A NEW MASK-BASED OBJECTIVE MEASURE FOR PREDICTING THE INTELLIGIBILITY OF BINARY MASKED SPEECH

Chengzhu Yu, Kamil K. Wójcicki, P. C. Loizou, John H. L. Hansen

Dept. of Electrical Engineering
Erik Jonsson School of Enigneering and Computer Science
University of Texas at Dallas, Richardson, TX 75080
chengzhu.yu@utd.edu

ABSTRACT

Mask-based objective speech-intelligibility measures have been successfully proposed for evaluating the performance of binary masking algorithms. These objective measures were computed directly by comparing the estimated binary mask against the ground truth ideal binary mask (IdBM). Most of these objective measures, however, assign equal weight to all time-frequency (T-F) units. In this study, we propose to improve the existing mask-based objective measures by weighting each T-F unit according to its target or masker loudness. The proposed objective measure shows significantly better performance than two other existing mask-based objective measures.

Index Terms— Speech intelligibility, objective measure, ideal binary mask, speech separation

1. INTRODUCTION

Understanding speech in the presence of background noise is one of the most challenging tasks for listeners with hearing loss. Recent studies have shown that speech separation techniques based on the concept of ideal binary mask has the potential for restoring intelligibility of speech corrupted by competing noise both for normal hearing and hearing impaired persons [1, 2, 3, 4].

Binary masking is a strategy for applying binary gains on a T-F representation. Ideal binary mask (IdBM) was defined by comparing the local signal-to-noise ratio (SNR) of each T-F unit against a fixed threshold [1]. T-F units with local SNR higher than the threshold are defined as target-dominated T-F units and are kept, while others as masker-dominated T-F units and are discarded.

Motivated by the above studies, increasing effort has been put on designing algorithms that could accurately predict the IdBM [5, 6, 7, 8]. Objective metrics for evaluating the performance of these binary masking algorithms are of great interest, since subjective tests can be time-consuming [5, 9, 10].

Mask-based objective speech-intelligibility measures such as hit rate minus false alarm rate (HIT-FA) [5] and ideal binary mask ratio (IBMR) [9] were proposed and frequently used as metrics to evaluate the performance of binary masking techniques. Those mask-based objective intelligibility measures are often obtained by counting the mismatched T-F units between estimated binary mask and IdBM. Since the calculation of mask-based objective measures does not require the resynthesized output, they are robust to many convolutional distortions that are not generated by the binary masking algorithm itself [9]. While those measures have been shown to have modestly high correlation with subjective scores, in its calculation the contribution of individual T-F units are equally weighted. However, it turns out that mask errors localized in louder T-F units are more harmful to speech intelligibility than those in quieter T-F units (see Fig. 2).

In this study, we propose a new mask-based objective intelligibility measure, loudness weighted hit-false (LWHF), in which each T-F unit is weighted according to the loudness of its target or masker content.

2. LOUDNESS WEIGHTED HIT-FALSE MEASURE

In order to associate the appropriate weight to each T-F unit, we categorize T-F units into two classes: target-present T-F units and target-absent T-F units as in [2]. A previous study [2] has demonstrated that target-present T-F units have differential contributions to speech intelligibility compared to target-absent T-F units. In particular, target-present T-F units have a positive contribution to speech intelligibility, while target-absent T-F units incline to distort speech intelligibility. It could be further expected that the positive contribution of target-present T-F units comes from the underlying target component, while the negative contribution of target-absent T-F units is caused by its masker component. In addition, we will define mask errors occurring in target-present T-F units as miss errors and those occurring in target-absent T-F units as false alarm error. Finally, in our proposed method, miss errors are weighted according to the loudness degree of its target component, while false alarm errors are weighted ac-

cording to the loudness degree of its masker content.

2.1. Loudness Spectrogram Computation

Let $\mathrm{Y(n)} = \mathrm{X(n)} + \mathrm{d(n)}$ be the mixture signal, with $\mathrm{X(n)}$ denoting the target signal and $\mathrm{d(n)}$ denoting the masker signal. Signals ($\mathrm{Y(n)}$, $\mathrm{X(n)}$ and $\mathrm{d(n)}$) are first segmented in time using Hamming window (20ms) with 50% overlap between segments. A fast Fourier transform (FFT) is then applied to each segment. T-F analyzed signals ($\mathrm{Y(t,f)}$,$\mathrm{X(t,f)}$ and $\mathrm{d(t,f)}$) are pre-emphasized by an equal-loudness curve, simulating the perceptual sensitivity of the human ear to the intensity of sound at different frequency locations [11].

$$\overline{\mathrm{Y(t,f)}} = \mathrm{Y(t,f)E(f)} \quad (1)$$

$$\overline{\mathrm{X(t,f)}} = \mathrm{X(t,f)E(f)} \quad (2)$$

$$\overline{\mathrm{d(t,f)}} = \mathrm{d(t,f)E(f)} \quad (3)$$

$\mathrm{E(f)}$ is an approximation of equal loudness contour (valid up to 5000 Hz) and is given by

$$\mathrm{E(f)} = \frac{[(f^2 + 56.8 \times 10^6)f^4]}{[(f^2 + 6.3 \times 10^6)^2 \times (f^2 + 0.38 \times 10^9)}. \quad (4)$$

After multiplying by the equal-loudness contour, the loudness spectrogram is calculated by applying a cubic root amplitude compression [11].

$$\mathrm{L_Y(t,f)} = (\overline{\mathrm{Y(t,f)}})^{0.33}, \quad (5)$$

$$\mathrm{L_X(t,f)} = (\overline{\mathrm{X(t,f)}})^{0.33}, \quad (6)$$

$$\mathrm{L_d(t,f)} = (\overline{\mathrm{d(t,f)}})^{0.33} \quad (7)$$

where $\mathrm{L_Y(t,f)}$,$\mathrm{L_X(t,f)}$, and $\mathrm{L_d(t,f)}$ indicate the loudness spectrogram of the mixture, target and masker signals, respectively.

2.2. Loudness weighted miss error

The loudness weighted miss error ($\mathrm{T_1}$) of the binary masked speech is defined as follows:

$$\mathrm{T_1} = \sum \mu(\mathrm{t,f}) \times \mathrm{MISS(t,f)}, \quad (8)$$

where $\mathrm{MISS(t,f)}$ is the binary indication of miss error of each T-F unit, and $\mu(\mathrm{t,f})$ is weight value associated with each miss error. Since miss errors occur only in target-present T-F units, $\mu(\mathrm{t,f})$ is related to the loudness of the local target component. Thus, we define $\mu(\mathrm{t,f})$ as follows:

$$\mu(\mathrm{t,f}) = \mathrm{g(L_X(t,f))}, \quad (9)$$

where $\mathrm{g(.)}$ is a sigmoid function for mapping each target-present T-F unit to the perceptual weight according to its target loudness,

$$\mathrm{g(x)} = \frac{1}{1 + \exp(\alpha_1(\mathrm{x} - \beta_1))}. \quad (10)$$

Values of $\alpha_1 = -10$ and $\beta_1 = 0.7$ yielded the best correlation for our test material.

2.3. Loudness weighted false alarm error

The loudness weighted false alarm error ($\mathrm{T_2}$) of the binary masked speech is defined as follows:

$$\mathrm{T_2} = \sum \nu(\mathrm{t,f}) \times \mathrm{FA(t,f)}, \quad (11)$$

where $\mathrm{FA(t,f)}$ is the binary indication of false alarm error of each T-F unit, and $\nu(\mathrm{t,f})$ is the weight value associated with each false alarm error . Since false alarm errors occur only in target-absent T-F units, $\nu(\mathrm{t,f})$ is related to the loudness of local masker component. Thus, we define $\nu(\mathrm{t,f})$ as follows:

$$\nu(\mathrm{t,f}) = \mathrm{h(L_d(t,f))}, \quad (12)$$

where $\mathrm{h(.)}$ is a sigmoid function used for mapping each speech-absent T-F units to perceptual weight according to its masker loudness,

$$\mathrm{h(x)} = \frac{1}{1 + \exp(\alpha_2(\mathrm{x} - \beta_2))}. \quad (13)$$

Values of $\alpha_2 = -10$ and $\beta_2 = 0.8$ yielded best correlation for our test material.

2.4. Proposed objective intelligibility measure

In order to incorporate the perceptual effect of miss errors and false alarm errors on speech intelligibility simultaneously, we propose a new objective intelligibility measure for binary masked speech called Loudness weighted Hit-False (LWHF) as an improvement over the previous HF measure. LWHF is defined as follows:

$$\mathrm{LWHF} = \frac{\mathrm{T} - \mathrm{T_1} - \mathrm{T_2}}{\mathrm{T}}, \quad (14)$$

where T indicates the loudness weighted sum of target-present T-F units, and

$$\mathrm{T} = \sum \mu(\mathrm{t,f}) \times \mathrm{L_X(t,f)}. \quad (15)$$

3. EVALUATION AND COMPARISON

3.1. Subjective data

Speech sentences were taken from the IEEE database (1969). A 20-talker simulated babble noise was used as the masker to corrupt the sentences at -5 dB SNR. For each sentence, we separate T-F units into two classes, target-present and target-absent T-F units, and compute IdBM as in [2].

Target-present T-F units are further categorized into four groups L_1, L_2, L_3 and L_4, by increasing target loudness. Each group includes a fixed percentage of the target-present T-F units. For example, L_1 consists of the target-present T-F units having target loudness in the lowest level, while L_4 consists of the target-present T-F units having target loudness in

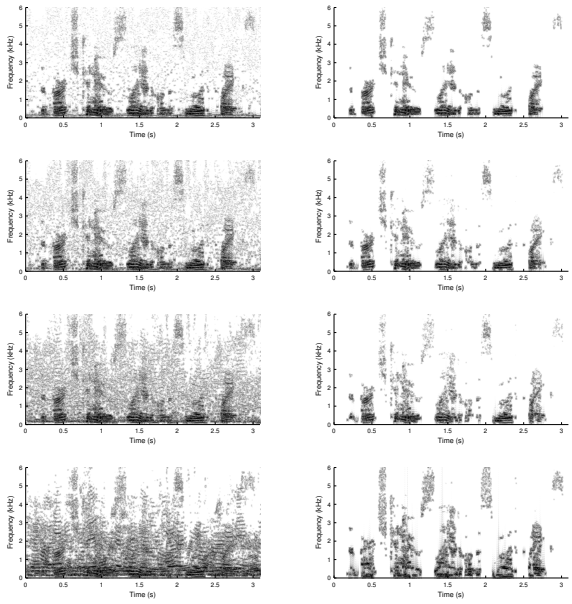

Fig. 1. The left column shows spectrograms of sentences synthesized from four new binary masks derived by masking target-absent T-F units belonging to T_1, T_2, T_3 and T_4, respectively from top to bottom. The right column shows the spectrograms of the same sentences synthesized from new binary masks derived by masking target-present T-F units belonging to L_1, L_2, L_3 and L_4, respectively from top to bottom.

the highest level. Similarly, we also categorized target-absent T-F units into another four groups T_1, T_2, T_3 and T_4, but according to masker loudness.

Next we introduces mask errors to all T-F units of a given loudness group, while no errors were introduced to other T-F units. We repeated this process separately for each of above eight groups, L_1, L_2, L_3, L_4, T_1, T_2, T_3 and T_4, to create eight new binary masks. Since each of the four target-present groups contains the same number of T-F units, the new derived binary masks have the same miss error rate, but concentrated asymmetrically in T-F units belongs to different loudness levels. Similarly, the four new binary masks derived from four target-absent groups have the same false alarm rate, but concentrated asymmetrically in T-F units belongs to different loudness levels. The eight new derived binary masks were applied to mixture signals to produce stimuli for our test. Fig. 1 shows an example of the stimuli spectrograms.

For the listening experiments, eight normal-hearing listeners participated in the experiments. The participants were all native speakers of American English. Subjects participated in a total of 8 conditions (=4 target-present groups+ 4 target-absent groups). Each condition used two lists of non-repeated sentences (i.e., 20 sentences). The order of the test conditions was produced randomly for each subjects.

3.2. Results and discussion

In order to evaluate the proposed method, we compare it against two other existing mask-based objective measures, HIT-FA and IBMR, based on the stimuli produced previously. Results are shown in Fig. 2.

The first column of Fig. 2 indicates the average subjective listening scores on stimuli from the 8 different conditions outlined in Sec.3.1. Minor degradation in the intelligibility is observed when mask errors were introduced to target-present T-F units having lower target loudness (L_1, L_2, L_3), while performance drops significantly when the same amount of errors were introduced to target-present T-F units having highest target loudness (L_4). A similar tendency is observed when mask errors were introduced to target-absent T-F units. A gradual drop in performance is observed as the location of mask errors shifts from T1 to T3. Dramatic degradation in performance occurred when mask errors were introduced to T-F units belonging to T4. This demonstrates the fact that the importance of each T-F units varies in accordance with loudness of its signal content.

It is clear from the Fig. 2 that existing mask-based objective measures, HIT-FA and IBMR, could not provide consistent prediction on the stimuli created from binary masks having asymmetric mask errors. This is due to the fact that HIT-FA and IBMR assume that each T-F unit has an equal contribution to speech intelligibility. On the other hand, prediction from the proposed mask-based objective measure (LWHF) is in general consistent with subjective listening scores.

4. CONCLUSION

While existing mask-based objective measures, such as HIT-FA and IBMR have shown modestly good correlation with subjective intelligibility scores in some conditions, the consistency is not always the case when mask errors were distributed asymmetrically in T-F units of different loudness levels. This is due to the fact that HIT-FA and IBMR were based on the simple assumption that each T-F unit has the same contribution to speech intelligibility. This study has proposed a new mask-base objective measure in which each T-F unit is weighted according to the loudness of its speech content. The proposed metric shows significantly better performance than two other well established previous mask-based objective measures.

Acknowledgments

This research was supported by Grant No. R01 DC010494 from the National Institute of Deafness and other Communication Disorders (NIDCD), National Institutes of Health (NIH).

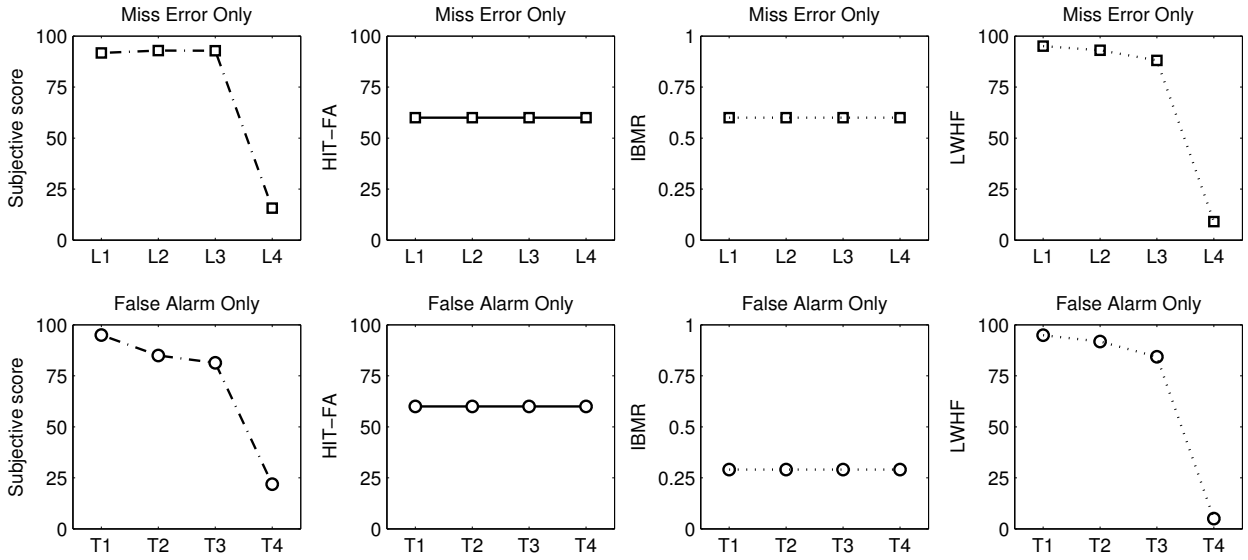

Fig. 2. Comparison of proposed method (LWHF) with two existing mask-based objective speech intelligibility measures, namely Hit-False (HIT-FA) and ideal binary mask ratio (IBMR), on speech stimuli produced from binary masks having asymmetric mask errors. Subjective performance is used as reference for comparison.

5. REFERENCES

[1] D. Wang, "On ideal binary mask as the computational goal of auditory scene analysis," in *Speech Separation by Humans and Machines*, P. Divenyi, Ed., pp. 181–187. Kluwer Academic, Dordrecht, Netherlands, 2005.

[2] M.C. Anzalone, L. Calandruccio, K.A. Doherty, and L.H. Carney, "Determination of the potential benefit of time-frequency gain manipulation.," *Ear and Hear.*, vol. 27, no. 5, pp. 480–492, Oct 2006.

[3] Ning Li and Philipos C. Loizou, "Factors influencing intelligibility of ideal binary-masked speech: Implications for noise reduction," *J. Acoust. Soc. Amer.*, vol. 123, no. 3, pp. 1673–1682, 2008.

[4] DeLiang Wang, Ulrik Kjems, Michael S. Pedersen, Jesper B. Boldt, and Thomas Lunner, "Speech perception of noise with binary gains," *J. Acoust. Soc. Amer.*, vol. 124, pp. 2303–2307, 2008.

[5] Gibak Kim, Yang Lu, Yi Hu, and Philipos C. Loizou, "An algorithm that improves speech intelligibility in noise for normal-hearing listeners," *J. Acoust. Soc. Amer.*, vol. 126, no. 3, pp. 1486–1494, 2009.

[6] Kun Han and Deliang Wang, "An SVM based classification approach to speech separation," in *ICASSP-11*, 2011, pp. 4632–4635.

[7] Yi Hu and Philipos C. Loizou, "Environment-specific noise suppression for improved speech intelligibility by cochlear implant users," *J. Acoust. Soc. Amer.*, vol. 127(6), pp. 3689–3695, 2010.

[8] Yuxuan Wang, Kun Han, and DeLiang Wang, "Acoustic features for classification based speech separation," in *Proc. ISCA Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*, Portland, Oregon, 2012.

[9] Christopher Hummersone, Russell Mason, and Tim Brookes, "Ideal binary mask ratio: a novel metric for assessing binary-mask-based sound source separation algorithms," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 7, pp. 2039–2045, 2011.

[10] Cees H. Taal, Richard C. Hendriks, Richard Heusdens, and Jesper Jensen, "An algorithm for intelligibility prediction of time-frequency weighted noisy speech," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 7, pp. 2125–2136, Sept 2011.

[11] H. Hermansky, "Perceptual linear predictive (plp) analysis of speech," *J. Acoust. Soc. Amer.*, vol. 87, no. 4, pp. 1738–1752, 1990.