Outcome measures based on classification performance fail to predict the intelligibility of binary-masked speech

Kressner, Abigail Anne; May, Tobias; Rozell, Christopher J.

Published in:
Journal of the Acoustical Society of America

Link to article, DOI:
10.1121/1.4952439

Publication date:
2016

Document Version
Publisher's PDF, also known as Version of record

Outcome measures based on classification performance fail to predict the intelligibility of binary-masked speech (L)

Abigail Anne Kressner[a] and Tobias May
Hearing Systems Group, Department of Electrical Engineering, Technical University of Denmark, DK-2800 Kongens Lyngby, Denmark

Christopher J. Rozell
School of Electrical and Computer Engineering, 777 Atlantic Drive NW, Georgia Institute of Technology, Atlanta, Georgia 30332, USA

(Received 25 January 2016; revised 7 May 2016; accepted 11 May 2016; published online 1 June 2016)

To date, the most commonly used outcome measure for assessing ideal binary mask estimation algorithms is based on the difference between the hit rate and the false alarm rate (H-FA). Recently, the error distribution has been shown to substantially affect intelligibility. However, H-FA treats each mask unit independently and does not take into account how errors are distributed. Alternatively, algorithms can be evaluated with the short-time objective intelligibility (STOI) metric using the reconstructed speech. This study investigates the ability of H-FA and STOI to predict intelligibility for binary-masked speech using masks with different error distributions. The results demonstrate the inability of H-FA to predict the behavioral intelligibility and also illustrate the limitations of STOI. Since every estimation algorithm will make errors that are distributed in different ways, performance evaluations should not be made solely on the basis of these metrics. [http://dx.doi.org/10.1121/1.4952439]

[ZHM] Pages: 3033–3036

I. INTRODUCTION

The ideal binary mask (IBM) algorithm improves speech intelligibility outcomes in the frameworks of both noise reduction and cochlear implant channel selection (e.g., Roman *et al.*, 2003; Wang, 2005; Anzalone *et al.*, 2006; Brungart *et al.*, 2006; Hu and Loizou, 2008). The general approach is to generate a matrix of binary gain values in the time-frequency (T-F) domain based on the local signal-to-noise ratio (SNR) within each T-F unit. When *a priori* knowledge of the target and the interferer is available, the local SNRs can be computed using the T-F representation of each signal individually. Mask units that are dominated by the target are assigned a value of one and zero otherwise. Without *a priori* knowledge, the mask values are estimated, often by reformulating the mask estimation problem as a classification problem and using machine learning techniques to perform the classification. In the final stage of the binary-masking approach, the mask (ideal or estimated) is applied to the noisy mixture to segregate the target from the interfering signal.

Estimation algorithms can make one of two types of errors: false positive (i.e., type-I or false alarm) errors occur when interferer-dominated units are incorrectly labeled target-dominated, and false negative (i.e., type-II or miss) errors occur when target-dominated units are incorrectly labeled interferer-dominated. To investigate the influence of different distributions of these errors on speech intelligibility outcomes, Kressner and Rozell (2015) and Kressner *et al.* (2016) scored normal hearing (NH) listeners and cochlear implant (CI) recipients, respectively, on their word recognition of noisy speech that had been processed with binary masks containing different distributions of errors. These studies together demonstrate that the impact of false positive and false negative error rates on speech intelligibility scores is highly dependent on how the errors are distributed.

To date, however, the most commonly used outcome measure for assessing segregation performance is the hit-minus-false-alarm (H-FA) metric, which is the difference between the *hit rate* (i.e., the percentage of correctly classified target-dominated T-F units) and the *false alarm rate* (i.e., the percentage of incorrectly classified interferer-dominated T-F units). The prevalence of this metric emerged after Kim *et al.* (2009) reported a correlation ($r = 0.80$) between H-FA and speech intelligibility in their listener study. However, Kim *et al.* (2009) conducted their listener study with masks that were estimated with only one algorithm. Since their algorithm likely makes errors in similar ways in all of the masks it estimates, error distribution was not a factor in their analysis. When developing and optimizing algorithms that estimate the IBM though, more than one algorithm or design of an algorithm is being compared, and each of these algorithms will make errors in different ways. Thus, it is important to consider whether H-FA can predict the intelligibility outcomes for binary masks with different error distributions.

Alternatively to H-FA, binary-masked speech has also been evaluated in literature with the short-time objective intelligibility (STOI) metric (Taal *et al.*, 2011). STOI was specifically designed to be able to predict, among other things, intelligibility outcomes for binary-masked speech using ideal

[a]Electronic mail: aakress@elektro.dtu.dk

masks and masks with artificially induced, uniformly random errors. Given that STOI evaluates the reconstructed signals as a whole rather than each classification decision independently, it holds promise for being able to predict outcomes for masks with different error distributions because it can take into account the perceptual relevance of the errors.

Several other metrics have been proposed to assess sound source separation algorithms, such as the loudness-weighted H-FA (Yu *et al.*, 2014), the IBM ratio (Hummersone *et al.*, 2011), and the intelligibility metric based on an auditory preprocessing model (Christiansen *et al.*, 2010). However, these metrics have gained limited traction due to either lacking generalizability or accessibility. Therefore, this study investigates the ability of the two most commonly used metrics, H-FA and STOI, to predict behavioral speech intelligibility outcomes for masks with varying distributions of errors.

II. METHODS

The objective measures H-FA and STOI were assessed on their ability to predict the intelligibility scores from the listener studies in Kressner and Rozell (2015). In these experiments, speech mixed with babble was processed with binary masks generated from a statistical model that artificially introduced errors with parametrically controlled distributions. NH listeners were then scored on how many words they could correctly identify in the processed sentences for a variety of error distributions. In the first two experiments, the masks contained varying rates of either false positive or false negative errors (α or β, respectively) that were distributed either randomly (i.e., uniform distribution) or with a varying amount of clustering. The clustering parameter γ defined how much more likely neighboring T-F units were to have the same gain values than different gain values. Thus, binary masks with a higher γ were more likely to contain errors that were clustered together in time and frequency. The third listener experiment addressed the more realistic scenario where the masks contained both false positive and false negative errors. These errors were then either random (i.e., unstructured, $\gamma = 1.0$) or clustered with $\gamma = 2.0$.

The masks and mixture signals were regenerated and processed for each of the three experiments using the same procedures as in Kressner and Rozell (2015). Then H-FA and STOI were computed for each individual sentence. For H-FA, each mask was compared to its ideal version, the true positive and false positive rates were calculated, and then H-FA was computed. For STOI, the procedure from Taal *et al.* (2011) was followed and the STOI scores were converted to word recognition predictions using the database-specific mapping. Means were taken for both H-FA and STOI across all sentences to obtain an overall prediction for each condition.

III. RESULTS

Figure 1 shows the behavioral results and predicted scores for the first two experiments of Kressner and Rozell (2015). Figure 1(a) shows the behavioral results when false positive errors are introduced (i.e., more energy from the interferer-dominated T-F units is erroneously retained), whereas Fig. 1(b) shows the behavioral results when false negative errors are introduced (i.e., fewer of the target-dominated T-F units are retained than in the IBM). Figures 1(c) and 1(d) show the predicted speech intelligibility using H-FA, and Figs. 1(e) and 1(f) show the predicted speech intelligibility using STOI.

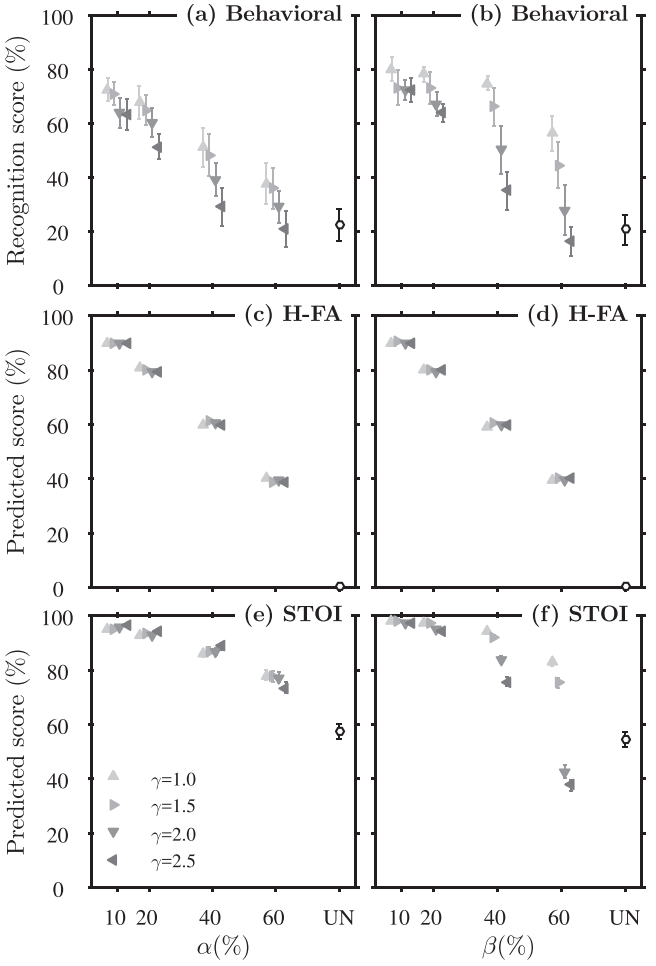

FIG. 1. Mean behavioral word recognition scores from Kressner and Rozell (2015) for binary-masked speech that was processed with masks containing either (a) only false positive errors (α) or (b) only false negative errors (β). Corresponding mean H-FA scores are shown in (c) and (d), respectively, and corresponding mean STOI scores are shown in (e) and (f), respectively. The amount of error clustering (γ) in the masks is indicated by the different symbols. Error bars indicate 95% confidence intervals. UN is the unity mask control condition.

The behavioral results suggest that false negative errors can be as detrimental to speech intelligibility as false positive errors if they are clustered. However, H-FA fails to predict the impact of the distribution of errors, and instead, predicts that all masks with the same error rates yield the same intelligibility outcome. Thus, even though the correlation between mean H-FA and behavioral scores for conditions with $\gamma = 2.0$ (i.e., the conditions with an error distribution that most closely match the error distribution of the estimated masks of Kim *et al.*; Kressner and Rozell, 2015) is high ($r = 0.97$), H-FA is unable to account for the differences in the behavioral scores that arise when masks contain errors that are distributed differently.

In contrast to H-FA, STOI is able to qualitatively predict the trends in the behavioral data when false negative errors are presented, as demonstrated by the similarities between Figs. 1(f) and 1(b). Additionally, STOI is also able to predict the trends in the behavioral data for both false positive and false negative errors when the errors are unstructured ($\gamma = 1.0$). These $\gamma = 1.0$ conditions contain unstructured errors in the same way as the masks from Li and Loizou (2008), and since Taal *et al*. (2011) used the data from the Li and Loizou (2008) study to develop STOI, it is not surprising that STOI is able to predict intelligibility well for these conditions. Nevertheless, STOI is unable to predict the influence of clustering on the effect of false positive errors [compare Fig. 1(e) with Fig. 1(a)]. It is clear that these objective measures are not capturing the effect of structured mask errors even in the relatively simple cases of single error types. Unfortunately, the real situation is even more complex because estimation algorithms are unlikely to make only false positive or false negative errors.

The final listener study in Kressner and Rozell (2015) addresses this more realistic scenario with interacting false positive and false negative errors. Figure 2(a) shows a contour plot based on the behavioral word recognition for both unstructured ($\gamma = 1.0$) and more realistic, clustered errors ($\gamma = 2.0$). Based on this contour plot, if the errors in the masks are unstructured, all combinations of α and β that fall on or below the solid contour line marked 50%, for example, would lead to mean word recognition scores of 50% or better. In contrast, if the errors in the masks are clustered with $\gamma = 2.0$, only combinations of false positive rates and false negative rates that fall on or below the *dashed* contour line marked 50% would lead to mean word recognition scores of 50% or better.

There are two salient features in the contour plot of the behavioral data. First, there is a shift of the $\gamma = 2.0$ contour lines towards the origin compared to the respective $\gamma = 1.0$ contour lines, which suggests that masks with higher amounts of clustering must achieve higher accuracy rates in order to yield the same intelligibility outcomes. Furthermore, there is a change in the slopes of the $\gamma = 2.0$ contour lines compared to the $\gamma = 1.0$ contour lines. Because the slopes of the $\gamma = 1.0$ contour lines in Fig. 2(a) are nearly equal to -1, masks containing unstructured errors appear to be equally influenced by false positive and false negative errors. In contrast, the $\gamma = 2.0$ contour lines are more steeply sloping, which suggests that high false negative error rates (β) are more detrimental to intelligibility outcomes than high false positive error rates (α) when the errors are clustered.

Figure 2(b) shows contours based on the intelligibility outcomes H-FA predicts. The general qualitative relationship between intelligibility and different combinations of false positive and false negatives rates are predicted well for the conditions with unstructured errors ($\gamma = 1.0$), as demonstrated particularly by the fact that the $\gamma = 1.0$ contour lines in Fig. 2(b) are placed in approximately the same location as the respective $\gamma = 1.0$ contour lines in Fig. 2(a), as well as by the fact that the $\gamma = 1.0$ contour lines in Fig. 2(b) all have approximate slopes of -1. However, H-FA fails to predict the negative impact that the clustering of the errors has on intelligibility, as demonstrated by the lack of shift of the $\gamma = 2.0$ contour lines as well as the lack of increased steepness in the $\gamma = 2.0$ contour lines. In contrast to H-FA, STOI in Fig. 2(c) successfully predicts the qualitative trends in Fig. 2(a) relating to the shift of the $\gamma = 2.0$ contour lines and the change in slope. However, it tends to overpredict the intelligibility outcomes in general, and it underpredicts the effect of error clustering.

IV. DISCUSSION

Estimation algorithms will likely produce masks with errors that are distributed in different ways depending on the design of the algorithm. For example, one algorithm might include a spectro-temporal integration stage to incorporate contextual information (Healy *et al*., 2013; May and Dau, 2013, 2014) and consequently increase clustering in the masks. Alternatively, another algorithm may use a classifier that, for example, consistently mislabels the high frequency channels or the acoustic onsets. Although H-FA can predict outcomes relatively well among masks with the same error distributions, this study has demonstrated that it fails to predict the differences in intelligibility that arise when masks contain different error distributions. Thus, it is an unreliable metric to use when evaluating estimation algorithms. In addition to using H-FA for evaluation, many supervised learning approaches in the

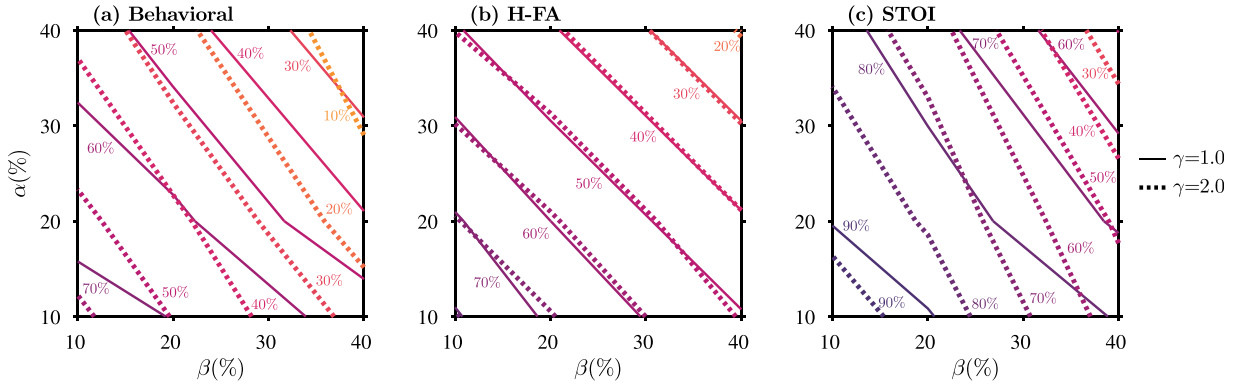

FIG. 2. (Color online) Contours of (a) behavioral word recognition from Kressner and Rozell (2015) (redrawn here in percent correct rather than as a relative score) and predicted scores using (b) H-FA and (c) STOI for speech processed with binary masks that contain a range of false positive (α) and false negative (β) rates and two levels of clustering (γ). Masks with unstructured errors ($\gamma = 1.0$) are indicated with solid contour lines, whereas masks with clustered errors ($\gamma = 2.0$) are indicated with dashed contour lines.

literature have used H-FA as a design objective (e.g., Han and Wang, 2012; May and Dau, 2014). Since a higher H-FA score does not necessarily produce a higher intelligibility score, H-FA may also be unfit as a cost function for algorithm design.

Yu *et al*. (2014) tried to address some of the limitations of H-FA when they proposed the loudness-weighted H-FA, a mask-based metric that takes into account the relative importance of each error. However, the importance weights in the metric were fit to masks that employ an alternate definition of the IBM (i.e., the "target binary mask"; Kjems *et al*., 2009) and that use an FFT-based frequency decomposition. Furthermore, the weights were fit only to the behavioral scores for their own listener study, which introduced either only false positive errors or only false negative errors to each mask. Since their metric is not directly applicable to masks that employ a different mask definition than the "target binary mask," make use of a different T-F decomposition, or contain both false positive and false negative errors, it is not generalizable enough in its current form for widespread use.

In contrast to H-FA, STOI is able to qualitatively predict the effects of clustering on speech intelligibility outcomes. It is therefore a potential alternative to H-FA. However, STOI tended to overpredict intelligibility, which is consist with the findings in Healy *et al*. (2015). Furthermore, it is unclear how STOI's underprediction of the effect of clustering will impact its ability to compare different estimation algorithms. To give an illustrative example of how this can be problematic, suppose that a hypothetical estimation algorithm tends to make errors that are randomly distributed (i.e., $\gamma = 1.0$) with $\alpha = 10\%$ and $\beta = 35\%$. Then Fig. 2(a) suggests that listeners would on average recognize about 61% of words in sentences processed with masks from that algorithm. Figure 2(c), on the other hand, suggests that STOI would predict a score of about 82% correct. Next, suppose that a second hypothetical algorithm makes errors that tend to cluster together such that $\gamma = 2.0$, and on average, the algorithm makes errors such that $\alpha = 15\%$ and $\beta = 10\%$. Figure 2(a) suggests that listeners would recognize about 58% of words in the sentences processed by this second algorithm, which is slightly less than the first algorithm. However, Fig. 2(c) suggests that STOI would predict a score of about 92%, which is better than the first algorithm. Thus, because STOI underpredicts the effect of clustering, it would incorrectly predict that the second algorithm would elicit higher intelligibility than the first algorithm. This hypothetical example is informative, but further investigation is of course needed in order to fully understand how the actual error distributions in estimated binary masks (as opposed to systematically generated error distributions) impact intelligibility outcomes, and furthermore, whether or not STOI is able to predict the outcomes. It is clear, however, that the performance of estimation algorithms should not be evaluated solely on the basis of H-FA since it ignores error distributions altogether.

ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation Graduate Research Fellowship under Grant No. DGE-1148903, a Det Frie Forskningsråd (DFF) Individual Postdoctoral Grant, and the EU FET grant TWO!EARS, No. ICT-618075.

Anzalone, M. C., Calandruccio, L., Doherty, K. A., and Carney, L. H. (**2006**). "Determination of the potential benefit of time-frequency gain manipulation," Ear Hear. **27**, 480–492.

Brungart, D. S., Chang, P. S., Simpson, B. D., and Wang, D. (**2006**). "Isolating the energetic component of speech-on-speech masking with ideal time-frequency segregation," J. Acoust. Soc. Am. **120**, 4007–4018.

Christiansen, C., Pedersen, M. S., and Dau, T. (**2010**). "Prediction of speech intelligibility based on an auditory preprocessing model," Speech Commun. **52**, 678–692.

Han, K., and Wang, D. (**2012**). "A classification based approach to speech segregation," J. Acoust. Soc. Am. **132**, 3475–3483.

Healy, E. W., Yoho, S. E., Chen, J., Wang, Y., and Wang, D. (**2015**). "An algorithm to increase speech intelligibility for hearing-impaired listeners in novel segments of the same noise type," J. Acoust. Soc. Am. **138**, 1660–1669.

Healy, E. W., Yoho, S. E., Wang, Y., and Wang, D. (**2013**). "An algorithm to improve speech recognition in noise for hearing-impaired listeners," J. Acoust. Soc. Am. **134**, 3029–3038.

Hu, Y., and Loizou, P. C. (**2008**). "A new sound coding strategy for suppressing noise in cochlear implants," J. Acoust. Soc. Am. **124**, 498–509.

Hummersone, C., Mason, R., and Brookes, T. (**2011**). "Ideal binary mask ratio: A novel metric for assessing binary-mask-based sound source separation algorithms," IEEE Trans. Audio Speech Lang. Process. **19**, 2039–2045.

Kim, G., Lu, Y., Hu, Y., and Loizou, P. C. (**2009**). "An algorithm that improves speech intelligibility in noise for normal-hearing listeners," J. Acoust. Soc. Am. **126**, 1486–1494.

Kjems, U., Boldt, J. B., Pedersen, M. S., Lunner, T., and Wang, D. (**2009**). "Role of mask pattern in intelligibility of ideal binary-masked noisy speech," J. Acoust. Soc. Am. **126**, 1415–1426.

Kressner, A. A., and Rozell, C. J. (**2015**). "Structure in time-frequency binary masking errors and its impact on speech intelligibility," J. Acoust. Soc. Am. **137**, 2025–2035.

Kressner, A. A., Westermann, A., Buchholz, J. M., and Rozell, C. J. (**2016**). "Cochlear implant speech intelligibility outcomes with structured and unstructured binary mask errors," J. Acoust. Soc. Am. **139**, 800–810.

Li, N., and Loizou, P. C. (**2008**). "Factors influencing intelligibility of ideal binary-masked speech: Implications for noise reduction," J. Acoust. Soc. Am. **123**, 1673–1682.

May, T., and Dau, T. (**2013**). "Environment-aware ideal binary mask estimation using monaural cues," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust. (WASPAA)*, pp. 1–4.

May, T., and Dau, T. (**2014**). "Computational speech segregation based on an auditory-inspired modulation analysis," J. Acoust. Soc. Am. **136**, 3350–3359.

Roman, N., Wang, D., and Brown, G. J. (**2003**). "Speech segregation based on sound localization," J. Acoust. Soc. Am. **114**, 2236–2252.

Taal, C. H., Hendriks, R. C., Heusdens, R., and Jensen, J. (**2011**). "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," IEEE Trans. Audio Speech Lang. Process. **19**, 2125–2136.

Wang, D. (**2005**). "On ideal binary mask as the computational goal of auditory scene analysis," in *Speech Separation by Humans and Machines*, edited by P. Divenyi (Springer, New York), pp. 181–197.

Yu, C., Wójcicki, K. K., Loizou, P. C., Hansen, J. H., and Johnson, M. T. (**2014**). "Evaluation of the importance of time-frequency contributions to speech intelligibility in noise," J. Acoust. Soc. Am. **135**, 3007–3016.