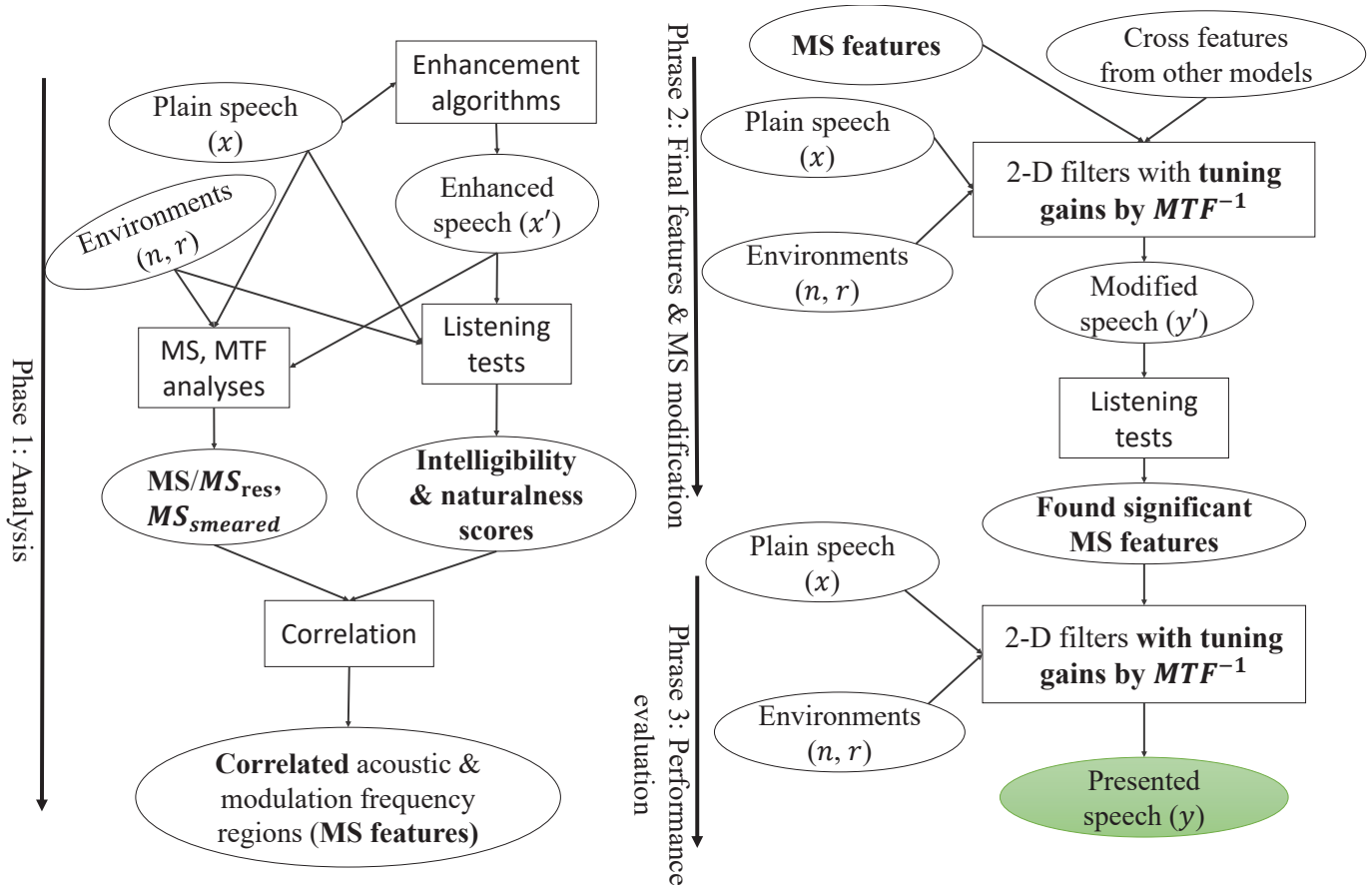


Figure 2: Methodology based on the proposed concept of this study

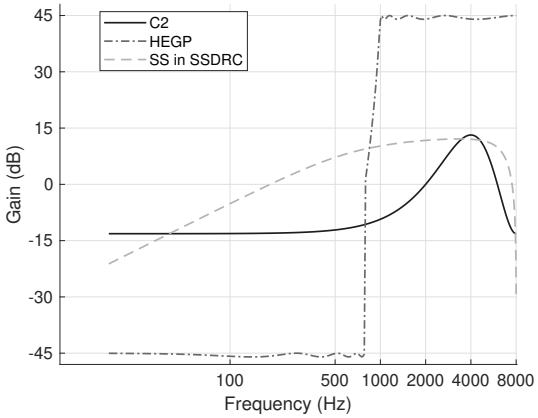


Figure 3: Three basic spectral shaping of the enhancement algorithms used in the present study

Seventeen native Japanese (14 males and three females) aged mean 23.5 and standard deviation 1.7 with no report of hearing problems joined the tests and evaluated all speech types in all noise at two SNRs at 80 dB sound pressure level (SPL) in several sections (The SPL of noise was set at 80 dB). Within one section, one listener listened to 60 unique words at a noise type × an SNR level. The listeners did the intelligibility test and the naturalness test in sequence as follows.

- Intelligibility: During this task, the stimulus was played only one time. The listeners were asked to write down the word they heard by using a keyboard. They clicked on the next button to continue.
- Naturalness: During this task, the stimulus could be played again. The listeners were asked to evaluate their feeling of naturalness (human voices) in four scales (1: unnatural, 2: rather unnatural, 3: rather natural, 4: natural) by clicking on the correspondent buttons. We did not include the distortion or colorization to define the naturalness. There were some instructions for listeners in advance for how to evaluate naturalness. They were told to evaluate according to the feeling of human voices or not. They should not care about the distortion or any meanings of the words.

  The next stimulus would be played immediately after that.
- Configuration: We conducted the experiment in a soundproof room with a high-quality headphone (STAX SL51-2216) connected with a desktop computer via an amplifier (STAX SRM-1/MK-2). We used an amplifier to set an exact noise level of 80 dB SPL for the test, which was measured by a calibrated sound level meter (a hand-held analyzer type 2250 Bruel. & Kjar).

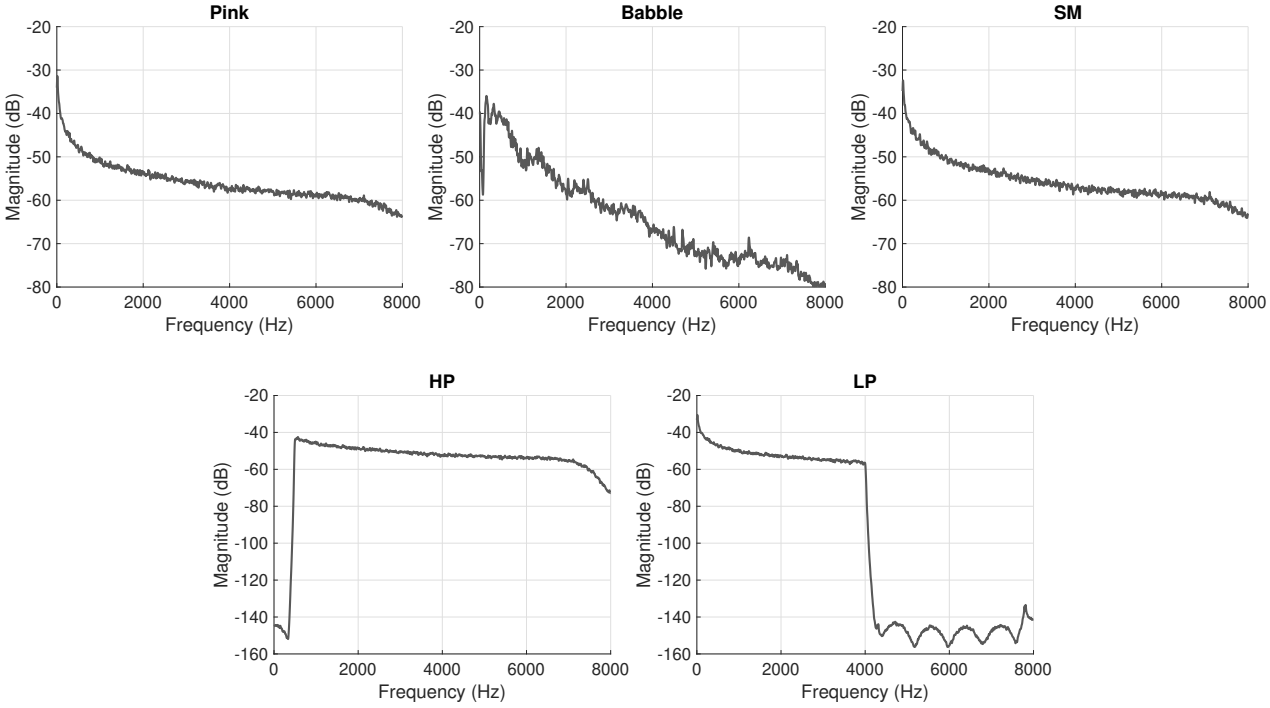


Figure 4: Long-term average spectra of the noise maskers used in the experiment for analyzed speech and in the creation of stimuli for evaluation of significantly effective features

The intelligibility and naturalness scores of these speech in the presence of five different noise maskers at two SNR levels were shown in Figs. 5 and 6 respectively. It could be seen that there was differences among speech types in both intelligibility and naturalness scores. it was lots of increases for SS, SSFS, and SSDRC, and slightly increases or decreases for C2 and HEGP. DRC are often expected to add much improvement. However, in this study the speech materials were three-mora Japanese words with short durations and, perhaps a quite balanced phoneme structure in a mora (the consonant and the vowel in a mora seem have balanced power envelopes). Therefore, the DRC was still beneficial but lightly presented within short durations and such the phoneme structure when aiming to emphasize abrupt regions as voice onset, offset and consonant parts. This result showed that the enhancement algorithms with their modified frequency regions might present **diverse relations** for intelligibility and naturalness. Pearson correlation coefficient between intelligibility and naturalness scores was 0.61, which was only quite highly correlated. This correlation indicated that these scores might be unable to imply each other. Both intelligibility and naturalness scores thus still needed to be investigated in further analyses to ensure accuracy. Next, the MS/MS_{res}, MTF, and $MS_{smeared}$ of clean speech and noise used in the experiment were calculated by the methods in the next section.

3.1.3. Analyses of modulation spectrum, modulation transfer function and formation of smeared modulation spectrum

The estimation of MS/MS_{res}, MTF and $MS_{smeared}$ were based on modulation filtering techniques in both MS and MTF analyses. For performing this analysis ideally, we used the designated source signals of clean speech and noise in the listening tests.

a. Modulation filtering

Numerous methods have been reported for the estimation of speech MS (Ivanov and Chen, 2012, Moro-Velázquez et al., 2016). A study of modern psychophysical models of temporal processing indicated that temporal amplitude envelope is processed by a modulation filter bank with multi-resolution frequencies of the temporal envelope (Zhu et al., 2018, Jørgensen et al., 2013). As was used in Unoki and Zhu's study (Unoki and Zhu, 2020b) and Zhu *et al.*'s study (Zhu et al., 2018), we used a modulation filtering technique to estimate the MS/MS_{res} of the plain and enhanced speech. This modulation filtering was performed using an acoustic filter bank concatenated with a modulation filter bank. The former was a bank of 18 filters: $1/3^{rd}$ octave band-pass filters with bandwidths of 160-8000 Hz, which followed the SII specifications. The latter was also a bank of 18 filters: a low-pass filter (2nd order Butterworth filter) with cutoff frequency $Fc = 0.4$ Hz and 17 $1/3^{rd}$ octave band-pass filters with bandwidths of 0.5-20 Hz. Houtgast et al. (Houtgast and Steeneken, 1985) also used 0.5-20 Hz band-pass filters to calculate the speech power envelope spectrum. Our estimated MS/MS_{res} thus contained 0 Hz modulation and showed as acoustic frequency features (also as AF features). The time or modulation frequency features (also as MF features) were above 0 Hz up until a modulation frequency of 20 Hz.

Furthermore, it was thus needed to discuss the capability in a

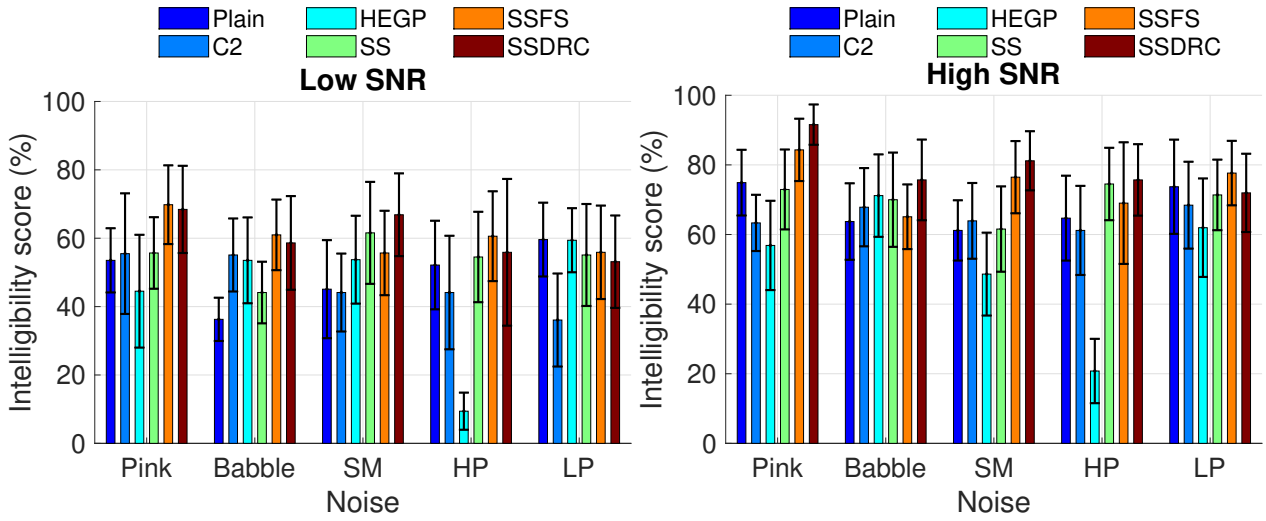


Figure 5: Intelligibility scores (percentage of correctly identified mora in a word) of analyzed speech in the presence of making noise at low and high SNRs.

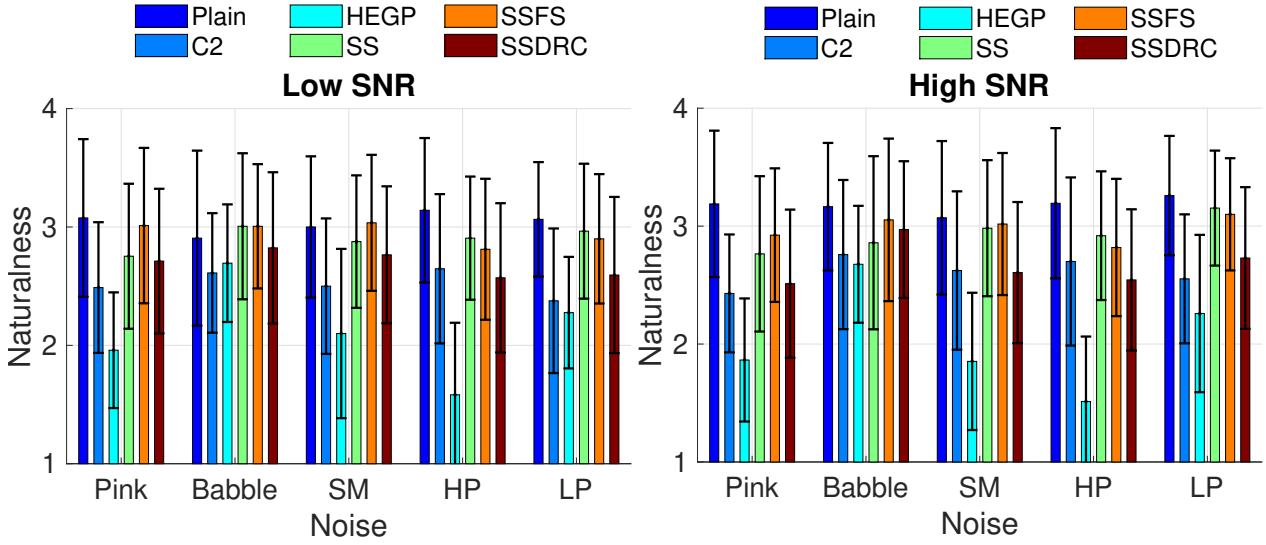


Figure 6: Naturalness scores of analyzed speech in the presence of making noise at low and high SNRs.

representation of both local and global time-frequency features in the MS/MS_{res}. As $1/3^{rd}$ octave bands were used, the bandwidths were suitable to capture both global frequency features, which might be from changes in spectral tilt and large frequency regions, and local features, which might be from changes in formants or narrow frequency regions. Also, as indicated in the calculation of STOI (short-time objective intelligibility) (Taal et al., 2010), to have a proper modulation frequency resolution, an appropriate length of the analysis time frame could be 300-500 ms, which can cover modulation frequencies until 2 to 3 Hz (or 1/500 ms to 1/300 ms). This time frame was also the same as the average duration of the plain speech and enhanced speech used in our listerning tests for analysis.

b. Modulation transfer function

The MTF is fully determined mathematically for stationary noise by the signal-to-noise ratio (Houtgast and Steeneken, 1985, Unoki et al., 2009). For each band-limited acoustic frequency, i.e, f_a, the MTF is independent of the band-limited modulation frequency, i.e., f_m and defined as

$$m_N(f_{a_k}, f_m) = \frac{1}{1 + 10^{\frac{-SNR_{f_{a_k}}}{10}}} \tag{3}$$

where $SNR_{f_{a_k}} = 10\log 10\left(\overline{e^2_{x_k}}/\overline{e^2_{n_k}}\right)$. x_k and n_k were filtered speech and noise at the k^{th} f_a and e^2 was power envelope. Without explicitly showing the ordered f_a, we obtained the MTF in noise as $m_N(f_a, f_m)$.

c. Smeared modulation spectrum

From Eqs. 1 and 3, $MS_{smeared}$ for each acoustic band f_a and each modulation band f_m can be calculated using

$$MS_{smeared}(f_a, f_m) = MS(f_a, f_m) \times m_N(f_a, f_m). \tag{4}$$

In short, $MS_{smeared}$ at 0 Hz modulation shows the effect of noise on the frequency features. The $MS_{smeared}$ at over 0-20 Hz modulation shows the effect of the reverberation on the time features.

3.1.4. Extraction of modulation spectral features

So far, we had collected several speech enhancement methods and applied them to increase intelligibility, then investigated the properties of the enhanced speech. Each method was used to mainly modify different acoustic and modulation frequency regions. The resulting intelligibility and naturalness scores differed. Meanwhiles, we calculated the MS/MS_{res}, MTF, and $MS_{smeared}$ of speech and noise used in the intelligibility and naturalness evaluation. Then, we identified **correlated** acoustic and modulation frequency regions or correlated MS features to modify the MS of plain speech more by using correlation between these $MS_{res}/MS_{smeared}$ and these intelligibility and naturalness scores.

The strategy for us to decide the boundaries for the extracted acoustic and modulation frequency regions was by a relative consideration (not precise) over the lower and upper frequencies of the $1/3^{rd}$ octave bands used by the modulation filtering. We directed the boundaries toward the patterns appeared in the analysis results and connected them with the previous knowledge about formants, consonant bursts, vocal fold movements, and the dominant modulation frequencies and their selective regions for intelligibility and prosody.

a. MS features: AF components

The $MS_{smeared}$ at 0 Hz modulation showed the effect of noise on the frequency features. Figure 7 shows the MS_{res} and $MS_{smeared}$ at 0 Hz modulation of the analysis speech with different levels of intelligibility. It could be seen that MS (Fig. 7a) only reflected the increased modification made by the spectral shaping. However, in Figs 7b and 7c, $MS_{smeared}$ presented about what was increased in the $MS_{smeared}$ of the enhanced speech comparing to the $MS_{smeared}$ of plain speech to reach to the MS of plain speech that might contribute to intelligibility shown in Fig. 5. The intelligibility was better for SS and its family than C2, HEGP, and plain speech. This better intelligibility might be because the $MS_{smeared}$ indexes of the frequency regions around 500 Hz, 1.25 kHz, 2.5-3 kHz, and 5-6 kHz were increased as can be seen in Figs 7b and 7c. These frequencies seemed to relate to vowel formants (F_1 and F_2, which were about 500 Hz, and 1.5 kHz), consonant bursts, and vocal fold movements (around 4-6 kHz). The heavily decreased $MS_{smeared}$ for around 500 Hz might cause decreasing the intelligibility and naturalness of HEGP.

b. MS features: MF components

The MS_{res} obtained using DRC, which modified the time features, was the difference between the MS_{res} of SSDRC and SSFS for 0-20 Hz modulation.

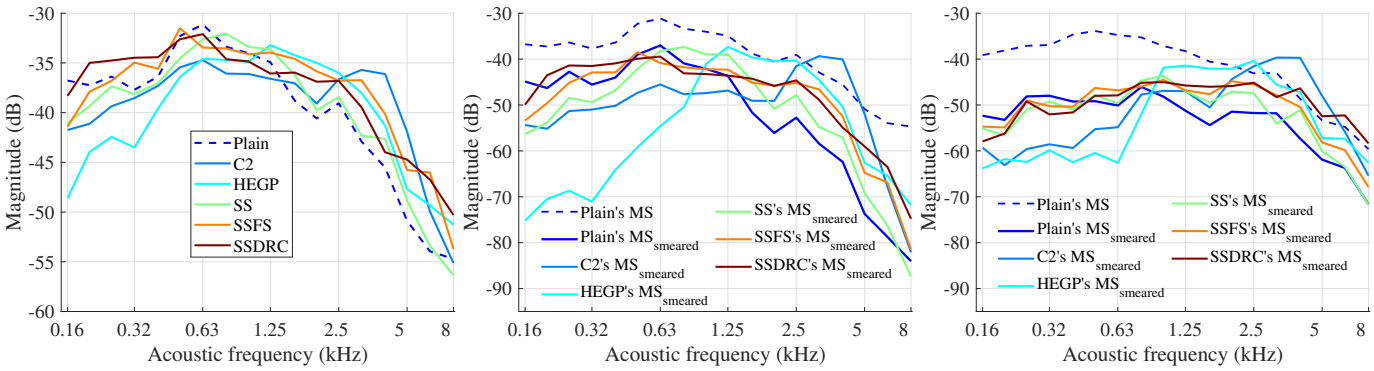
As shown in Figs. 5 and 6, SSDRC got higher intelligibility and lower naturalness than SSFS. This could be because DRC increased MS_{res} at two regions: around 4 Hz and above 8 Hz modulation (Figs. 8a and 8b). As shown in Fig. 8b, this MS_{res} had two peaks, one at around 4 Hz and one at around 20 Hz. It could be noted that as we analyzed on the data of various words and different male and feature speakers, the resulting 4 Hz and 20 Hz modulation frequency can be considered not specific for data, speaker, or gender. Furthermore, the frequencies seem align with previous findings on dominant frequencies in modulation spectrum for speech intelligibility (4-6 Hz) and prosody (6-8 Hz). Meanwhile, 20 Hz seems to be an upper frequency for the modification still affecting. Therefore, these time features were coincident with the typical peak at 4 Hz of speech MS for speech intelligibility and above 8 Hz for speech prosody.

c. Correlation

The Pearson correlations between $MS_{smeared}$ and MS_{res} for each acoustic frequency band over three modulation frequency bands and the intelligibility/naturalness scores were calculated. Each acoustic frequency band for $MS_{smeared}$ at 0 Hz modulation was used in the calculation. Given that the environment was noise only, the effects of the environment on $MS_{smeared}$ for 0-20 Hz modulation were equal; we thus needed to consider the MS_{res} for 0-20 Hz, i.e., the MS_{res} by DRC. Also, due to the two peaks around 4 Hz and 8-20 Hz modulation, we used each acoustic frequency band of MS_{res} for the modulation frequency bands of 0-8 and 8-20 Hz of DRC to entirely cover these two peaks in the calculation. The correlation could be interpreted like that the more positive correlation coefficients, the more agreement between increasing $MS_{smeared}/MS_{res}$ indexes and increasing intelligibility and vice versa.

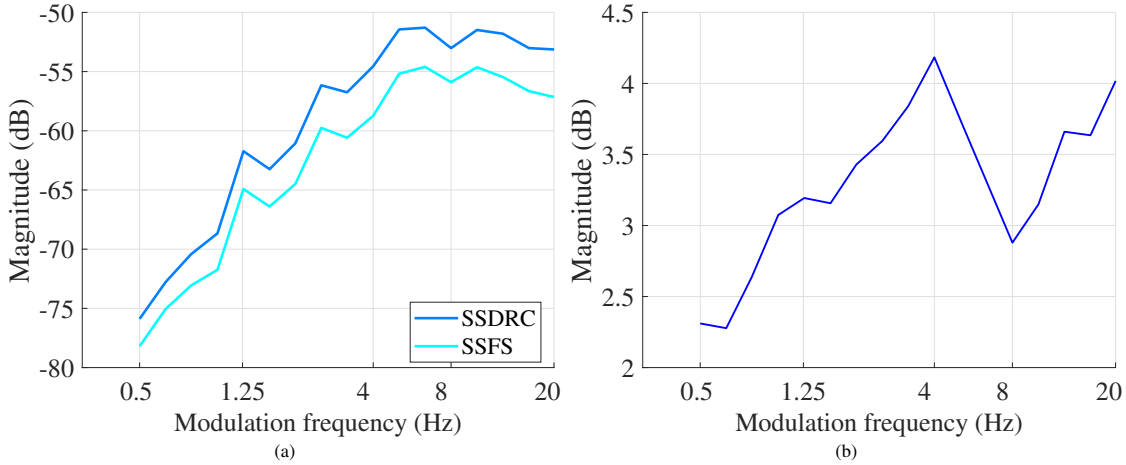
Figure 9 shows correlations between $MS_{smeared}$ and MS_{res} for each acoustic frequency band (0, 0-8, 8-20 Hz modulation bands) and intelligibility scores for analyzed speech in noise in Fig. 5. In the stationary noise of Pink, SM, HP, and LP noise (Fig. 9a), it was shown that some frequency regions below 500 Hz still had high correlation with the intelligibility scores. It seemed suitable to choose 250 Hz to include F1 for vowels /i/ and /u/. Thus the highly correlated frequencies with intelligibility were *relatively* 250/500-2250 Hz, 4.5-6.5 kHz, and around 4 Hz and above 8 Hz modulation in the acoustic spectra of 300-750, 1250-2250 Hz and 4.5-6.5 kHz. These acoustic spectra were also the regions for formants, consonant bursts, and vocal fold movements. The region between 750 and 1250 Hz with around 4 Hz and 8 Hz modulation got a negative correlation while they were positive at 0 Hz modulation. It could be seen that this result seemed reasonable because the 750-1250 Hz region could be the dip between formants F_1 and F_2, when making so many fluctuations by increasing the MS indexes above 0 Hz within this region, it might affect to reduce the prominence of formants. Therefore, we should avoid increasing the MS indexes at this region for the time features.

In the non-stationary noise of the babble noise (Fig. 9b), it seemed that the highly correlated frequency regions with intelligibility could be started from 500 Hz or 1 kHz and con-



(a) MS (plain) and MS_{res} (others) (b) $MS_{smeared}$ (Pink, low SNR) (c) $MS_{smeared}$ (Babble, low SNR)

Figure 7: MS, MS_{res} and $MS_{smeared}$ at 0 Hz modulation of analyzed speech in the presence of some noise at some low and high SNRs.



(a) (b)

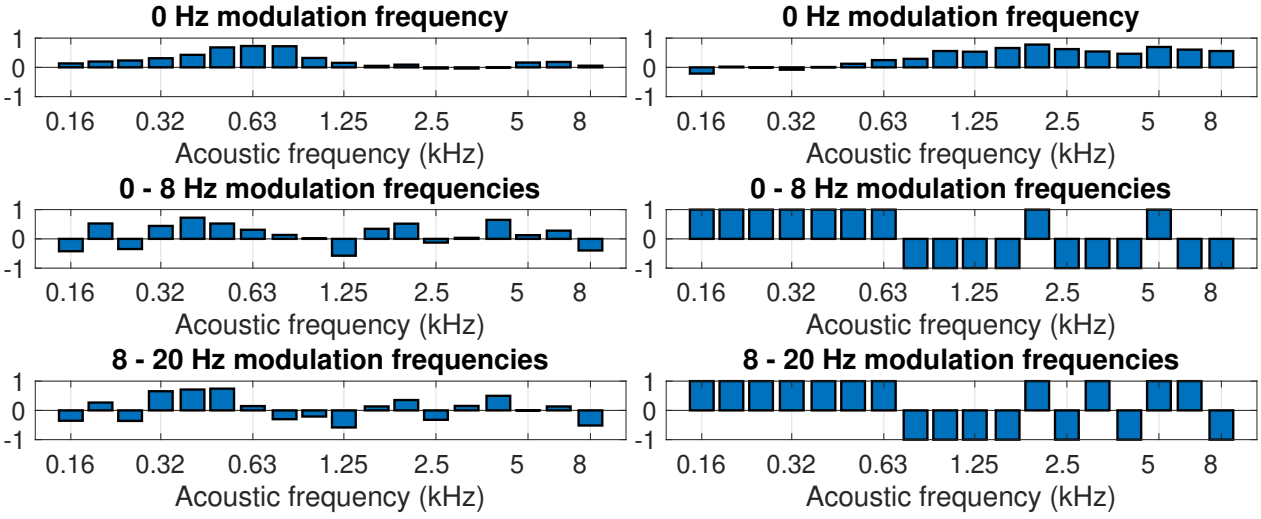
Figure 8: (a) MS_{res} of SSDRC and SSFS and (b) their difference (MS_{res} by DRC) over 0-20 Hz modulation in the acoustic spectrum of 5 kHz in the presence of SM noise at high SNR.

tinued expanding the entire higher frequency regions. This correlation meant that the frequency features could also be increased in MS above 500 Hz or above 1 kHz. These features seemed to be coincident with the arguments from previous studies (Tang and Cooke, 2018), whether to increase the spectrum from 500 Hz or 1 kHz. Also, it could be seen that the time features in this babble noise were relatively as same as the time features in the stationary noise.

Figure 10 shows correlations between $MS_{smeared}$ and MS_{res} for each acoustic frequency band (0, 0-8, 8-20 Hz modulation bands) and naturalness scores for analyzed speech in noise in Fig. 6. In both stationary and non-stationary noises, in the $MS_{smeared}$ at 0 Hz modulation band, highly correlated frequencies with naturalness was the acoustical bands below 1 kHz. The increases in the MS at higher acoustical bands than 1 kHz harmed naturalness. However, it might be important to increase MS at these regions for speech intelligibility. Thus, we still considered increasing them with particular concerns with naturalness in the evaluation. MS_{res} for acoustical bands around 500, 1250-2250 Hz, and 4.5-6.5 kHz (0-8, 8-20 Hz modulation bands) seemed also highly correlated with naturalness.

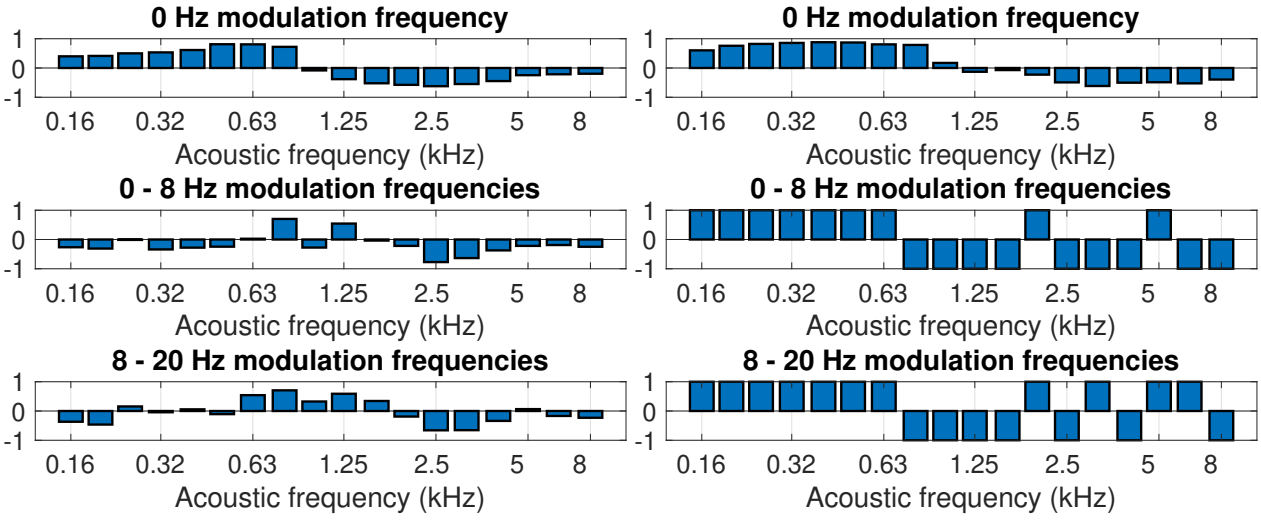
The effects on the time features for naturalness were relatively the same as these for intelligibility. These shared features might come from a similar pattern between intelligibility and naturalness scores (as shown in Figs. 5 and 6). Therefore, we used the time-frequency features extracted by the correlation with intelligibility scores as the final features. Furthermore, the time features seemed to depend on the frequency features due to MS_{res} for only specific acoustic frequency bands (0-8, 8-20 Hz modulation bands) positively correlating with intelligibility/naturalness scores.

From the explanation above, we tentatively selected the correlated MS features as follows. Based on the correlated frequencies regions, which reflected formants, consonant bursts, and vocal fold movements, the frequency features could be sparsely increased in the MS around 250/500-2250 Hz (AF region 1) and 4.5 - 6.5 kHz (AF region 2) acoustic frequencies or continuously increasing in the MS above 500 Hz or above 1 kHz, so-called AF features. Because there existed the modulation selective process among modulation frequency regions (Jørgensen and Dau, 2011), our multi-resolution frequency analysis over the $1/3^{rd}$ octave scales, and the dominant frequencies around between 2-20 Hz in overall with the peak at 4 Hz for speech intelligibility and 8-20 Hz for speech prosody, we made the two separated regions as 2-6 Hz and 8-20 Hz modulation. There-



(a) In the stationary noise (b) In the babble noise

Figure 9: Pearson correlation between $MS_{smeared}$ and MS_{res} for each acoustic frequency band (0, 0-8, 8-20 Hz modulation bands) and intelligibility scores for analyzed speech in noise in Fig. 5.



(a) In the stationary noise (b) In the babble noise

Figure 10: Pearson correlation between $MS_{smeared}$ and MS_{res} for each acoustic frequency band (0, 0-8, 8-20 Hz modulation bands) and naturalness scores for analyzed speech in noise in Fig. 6.

fore, the MF features could be increased in MS around 2-6 Hz (MF region 1) and 8-20 Hz (MF region 2) modulation in the acoustic spectra of 300-750, 1250-2250 Hz and 4.5-6.5 kHz, so-called MF features.

3.2. Phase 2: Final features & MS modification

We proposed control method for MS modification based on 2-D filters, which was described in the following sections. Then, by doing experiments to examine single and combined feature contributions to intelligibility and naturalness, we finalized the significant MS features among the feature sets described in feature sets in the Sec. 3.1.4 to increase intelligibility and preserve naturalness under noise. They were two sparse features: A500s and M as our definition, significantly addictively contributing to increase intelligibility and preserve naturalness for the presented speech under noise.

- o A500s: increasing modulation spectra at AF regions 500-2250 Hz by 5-15 dB and 4.5 - 6.5 kHz by 15-25 dB. It can be seen that increasing these frequency regions as this feature shows a more emphasis on magnitude at the regions of vowel formants and consonat bursts
- o M: increasing modulation spectra at MF regions 2-6 Hz and 8-20 Hz by 6-10 dB in the acoustic spectra of 300-750, 1250-2250 Hz and 4.5-6.5 kHz. This feature also show a more emphasis on magnitude at the common modulation spetral peaks for speech intelligiblity and speech prosody at the acoustic frequencies of vowel formants and consonant bursts.

We only briefly summarized these features like this to more concentrate on how to modify speech MS for intelligibility and naturalness. Specifically, the proposed 2-D filter design and MS modification based on these features to obtain the best perfor-

mance were as follows. To synthesize MS-modified speech, an analysis-synthesis method based on a multi-rate signal processing technique (Milic, 2009) was developed to modify MS of plain speech by amplifying acoustical and modulation bands as described in the extracted MS features. The amplified values were the averaged MTF^{-1} calculated (simply, an inverse operation) from the MTF of the reference noise by Eq. 3 within specific acoustic and modulation frequency regions of the correlated MS features with a correction in limited ranges to preserve the voice quality of the plain speech. The AF regions 1 and 2 of the sparse frequency features and the frequency regions of the continuous frequency features were all empirically limited to 10-15 dB, i.e., a mostly flat response for a fair comparison of intelligibility. The derivation of these amplifications as 2-D filters is illustrated in Fig. 11. It should be noticed that this developed method was for synthesis based on a multi-rate signal processing technique. It was different from the modulation filtering technique described in Sec. 3.1.3, which was mainly used for the estimation of MS/MS_{res}, MTF, and $MS_{smeared}$ in Sec. 3.1. In this synthesis of MS-modified speech, we only used the modulation filtering technique for the estimation of MTF.

3.2.1. 2-D filter design

Figure 11 shows a three-step estimation of the 2-D filter for MS modification. Figure 11 (a) presents a typical MS estimated from noise and reverbeation. As was described before, the MTF is fully determined mathematically for stationary noise by the signal-to-noise ratio (Houtgast and Steeneken, 1985, Unoki et al., 2009). For each band-limited acoustic frequency, i.e, f_a, the MTF is independent of the modulation frequency, i.e., f_m and defined as $m_N(f_a, f_m)$. In this estimation, if the reverberation is involved, the MTF for reverberation noise was defined as $m_R(f_a, f_m)$ using the modulation filtering technique for a provided delivered room impulse response (RIR). The MTF under noisy reverberant conditions was calculated using

$$m(f_a, f_m) = m_N(f_a, f_m) \times m_R(f_a, f_m). \quad (5)$$

From (1) and (5), $MS_{smeared}$ can be calculated using

$$MS_{smeared}(f_a, f_m) = MS(f_a, f_m) \times m(f_a, f_m). \quad (6)$$

In Fig. 11 (b), the regions demonstrate the A500s and M features. The average is logically interpreted like the way the MTF is constructed. That is, average was taken for MTF in noise and reverberation separately. Firstly, it was taken at 0 Hz for frequency features (A500s) with MTF in noise and is also going to apply for all modulation frequencies in the amplified process. Secondly, it was taken above 0 Hz (specifically, modulation frequency regions in their accompanied frequency regions specified in time features, i.e., M) with MTF in reverberation and is going to apply for only these modulation frequency regions in their accompanied frequency regions in the amplified process.

Figure 11 (c) partially depictsthe filter shape with a portion of A500s with the region of 500-2250 Hz overlapped with a portion of M (2-6 Hz and 8-20 Hz modulation in 1250-2250 Hz). As can be seen that the gain by MTF in noise is made up for all modulation frequencies at the portion of A500s and the gain by MTF in reverberation is added afterward at the portion of M. Furthermore, these inverse MTFs are also limited within ranges for a safer modification to avoid over-modifying these features as previously mentioned in the A500s and M.

3.2.2. Modification of modulation spectrum

Specifically, to imitate the MS analysis, enabling reconstruction, we used a multi-rate signal processing technique for the MS analysis, modification, and synthesis steps (Fig. 12a). An acoustical analysis bank was used to filter plain speech into band-limited signals, and then the power envelopes of the band-limited signals were extracted. Next, to avoid modifying non-speech segments (modifying them might cause noise), VAD was used to mask the speech-absent portions of these power envelopes with a silence threshold of 0.005 on the max-value-normalized power envelopes. A modulation analysis bank was then applied to each speech-segment power envelope. Afterward, a processing unit with gain control amplified specific acoustical bands and modulation bands (as mentioned in settings in the A500s and M and in Fig. 11). The processing unit sequentially applied gains by the 2-D filters to the acoustic frequency regions (all 0-20 Hz modulation regions) and the modulation frequency regions (0-20 Hz, excluding 0 Hz). Finally, to obtain the modified speech, the reconstruction was processed in inverse order from the analysis with modulation synthesis bank, VAD, and acoustical synthesis bank. The VAD implemented for the synthesis process is not the same as the VADs in the analysis process. The VADs in the analysis is for extract the voiced regions for modification. The VADs in the synthesis process is to provide location information, which was used to realign the voiced and unvoiced regions correctly as was in the analysis. It is also used to eliminate any unexpected calculation affecting the unvoiced regions. All these processes are illustrated in Figs 12b and 12c, which detail the low-level architecture of the high-level architecture in Fig. 12a. Further information for designing the acoustical analysis/synthesis banks and the modulation analysis/synthesis banks was described as follows. We designed equal-bandwidth filter banks with a perfect reconstruction of signals using a tree-based model (Milic, 2009, Chapter 12) [1]. Starting from the basis of the analysis/synthesis orthogonal bank with two FIR filters (implemented by the function *firpr2chfb* in Matlab), expanding to deeper levels, i.e., five levels was to construct 32 (i.e., 2^5) filters for each acoustical analysis/synthesis bank. Also, expanding seven levels was to create 128 (i.e., 2^7) filters for each modulation analysis/synthesis bank. Thirty-two filters or bands were chosen because the sampling frequency of plain speech was at 16 kHz, and it was to obtain a bandwidth of 250 Hz on each acoustical band. This bandwidth was suitable to represent the extracted MS features

[1] This model is conventional in multi-rate signal processing for the modification and synthesis. The basic knowledge is mentioned in the chapter 12 of the book "Multirate Filtering for Digital Signal Processing: MATLAB s". It is recommended to go through some implementation exercises of the chapter 12 (Milic et al.) to understand and further develop specialized tools as our proposed modification-synthesis method.

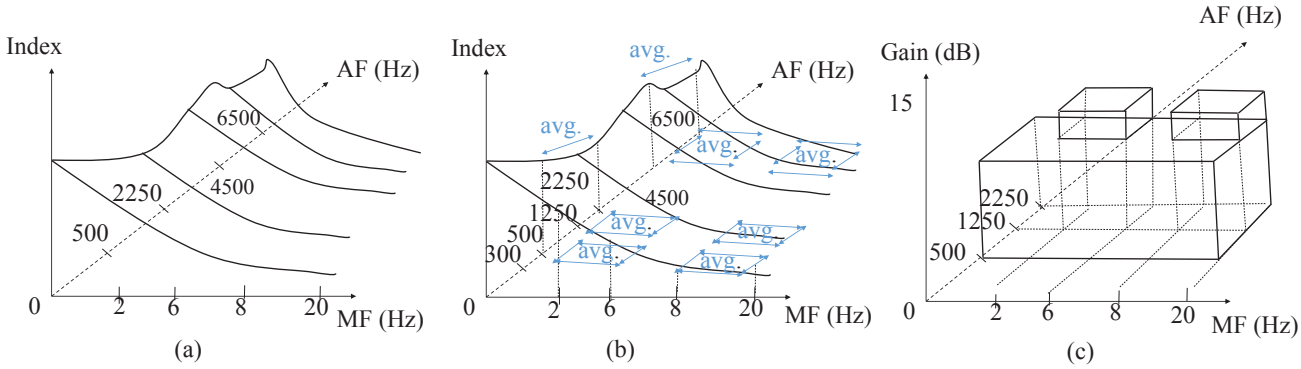


Figure 11: Gain derivation from MTF by environments: (a) MTF estimated from SNRs and a room impulse response (assuming the reverberation exists); (b) MTF with specific acoustic and modulation frequency regions indicated in the investigated features that were taken average. (c) A construction of gain by inverting the averaged MTF.

Table 1: SNR (decibels, dB) under various conditions used in HC 2.0 listening tests.

Reverberation	**SNR**	**German**	**English**	**Spanish**
near (1 m)	low	−15.0	−13.0	−17.0
	mid	−12.5	−8.5	−14.5
	high	−10.0	−4.0	−11.5
mid (2.5 m)	low	−13.0	−11.0	−17.0
	mid	−10.0	−5.0	−14.0
	high	−7.0	−1.0	−11.0
far (4 m)	low	−13.0	−10.0	−18.0
	mid	−9.0	−4.0	−14.0
	high	−5.0	2.0	−10.0

as frequency features. Due to the bandwidth of 250 Hz of the acoustical band, the modulation analysis/synthesis banks thus needed 128 filters to obtain a bandwidth of 2 Hz for each modulation band to enable us to modify the extracted MS features as time features. These banks of acoustical and modulation analyses/syntheses can obtain a perfect reconstruction of signals with only some delay. The delay was removed by shifting the reconstructed signal.

3.3. Phase 3: Performance evaluation

We aimed to efficiently control speech MS for improving speech intelligibility and naturalness in adverse conditions. So far, the features A500s plus M of the present study (a time-frequency feature pair) were evaluated in noisy reverberant environments and languages by Hurricane challenge 2.0 (HC 2.0) (Rennies-Hochmuth et al., 2020, Rennies et al., 2020), which was referred to as **MS500** (Van Ngo et al., 2020). The speech material was in German and Spanish (100 sentences each) and English (90 sentences) as recorded by native male speakers and was used as plain speech. The evaluation was performed by HC 2.0 with about 180 listeners. They created stimuli for the experiment using our provided MS500 speech, the other speech from other methods and plain speech, babble noise, and the RIR. The clean speech was filtered using the RIR, and then the noise was added to obtain the targeted global SNR. There should not have any differences statistically in experimental conditions among the proposed and other methods. The design and configuration of the experiments were performed under the regulation of HC 2.0. Some documents related to their design were provided to us (URL: https://www.cstr.ed.ac.uk/projects/hurricane/2/) and any other participants with clear explanation about settings, listeners and conditions. Table 1 shows the evaluated SNRs and reverberation conditions in terms of the distance between the loudspeaker and the listener. Figure 13 shows the waveforms and spectrograms of plain speech and our MS500 speech for the English words ”four large rings” produced in low SNR and far reverberation conditions.

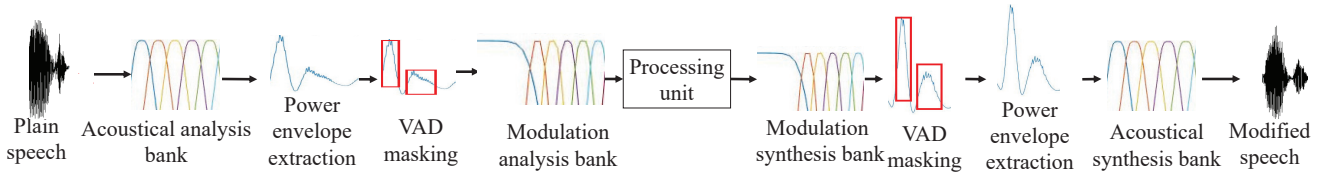
The methods submmited to the HC 2.0 included ACO (Bederna et al., 2020), ASE (Chermaz and King, 2020), exactMaxSII, DeepSSC-Lomb, DSSC-L/eMSII, iMetricGAN (Li et al., 2020), MS500, IISPA (Schädler, 2020), and SSDRC [see (Rennies et al., 2020) for more details about those methods]. Figure 14 shows the differences between intelligibility scores by these methods and the plain speech under the tested conditions.

(-) The best method (ASE) could get upto 60% better recognition rates than the plain speech in all conditions. This method applied knowledge from sound engineering to control auditory band signals. Perhaps, this existing knowledge could be a competitve method for the noval research.

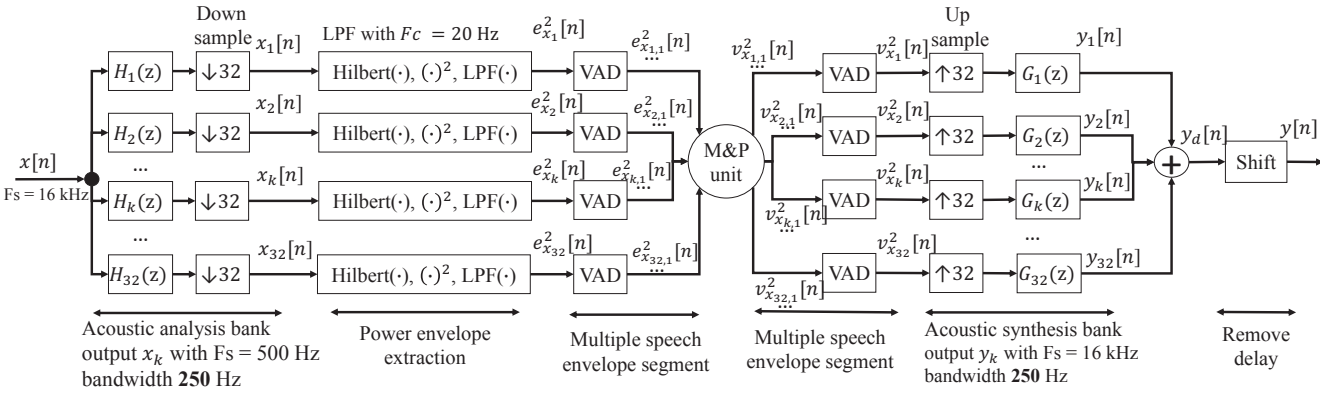
(-) The method with the lowest performance in most of all condition is DeepSSC-Lomb. This method controled the duration of speech signal, which might cause some problems when presenting speech under reverberant conditions.

(-) MS500, ACO and IISPA seems to share a similar trend about their performance over languages. Each of them performed worse in one language while better in the others.

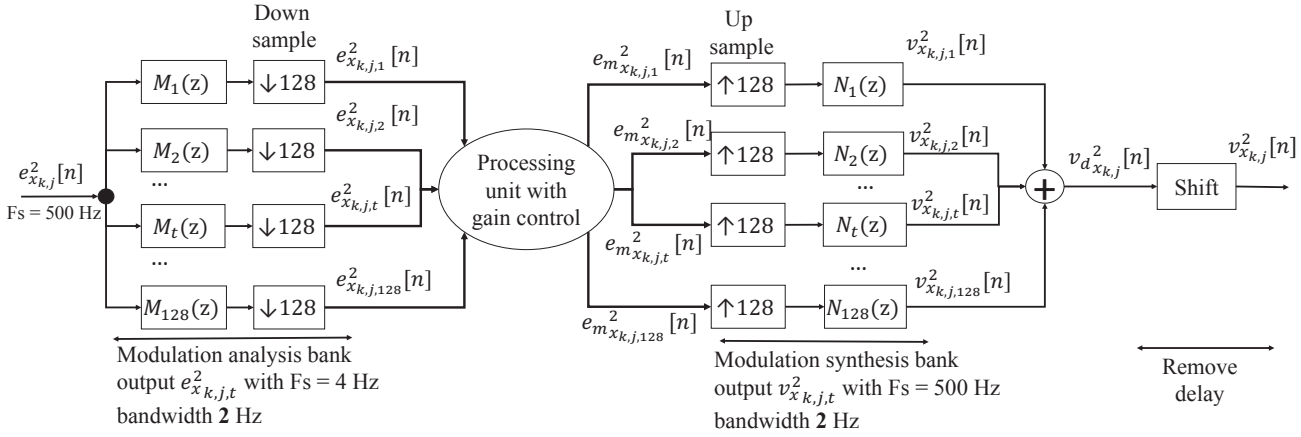
(-) MS500 only got an averaged result among them. Specifically, the MS500 obtained an improvement of intelligibility for English and Spanish in all conditions but not for Ger-



(a) Block diagram of process for converting plain speech into MS-modified speech using multi-rate signal processing technique.



(b) Acoustic analysis and synthesis banks, power envelope extraction, and power envelope masking with voice activity detection (VAD)



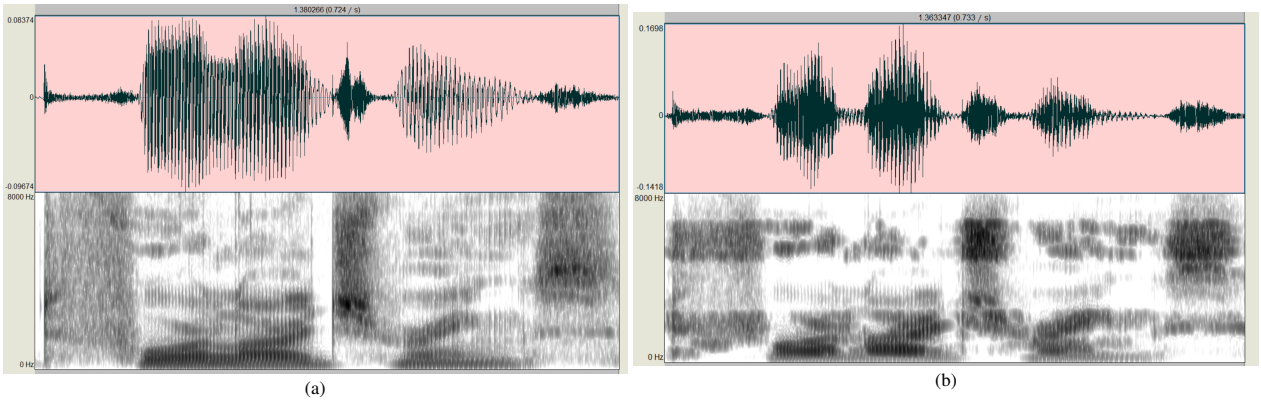
(c) Modulation analysis and synthesis banks (**M&P unit**)

Figure 12: Conversion of plain speech into MS-modified speech using multi-rate signal processing technique.

man in some conditions. An improvement of about 4 – 18% in recognition rates from plain speech was obtained.

Comparing to the ASE method (Chermaz and King, 2020), which was designed and tuned by professional acoustic engineers for the specific noise and reverberant conditions. On the other hand, our method was tuned by the estimated MTF. This makes our method applied in any environmental conditions.

MS500 was a result extracted from studying enhanced speech under designed noise. MS500 came up with the systematic concept to modify speech MS based on the MTF and MS concepts. Besides, MS500 came with more flexiblility and stability. MS500 i.e., A500s plus M were controlled in such a way that used inverse of an estimation of MTF by given reference noises to increase spectral amplitude in the specific acoustic and modulation frequency regions. Specifically, comparing to MTF^{-1}-based methods, we were not going to eliminate all effect of MTF over all frequencies, which can be dangerous if some important speech features for intelligibility are broken. Instead, we focused on the frequency regions of A500s and M, which might be more appropriate because only important parts for speech intelligibility were increased. The MTF within the frequency regions specified in A500s and M was averaged, inversed, and limited to a range of 5-15 dB for the region between 500-2250 Hz and 15-25 dB for the region between 4500-6500 Hz. The limited range for modulation frequency regions was approached to 6-10 dB. These limited ranges were the results of doing several pilot tests for checking intelligibility improvements on various limitation. They were wider than the ranges used in the verification experiments of this study, perhaps, to allow more freedom in response to low SNR conditions. In short, the advanced control strategy that is tuning the MS amplitudes with different amplification of these significant frequency re-



(a) (b)

Figure 13: Waveforms and spectrograms of plain speech (a) and MS500 speech (b) for the English words "four large rings" produced in low SNR and far reverberation conditions.

gions to identify the best increase in speech intelligibility. As a result, this strategy improved speech intelligibility and became robust to varied noisy conditions. It could be considered as our recently proposed efficient RMS or MS_{res} and will be improved in near future. Otherwise, the limited increases in intelligibility could be due to our modification techniques. We only modified the MS magnitute while the phase information should be also modified to avoid mismatching problems between magnitude and phase. German words often contain many plosive consonants, which are delicated to modify their MS. In future work, we intend to improve our modification techniques more based to resolve their limitation.

4. General discussion

Our methodology can be grouped into reinforcement learning to extract typical features and a set of rules. In this study, a correlation method was applied to explicitly reinforce the learning rules. A deep reinforcement can be applied but it seems problematic that the deep reinforcement learning could be appropriate as we might need huge data and a complex model design. Moreover, in our situation, the deep leaning model can be over complicated and implicit comparing to the correlation with explicit analysis results for a dynamical modification in the next steps in our methodology.

Regarding the commonalities and differences from previous methods (Wang et al., 2017, Hermansky and Morgan, 1994, Kanedera et al., 1997, Lee et al., 2018), especially the method by Wang et al. (2017) known as DSR, both our proposed method and the DSR method altered or kept the low modulation frequency regions. For differences, our method emphasized two dominant low frequencies 4 Hz and 20 Hz and their vicinity, while the DSR might keep all low modulation frequency regions and disregard higher frequencies. Our method is 2D filtering on both acoustic and modulation frequency, the DSR might only be 1D filtering on modulation frequencies.

Intelligibility is also related to the context (Hawkins and Warren, 1994) and duration (Cooke and Aubanel, 2017) of utterances. In this situation, rather than modifying the modulation spectrum, some other modifications may improve the intelligibility. We tried to exclude contextual predictability in our study to fairly investigate other features as we experimented with the plain speech of short words with no repetition and low familiarness in the listerning tests for our analysis. Duration can contribute to intelligibility in fluctuating noise and seems not contribute to intelligibility in stationary noise. Therefore, duration might not be a stable feature for any kinds of noise and/or reverberation. Duration was not modified in this study.

5. Conclusion

In conclusion, this study have presented an approach to control speech modulation spectrum using environmental influences efficiently by two-dimensional filtering to improve speech intelligibility and naturalness under adverse conditions. To do this, firstly, basing on the basic concepts of modulation spectrum and modulation transfer function, we performed an analysis on correlated relations of smeared modulation spectrum of different enhanced speech for their intelligibility and naturalness scores under noise by listening tests to extract correlated acoustic and modulation frequency regions of speech modulation spectrum, briefly called correlated MS features. Then, the correlated MS features were investigated with any of their single/jointly combinations for intelligibility and naturalness benefits to identify the significant ones. Meanwhile, the MS modification technique based on multi-rate signal processing was proposed. A method to derive the gains for two-dimensional filters, which was applied for these correlated MS features, was invented with the consultation of the inverse modulation transfer function by environments. Finally, the efficient resistant modulation spectrum was obtained with the design of the two-dimensional filters as AF region 1 from 500-2500 Hz, AF region 2 from 4500-6500 Hz with filter gains derived by consulting MTF^{-1} and limited within 5-15 dB for AF region and 15-25 dB for AF region 2; MF region 1 from 2-6 Hz modulation, MF region 2 from 8-20 Hz modulation in the spectra of 300-750, 1250-2500, and 4000-8000 Hz with filter gains derived by consulting MTF^{-1} and limited within 6-10 dB. The results were purely from enhancement algorithms for noisy conditions, however obtained

		Noise-dep.?	Reverb-dep.?	Reverb near; SNR low	Reverb near; SNR mid	Reverb near; SNR high	Reverb mid; SNR low	Reverb mid; SNR mid	Reverb mid; SNR high	Reverb far, SNR low	Reverb far, SNR mid	Reverb far, SNR high
German	*Plain speech score (LSD)*			*11.1 (6.7)*	*40.8 (6.5)*	*64.9 (6.5)*	*15.4 (6.7)*	*44.6 (6.5)*	*70.3 (6.8)*	*12.3 (6.7)*	*41.0 (6.5)*	*76.7 (6.8)*
	ACO	✓	✓	2.5	-0.8	0.0	9.3	10.7	6.4	9.0	22.6	8.2
	ASE	✗	✗	**50.0**	**45.9**	**29.7**	**43.8**	**44.3**	**26.7**	**31.8**	**44.4**	**20.7**
	exactMaxSII	✓	✗	41.5	30.8	14.4	31.5	13.6	2.6	16.6	14.9	11.5
	DeepSSC-Lomb	✗	✗	-5.7	-31.5	-37.2	-10.2	-31.5	-36.6	-8.9	-23.4	-24.8
	DSSC-L/eMSII	✓	✗	25.6	7.4	-10.5	-2.0	-22.6	-27.2	-5.1	-18.5	-33.6
	iMetricGAN	✓	✗	47.0	33.6	21.8	25.7	29.3	19.2	13.6	23.0	8.7
	MS500	✓	✓	13.1	-2.8	-8.9	2.3	-7.5	-3.4	-4.1	-5.7	-4.8
	IISPA	✓	✓	43.6	31.3	20.3	27.5	21.8	6.6	17.5	12.0	-1.1
	SSDRC	✗	✗	47.0	42.1	29.3	36.4	39.3	21.5	20.8	33.9	16.7
English	*Plain speech score (LSD)*			*7.3 (4.1)*	*18.5 (6.6)*	*50.5 (6.8)*	*13.8 (5.7)*	*43.8 (7.1)*	*73.5 (5.4)*	*18.0 (6.2)*	*42.7 (6.9)*	*75.8 (5.4)*
	ACO	✓	✓	-0.5	8.3	17.8	12.8	26.2	14.0	17.5	27.5	15.8
	ASE	✗	✗	6.8	**42.8**	**40.5**	**27.0**	**42.7**	**23.2**	**22.8**	**42.0**	**18.8**
	exactMaxSII	✓	✗	**10.0**	22.8	18.3	10.3	21.8	12.0	4.0	19.0	9.0
	DeepSSC-Lomb	✗	✗	-4.3	-10.0	-14.2	-4.2	-12.7	-8.7	-7.0	-6.0	-12.3
	DSSC-L/eMSII	✓	✗	2.0	7.8	-1.3	-2.0	-4.3	1.8	-6.3	1.3	-2.7
	iMetricGAN	✓	✗	6.7	27.8	27.5	18.5	26.8	14.8	13.5	34.0	13.5
	MS500	✓	✓	2.3	12.0	11.0	9.3	15.3	8.7	7.5	16.2	6.5
	IISPA	✓	✓	3.7	13.7	9.3	1.0	1.8	-2.5	-2.8	1.5	-9.2
	SSDRC	✗	✗	9.7	31.3	36.7	21.3	31.2	19.5	18.2	34.3	17.7
Spanish	*Plain speech score (LSD)*			*14.8 (6.2)*	*43.2 (6.8)*	*66.3 (5.9)*	*12.7 (6.4)*	*25.5 (6.8)*	*52.5 (5.9)*	*6.8 (6.2)*	*27.5 (6.8)*	*55.0 (5.9)*
	ACO	✓	✓	4.3	-2.7	6.3	1.2	13.0	12.8	0.8	11.0	19.8
	ASE	✗	✗	**58.8**	43.7	**29.0**	**46.7**	**56.5**	**41.2**	**30.2**	**45.7**	**38.2**
	exactMaxSII	✓	✗	23.7	17.2	19.2	32.7	34.2	27.5	22.2	17.7	25.2
	DeepSSC-Lomb	✗	✗	-11.2	-36.8	-46.0	-10.8	-16.3	-34.7	-5.7	-21.7	-32.5
	DSSC-L/eMSII	✓	✗	2.8	-0.2	-6.3	11.2	10.7	-2.2	7.5	-4.5	-6.8
	iMetricGAN	✓	✗	42.2	34.0	23.5	28.0	37.0	23.2	15.2	23.0	15.8
	MS500	✓	✓	17.7	4.5	12.3	7.5	17.8	12.3	4.2	9.0	11.8
	IISPA	✓	✓	41.7	35.5	20.0	30.2	37.5	25.0	17.7	18.5	14.5
	SSDRC	✗	✗	49.0	**44.3**	27.8	39.0	49.2	39.8	15.0	32.0	28.0

Figure 14: Differences from reference plain baseline scores (given in italics) in percentage points for all methods under various noisy reverberant conditions [taken from (Rennies et al., 2020, Table 2)]

up to 20 % intelligibility more than plain speech under noise and reverberation. Further investigation with more appropriate enhanced speech like intelligible speech produced by humans under noisy reverberant conditions (the same principle as Lombard speech) in broader modulation regions is expected to find out more suited efficient MS modification based on our proposed concept and methodology. Finally, this study provides essential and necessary information for speech enhancement under adverse conditions and applications to public announcement systems.

6. Acknowledgments

This study was supported by SECOM Science and Technology Foundation, JST-Mirai Program of Japan Science and Technology Agency (Grant Number: JPMJMI18D1), and SCOPE Program of Ministry of Internal Affairs and Communications (Grant Number: 201605002).

References

ANSI ANSI. S3. 5-1997, methods for the calculation of the speech intelligibility index. *New York: American National Standards Institute*, 19:90–119, 1997.

Babble-Noise. *Noisex*. NOISE-ROM-0, NATO: AC243/(Panel 3)/RSG10, 1990.

Felicitas Bederna, Henning Schepker, Christian Rollwage, Simon Doclo, Arne Pusch, Jörg Bitzer, and Jan Rennies. Adaptive compressive onset-enhancement for improved speech intelligibility in noise and reverberation. In *Proceedings of Interspeech*, 2020.

Hans R Bosker and Martin Cooke. Enhanced amplitude modulations contribute to the lombard intelligibility benefit: evidence from the nijmegen corpus of lombard speech. *The Journal of the Acoustical Society of America*, 2020.

Ann R Bradlow and Jennifer A Alexander. Semantic and phonetic enhancements for speech-in-noise recognition by native and non-native listeners. *The Journal of the Acoustical Society of America*, 121(4):2339–2349, 2007.

Carol Chermaz and Simon King. A sound engineering approach to near end listening enhancement. In *Proceedings of Interspeech*, 2020.

PRICE CODE. Sound system equipment–part 16: Objective rating of speech intelligibility by speech transmission index. 2003.

Martin Cooke and Vincent Aubanel. Effects of linear and nonlinear speech rate changes on speech intelligibility in stationary and fluctuating maskers. *The Journal of the Acoustical Society of America*, 141(6):4126–4135, 2017.

Suradej Duangpummet, Jessada Karnjana, Waree Kongprawechnon, and Masashi Unoki. A robust method for blindly estimating speech transmission index using convolutional neural network with temporal amplitude envelope. In *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 1208–1214. IEEE, 2019.

EQ. Equalization (audio). URL `https://en.wikipedia.org/wiki/Equalization_(audio)`.

John HL Hansen, Jaewook Lee, Hussnain Ali, and Juliana N Saba. A speech perturbation strategy based on "lombard effect" for enhanced intelligibility for cochlear implant listeners. *The Journal of the Acoustical Society of America*, 147(3):1418–1428, 2020.

Sarah Hawkins and Paul Warren. Phonetic influences on the intelligibility of conversational speech. *Journal of Phonetics*, 22(4):493–511, 1994.

Hynek Hermansky. Modulation spectrum in speech processing. In *Signal Analysis and Prediction*, pages 395–406. Springer, 1998.

Hynek Hermansky and Nelson Morgan. Rasta processing of speech. *IEEE transactions on speech and audio processing*, 2(4):578–589, 1994.

T Houtgast and H JMi Steeneken. The modulation transfer function in room acoustics as a predictor of speech intelligibility. *Acta Acustica United with Acustica*, 28(1):66–73, 1973.

Tammo Houtgast and Herman JM Steeneken. A review of the MTF concept in room acoustics and its use for estimating speech intelligibility in auditoria. *The Journal of the Acoustical Society of America*, 77(3):1069–1077, 1985.

Alexei Ivanov and Xin Chen. Modulation spectrum analysis for speaker personality trait recognition. In *INTERSPEECH*, 2012.

Søren Jørgensen and Torsten Dau. Predicting speech intelligibility based on the signal-to-noise envelope power ratio after modulation-frequency selective processing. *The Journal of the Acoustical Society of America*, 130(3):1475–1487, 2011.

Søren Jørgensen, Stephan D Ewert, and Torsten Dau. A multi-resolution envelope-power based model for speech intelligibility. *The Journal of the Acoustical Society of America*, 134(1):436–446, 2013.

Noboru Kanedera, Takayuki Arai, Hynek Hermansky, and Misha Pavel. On the importance of various modulation frequencies for speech recognition. In *Fifth European Conference on Speech Communication and Technology*, 1997.

Maria Koutsogiannaki and Yannis Stylianou. Modulation enhancement of temporal envelopes for increasing speech intelligibility in noise. In *Interspeech*, pages 2508–2512, 2016.

Akira Kurematsu, Kazuya Takeda, Yoshinori Sagisaka, Shigeru Katagiri, Hisao Kuwabara, and Kiyohiro Shikano. ATR japanese speech database as a tool of speech recognition and synthesis. *Speech communication*, 9(4):357–363, 1990.

Akiko Kusumoto, Takayuki Arai, Keisuke Kinoshita, Nao Hodoshima, and Nancy Vaughan. Modulation enhancement of speech by a pre-processing algorithm for improving intelligibility in reverberant environments. *Speech Communication*, 45(2):101–113, 2005.

Shih-kuang Lee, Syu-Siang Wang, Yu Tsao, and Jeih-weih Hung. Speech enhancement based on reducing the detail portion of speech spectrograms in modulation domain via discretewavelet transform. In *2018 11th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, pages 16–20. IEEE, 2018.

Haoyu Li, Szu-Wei Fu, Yu Tsao, and Junichi Yamagishi. imetricgan: Intelligibility enhancement for speech-in-noise using generative adversarial network-based metric learning. *arXiv preprint arXiv:2004.00932*, 2020.

Yang Liu, Shota Morita, and Masashi Unoki. MTF-based kalman filtering with linear prediction for power envelope restoration in noisy reverberant environments. *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, 99(2):560–569, 2016.

Etienne Lombard. Le signe de l'elevation de la voix. *Ann. Mal. de L'Oreille et du Larynx*, pages 101–119, 1911.

Ljiljana Milic. *Multirate Filtering for Digital Signal Processing: MATLAB Applications*. IGI Global, 2009.

Ljiljana Milic, Jelena Certi, and Irena Jankovic. Chapter xii: Examples of multirate filter banks - exercises. URL `http://home.etf.rs/~milic/Solution_Manual/Chapter_12_exercises/Chapter_12_exercises.html`.

Shota Morita, Xugang Lu, Masashi Unoki, and Masato Akagi. Method of estimating signal-to-noise ratio based on optimal design for sub-band voice activity detection. *Journal of Information Hiding and Multimedia Signal Processing*, 8(6):1446–1459, 2017.

Laureano Moro-Velázquez, Jorge Andrés Gómez-García, and Juan Ignacio Godino-Llorente. Voice pathology detection using modulation spectrum-optimized metrics. *Frontiers in bioengineering and biotechnology*, 4:1, 2016.

Thuanvan Ngo, Masato Akagi, and Peter Birkholz. Effect of articulatory and acoustic features on the intelligibility of speech in noise: An articulatory synthesis study. *Speech Communication*, 117:13–20, 2020a.

Thuanvan Ngo, Rieko Kubo, and Masato Akagi. Mimicking Lombard Effect: An analysis and reconstruction. *IEICE Transactions on Information and Systems*, E103.D(5):1108–1117, 2020b. doi: 10.1587/transinf.2019EDP7260.

Markus Niermann, Peter Jax, and Peter Vary. Near-end listening enhancement by noise-inverse speech shaping. In *2016 24th European Signal Processing Conference (EUSIPCO)*, pages 2390–2394. IEEE, 2016.

Gaurang Parikh and Philipos C Loizou. The influence of noise on vowel and consonant cues. *The Journal of the Acoustical Society of America*, 118(6):3874–3888, 2005.

Michael A Picheny, Nathaniel I Durlach, and Louis D Braida. Speaking clearly for the hard of hearing ii: Acoustic characteristics of clear and conversational speech. *Journal of Speech, Language, and Hearing Research*, 29(4):434–446, 1986.

Pink-Noise. Various - audio test CD-1 - 91 test signals for home and laboratory use, 1984. URL `https://www.discogs.com/`.

Alexander Raake. Speech quality of voip. *Assessment and Prediction*, 2006.

Jan Rennies, Henning Schepker, Cassia Valentini-Botinhao, and Martin Cooke. Intelligibility-enhancing speech modifications–the hurricane challenge 2.0. *Proc. Interspeech, Shanghai, China*, 2020.

Jan Rennies-Hochmuth, Martin Cooke, and Cassia Valentini-Botinhao. The hurricane challenge, 2020. URL `https://hurricane-challenge.inf.ed.ac.uk/`.

Bastian Sauert and Peter Vary. Near end listening enhancement optimized with respect to speech intelligibility index and audio power limitations. In *2010 18th European Signal Processing Conference*, pages 1919–1923. IEEE, 2010.

Marc René Schädler. Optimization and evaluation of an intelligibility-improving signal processing approach (iispa) for the hurricane challenge 2.0 with fade. In *Proceedings of Interspeech*, 2020.

Cees H Taal and Jesper Jensen. Sii-based speech preprocessing for intelligibility improvement in noise. In *INTERSPEECH*, pages 3582–3586, 2013.

Cees H Taal, Richard C Hendriks, Richard Heusdens, and Jesper Jensen. A short-time objective intelligibility measure for time-frequency weighted noisy speech. In *2010 IEEE international conference on acoustics, speech and signal processing*, pages 4214–4217. IEEE, 2010.

Cees H Taal, Richard C Hendriks, and Richard Heusdens. Speech energy redistribution for intelligibility improvement in noise based on a perceptual distortion measure. *Computer Speech & Language*, 28(4):858–872, 2014.

Yan Tang and Martin Cooke. Learning static spectral weightings for speech intelligibility enhancement in noise. *Computer Speech & Language*, 49:1–16, 2018.

Yan Tang, Martin Cooke, et al. Glimpse-based metrics for predicting speech intelligibility in additive noise conditions. In *INTERSPEECH*, pages 2488–2492, 2016.

Masashi Unoki and Sota Hiramatsu. Mtf-based method of blind estimation of reverberation time in room acoustics. In *2008 16th European Signal Processing Conference*, pages 1–5. IEEE, 2008.

Masashi Unoki and Zhi Zhu. Relationship between contributions of temporal amplitude envelope of speech and modulation transfer function in room acoustics to perception of noise-vocoded speech. *Acoustical Science and Technology*, 41(1):233–244, 2020a. doi: 10.1250/ast.41.233.

Masashi Unoki and Zhi Zhu. Relationship between contributions of temporal amplitude envelope of speech and modulation transfer function in room acoustics to perception of noise-vocoded speech. *Acoustical Science and Technology*, 41(1):233–244, 2020b.

Masashi Unoki, Masakazu Furukawa, Keigo Sakata, and Masato Akagi. An improved method based on the MTF concept for restoring the power envelope from a reverberant signal. *Acoustical science and technology*, 25(4):232–242, 2004.

Masashi Unoki, Yutaka Yamasaki, and Masato Akagi. Mtf-based power envelope restoration in noisy reverberant environments. In *2009 17th European Signal Processing Conference*, pages 228–232. IEEE, 2009.

Masashi Unoki, Akikazu Miyazaki, Shota Morita, and Masato Akagi. Method of blindly estimating speech transmission index in noisy reverberant environments. *Journal of Information Hiding and Multimedia Signal Processing*, 8(6):1430–1445, 2017.

Thuan Van Ngo, Tuan Vu Ho, Masashi Unoki, Rieko Kubo, and Masato Akagi. Enhancement of speech intelligibility under noisy reverberant conditions based on modulation spectrum concept. In *2020 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 753–758. IEEE, 2020.

Syu-Siang Wang, Payton Lin, Yu Tsao, Jeih-Weih Hung, and Borching Su. Suppression by selecting wavelets for feature compression in distributed speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(3):564–579, 2017.

Nils Westerlund, Mattias Dahl, and Ingvar Claesson. Adaptive gain equalizer for speech enhancement, 2002.

Danying Xu, Fei Chen, Fan Pan, and Dingchang Zheng. Factors affecting

the intelligibility of high-intensity-level-based speech. *The Journal of the Acoustical Society of America*, 146(2):EL151–EL157, 2019.

Zhi Zhu, Ryota Miyauchi, Yukiko Araki, and Masashi Unoki. Contributions of temporal cue on the perception of speaker individuality and vocal emotion for noise-vocoded speech. *Acoustical Science and Technology*, 39(3):234–242, 2018.

Tudor-Catalin Zorila, Varvara Kandia, and Yannis Stylianou. Speech-in-noise intelligibility improvement based on spectral shaping and dynamic range compression. In *INTERSPEECH*, pages 635–638, 2012.