The role of temporal resolution in modulation-based speech segregation

Tobias May, Thomas Bentsen and Torsten Dau

Hearing Systems Group,
Technical University of Denmark,
DK-2800 Kgs. Lyngby, Denmark
{tobmay, thobe, tdau}@elektro.dtu.dk

10.21437/Interspeech.2015-78

Abstract

This study is concerned with the challenge of automatically segregating a target speech signal from interfering background noise. A computational speech segregation system is presented which exploits logarithmically-scaled amplitude modulation spectrogram (AMS) features to distinguish between speech and noise activity on the basis of individual time-frequency (T-F) units. One important parameter of the segregation system is the window duration of the analysis-synthesis stage, which determines the lower limit of modulation frequencies that can be represented but also the temporal acuity with which the segregation system can manipulate individual T-F units. To clarify the consequences of this trade-off on modulation-based speech segregation performance, the influence of the window duration was systematically investigated.

Index Terms: speech segregation, ideal binary mask, amplitude modulation spectrogram features, temporal resolution

1. Introduction

Despite substantial research efforts that focused on the development of noise reduction algorithms over the past decades, the improvement of speech intelligibility in noisy conditions remains a challenging task [1, 2]. Assuming *a priori* knowledge about the target speech and the interfering noise, it is possible to construct an ideal binary mask (IBM) which separates the time-frequency (T-F) representation of noisy speech into target-dominated and masker-dominated T-F units. The IBM has been shown to significantly improve speech perception in noisy conditions [3, 4, 5]. The IBM produces intelligible speech when a resolution of about 12 - 16 frequency channels is used [4, 6]. At the same time, the manipulation of individual T-F units should be performed with a temporal resolution of at least 15 ms, in order to produce significant speech reception threshold (SRT) improvements [3].

Unfortunately, the IBM is not available in practice and, hence, needs to be estimated based on the noisy speech. In that context, the aforementioned requirements regarding the spectral and temporal resolution determine the bandwidth and the window size with which an estimated binary mask (EBM) should be obtained. In contrast to IBM processing, where the T-F manipulation can be performed at an arbitrarily high temporal resolution (e.g. on a sample-by-sample basis [3]), algorithms which try to derive an EBM typically operate on window durations between 20 ms [7] and 90 ms [8].

This work was supported by the EU FET grant TWO!EARS, ICT-618075 and by the Centre for Applied Hearing Research (CAHR).

Several previous studies have employed the extraction of amplitude modulation spectrogram (AMS) features with linearly-scaled modulation filters [7, 9, 10, 11]. Recently, it has been shown that a speech segregation system based on logarithmically-scaled AMS features, inspired by auditory processing principles, is superior to the linear AMS feature representation and can estimate the IBM with high accuracy [12]. One critical parameter is the window duration in the AMS feature representation. Modulation-based processing commonly involves longer analysis windows to fully resolve a period of low-frequency modulations within a single analysis window (e.g. 250 ms to analyze one period of 4 Hz modulations). This seems important for the ability to estimate speech-dominated T-F units, since it is known that low-frequency modulations are important for speech perception in the presence of stationary background noise [13]. In addition, a longer analysis window may also improve the accuracy of the EBM, since more information can be extracted from the noisy speech. However, a longer analysis window will introduce temporal smearing, which, in turn, may limit the effectiveness of manipulating individual T-F units.

Furthermore, many computational segregation systems exploit contextual information, either implicitly through the use of *delta* features [7, 9], or explicitly, by incorporating a spectro-temporal integration stage [10, 12, 14]. However, the interaction between the window duration and the spectro-temporal integration stage and its impact on speech segregation performance has not yet been clarified.

The goal of the present study is, therefore, to investigate the influence of the window duration on computational speech segregation based on auditory-inspired modulation features. Specifically, the interaction between window duration, estimation accuracy of the EBM and predicted speech intelligibility is analyzed. Moreover, the influence of a spectro-temporal integration stage is examined. The estimation accuracy of the EBM is measured using a technical classification measure (the hit rate minus false alarm rate) [9]. In addition, the predicted intelligibility of the reconstructed target speech is evaluated using the short-time objective intelligibility (STOI) metric [15].

2. Computational speech segregation

The segregation system consisted of a Gammatone-based analysis and synthesis stage. In the analysis stage, the noisy speech was sampled at a rate of 16 kHz and decomposed into 31 frequency channels using a Gammatone filterbank. The center frequencies were equally spaced on the equivalent rectangular bandwidth (ERB) scale between 80 and 7642 Hz. The envelope in each frequency channel was extracted by half-wave rectification and further smoothed by a second-order low-pass filter

with a cutoff frequency of 1 kHz to roughly simulate the loss of phase-locking in the auditory system towards higher frequencies. Based on this auditory spectrogram-like representation, a set of AMS features was extracted. A two-layer segregation stage, as further described in Sec. 2.2, was trained to discriminate between speech-dominated and noise-dominated T-F units by exploiting *a priori* knowledge about the AMS feature distribution corresponding to speech and noise activity [12]. This segregation stage produced an EBM that was applied to the individual subbands of the noisy speech in the synthesis stage in order to attenuate noise-dominated T-F units.

2.1. AMS feature extraction

Prior to the AMS feature extraction, each subband envelope was normalized by its median computed over the entire sentence. This normalization stage was shown to be crucial in order to deal with effects of room reverberation, spectral distortions and unseen signal-to-noise ratios (SNRs) [11, 12].

Each normalized subband was then analyzed by a modulation filterbank, consisting of a first-order low-pass filter and second-order band-pass filters whose center frequencies were logarithmically spaced up to 1024 Hz [12]. The bandpass filters were assumed to have a constant Q-factor of 1, inspired by findings in auditory modeling [16]. The cutoff frequency of the modulation low-pass filter f_{LP} was set to the inverse of the window duration T_w, to ensure that at least one period of the modulation frequency was included in the analysis window. The modulation power was measured for each frequency channel by computing the root mean square (RMS) value within each time window at the output of each modulation filter.

2.2. Segregation stage

In order to discriminate between speech-dominated and noise-dominated T-F units, a two-layer segregation stage was employed, which consisted of a Gaussian mixture model (GMM) classifier combined with a spectro-temporal integration stage based on a support vector machine (SVM) classifier [12]. First, a GMM classifier was trained for each individual frequency channel f to model the AMS feature distribution of speech-dominated and noise-dominated T-F units, denoted by $\lambda_{1,f}$ and $\lambda_{0,f}$. Given the AMS feature vector $\mathbf{X}(t,f)$ for a particular time frame t and frequency channel f, the *a posteriori* probability of speech and noise presence was computed by

$$P(\lambda_{1,f}|\mathbf{X}(t,f)) = \frac{P(\lambda_{1,f})P(\mathbf{X}(t,f)|\lambda_{1,f})}{P(\mathbf{X}(t,f))}, \quad (1)$$

$$P(\lambda_{0,f}|\mathbf{X}(t,f)) = \frac{P(\lambda_{0,f})P(\mathbf{X}(t,f)|\lambda_{0,f})}{P(\mathbf{X}(t,f))}, \quad (2)$$

where the two *a priori* probabilities $P(\lambda_{0,f})$ and $P(\lambda_{1,f})$ were computed by counting the number of feature vectors during training. The EBM without spectro-temporal integration was estimated by comparing the two *a posteriori* probabilities of speech and noise presence for each individual T-F unit

$$\mathcal{M}(t,f) = \begin{cases} 1 & \text{if } P(\lambda_{1,f}|\mathbf{X}(t,f)) > P(\lambda_{0,f}|\mathbf{X}(t,f)) \\ 0 & \text{otherwise.} \end{cases} \quad (3)$$

In the second layer, the *a posteriori* probability of speech presence $P(\lambda_{1,f})$ was considered as a new feature spanning across a spectro-temporal integration window, and subsequently learned by a SVM classifier [12]. The output of this second classification layer represented the EBM with spectro-temporal integration.

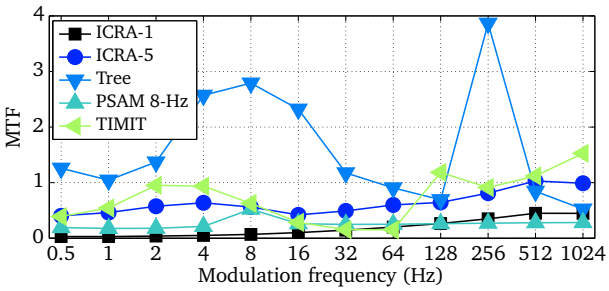

Figure 1: MTF of the four different background noises and the speech material from the TIMIT database.

2.3. Waveform synthesis

Before the EBM was applied to the noisy speech, a lower limit β was incorporated. This flooring limited the amount of noise attenuation, but reduced the impact of distortions (musical noise) caused by the binary processing [3]. A flooring value of $\beta = 0.1$, corresponding to 20 dB attenuation, was considered appropriate. This frame-based EBM was then interpolated to a sample-based EBM. Transitions in the EBM from speech-dominated to noise-dominated units or noise-dominated to speech-dominated units were smoothed by a raised-cosine window [17]. Then, the sample-based EBM was applied to the subband signals of the noisy speech. To remove across-frequency phase differences, the weighted subband signals were time-reversed, passed through the corresponding Gammatone filter, and time reversed again [17, 18]. Finally, the target signal was reconstructed by summing up the weighted and phase-aligned subband signals across all frequency channels.

3. Evaluation

3.1. Stimuli

Noisy speech was created by corrupting randomly selected male and female sentences from the TIMIT corpus with one of four different noise signals, from which a random segment was selected for each sentence. The noise was switched on 250 ms before the speech onset and was switched off 250 ms after the speech offset. The following noise types were used: two types of speech-shaped noise (SSN) (stationary ICRA1-noise and non-stationary, speech-modulated ICRA5-noise; [19]), 8-Hz amplitude-modulated pink noise and a recording of a cracking oak tree with wind noise[1]. The noise signals were split in two halves of equal size to prevent any overlap between the signals used during training and testing, which would result in an overly optimistic segregation performance [20].

An analysis of the broadband envelope fluctuations of all four noise types and the speech material from the TIMIT corpus is presented in Fig. 1, where the modulation transfer function (MTF) is shown for various modulation frequencies [19, 21]. The envelope fluctuations of the ICRA-5 noise are concentrated at low-frequency modulations with a peak around 4 Hz, and the general shape of the MTF is quite similar to the TIMIT speech material. In contrast, the MTF of the stationary ICRA-1 noise is pretty flat. Moreover, the cracking tree noise has strong contribution both at low and high modulation frequencies, whereas the MTF of the amplitude-modulated pink noise peaks at 8 Hz.

[1]Recording taken from `www.freesound.org/people/klankbeeld/sounds/211776/`

3.2. Model training

The GMM classifier described in Sec. 2.2 was trained with randomly selected sentences from the training set of the TIMIT corpus [22] that were corrupted with one of the four background noises at $-5, 0$ and $5\,\mathrm{dB}$ SNR. As explained in Sec. 3.4, the number of sentences involved in the training depends on the AMS feature configuration (see Tab. 1). A local criterion (LC) of $-5\,\mathrm{dB}$ was applied to the *a priori* SNR in order to separate the AMS feature distribution into speech-dominated and noise-dominated T-F units. The SVM-based spectro-temporal integration stage consisted of a causal, plus-shaped integration window spanning across 9 adjacent frequency channels and 3 time frames [12]. A linear SVM classifier [23] was trained with 10 sentences mixed at $-5, 0$ and $5\,\mathrm{dB}$ SNR. Afterwards, new SVM decision thresholds were obtained that maximized the hit minus false alarm (HIT - FA) rate [7] on a validation set of 10 sentences mixed at $-5, 0$ and $5\,\mathrm{dB}$ SNR. A separate GMM and SVM classifier was trained for each noise type.

3.3. Model evaluation

The segregation system was evaluated with 60 randomly selected sentences from the testing set of the TIMIT corpus mixed with the four different background noises at $-5, 0$ and $5\,\mathrm{dB}$ SNR. The segregation performance was assessed by comparing the EBM with the IBM. Specifically, the hit rate (HIT; percentage of correctly identified speech-dominated T-F units) minus the false alarm rate (FA; percentage of erroneously classified noise-dominated T-F units) was reported. In addition, the predicted intelligibility of the reconstructed speech signal was compared to the clean speech signal using the STOI metric [15], which has been shown to correlate with subjectively-measured speech intelligibility scores. For the STOI evaluation, the $250\,\mathrm{ms}$ noise-only segments at the beginning and the end of each sentence were discarded.

Moreover, the segregation system was compared to an short-time discrete Fourier transform (STFT)-based speech enhancement algorithm in Sec. 4.2. Specifically, the log-minimum mean square error (MMSE) noise reduction algorithm[2] [24] combined with the MMSE-based noise power estimation algorithm[2] [25] was used. The complete $250\,\mathrm{ms}$ noise-only segments before speech onset were used to properly initialize the noise power estimation.

3.4. Experimental setup

The segregation system was trained with AMS features based on 7 different window durations T_w, as shown in Tab. 1. Accordingly, the cutoff frequency of the modulation low-pass filter f_{LP} varied between $4\,\mathrm{Hz}$ (9 AMS features) and $256\,\mathrm{Hz}$ (3 AMS features). The frame shift was always set to $T_s = T_w/4$. As a result, the number of feature vectors available during training was higher for the AMS features with shorter window durations compared to longer window durations. To compensate for this, the number of TIMIT sentences used to train the GMM classifier was adjusted for window durations above $32\,\mathrm{ms}$ according to Tab. 1. 6 To investigate the influence of exploiting contextual information, two different segregation systems were trained: a single-layer GMM-based segregation system and a two-layer GMM-SVM segregation system including the spectro-temporal integration stage, both of which are described in Sec. 2.2.

[2]Matlab implementations were taken from the Voicebox toolbox provided by M. Brookes: `www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html`

Table 1: AMS feature settings.

T_w	T_s	f_{LP}	# dim.	# sentences
$256\,\mathrm{ms}$	$64\,\mathrm{ms}$	$4\,\mathrm{Hz}$	9	960
$128\,\mathrm{ms}$	$32\,\mathrm{ms}$	$8\,\mathrm{Hz}$	8	480
$64\,\mathrm{ms}$	$16\,\mathrm{ms}$	$16\,\mathrm{Hz}$	7	240
$32\,\mathrm{ms}$	$8\,\mathrm{ms}$	$32\,\mathrm{Hz}$	6	120
$16\,\mathrm{ms}$	$4\,\mathrm{ms}$	$64\,\mathrm{Hz}$	5	120
$8\,\mathrm{ms}$	$2\,\mathrm{ms}$	$128\,\mathrm{Hz}$	4	120
$4\,\mathrm{ms}$	$1\,\mathrm{ms}$	$256\,\mathrm{Hz}$	3	120

4. Experimental results

4.1. Effect of the window duration

The performance of the AMS-based segregation system is shown in Fig. 2 as a function of the window duration for the four different background noises. The top panel in each of the four subplots shows the STOI improvement relative to the unprocessed noisy speech for the IBM as well as the EBM with and without the SVM-based spectro-temporal integration stage. In addition, the corresponding HIT - FA rates of the two EBM systems are shown in the bottom panel.

It can be seen that the IBM produced the highest STOI improvements due to the availability of *a priori* information and the performance increased monotonically with increasing temporal resolution. Despite the fact that the HIT - FA rates of both EBM systems almost continuously increased with increasing window durations for all the noise types, the STOI improvement showed a plateau for window durations between $32 - 64\,\mathrm{ms}$, and the performance was lower for shorter and longer window durations. Considering the ICRA-5 noise, there was a considerable improvement in the HIT - FA rates when increasing the window duration from $16\,\mathrm{ms}$ to $32\,\mathrm{ms}$, which also led to a larger STOI improvement.

Overall, the EBM system with the SVM-based spectro-temporal integration stage produced substantially higher HIT - FA rates, which was also reflected in larger STOI improvements. In addition, the SVM-based integration of contextual information seemed to reduce the required window size. This was most noticeable for the PSAM 8-Hz noise, for which the EBM-GMM system with a window duration of $128\,\mathrm{ms}$, required to resolve a full period of $8\,\mathrm{Hz}$, produced the largest STOI improvements. The same performance was obtained with the EBM with the spectro-temporal integration stage using a window size of $32\,\mathrm{ms}$.

4.2. Comparison with noise reduction algorithm

Inspired by the analysis presented in [8], Fig. 3 shows the sentence-based STOI predictions for the unprocessed noisy speech in relation to the measured STOI improvement for the following three systems: a) the EBM with the spectro-temporal integration stage, b) the log-MMSE noise reduction algorithm and c) the IBM. In addition, a least-squares fit is shown for each noise type. Based on the analysis in the previous section, all algorithms operated on a window size of $32\,\mathrm{ms}$.

As expected, the IBM-based system produced the largest STOI improvements across all noise types. Also the EBM system improved the predicted speech intelligibility, in particular for conditions where the STOI values of the noisy speech were below 0.7. Whereas the STOI improvements were moderate for the IRCA-1 noise and the PSAM 8-Hz, a larger benefit was observed for the ICRA-5 noise and the tree noise.

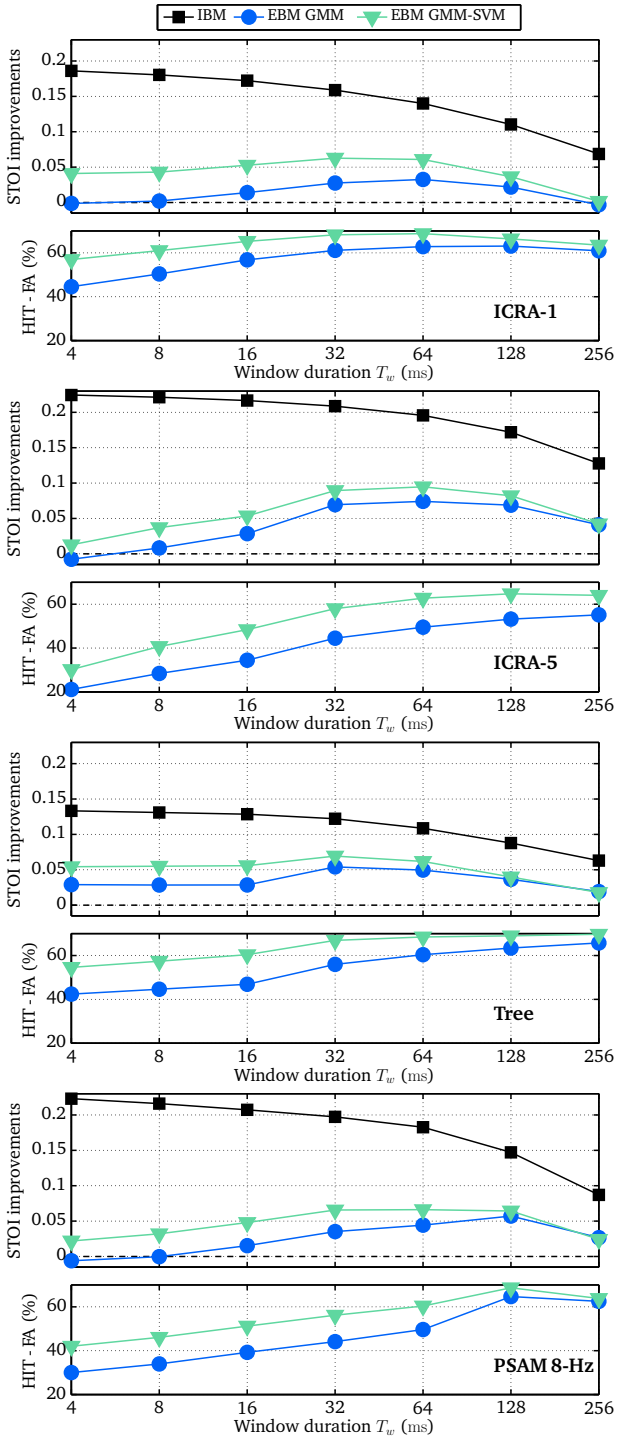

Figure 2: STOI improvements for the IBM and two EBM systems along with their corresponding HIT - FA rates averaged across all sentences and SNRs. The results are shown separately for each of the four noise types.

The log-MMSE-based noise reduction system showed minor improvement for the ICRA-1 noise, presumably because the stationary background noise could be reasonably well estimated. However, in case of the other non-stationary noises, it appeared that the rapid fluctuations could not be predicted by

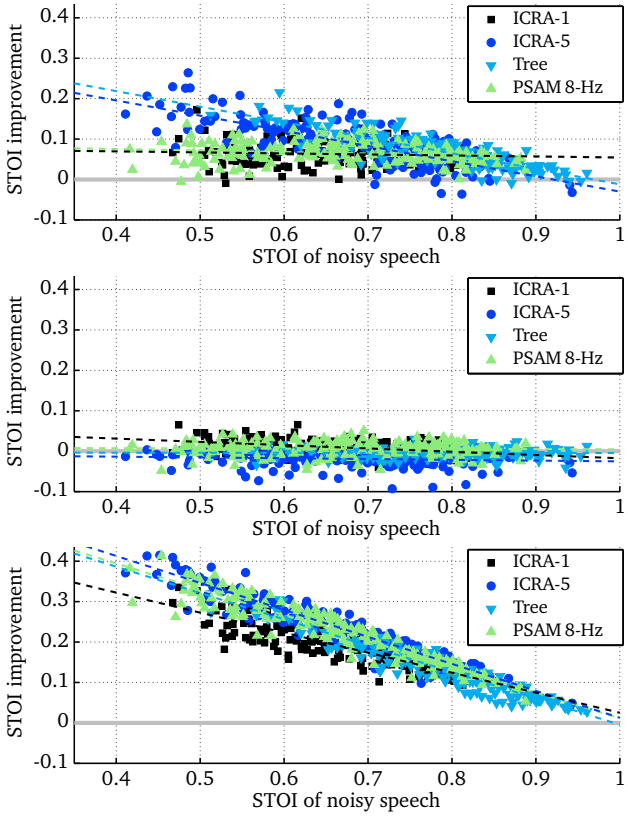

Figure 3: STOI predictions for the EBM including the spectro-temporal integration stage (top panel), log-MMSE noise reduction (middle panel) and IBM processing (bottom panel).

the noise estimation algorithm. As a consequence, the predicted intelligibility improvements were around zero or even negative, which is in line with previous studies [1, 2, 8]

5. Discussion and conclusion

The choice of a window duration in modulation-based speech segregation constitutes a trade-off between the ability to resolve low-frequency modulations and the temporal resolution with which the segregation system can manipulate individual T-F units. This choice is only moderately affected by the modulation content of the interfering noise. In general, a window size of 32 ms seems to represent a good compromise. It is conceivable that the modulation analysis could be performed at multiple time constants, as implemented in [26], and that the decision about speech and noise activity is combined across various decision streams based on different time constants.

The spectro-temporal integration stage effectively improves the ability of the segregation system to analyze low-frequency modulations by combining contextual knowledge about the speech presence probability across neighboring T-F units, thereby reducing the required window duration. However, a high performance in terms of the frequently-used performance metric, the HIT - FA rate, does not necessarily lead to improvements in predicted speech intelligibility, if the T-F manipulation is not performed with a sufficiently high temporal resolution. Finally, the segregation system has been evaluated using a technical performance measure and model predictions. The next step is to confirm these findings with behavioral listening tests.

6. References

[1] Y. Hu and P. C. Loizou, "A comparative intelligibility study of single-microphone noise reduction algorithms," *J. Acoust. Soc. Amer.*, vol. 122, no. 3, pp. 1777–1786, 2007.

[2] G. Hilkhuysen, N. Gaubitch, M. Brookes, and M. Huckvale, "Effects of noise suppression on intelligibility: Dependency on signal-to-noise ratios," *J. Acoust. Soc. Amer.*, vol. 131, no. 1, pp. 531–539, 2012.

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

[4] D. L.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.

[5] U. Kjems, J. B. Boldt, M. S. Pedersen, T. Lunner, and D. L. 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.

[6] N. Li and P. C. Loizou, "Effect of spectral resolution on the intelligibility of ideal binary masked speech," *J. Acoust. Soc. Amer.*, vol. 123, no. 4, pp. EL59–EL64, 2008.

[7] K. Han and D. L. Wang, "An SVM based classification approach to speech separation," in *Proc. ICASSP*, 2011, pp. 4632–4635.

[8] S. Gonzalez and M. Brookes, "Mask-based enhancement for very low quality speech," in *Proc. ICASSP*, 2014, pp. 7029–7033.

[9] 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.

[10] T. May and T. Dau, "Environment-aware ideal binary mask estimation using monaural cues," in *Proc. WASPAA*, 2013, pp. 1–4.

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

[12] 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.

[13] R. Drullman, J. M. Festen, and R. Plomp, "Effect of temporal envelope smearing on speech preception," *J. Acoust. Soc. Amer.*, vol. 95, no. 2, pp. 1053–1064, 1994.

[14] 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.

[15] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. 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, 2011.

[16] S. D. Ewert and T. Dau, "Characterizing frequency selectivity for envelope fluctuations," *J. Acoust. Soc. Amer.*, vol. 108, no. 3, pp. 1181–1196, 2000.

[17] D. L. Wang and G. J. Brown, Eds., *Computational Auditory Scene Analysis: Principles, Algorithms and Applications*. Wiley/IEEE Press, 2006.

[18] G. J. Brown and M. Cooke, "Computational auditory scene analysis," *Comp. Speech and Lang.*, vol. 8, no. 4, pp. 297–336, 1994.

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

[21] T. Houtgast and H. J. M. Steeneken, "The modulation transfer function in room acoustics as a predictor of speech intelligibility," *Acustica*, pp. 28–66, 1973.

[22] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, "DARPA TIMIT Acoustic-phonetic continuous speech corpus CD-ROM," *National Inst. Standards and Technol. (NIST)*, 1993.

[23] 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.

[24] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 33, no. 2, pp. 443–445, 1985.

[25] T. Gerkmann and R. C. Hendriks, "Unbiased MMSE-based noise power estimation with low complexity and low tracking delay," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 4, pp. 1383–1393, 2012.

[26] S. Jørgensen, S. D. Ewert, and T. Dau, "A multi-resolution envelope-power based model for speech intelligibility," *J. Acoust. Soc. Amer.*, vol. 134, no. 1, pp. 1–11, 2013.