# Enhancement of speech intelligibility under noisy reverberant conditions based on modulation spectrum concept

Thuan Van Ngo*, Tuan Vu Ho*, Masashi Unoki*, Rieko Kubo†, and Masato Akagi*
* Japan Advanced Institute of Science and Technology, Japan
E-mail: {vanthuanngo,tuanvu.ho,unoki,akagi}@jaist.ac.jp
† National Institute of Information and Communications Technology, Japan
E-mail: rkubo@nict.go.jp

***Abstract*—This study focuses on identifying effective features for controlling speech to increase speech intelligibility under adverse conditions. Previous methods either reduce noise and reverberation throughout speech presentation or enhance speech before presenting it by controlling its intensity and/or spectral properties to increase intelligibility. Among them, a method based on modulation transfer function theory, in which the environmental effects are inverted to anticipate attenuation of the modulation spectrum of speech, shows excellent potential due to its systematic and explicit derivation of intelligibility enhancement against environmental smears. However, directly obtaining that inversion requires estimating the modulation transfer function. The estimate seems complicated and tolerant under realistic variable conditions. 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 $20\%$, which demonstrates that our concept is valid.**

**Index Terms**: Speech intelligibility, modulation spectrum, modulation transfer function, smeared modulation spectrum.

## I. Introduction

The presence of noise and reverberation in public announcements in train stations and airports often smears the speech spectra, thus making it hard for listeners to understand the announcement. Speech intelligibility could be maintained by reducing noise and reverberation throughout the presentation. However, this is impractical due to the complex architectures of such locations and the cost of installing the necessary devices. A more practical and efficient approach is to enhance the speech before presentation to compensate for degradation in intelligibility due to smearing.

Models such 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. Speech intelligibility has been improved in noisy environments [1], [2], [3], [4] by optimizing the index of the perceptual model used for intelligibility measurement such as the Speech Intelligibility Index (SII) [5], the Speech Transmission Index [6], and the high energy glimpse portion (HEGP) [7]. Further analyses of the speech after index optimization indicated that increasing the spectrum above 1 kHz increases intelligibility [4]. Compressing models as an typical operation of dynamic range compression (DRC) [8] has been used to reduce the speech amplitude on the basis of an input-output energy curve. Different configurations of the curve make the DRC act differently. The DRC emphasizes loudness in the voice onsets and offsets and in the stops and nasals, thereby increasing intelligibility. 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 systematic and explicit derivation to enhance speech intelligibility against environmental smears.

In the MTF concept, which was proposed by Houtgast and Steeneken [9], [10], 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. 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 [9], [11], [12]. The higher the MS index in these modulation frequencies, the better the intelligibility. That is, speech is intelligibly presented if its MS resists smearing of the MTF by the environments. 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 ($MS_{res}^{opt}$) 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

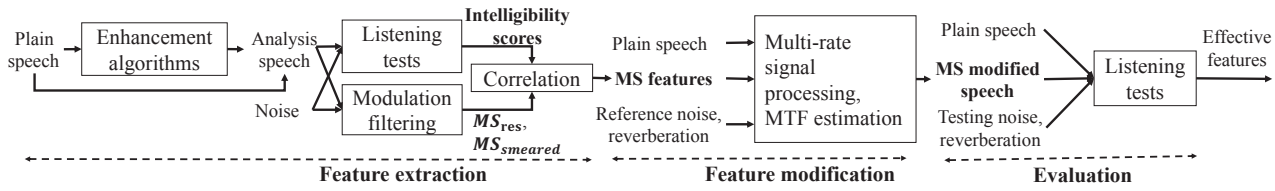



Fig. 1: Proposed concept and implementation of present study

intelligible MS. Several studies tried to estimate $MTF^{-1}$ [12], [13]. However, directly obtaining $MTF^{-1}$ is complicated, especially in realistic environments where backgrounds are diverse and varying because it requires estimating MTF. This estimation is unsuited to realistic environments.

The present study aimed to extract effective features to modify the MS of the original speech by analyzing the relations of $MS_{smeared}$ for intelligibility. The formation of an $MS_{smeared}$ and our proposed concept and implementation are described in the following sections.

## II. Formation of smeared modulation spectrum

This section explains the methods used to estimate $MS$, $MS_{res}$, $MTF$, and $MS_{smeared}$ with provided source signals of clean speech, noise and reverberation for the proposed concept and implementation of this study, especially in feature extraction.

As was used in Zhu *et al.*'s studies [14], [15], we used a modulation filtering technique to estimate the $MS$/$MS_{res}$ of speech. The filtering was done 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 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. [10] also used 0.5-20 Hz band-pass filters to estimate the speech power envelope spectrum. Our estimated $MS$/$MS_{res}$ thus contained 0 Hz modulation and showed as frequency features. The time features were above 0 Hz up until a modulation frequency of 20 Hz.

The MTF is fully determined mathematically for stationary noise by the signal-to-noise ratio [10], [16]. 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 study, 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

$$m(f_a, f_m) = m_N(f_a, f_m) \times m_R(f_a, f_m). \quad (3)$$

From (1) and (3), $MS_{smeared}$ can be calculated using

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

In general, the $MS_{smeared}$ 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 reverberation on the time features.

## III. Proposed concept and implementation

As shown in Fig. 1 our concept is based on three steps: feature extraction, feature modification, and evaluation.

### A. Extraction of modulation spectral features

As shown in the feature extraction portion of Fig. 1, **aiming to capture different information in $MS_{res}$/$MS_{smeared}$**, we collected variable speech enhancement methods and applied them to increase intelligibility, then investigated the properties of the enhanced speech. Each method mainly modified different acoustic and modulation frequency regions. The resulting intelligibility scores differed. Then, we identified significant acoustic and modulation frequency regions to modify the MS more by using correlation, as described below.

First, we synthesized enhanced speech from plain speech using the enhancement methods of Ngo *et al.* [17] (increasing spectral regions from 2-6 kHz by 13 dB as C2 speech), Tang *et al.* [4] (increasing spectral regions above 1 kHz by 45 dB, decreasing regions below 1 kHz by 45 dB as HEGP speech), and Zorila *et al.* [8] (increasing spectra from 1-4 kHz by 12 dB with pre-emphasis, formant sharpening, and DRC, cumulatively defined as SS, SSFS, SSDRC speech). The plain speech was three-mora Japanese words (about 350-450 ms for each), taken from the male speech of A-set in ATR dataset [18]. Then, we conducted listening tests in a noise environment that followed the same designed as in Tang *et al.*'s study [4] with seven-teen native Japanese people to obtain the intelligibility scores of the plain and enhanced speech (as shown in Fig. 2a). As can be seen that, different enhanced methods yielded variable intelligibility improvements in different noise. Perhaps, 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. Next, we estimated $MS$/$MS_{res}$, $MTF$ to

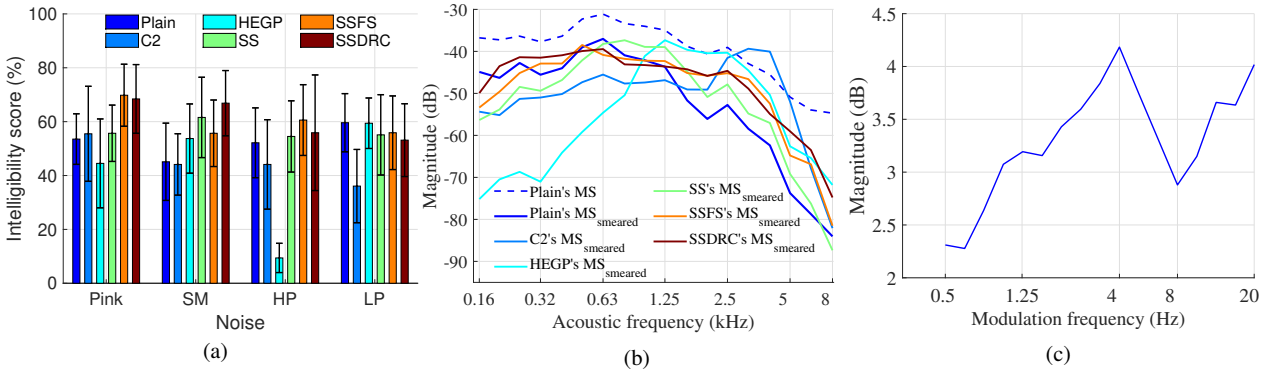



Fig. 2: (a) Intelligibility scores of analyzed speech for pink noise at $-9.5$ dB SNR, speech-modulated (SM) noise at $-10.5$ dB SNR, high-pass (HP, the high-pass filtered pink noise with cutoff frequency 500 Hz) noise at $-12$ dB SNR, and low-pass (LP, the low-pass filtered pink noise with cutoff frequency 4 kHz) noise at $-13$ dB SNR, (b) $MS_{smeared}$ at 0 Hz modulation of analyzed speech for pink noise, and (c) $MS_{res}$ over 0-20 Hz modulation using DRC (difference between $MS_{res}$ of SSDRC and SSFS) for SM noise at the acoustic frequency of 5 kHz.

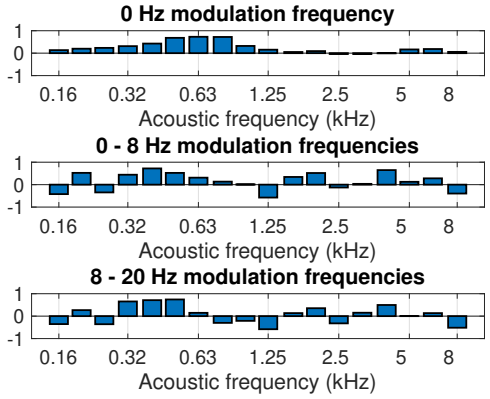



Fig. 3: Pearson correlation between $MS_{smeared}$ and $MS_{res}$ for each acoustic frequency band (0, 0-8, 8-20 Hz modulation bands) and intelligibility scores for analyzed speech in all tested noise.

calculate $MS_{smeared}$ from the speech and noise in the tests by the methods described in Sec. II. We decomposed the enhancements on the basis of the relationships between $MS_{res}$ and intelligibility and between $MS_{smeared}$ and intelligibility to unfold their essential characteristics in "MS features."

(i) The $MS_{smeared}$ at 0 Hz modulation showed the effect of noise on the frequency features. As shown in Fig. 2b, $MS_{smeared}$ presented a lot of information about what was increased in the $MS_{smeared}$ of enhanced speech from the $MS_{smeared}$ of plain speech and reached to the MS of plain speech to contribute to intelligibility Fig. 2a. In general, the better intelligibility for the enhanced speech of SS, SSFS and SSDRC (Fig. 2a) might be because their $MS_{smeared}$ indexes of the frequency regions around 500 Hz, around 2 kHz, and around 5-6 kHz were increased.

(ii) 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 typically shown in Fig. 2c, it had two peaks, one at around 4 Hz and one at around 20 Hz.

(iii) The Pearson correlations between $MS_{smeared}$ and $MS_{res}$ for each acoustic frequency band over three modulation frequency bands and the intelligibility 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 $MSres$ by DRC. Also, due to the two peaks, we used each acoustic frequency band of $MS_{res}$ for the modulation frequency bands of 0-8 and 8-20 Hz of DRC in the calculation. The correlation values could be interpreted like that the more positive correlation coefficient, the more correlation between increasing $MS_{smeared}$/$MS_{res}$ indexes and increasing intelligibility and vice versa. From the positively correlated regions (Fig. 3) and the typical $MS_{smeared}$ and $MS_{res}$ properties of the analyzed speech [as described in (i) and (ii)], we tentatively used two MS features. The frequency features could be increased in MS at 500-2250 Hz (acoustic frequency (AF) region 1) and 4.5-6.5 kHz (AF region 2). The time features could be increased in MS at 2-6 Hz (modulation frequency (MF) region 1) and 8-20 Hz (MF region 2) in the acoustic spectra of 300-750, 1250-2250, and 4500-6500 Hz. In the time features, the region between 750 and 1250 Hz got a negative correlation while they were positive in the frequency features. We thought that this was 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.

### *B. Modification of modulation spectrum*

As shown in Fig. 1, to synthesize MS-modified speech, an analysis-synthesis method was developed to modify the MS of

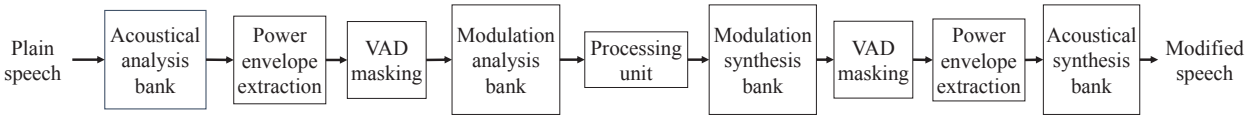
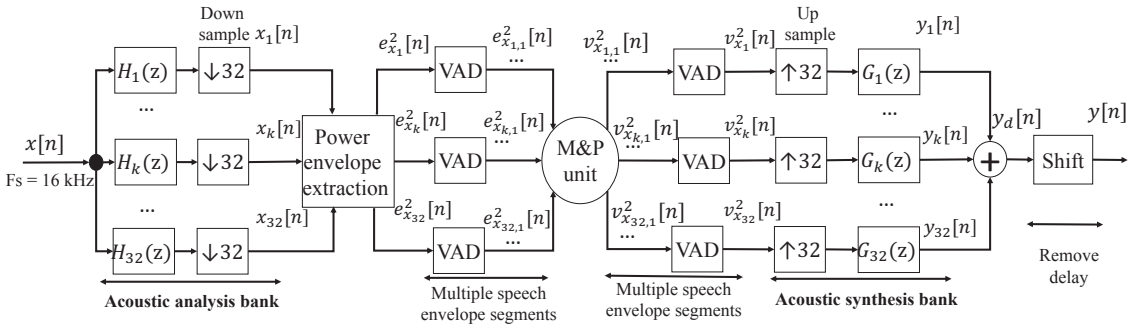
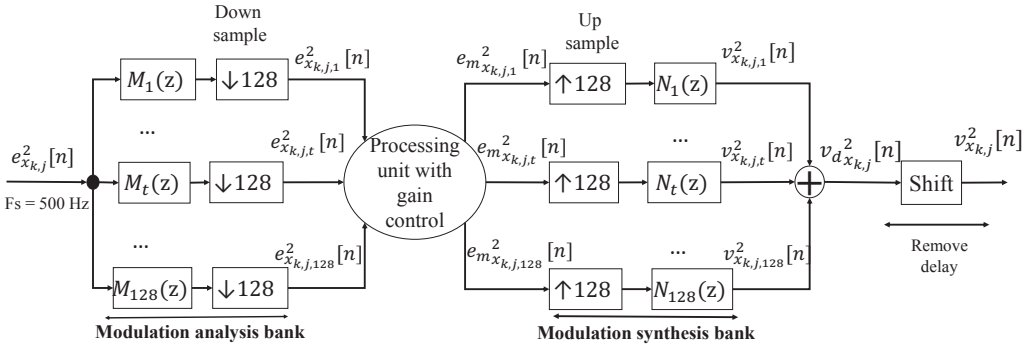



Fig. 4: Block diagram of process for converting plain speech into MS-modified speech using multi-rate signal processing technique.




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




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

Fig. 5: Banks of acoustical and modulation analyses and syntheses used in conversion of plain speech into MS-modified speech.

plain speech. 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. II, which was mainly used for the estimation of $MS/MS_{res}$, $MTF$, and $MS_{smeared}$ in Sec. III-A. In this modification, we only used the modulation filtering technique for the estimation of MTF. The details are described as follows.

To imitate the MS analysis, enabling reconstruction, we used a multi-rate signal processing technique [19] for the MS analysis, modification, and synthesis steps (Fig. 4). 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 described in Sec. III-A). It sequentially applied gains to the acoustic frequency regions (all 0-20 Hz modulation regions) and the modulation frequency regions (0-20 Hz, excluding 0 Hz).

The amplified values were consulted from the averaged $MTF^{-1}$ calculated from the estimated MTF of the provided reference noise and RIR by using Eq. (3) within these acoustic and modulation frequency regions of the MS features. As was indicated in the introduction section, our method aimed to tackle the variable environments. In realistic situation, the estimation of MTF was always tolerant in changing environments, which led the estimation of $MTF^{-1}$ also to be tolerant. We thus corrected these averaged estimated $MTF^{-1}$ within limited ranges to preserve the voice quality of the plain speech. AF regions 1 and 2 were limited to 5-15 dB and 15-

TABLE I: 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 |

25 dB, respectively (an incremental response as the analyzed $MS_{smeared}$ of SSDRC). MF regions 1 and 2 were both limited to 6-10 dB.

Finally, to obtain the modified speech, the reconstruction was processed in reverse order from the analysis with the modulation synthesis bank, VAD, and acoustical synthesis bank. The processes are illustrated in Figs 5a and 5b.

### C. Evaluation

As shown in Fig. 1, we evaluated the effects of the extracted MS features on intelligibility under various conditions for three languages. The speech material was provided by Hurricane Challenge 2.0 (HC 2.0) [20] in German and Spanish (100 sentences each) and English (90 sentences) as recorded by native male speakers and was used as plain speech. Two speech types were used: the plain speech and speech modified using the MS features described in Sec. III-A modified using the technique described in Sec. III-B. The latter type is referred to as "MS500 speech." The evaluation was performed by HC 2.0 with about 180 listeners. They created stimuli for the experiment using the MS500 speech and their 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. Table I shows the evaluated SNRs and reverberation conditions in terms of the distance between the loudspeaker and the listener.

## IV. Results and discussion

The results of the listening tests are shown in Figure 6. They indicate that the MS500 speech had better intelligibility than the plain speech under all noisy reverberant conditions for English and Spanish (5-20%). The higher first and third quartiles for MS500 further support the substantial improvement obtained. However, intelligibility was not improved in other cases for the German dataset. So far, one possible reason could be that German words often contain plosive consonants, and modifying their MS is a delicate operation. It might be also due to the criteria to proceed VAD during the MS modification could not be well suited to all Germain speech. Moreover, all the noise used for evaluating enhanced speech were stationary, perhaps more types of noise should be added to investigate $MS_{res}$ and $MS_{smeared}$ better. Furthermore, in recent study, because the experiments were carried out by HC 2.0, they might have their own comparison results of our methods with others. In future, we will aim to improve all the remaining aspects in our techniques more and compare with other methods to justify our methods more.

## V. Conclusion

This study presented a method for improving the intelligibility of speech under noisy reverberant conditions. Smeared modulation spectra of the speech signals by the environments were derived on the basis of the modulation spectrum and modulation transfer function concepts. Concerning listening tests, correlated acoustic, and modulation frequencies in the MS with intelligibility under noise were extracted from differently enhanced speech and considered as modulation spectral features. We then modified plain speech by controlling these features. Listening tests under various conditions demonstrated that the extracted features effectively increased the intelligibility of the modified speech. In other words, our proposed concept, obtaining significant characteristics from different enhancement methods by relations of smeared modulation spectra for intelligibility, is valid. Building on these findings for stationary noise, we can investigate additional features contributing to intelligibility for all kinds of noise. In addition to investigating intelligibility, our method can be used to investigate other speech qualities.

## VI. Acknowledgements

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

[1] B. Sauert and P. Vary, "Near end listening enhancement optimized with respect to speech intelligibility index and audio power limitations," in *EUSIPCO*, 2010, pp. 1919–1923.

[2] C. H. Taal and J. Jensen, "SII-based speech preprocessing for intelligibility improvement in noise." in *INTERSPEECH*, 2013, pp. 3582–3586.

[3] C. H. Taal, R. C. Hendriks, and R. Heusdens, "Speech energy redistribution for intelligibility improvement in noise based on a perceptual distortion measure," *Computer Speech & Language*, vol. 28, no. 4, pp. 858–872, 2014.

[4] Y. Tang and M. Cooke, "Learning static spectral weightings for speech intelligibility enhancement in noise," *Computer Speech & Language*, vol. 49, pp. 1–16, 2018.

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

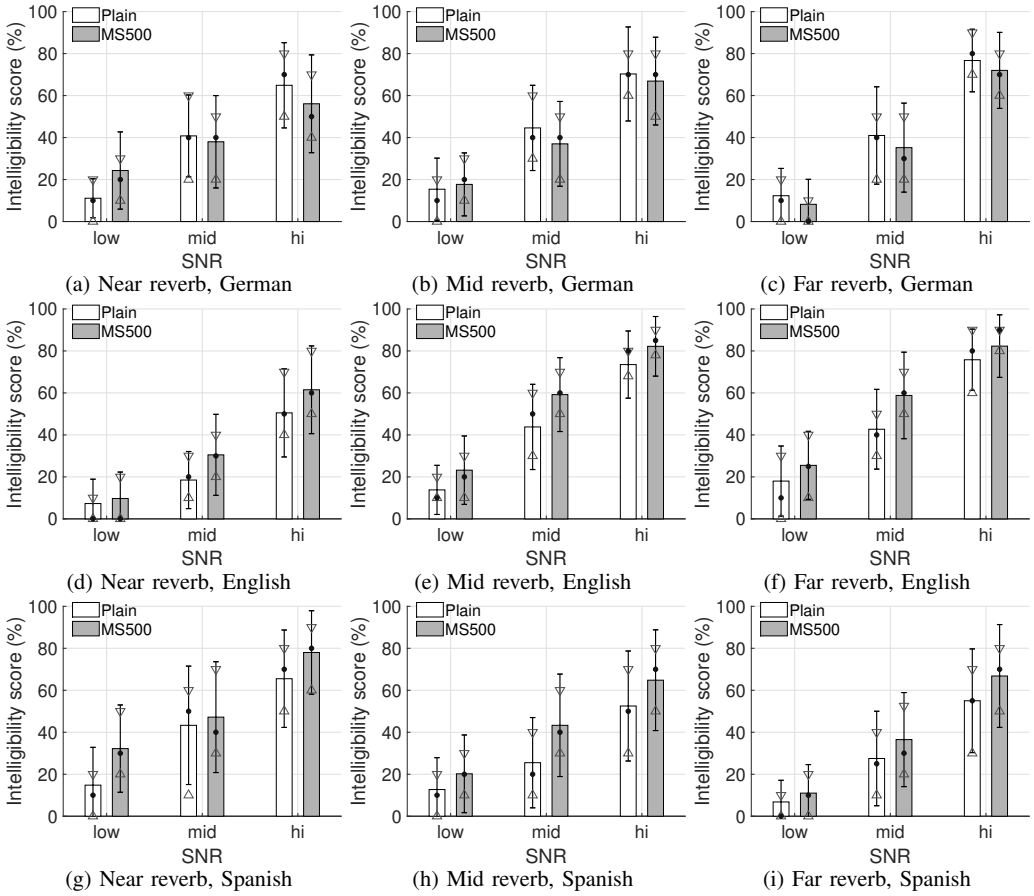
[6] P. CODE, "Sound system equipment–part 16: Objective rating of speech intelligibility by speech transmission index," 2003.

[7] Y. Tang, M. Cooke *et al.*, "Glimpse-based metrics for predicting speech intelligibility in additive noise conditions." in *INTERSPEECH*, 2016, pp. 2488–2492.

[8] T.-C. Zorila, V. Kandia, and Y. Stylianou, "Speech-in-noise intelligibility improvement based on spectral shaping and dynamic range compression," in *INTERSPEECH*, 2012, pp. 635–638.

[9] T. Houtgast and H. J. Steeneken, "The modulation transfer function in room acoustics as a predictor of speech intelligibility," *Acta Acustica United with Acustica*, vol. 28, no. 1, pp. 66–73, 1973.

[10] T. Houtgast and H. J. 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*, vol. 77, no. 3, pp. 1069–1077, 1985.




(a) Near reverb, German (b) Mid reverb, German (c) Far reverb, German

(d) Near reverb, English (e) Mid reverb, English (f) Far reverb, English

(g) Near reverb, Spanish (h) Mid reverb, Spanish (i) Far reverb, Spanish

Fig. 6: Intelligibility scores (percentage of correctly identified words) of plain and MS-modified speech under 3 SNR noise levels (low, mid, hi) $\times$ 3 reverberation (reverb) conditions (near, mid, far) $\times$ 3 languages (German, English, Spanish). Bars indicate mean and standard deviation. Triangles, inverse triangles, and circles indicate first quartile, third quartile, and median respectively.

[11] H. Hermansky, "Modulation spectrum in speech processing," in *Signal Analysis and Prediction*. Springer, 1998, pp. 395–406.

[12] A. Kusumoto, T. Arai, K. Kinoshita, N. Hodoshima, and N. Vaughan, "Modulation enhancement of speech by a pre-processing algorithm for improving intelligibility in reverberant environments," *Speech Communication*, vol. 45, no. 2, pp. 101–113, 2005.

[13] M. Koutsogiannaki and Y. Stylianou, "Modulation enhancement of temporal envelopes for increasing speech intelligibility in noise," in *Interspeech*, 2016, pp. 2508–2512.

[14] Z. Zhu, R. Miyauchi, Y. Araki, and M. Unoki, "Contributions of temporal cue on the perception of speaker individuality and vocal emotion for noise-vocoded speech," *Acoustical Science and Technology*, vol. 39, no. 3, pp. 234–242, 2018.

[15] M. Unoki and Z. Zhu, "Relationship between contributions of temporal amplitude envelope of speech and modulation transfer function in room acoustics to perception of noise-vocoded speech," *Acoustical Science and Technology*, vol. 41, no. 1, pp. 233–244, 2020.

[16] M. Unoki, Y. Yamasaki, and M. Akagi, "MTF-based power envelope restoration in noisy reverberant environments," in *EUSIPCO*, 2009, pp. 228–232.

[17] T. V. NGO, R. KUBO, and M. AKAGI, "Mimicking Lombard Effect: An analysis and reconstruction," *IEICE Transactions on Information and Systems*, vol. E103.D, no. 5, pp. 1108–1117, 2020.

[18] A. Kurematsu, K. Takeda, Y. Sagisaka, S. Katagiri, H. Kuwabara, and K. Shikano, "Atr japanese speech database as a tool of speech recognition and synthesis," *Speech communication*, vol. 9, no. 4, pp. 357–363, 1990.

[19] L. Milic, *Multirate Filtering for Digital Signal Processing: MATLAB Applications: MATLAB Applications*. IGI Global, 2009.

[20] J. Rennies-Hochmuth, S. Henning, M. Cooke, and C. Valentini-Botinhao, "The hurricane challenge." [Online]. Available: https://hurricane-challenge.inf.ed.ac.uk/