INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Comparing the influence of spectro-temporal integration in computational speech segregation

Thomas Bentsen, Tobias May, Abigail A. Kressner and Torsten Dau

Hearing Systems Group, Technical University of Denmark
DK-2800 Kgs. Lyngby, Denmark

{thobe, tobmay, aakress, tdau}@elektro.dtu.dk

Abstract

The goal of computational speech segregation systems is to automatically segregate a target speaker from interfering maskers. Typically, these systems include a feature extraction stage in the front-end and a classification stage in the back-end. A spectro-temporal integration strategy can be applied in either the front-end, using the so-called delta features, or in the back-end, using a second classifier that exploits the posterior probability of speech from the first classifier across a spectro-temporal window. This study systematically analyzes the influence of such stages on segregation performance, the error distributions and intelligibility predictions. Results indicated that it could be problematic to exploit context in the back-end, even though such a spectro-temporal integration stage improves the segregation performance. Also, the results emphasized the potential need of a single metric that comprehensively predicts computational segregation performance and correlates well with intelligibility. The outcome of this study could help to identify the most effective spectro-temporal integration strategy for computational segregation systems.

Index Terms: computational speech segregation, binary masks, supervised learning, spectro-temporal integration.

1. Introduction

Computational speech segregation systems attempt to automatically segregate a target signal from interfering noise. One frequently-used approach is to construct an ideal binary mask (IBM) by retaining only those time-frequency (T-F) units that are target-dominated [1]. Many studies have used the IBM to segregate a target speech signal from a noisy mixture and demonstrated large intelligibility improvements [2, 3, 4]. However, *a priori* knowledge about the target and interferer is rarely available in realistic conditions and therefore, the goal of computational speech segregation systems is to obtain an estimated binary mask (EBM) given the noisy speech. Despite high levels of interfering noise, speech-dominated T-F units tend to cluster in spectro-temporal regions, forming so-called *glimpses*, and the size of these glimpses has been shown to correlate well with speech intelligibility scores from normal-hearing listeners [5]. Consequently, several studies have tried to explore spectro-temporal context in computational segregation systems. One strategy is to exploit context in the front-end by using so-called delta features [6], which capture feature variations across time and frequency at the expense of a higher dimensional feature vector. Alternatively, spectro-temporal context can be exploited in the classification back-end by employing a two-layer segregation stage [7, 8]. Specifically, the posterior probability of speech presence obtained from a first classifier is learned by a second classifier across a spectro-temporal window, where the amount of integration can be controlled by the size of the window function [8].

To date, the effectiveness of computational segregation systems and the benefit of spectro-temporal integration strategies have been primarily evaluated using a technical metric, namely the H - FA, which quantifies segregation performance by calculating the difference between the percentage of correctly classified speech-dominated T-F units (hit rate, H) and the percentage of incorrectly classified noise-dominated T-F units (false alarm rate, FA) [6, 7, 8, 9, 10, 11]. However, there is evidence suggesting that speech intelligibility scores are highly dependent on the distribution of mask errors rather than the overall H - FA rate [12], and this questions the applicability of the H - FA as the sole metric to optimize or evaluate computational segregation systems. The clustering of the speech-dominated T-F units in glimpses suggests that a certain type of structure is inherently embedded in the IBM. However, depending on the choice of the spectro-temporal integration strategy in either the front-end or the back-end, it might have different consequences on the error distribution in the EBM.

The goal of the present study is, therefore, to systematically analyze the influence of spectro-temporal integration strategies in the front-end and the back-end of a speech segregation system using not only the H - FA, but also by considering the distribution of errors and the impact on predicted speech intelligibility using the short-term objective intelligibility (STOI) metric [13]. In previous studies [6, 7], the same short noise recording has been used for training and testing. In such experimental setups, a classification-based segregation system can then potentially capture all characteristics of the signals [11]. A second goal is, therefore, to analyze the potential influence of the noise duration on each of the spectro-temporal integration strategies.

2. The speech segregation system

The segregation system consisted of a feature extraction front-end and a classification back-end [14], as shown in Fig. 1. The target signal was reconstructed by applying the EBM to the sub-band signals of the noisy speech, as illustrated by the dashed line. Each processing stage is described in detail in the following.

2.1. Feature extraction front-end

The distinct characteristics of speech and noise components were captured by amplitude modulation spectrogram (AMS) features [6, 8, 14, 15]. To derive these, the noisy speech was

This work was supported by the Oticon Centre of Excellence for Hearing and Speech Sciences, the EU FET grant TWO!EARS, ICT-618075 and by the Danish Council for Independent Research (DFF) with grant number DFF-5054-00072.

http://dx.doi.org/10.21437/Interspeech.2016-1025

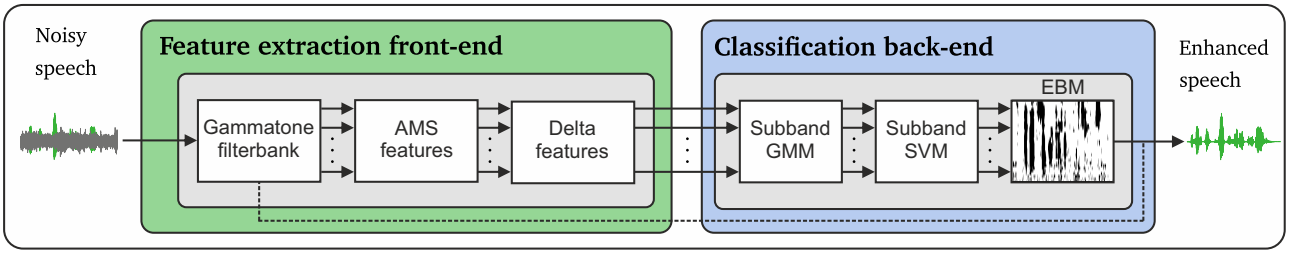

Figure 1: Block diagram of the segregation system that shows the main blocks of the feature extraction front-end and the classification back-end. The dashed line illustrates the reconstruction of the target by applying the EBM to the subband signals of the noisy speech.

sampled at a rate of 16 kHz and decomposed into 31 frequency channels by a Gammatone filterbank, whose center frequencies were equally spaced on the equivalent rectangular bandwidth (ERB) scale between 80 and 7642 Hz. The envelope in each subband was extracted by half-wave rectification and low-pass filtering with a cutoff frequency of 1 kHz. Then, each envelope was normalized by its median that was computed over the entire signal, which was shown to improve the generalization to unseen acoustic conditions (e.g., signal-to-noise ratios (SNRs) and room reverberation) [8, 16]. The normalized envelopes were then processed by a modulation filterbank that consisted of one first-order low-pass and five band-pass filters with logarithmically spaced center frequencies and a constant Q-factor of 1. The root mean square (RMS) value of each modulation filter was then calculated across time frames corresponding to 32 ms with 75 % overlap, resulting in a 6-dimensional feature vector for each T-F unit $\boldsymbol{A}(t,f) = \{M_1(t,f), \ldots, M_6(t,f)\}^T$.

Context was explored in the front-end by appending delta features across time (Δ_T) and frequency (Δ_F) [6, 9, 10]. The final feature vector for each individual T-F unit at time frame t and frequency channel f consisted of $\boldsymbol{X}(t,f) = [\boldsymbol{A}(t,f), \Delta_T\boldsymbol{A}(t,f), \Delta_F\boldsymbol{A}(t,f)]$, where:

$$\Delta_T\boldsymbol{A}(t,f) = \begin{cases} \boldsymbol{A}(2,f) - \boldsymbol{A}(1,f), & \text{if } t = 1 \\ \boldsymbol{A}(t,f) - \boldsymbol{A}(t-1,f), & \text{otherwise,} \end{cases} \quad (1)$$

$$\Delta_F\boldsymbol{A}(t,f) = \begin{cases} \boldsymbol{A}(t,2) - \boldsymbol{A}(t,1), & \text{if } f = 1 \\ \boldsymbol{A}(t,f) - \boldsymbol{A}(t,f-1), & \text{otherwise.} \end{cases} \quad (2)$$

The size of the feature vector including delta features then increased from 6 dimensions to 18 dimensions.

2.2. Classification back-end

The classification back-end consisted of a two-layer segregation stage [8, 14]. In the first layer, a Gaussian mixture model (GMM) classifier was trained to represent the speech and noise-dominated AMS feature distributions ($\lambda_{1,f}$ and $\lambda_{0,f}$) for each subband f. To separate the feature vector into speech- and noise-dominated T-F units, a local criterion (LC) was applied to the *a priori* SNR. The GMM classifier output was given as the posterior probability of speech and noise $P(\lambda_{1,f}|\boldsymbol{X}(t,f))$ and $P(\lambda_{0,f}|\boldsymbol{X}(t,f))$, respectively. The second layer consisted of a linear support vector machine (SVM) classifier [17], which considered the posterior probability of speech $P(\lambda_{1,f}|\boldsymbol{X}(t,f))$ across a spectro-temporal integration window $\mathcal{W}$ for each subband [8]:

$$\bar{\boldsymbol{X}}(t,f) := \{P(\lambda_{1,u}|\boldsymbol{X}(u,v)) : (u,v) \in \mathcal{W}(t,f)\}. \quad (3)$$

According to [8], a causal and plus-shaped window function $\mathcal{W}$ was used here, whereas the window size with respect to time and frequency was controlled by Δt and Δf, respectively.

3. Evaluation

3.1. Stimuli

The speech material was taken from the Danish conversational language understanding evaluation (CLUE) database [18], which consists of 70 sentences for training and 180 sentences for testing. Noisy speech mixtures with an average duration of 2 s were created by mixing individual sentences with a stationary (ICRA1) and a fluctuating 6-talker (ICRA7) noise masker [19]. Both maskers had the same long term average spectrum (LTAS) as the CLUE corpus. A randomly-selected noise segment was used for each sentence and the noise segment started 250 ms before the speech onset and ended 250 ms after the speech offset.

3.2. Model training

The segregation system was trained for each of the two noise maskers. To investigate the influence of the noise duration, different models were trained with noise files that were limited to $5, 10, 50$ s or the total duration of the noise recording (60 s for ICRA1 and 600 s for ICRA7). The first layer of the classification back-end consisted of a GMM classifier with 16 Gaussian components and diagonal covariance matrices. The GMM classifier was trained with the 70 training sentences that were mixed three times with a randomly-selected noise segment at $-5, 0$ and 5 dB SNR. The subsequent SVM classifier was trained with only 10 sentences mixed at $-5, 0$ and 5 dB SNR. Afterwards, a re-thresholding procedure was applied [8, 9] using a validation set of 10 sentences. Both classifiers employed a LC of -5 dB.

3.3. Model evaluation

The segregation system was evaluated with 180 CLUE sentences that were not used during training. Each sentence was mixed with ICRA1 and ICRA7 noises at -5 and 0 dB SNR. To study the influence of the noise duration, the trained models were evaluated with the same noise recordings used during training. Similar to the training, the noise recordings were limited in duration to $5, 10, 50$ s or the total duration of the noise recording. In addition, a different noise recording of the same noise type was used to test the ability of the segregation system to generalize to unseen noise fluctuations of the same kind.

Three different metrics were used for evaluation, namely the H - FA, the clustering parameter γ and the STOI metric. The clustering parameter γ was estimated by the graphical model described in [12]. Given a binary mask, the graphical model predicts the amount of clustering γ as a single number, where $\gamma = 1.0$ reflects a mask with uniformly and randomly connected T-F units. Larger values (e.g., $\gamma = 2.0$) reflect binary masks with T-F units that are twice as likely to be in the same state as its neighboring units [12]. The STOI measure is based

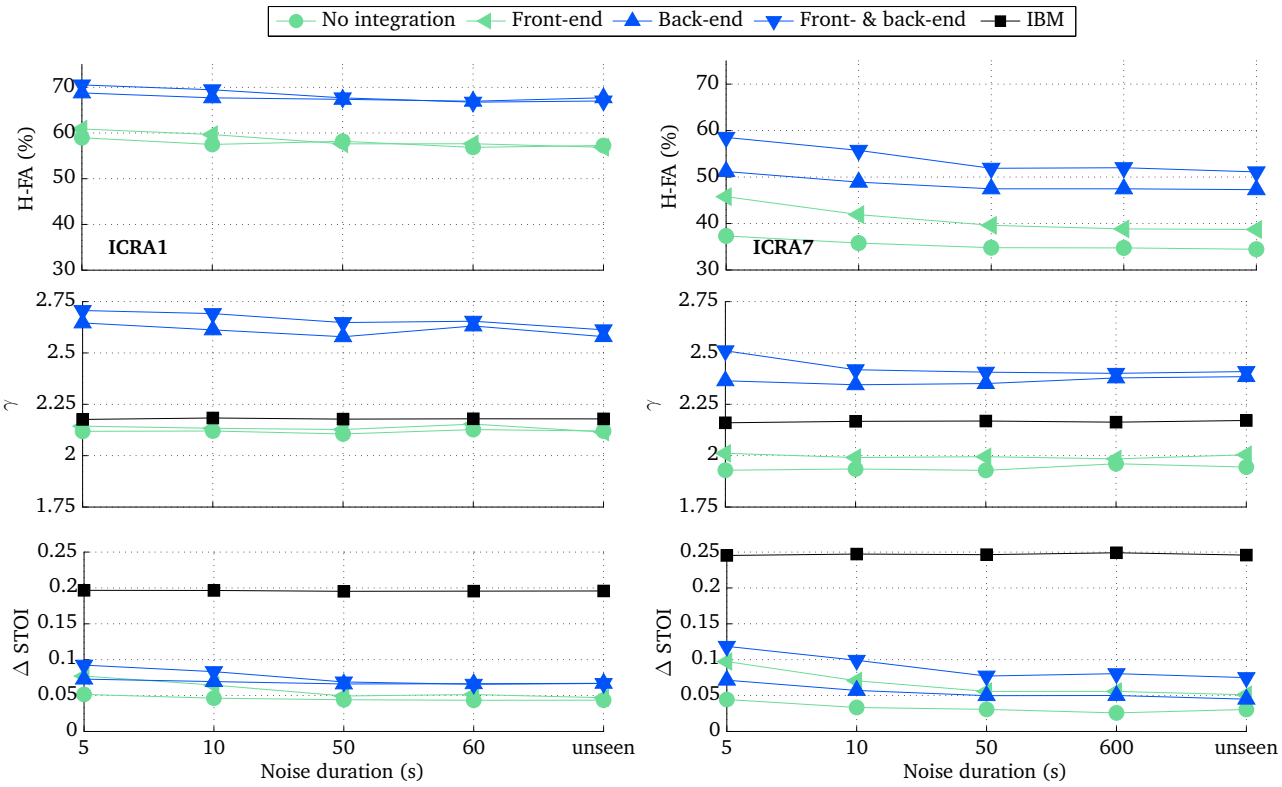

Figure 2: H-FA, γ and STOI improvements for the four models and the IBM averaged across 180 sentences and SNRs (-5 and 0 dB) for ICRA1 (left panels) and ICRA7 (right panels). Average STOI values of the unprocessed noisy speech were 0.66 (ICRA1) and 0.63 (ICRA7).

on a short-term correlation analysis between the clean and the degraded speech [13] mapped to a value between 0 and 1. In the current study, STOI improvements (Δ STOI) were reported as the relative difference between the predicted STOI values for the processed and the unprocessed noisy speech signal.

3.4. Experimental setup

To systemically analyze the influence of spectro-temporal integration in the front-end and the back-end, the following four segregation models were tested, as listed in Tab. 1. "No integration" denotes the model with no delta features in the front-end and no spectro-temporal integration in the back-end ($\Delta t = 1, \Delta f = 1$). "Front-end" includes the delta features. "Back-end" does not utilize delta features, but applies spectro-temporal integration in the back-end ($\Delta t = 3, \Delta f = 9$). "Front- & back-end" exploits both delta features in the front-end and spectro-temporal integration in the back-end ($\Delta t = 3, \Delta f = 9$).

Table 1: Configurations of the speech segregation system.

Model	Front-end		Back-end	
	Delta features	Feature dimension	$\mathcal{W}$ size Δt	$\mathcal{W}$ size Δf
No integration	no	6	1	1
Front-end	yes	18	1	1
Back-end	no	6	3	9
Front- & back-end	yes	18	3	9

4. Results

The performance of the four segregation models and the IBM is presented in Fig. 2 as a function of the noise duration for the two noise maskers ICRA1 (left panels) and ICRA7 (right panels). The three different panels on each side show the H - FA rate (top panels), the clustering parameter γ (middle panels) and the STOI metric (lower panels) averaged across 180 sentences and two SNRs (-5 and 0 dB).

In general, the segregation models produced higher H - FA rates in the presence of the stationary ICRA1 noise than for the ICRA7 noise, presumably because it was more difficult to separate the speech modulations from the non-stationary 6-talker babble noise. For both noise maskers, the lowest H - FA rates were observed for the "No integration" model and the highest H - FA rates for "Front- & back-end". Also, larger H - FA rates were obtained for the "Back-end" than the "Front-end" model. Each spectro-temporal integration strategy has previously been shown to improve H - FA rates separately [6, 8, 10, 11]. These previous results can be confirmed here for the ICRA7 noise by comparing both the "Back-end" and "Front-end" models with the "No integration" model.

The middle panels reveal that the IBM itself contains a certain amount of structure, presumably due to the compact representation of speech-dominated T-F units forming glimpses of the target signal. Also, reported values of γ from the model "No integration" are consistent with previous results [12, 20]. Most importantly, the γ values from models that exploited spectro-temporal context through the SVM classifier in the back-end (models "Back-end" and "Front- & back-end") are consistently larger than those from models where the SVM classifier did

not incorporate contextual information across adjacent T-F units (models "No integration" and "Front-end"). On the contrary, the delta features alone do not seem to increase the amount of clustering in the mask.

In the bottom panels, the STOI improvement of the IBM indicates the largest possible intelligibility improvement that the segregation models can achieve. The model "Front-end" produced larger STOI improvements than "Back-end" for the ICRA7 noise. Overall, the largest improvements were predicted for the model "Front- & back-end". In general, STOI predicted larger intelligibility improvements for ICRA7 than ICRA1.

Furthermore, Fig. 2 demonstrates that the segregation system can capture all relevant signal characteristics when the same noise recording was used for training and testing, resulting in high H - FA rates and large STOI improvements for short noise durations. This trend was more pronounced for the non-stationary ICRA7 noise and decreased with longer noise duration. However, a moderate classifier complexity was chosen here (16 Gaussian components with diagonal covariance matrices), which was shown to reduce the risk of over-fitting the segregation system [11]. As a result, the generalization ability was improved, indicated by a stable system performance in terms of H - FA rates and STOI improvements for noise durations of 50 s and beyond. In contrast to the H - FA rates and STOI, the γ values stayed almost constant across the noise duration range.

Figure 3 illustrates binary masks for one particular CLUE sentence mixed with ICRA7 noise at -5 dB SNR. Panel a) shows the IBM and panels b)-e) present the EBMs for the four tested models. The misclassified T-F units (misses and false alarms) are shown on top of the binary masks for a visualization of the error distributions. In addition, the evaluation metrics are shown in parenthesis. The effect of exploiting contextual knowledge in the back-end can be observed here. The panels d)-e) show masks with a larger amount of T-F clustering than the masks in panels b)-c). Obviously, the erroneous T-F units also become more structured.

5. Discussion and conclusion

Using the SVM classifier to exploit contextual knowledge in the back-end increased the H - FA rates but, at the same time, the amount of clustering (γ) in the masks was increased. In addition, the panels b)-e) in Fig. 3 revealed that the increased amount of clustering also led to an increased clustering of the two types of mask errors (miss and false alarm). Previously, it has been argued that clustering of the two types of errors reduces the intelligibility scores in comparison to the randomly distributed errors [12]. This is supported by the predictions of the intelligibility scores with STOI, where larger improvements using the delta features than exploiting contextual knowledge in the back-end alone are predicted for the ICRA7 noise. This also means that, for an increased γ, a higher H - FA rate is required to obtain the same intelligibility score. It therefore seems problematic to exploit context in the back-end using a SVM classifier, even though such a spectro-temporal integration stage improves the H - FA rate [7, 8]. The findings also suggest that using delta features might be a better spectro-temporal integration strategy in computational segregation systems, despite the fact that the H - FA rate does not increase as much as when exploiting contextual knowledge through a SVM classifier. However, it is necessary to confirm these findings with actual listening experiments.

In this study, both matched and unseen noise segments of the same noise type were used to evaluate classification-based

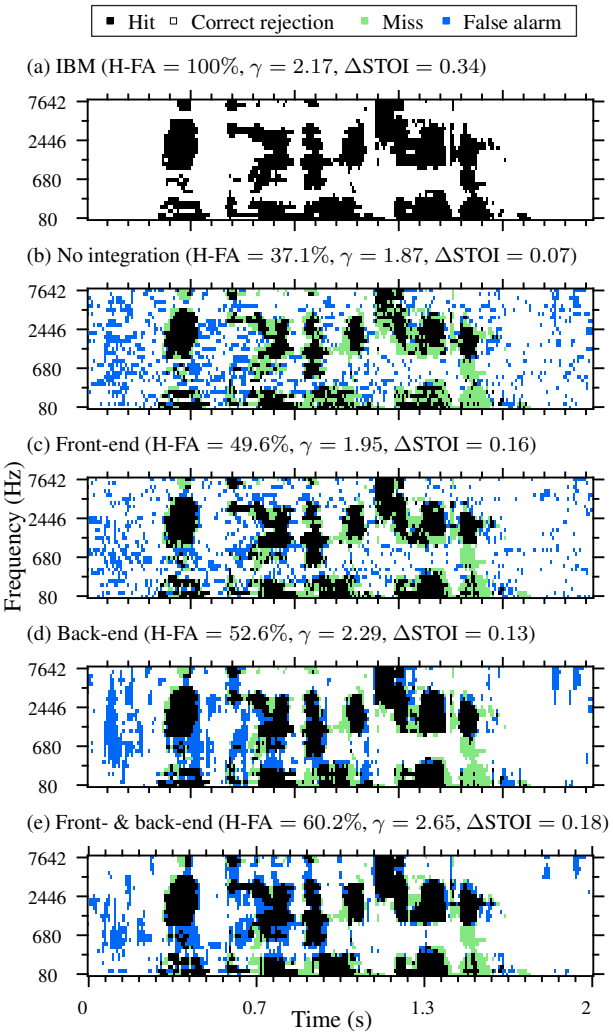

Figure 3: Binary masks for a CLUE sentence mixed with ICRA7 noise at -5 dB SNR. Misses (target-dominated T-F units erroneously labeled as masker-dominated) and false alarms (masker-dominated T-F units erroneously labeled as target-dominated) are shown on top of the masks.

segregation systems. As the ranking of the four models did not change with increasing noise durations, the findings of the influence of the spectro-temporal integration stage apply to both restricted and more realistic experimental setups with unseen noise segments of the same noise type. Future research will analyze the generalization ability of the segregation system to unseen noise types and will consider large-scale training [21].

A recent study highlighted potential limitations of STOI in predicting the intelligibility of binary-masked speech [22]. Two observations from this study support these findings. Firstly, a higher H - FA rate does not necessarily lead to a larger STOI improvement as seen by comparing the "Front-end" and "Back-end" models. Secondly, if the SVM-based integration strategy in the back-end indeed has a detrimental effect on the intelligibility scores, it would imply that STOI over-predicts the model "Front- & back-end". Thus, STOI alone would not account for all of the model differences described in this study. It emphasizes the potential need of a single metric that comprehensively predicts computational segregation performance and correlates well with intelligibility.

6. 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. Springer, 2005, pp. 181–197.

[2] D. S. Brungart, P. S. Chang, B. D. Simpson, and D. Wang, "Isolating the energetic component of speech-on-speech masking with ideal time-frequency segregation," *J. Acoust. Soc. Amer.*, vol. 120, no. 6, pp. 4007–4018, 2006.

[3] D. Wang, U. Kjems, M. S. Pedersen, J. B. Boldt, and T. Lunner, "Speech perception of noise with binary gains," *J. Acoust. Soc. Amer.*, vol. 124, no. 4, pp. 2303–2307, 2008.

[4] U. Kjems, J. B. Boldt, M. S. Pedersen, T. Lunner, and D. Wang, "Role of mask pattern in intelligibility of ideal binary-masked noisy speech," *J. Acoust. Soc. Amer.*, vol. 126, no. 3, pp. 1415–1426, 2009.

[5] M. Cooke, "A glimpsing model of speech perception in noise," *J. Acoust. Soc. Amer.*, vol. 119, no. 3, pp. 1562–1573, 2006.

[6] G. Kim, Y. Lu, Y. Hu, and P. 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.

[7] E. W. Healy, S. E. Yoho, Y. Wang, and D. L. Wang, "An algorithm to improve speech recognition in noise for hearing-impaired listeners," *J. Acoust. Soc. Amer.*, vol. 134, no. 6, pp. 3029–3038, 2013.

[8] T. May and T. Dau, "Computational speech segregation based on an auditory-inspired modulation analysis," *J. Acoust. Soc. Amer.*, vol. 136, no. 6, pp. 3350–3359, 2014.

[9] K. Han and D. L. Wang, "A classification based approach to speech segregation," *J. Acoust. Soc. Amer.*, vol. 132, no. 5, pp. 3475–3483, 2012.

[10] T. May and T. Dau, "Environment-aware ideal binary mask estimation using monaural cues," in *Proc. WASPAA*, New Paltz, NY, USA, 2013.

[11] ——, "Requirements for the evaluation of computational speech segregation systems," *J. Acoust. Soc. Amer.*, vol. 136, no. 6, pp. EL398–EL404, 2014.

[12] A. A. Kressner and C. J. Rozell, "Structure in time-frequency binary masking errors and its impact on speech intelligibility," *J. Acoust. Soc. Amer.*, vol. 137, no. 4, pp. 2025–2035, 2015.

[13] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of timefrequency weighted noisy speech," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 7, pp. 2125–2136, 2011.

[14] T. May, T. Bentsen, and T. Dau, "The role of temporal resolution in modulation-based speech segregation," in *Proc. Interspeech*, Dresden, Germany, 2015, pp. 170–174.

[15] J. Tchorz and B. Kollmeier, "SNR estimation based on amplitude modulation analysis with applications to noise suppression," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 11, no. 3, pp. 184–192, 2003.

[16] T. May and T. Gerkmann, "Generalization of supervised learning for binary mask estimation," in *Proc. IWAENC*, Juan les Pins, France, 2014, pp. 154–187.

[17] C. C. Chang and C. J. Lin, "LIBSVM: A library for support vector machines," Software is available at www.csie.ntu.edu.tw/~cjlin/libsvm, 2001.

[18] J. B. Nielsen and T. Dau, "Development of a Danish speech intelligibility test," *Int. J. Audiol.*, vol. 48, no. 10, pp. 729–741, 2009.

[19] W. A. Dreschler, H. Verschuure, C. Ludvigsen, and S. Westermann, "ICRA noises: Artificial noise signals with speech-like spectral and temporal properties for hearing instrument assessment," *Audiology*, vol. 40, no. 3, pp. 148–157, 2001.

[20] A. A. Kressner and C. J. Rozell, "Cochlear implant speech intelligibility outcomes with structured and unstructured binary mask errors," *J. Acoust. Soc. Amer.*, vol. 139, no. 2, pp. 800–810, 2016.

[21] J. Chen, Y. Wang, S. E. Yoho, D. Wang, and E. W. Healy, "Large-scale training to increase speech intelligibility for hearing-impaired listeners in novel noises," *J. Acoust. Soc. Amer.*, vol. 139, no. 5, pp. 2604–2612, 2016.

[22] A. A. Kressner, T. May, and C. J. Rozell, "Outcome measures based on classification performance fail to predict the intelligibility of binary-masked speech," *J. Acoust. Soc. Amer.*, vol. 139, no. 6, pp. 3033–3036, 2016.