Computational speech segregation based on an auditory-inspired modulation analysis

Tobias May[a] and Torsten Dau
Centre for Applied Hearing Research, Department of Electrical Engineering, Technical University of Denmark, DK-2800 Kongens Lyngby, Denmark

(Received 5 June 2014; revised 21 August 2014; accepted 3 November 2014)

A monaural speech segregation system is presented that estimates the ideal binary mask from noisy speech based on the supervised learning of amplitude modulation spectrogram (AMS) features. Instead of using linearly scaled modulation filters with constant absolute bandwidth, an auditory-inspired modulation filterbank with logarithmically scaled filters is employed. To reduce the dependency of the AMS features on the overall background noise level, a feature normalization stage is applied. In addition, a spectro-temporal integration stage is incorporated in order to exploit the context information about speech activity present in neighboring time-frequency units. In order to evaluate the generalization performance of the system to unseen acoustic conditions, the speech segregation system is trained with a limited set of low signal-to-noise ratio (SNR) conditions, but tested over a wide range of SNRs up to 20 dB. A systematic evaluation of the system demonstrates that auditory-inspired modulation processing can substantially improve the mask estimation accuracy in the presence of stationary and fluctuating interferers.
[http://dx.doi.org/10.1121/1.4901711]

PACS number(s): 43.72.Ar, 43.72.Dv [MAH] Pages: 3350–3359

I. INTRODUCTION

One of the most striking abilities of the human auditory system is the capability to focus on a desired target source and to segregate it from interfering background noise. Despite substantial progress in the field of computational auditory scene analysis (CASA) over the past decades, machine-based approaches that attempt to replicate human speech recognition abilities are still far away from being as robust as humans against the detrimental influence of competing sources and interfering noise. Even when considering a very restricted task, e.g., consonant or phoneme recognition, there is still a tremendous difference of ∼10–15 dB in performance when comparing machine-based recognition with the scores obtained by human listeners (e.g., Sroka and Braida, 2005; Meyer *et al.*, 2011).

Assuming *a priori* knowledge about the energy of the target source and all interfering sources in individual time-frequency (T-F) units, the concept of the ideal binary mask (IBM) has been introduced, where the T-F representation of noisy speech is classified into either target-dominated or interference-dominated T-F units (Wang, 2005). This classification is commonly derived by comparing the signal-to-noise ratio (SNR) in individual T-F units to a local criterion (LC). T-F units with an SNR above the predefined LC threshold are considered reliable and subsequently labeled as 1. All remaining T-F units are assumed to be dominated by noise and therefore labeled as 0. The resulting IBM can be interpreted as the ideal segregation and many studies have shown its potential for a wide range of applications, including speech intelligibility in noise (Brungart *et al.*, 2006; Li and Loizou, 2008; Kjems *et al.*, 2009; Wang *et al.*, 2009), automatic speech recognition (Cooke *et al.*, 2001), and speaker identification (May *et al.*, 2012a,b). However, the IBM is not available in practice and, therefore, its estimation in realistic scenarios is one of the key challenges of CASA, e.g., in connection to applications in hearing aids and communication devices.

Inspired by concepts of human auditory signal processing, several basic strategies of auditory grouping have been suggested to be involved in the segregation of speech from competing sources, e.g., based on the proximity in time and frequency, common onsets and offsets, as well as common amplitude modulation of a given source. Due to the increasing availability of computational resources and modern machine learning algorithms that are capable of dealing with high-dimensional data, recent studies have presented speech segregation systems that are based on various features, ranging between 51 and 90 dimensions (Han and Wang, 2012; Healy *et al.*, 2013; Wang and Wang, 2013; Wang *et al.*, 2013).

One drawback of using such a high-dimensional feature space is that the actual contribution of individual features is difficult to assess. The observation that an increasing number of studies have considered large feature sets may imply that the extraction of only one or a few feature types is not sufficient to provide reasonable speech segregation performance. However, it is possible that the way in which the individual features are computed limits the overall performance because important perceptual attributes relevant for speech segregation may not be captured appropriately. For example, several previous studies have employed amplitude modulation spectrogram (AMS) features with linearly scaled modulation filters (Kollmeier and Koch, 1994; Tchorz and Kollmeier, 2003; Kim *et al.*, 2009; Han and Wang, 2012; Healy *et al.*, 2013; May and Dau, 2013; Wang *et al.*, 2013;

[a]Author to whom correspondence should be addressed. Electronic mail: tobmay@elektro.dtu.dk

May and Dau, 2014; May and Gerkmann, 2014), which is not consistent with psychoacoustic data on modulation detection and masking in humans (Bacon and Grantham, 1989; Houtgast, 1989; Dau *et al.*, 1997a,b; Ewert and Dau, 2000). As demonstrated by Ewert and Dau (2000), the processing of envelope fluctuations can be described effectively by a second-order bandpass filterbank with logarithmically scaled modulation filters. Such a processing has recently also been successful in speech intelligibility prediction studies (Jørgensen and Dau, 2011; Jørgensen *et al.*, 2013), computational scene analysis (Christiansen *et al.*, 2014), and sound textures synthesis (McDermott and Simoncelli, 2011). Similar processing based on auditory coding principles might also be advantageous in computational speech segregation, but this has not yet been examined. Thus, investigating the role of individual grouping principles and specific auditory features in computational speech segregation would allow the analysis and verification of different feature implementations.

Computational speech segregation systems typically do not consider aspects of spectro-temporal *integration* of information in their decision stage. Apart from the use of so-called *delta features*, which attempt to capture feature variations across adjacent time and frequency units, the IBM is usually independently estimated in each T-F unit (Kim *et al.*, 2009; Han and Wang, 2012; Wang *et al.*, 2013). However, speech-dominant T-F units tend to occupy neighboring T-F units, resulting in so-called *glimpses* (Cooke, 2005, 2006). Accordingly, if speech activity is detected in one particular T-F unit, it is likely that speech information is also present in neighboring T-F units. Recently, it has been shown that considering the spectro-temporal *context* of the *a posteriori* probability of speech and noise presence at the classification stage can substantially improve the overall speech segregation performance without increasing the dimensionality of the feature space (May and Dau, 2013). Similarly, Healy *et al.* (2013) reported improvements in terms of speech segregation performance when considering the context of neighboring T-F units. By exploiting the feature context in the classification stage, the amount of integration across time and/or frequency could be directly specified by changing the size of the spectro-temporal integration window. However, the effect of the shape and the dimension of the spectro-temporal integration window on speech segregation performance has not yet been determined.

Another challenge in machine learning is the ability to generalize to acoustic conditions that have not been included in the training stage. Typically, this problem has been avoided by evaluating the speech segregation system only at those SNRs that have also been used during training (Kim *et al.*, 2009; Han and Wang, 2012; Healy *et al.*, 2013; May and Dau, 2013). Because it is impossible to train the system on all possible acoustic conditions that will be used for testing, it is important that the segregation system is functional over a wide range of SNRs, while the training has been limited to a restricted set of SNRs.

In the present study, the ideal segregation of noisy speech, as represented by the IBM, was estimated by only exploiting AMS features. In contrast to previous studies that employed linearly scaled modulation filters, an auditory-inspired modulation filterbank with logarithmically scaled modulation filters was used here and its influence on speech segregation performance was investigated. Moreover, the influence of spectro-temporal integration on speech segregation was analyzed by combining information present in neighboring T-F units. Specifically, the size and the shape of the spectro-temporal integration window in the classification stage were varied and analyzed in terms of their impact on speech segregation. The speech segregation system was trained with AMS features that were extracted for a limited set of low SNRs, but evaluated over a wide range of SNRs to analyze the ability of the system to generalize to unseen SNRs.

II. THE SPEECH SEGREGATION SYSTEM

The estimation of the IBM was accomplished in two stages: First, the AMS features were used to train a two-class Bayesian classifier, which estimated the *a posteriori* probability of speech and noise presence in individual T-F units. Second, these probabilities were considered across a spectro-temporal window of adjacent time and frequency units and the final mask estimation was obtained by comparing the probability of speech with the probability of noise presence. Both stages are described in more detail in the following.

A. AMS features

The noisy speech signal was sampled at a rate of 16 kHz and normalized according to its long-term root-mean-square value. Two different representations of the AMS features were compared: a "linear" representation based on linearly scaled modulation filters and a "logarithmic" representation where the modulation filters were scaled logarithmically, inspired by findings from auditory modeling.

1. Linearly scaled AMS features

The linear AMS feature representation was similar to that described in earlier studies (Tchorz and Kollmeier, 2003; Kim *et al.*, 2009; Han and Wang, 2012; Healy *et al.*, 2013; May and Dau, 2013; Wang *et al.*, 2013; May and Dau, 2014; May and Gerkmann, 2014). The noisy input was segmented into overlapping frames of 4 ms duration with a shift of 0.25 ms. Each frame was Hamming windowed and zero-padded to a length of 128 samples and a 128-point fast Fourier transform (FFT) was computed. The FFT magnitudes were multiplied by 25 bandpass filters, in the following referred to as "frequency channels," with center frequencies equally spaced on the mel-frequency scale between 80 Hz and 8000 Hz. The envelope in each frequency channel was then extracted by full-wave rectification, resulting in an auditory spectrogram-like representation.

Each frequency channel of the auditory spectrogram was further divided into overlapping segments of 32 ms duration with a shift of 16 ms. Each segment was Hamming windowed and zero-padded to a length of 256 samples and a 256-point FFT was applied to compute a modulation spectrogram for each frequency channel. Finally, the modulation spectrogram magnitudes were multiplied with 15 triangular-shaped modulation filters that were linearly spaced between 15.6 Hz and 400 Hz. Because the modulation spectrogram

had a frequency resolution of 15.6 Hz, each triangular filter contained modulation information derived from three adjacent FFT bins.

2. Logarithmically scaled AMS features

Each frequency channel of the auditory spectrogram was processed by a first-order low-pass filter with a cutoff frequency of 4 Hz and 8 second-order bandpass filters with center frequencies spaced logarithmically between 8 and 1024 Hz, altogether representing a modulation filterbank. The bandpass filters were assumed to have a constant-Q factor of 1 inspired by auditory modeling and speech intelligibility prediction studies (Ewert and Dau, 2000; Jørgensen and Dau, 2011; Jørgensen *et al.*, 2013). The absolute value of the output of each modulation filter was averaged across segments of 32 ms duration with a shift of 16 ms to produce the final set of nine logarithmically scaled AMS features for each frequency channel.

3. Normalization

Machine-learning-based segregation systems are typically trained with features that are extracted for a specific acoustic scenario, e.g., for a particular set of SNRs that were included in the training stage. The problem with these systems is that performance rapidly deteriorates as soon as a mismatch occurs between the acoustic conditions used for training and those used for testing. To alleviate the influence of the overall signal level on the AMS feature distribution, a normalization strategy was employed in the present study with the aim of improving the robustness of the system to mismatches between the SNRs in the training and the testing conditions. More specifically, the temporal envelope of the output of each frequency channel was normalized by its median prior to extracting the AMS features. The subband envelope signal is distributed between zero and an upper limit, leading to an asymmetric and skewed distribution. Therefore, a median-based normalization was chosen here and its influence is described in Sec. IV B.

B. Segregation stage

The segregation stage consisted of a Gaussian mixture model (GMM) classifier that was trained for each individual frequency channel, representing the AMS feature distributions of speech and noise-dominant T-F units (Kim *et al.*, 2009; May and Dau, 2013). Given the trained GMM models for speech and noise, denoted by λ_1 and λ_0, as well as the AMS feature vector, $\mathbf{X}(t, f)$, for a given time frame, t, and frequency channel, f, the *a posteriori* probabilities of speech and noise presence were given by

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

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

where the two *a priori* probabilities, $P(\lambda_{0,f})$ and $P(\lambda_{1,f})$, were determined by counting the number of feature vectors during training. Subsequently, the IBM was estimated by comparing the *a posteriori* probabilities of speech and noise presence for each individual T-F unit

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

C. Information integration across time and frequency

Instead of using the output of the Bayesian classifier directly to estimate the IBM according to Eq. (3), the *a posteriori* probability of speech presence, $P(\lambda_{1,f}|\mathbf{X}(t, f))$, was considered as a new feature spanning across the spectro-temporal integration window, $\mathcal{W}(t,f)$, and representing the centered T-F unit

$$\bar{\mathbf{X}}(t,f) := \{P(\lambda_{1,v}|\mathbf{X}(u,v)) : (u,v) \in \mathcal{W}(t,f)\}, \quad (4)$$

with the window function, $\mathcal{W}(t,f)$, defining the amount of spectro-temporal integration with respect to adjacent time and frequency units. The different window functions that were considered in the present study are shown in Fig. 1. Similar to Healy *et al.* (2013), this new feature vector, $\bar{\mathbf{X}}(t,f)$, was learned by a second classifier for speech- and noise-dominant T-F units. Depending on the size of the integration window, $\mathcal{W}(t,f)$, the dimensionality of this new feature vector could be quite large. Therefore, a support vector machine (SVM) classifier was employed here, capable of dealing with high-dimensional data and requiring only a little amount of training data.

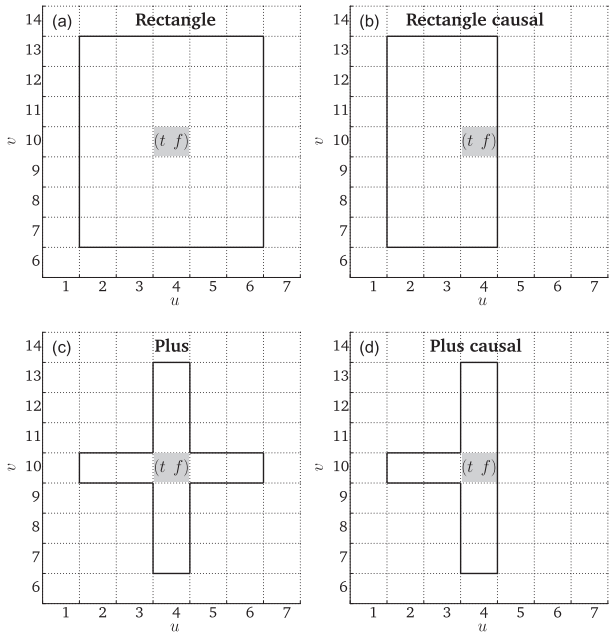

FIG. 1. Four different spectro-temporal integration functions, $\mathcal{W}(t,f)$, that are centered at $t=4$ and $f=10$, as indicated by the gray rectangle. The window size with respect to time and frequency was set to $\Delta t=5$ and $\Delta f=7$, where u and v denote the indices of the new feature space along time and frequency after spectro-temporal integration. The spectro-temporal integration was applied across all T-F units within the corresponding window shape (a) rectangle, (b) rectangle causal, (c) plus, and (d) plus causal. The causal windows only rely on past and present T-F units.

III. METHOD

A. Stimuli

Noisy speech mixtures were created by corrupting sentences from the Danish hearing in noise test (D-HINT; Nielsen and Dau, 2011) with seven different types of background noise. The sentences were spoken by a male talker at a conversational speech rate and sentences were down-sampled from 44.1 kHz to 16 kHz prior to processing. The following maskers were used: two types of speech-shaped noise (stationary ICRA1-noise and non-stationary, speech-modulated ICRA7-noise; Dreschler *et al.*, 2001), 8-Hz amplitude-modulated pink noise, non-stationary traffic noise with strong low-frequency modulations, classical music with strong harmonic components taken from Büchler (2002), as well as the more stationary destroyer noise and the highly fluctuating factory noise selected from the NOISEX database (Varga and Steeneken, 1993). These seven noise types and their overall duration are listed in Table I. Recently, it was shown that the amount of spectro-temporal noise variations that occur during training and testing determines the achievable segregation performance (May and Dau, 2014). Unlike in other studies where the same noise file was used for training and testing (Kim *et al.*, 2009; Healy *et al.*, 2013), the speech and the noise corpora were split in two halves of equal size in the present study. This ensured that there was no overlap between the signals used for training and evaluation. For each noisy speech mixture, a different randomly selected segment of the corresponding noise type was used. The noise was switched on 150 ms before the speech and was switched off 150 ms after the speech offset. Noisy speech mixtures had an average duration of 1.84 s.

B. Model training

The GMM-based segregation system was trained with 180 randomly selected sentences from the D-HINT corpus that were corrupted with one of the seven noise types at three different SNRs (−5, 0, and 5 dB SNR). The segregation models were separately trained for each noise type. To properly train the speech and the noise models, λ_1 and λ_0, the respective AMS features were selected by comparing the *a priori* SNR in individual T-F units with LC = −5 dB, which was used for all frequency channels. This choice was motivated by previous studies (Han and Wang, 2012; May and Dau, 2013). All GMM classifiers were restricted to only 16 Gaussian components with full covariance matrices to avoid over-fitting. The corresponding GMM parameters were initialized by 15 iterations of the *k*-means clustering algorithm and further refined using 5 iterations of the expectation-maximization algorithm. The spectro-temporal integration stage was based on linear SVMs (Chang and Lin, 2001) and the following two-step training procedure was performed. First, an SVM classifier was trained for each frequency channel with a small set of ten sentences mixed with each of the seven noise types at −5, 0, and 5 dB SNR. In the second step, a re-thresholding procedure was applied according to Han and Wang (2012), whereby new SVM decision thresholds were obtained that maximized the hit rate (HIT; percentage of correctly classified speech-dominant T-F units) minus the false alarm rate (FA; percentage of erroneously identified noise-dominant T-F units) using a validation set of ten sentences mixed with each of the seven noise types at −5, 0, and 5 dB SNR. Unlike the GMM-based segregation models that were separately trained for each noise type, the SVM-based integration stage was trained on all background noises. For a given spectro-temporal integration window, $\mathcal{W}(t,f)$, a separate SVM classifier was trained.

TABLE I. Types of background noises.

Noise type	Description	Duration
ICRA1	Stationary speech-shaped noise	120 s
ICRA7	Non-stationary six persons babble	1200 s
PSAM 8 Hz	Sinusoidal amplitude-modulated pink noise	∞ s
Traffic	Cars, trams, trucks, and trains passing by	360 s
Music	Classical music	570 s
Destroyer	Destroyer operations room noise	235 s
Factory	Factory floor noise inside a car factory	235 s

C. Model evaluation

The segregation system was evaluated with 60 randomly selected sentences from the D-HINT corpus that were different from those used during the training stage. Each sentence was mixed with the seven different background noises listed in Table I at −5, 0, 5, 10, 15, and 20 dB SNR.

In order to evaluate the speech segregation performance, the percentage of correctly identified T-F units was computed by comparing the estimated binary mask with the IBM. Specifically, HIT-FA was reported because this metric has been shown to correlate with human speech intelligibility (Kim *et al.*, 2009).

D. Analysis of modulation processing in computational speech segregation

The contribution of individual modulation filters to the overall speech segregation performance was analyzed. Therefore, the segregation system was first trained with an AMS feature vector that only consisted of the first (lowest) modulation filter. In the next steps, the feature vector was incrementally extended by incorporating higher modulation filters, until the full bank of modulation filters was used for training. The different AMS feature implementations and the effect of the normalization strategy on speech segregation performance were investigated.

Furthermore, the contribution of the spectro-temporal integration stage was considered. First, the effect of the size of the rectangular integration window, $\mathcal{W}$, on the speech segregation performance was analyzed by systematically changing the window dimensions with respect to time (Δt) and frequency (Δf) units: $\Delta t \in [1, 3, 5, 7]$ and $\Delta f \in [1, 3, 5, 7, 9, 11]$. Second, for a fixed window size, the influence of the following four window shapes was studied: "rectangle," "plus," "rectangle causal," and "plus causal." An example of the four window functions is shown in Fig. 1. The causal window shapes were limited to only rely on past and present T-F units, whereas all future T-F units with respect to the most central T-

F unit were removed. These causal window shapes have been considered to be particularly relevant for real-time applications, such as hearing aids. A final comparison was performed to quantify the overall contribution of spectro-temporal integration to computational speech segregation.

IV. RESULTS

A. Contribution of individual modulation filters

The contribution of individual modulation filters to the overall segregation performance is shown in Fig. 2 for noisy speech at −5 dB SNR. Figure 2(a) shows HIT-FA obtained with the linear ("lin AMS"; open symbols) and the logarithmically scaled ("log AMS," filled symbols) AMS features, respectively, as a function of the exploited modulation frequency range. In addition, the gray symbols represent results obtained with the linear AMS features when the upper modulation frequency range was extended from 400 Hz to 1025 Hz, i.e., employing 40 linearly scaled modulation filters. All segregation systems were first trained with an AMS feature vector that only contained the first modulation filter. The AMS feature vector was then gradually extended by higher modulation filters until the complete modulation filterbank was used.

The performance increase obtained with the linear AMS feature representation was found to saturate beyond the first four AMS features. The difference in performance when using the first 4 (33.5%) or 15 linear AMS features (35.4%) was only 2%. Even when using as many as 40 linear AMS features that exploit modulation frequencies up to 1025 Hz, only a minor improvement was achieved. The reason for this can be seen in Fig. 2(b), where the HIT rates and the FA rates are shown separately. The HIT rate, i.e., the percentage of correctly identified speech-dominant T-F units, increased with increasing modulation frequency range. However, the FA rate also increased at a similar rate. Thus, including higher-frequency modulation filters did not lead to an overall increase of the segregation performance. Since the linear AMS features were derived using an FFT-based modulation analysis with a frequency resolution of 15.6 Hz, each modulation filter was based on three FFT bins. Consequently, the linear AMS features are very frequency selective and, therefore, prone to the impact of noise. The classification results suggest that the linear AMS feature representation does not reliably capture speech or noise-specific modulations at high modulation frequencies.

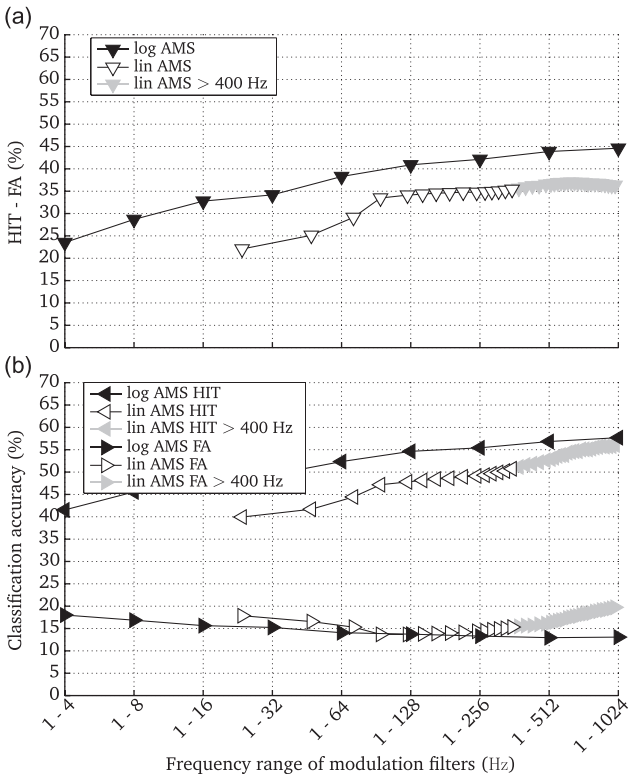

FIG. 2. Classification results of individual T-F units for noisy speech at −5 dB SNR as a function of the exploited modulation frequency range. Results are averaged across all noise types and frequency channels. (a) presents HIT-FA, whereas the corresponding HIT and FA rates are separately shown in (b). See main text for details.

In contrast, the segregation system trained with the logarithmically scaled AMS features showed a substantial performance improvement when higher frequency modulation filters were sequentially integrated. While the HIT rate increased systematically, the FA rate slightly decreased with increasing modulation frequency range, presumably because the increased bandwidth at higher modulation frequencies provided more robust modulation features. The benefit of analyzing modulation frequencies up to 1024 Hz indicates that it might not be sufficient to only consider modulation frequencies that are primarily relevant for speech. At the same time, it seems important to properly analyze noise-specific characteristics present at higher modulation frequencies, allowing the segregation system to identify the background noise and to distinguish it from the speech components.

Moreover, the logarithmically scaled modulation filters allow the analysis of low-frequency modulations, which are known to be important for speech perception in stationary noise (e.g., Drullman *et al.*, 1994). Although the frame size of 32 ms only resolves full periods of modulation frequencies >30 Hz, it is possible to analyze a fraction of modulation frequencies <30 Hz in one single frame, and this information can be successfully exploited by the segregation system. When considering only the first four logarithmically scaled modulation filters up to 32 Hz, a segregation performance of ∼35% was achieved. This corresponds to the maximum performance obtained with all 15 linear AMS features, and demonstrates the significance of low-frequency modulations for computational speech segregation. In addition, the systematic performance improvement with sequentially added modulation filters indicates the importance of modulation frequency selectivity.

The contribution of individual modulation filters to computational speech segregation was found to depend on the modulation characteristics of the interfering background noise. Figure 3 shows the HIT-FA rates obtained with the two AMS feature representations for the individual types of background noise. It can be seen that, in all noise conditions, the logarithmic AMS representation (filled symbols) achieved a higher classification performance than the linear AMS feature representation (open symbols). In general, the ability of the logarithmic AMS features to analyze lower

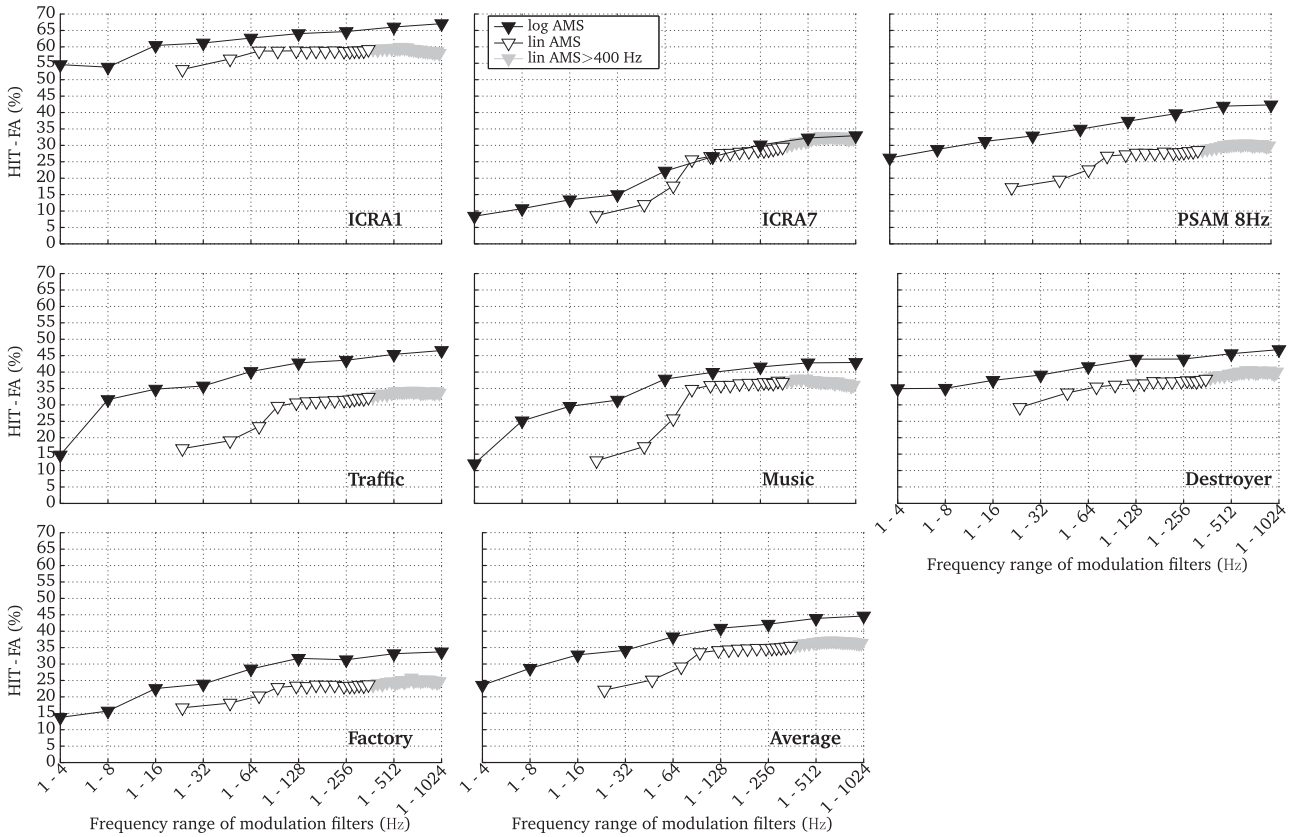

FIG. 3. Classification results of individual T-F units for noisy speech at −5 dB SNR as a function of the exploited modulation frequency range. Results are averaged across all frequency channels and shown separately for each background noise.

modulation frequencies seemed particularly beneficial when the background noise contained strong low-frequency modulations, such as in the case of traffic noise (middle left panel) and classic music (central panel). Moreover, the benefit of extracting higher-frequency modulations of the speech is evident for ICRA1 noise (upper left panel) and 8-Hz amplitude-modulated pink noise (upper right panel), which are stationary or contain only low-frequency modulations. Considering the speech-modulated ICRA7 noise (upper middle panel), the modulation characteristic of the noise is very similar to that of speech, which results in only a small difference between the two AMS feature representations. Hence, it seems that the benefit of the logarithmic AMS features is largest when the information in a particular modulation frequency range is dominated by either the speech or the noise.

B. Effects of modulation analysis, SNR, and noise type

Figure 4 shows the speech segregation performance obtained with the linear ("lin AMS"; open symbols) and the logarithmically scaled AMS features ("log AMS", filled symbols), respectively, as a function of the SNR [Fig. 4(a)], and for the different types of background noises [Fig. 4(b)]. For the three lowest SNRs (-5, 0, and 5 dB), the acoustic conditions matched those that were used during training. Therefore, segregation performance can be assumed to be mainly influenced by the two AMS feature representations. At the higher SNRs (10, 15, and 20 dB), the segregation system was evaluated with noisy speech for which it was not trained. Therefore, the influence of the normalization scheme was most pronounced at these SNRs.

The logarithmically scaled AMS features produced a considerably higher classification accuracy than the linear AMS features, although only 9 rather than 15 modulation filters were exploited. This performance difference was consistent across a wide range of SNRs and was ∼10%. Without normalization, the performance of both AMS feature representations decreased as soon as the segregation system was evaluated at SNRs that were outside the training range (i.e., >5 dB SNR), although the segregation task should become easier at higher SNRs. This decrease resulted from the SNR-dependent behavior of the AMS features. The fact that the performance for the linear AMS feature representation at 20 dB SNR was worse than at −5 dB SNR indicates that the linear AMS features are very sensitive to mismatches between the acoustic conditions used for training and testing. In contrast, when applying the normalization, the segregation performance continuously increased with increasing SNR, despite the fact that the system was only trained at lower SNRs. In other words, the median-based normalization strategy enables the segregation system to generalize to unseen SNRs. In addition, the median-based normalization also improves the segregation performance in the matched condition.

The classification performance averaged across all SNRs [Fig. 4(b)] was found to strongly depend on the characteristics of the background noise. For examples, for the stationary ICRA1 noise, the segregation performance for the logarithmic AMS feature representation was very high and >70%. In

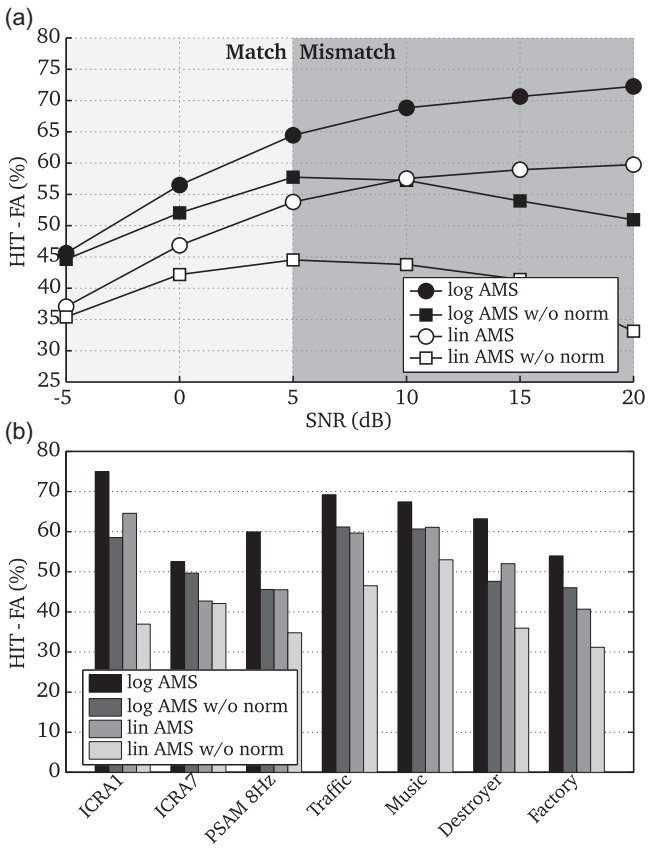

FIG. 4. Classification results of individual T-F units for various AMS feature implementations as a function of (a) the SNR averaged across all background noises and frequency channels and (b) the background noise averaged across all SNRs and frequency channels.

the presence of the non-stationary ICRA7 noise (six persons babble), the modulation characteristics of speech and noise were more similar, such that the performance decreased by ∼20%. Considering the 8-Hz amplitude-modulated pink noise, there is a substantial performance difference between the linear and the logarithmic AMS feature representations, suggesting that the 8-Hz component of the masker cannot be sufficiently captured by the linear AMS features.

C. Effect of spectro-temporal integration

The effect of a rectangular spectro-temporal integration window on speech segregation performance is shown in Fig. 5 using the logarithmically scaled AMS feature representation. The relative performance improvement (in %) due to spectro-temporal integration is shown as a function of the window size with respect to the dimensions time $\Delta t \in [1, 3, 5, 7]$, shown on the ordinate, and frequency $\Delta f \in [1, 3, 5, 7, 9, 11]$, indicated on the abscissa.

An improvement in segregation performance of up to 12% was obtained when the context of the *a posteriori* probability of speech presence was exploited across adjacent time and frequency units. Across-frequency integration appears to be more important than extending the integration window across adjacent time frames. When using a rectangular window over $\Delta t = 3$ adjacent time frames and $\Delta t = 9$ neighboring frequency channels, as indicated by the black

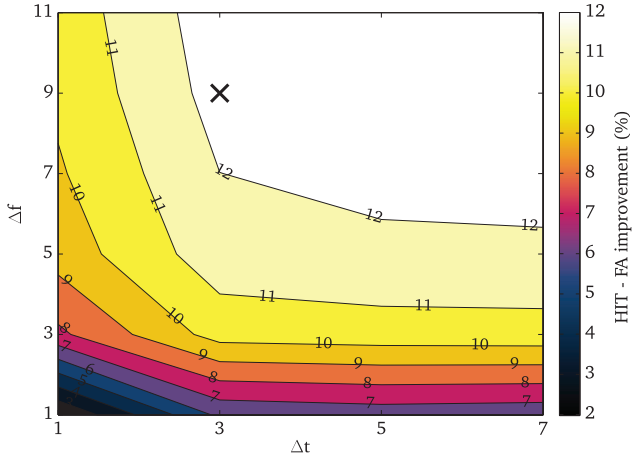

FIG. 5. (Color online) The effect of a rectangular spectro-temporal integration window on speech segregation performance for the logarithmically scaled AMS features. The relative performance improvement is averaged across three SNRs (−5, 0, and 5 dB SNR) and all noise types.

cross, the overall performance improvement saturated and no additional improvement was obtained when further extending the window across time or frequency.

The influence of the four different window shapes shown in Fig. 1 on speech segregation performance was investigated by keeping the overall window size fixed to $\Delta t = 3$ adjacent time frames and $\Delta f = 9$ neighboring frequency channels. The corresponding classification accuracies are averaged across three SNRs (−5, 0, and 5 dB SNR) and are shown in Table II for both, linear and logarithmic, AMS feature representations. In addition, for each window shape, the average number of T-F units involved in the spectro-temporal integration is specified in Table II.

Although the rectangular window contains more than twice the T-F units compared to the plus-shaped window, the classification accuracies were quite similar. This difference is particularly small when using the logarithmic AMS feature representation. Thus, for a given time frame, considering the probability of speech presence in neighboring frequency channels seems most important for the identification of speech-dominant T-F units. The evaluation of across-frequency information at past frames appears to be less important, presumably due to the sparse distribution of speech-dominant T-F units in the presence of noise. Moreover, no substantial performance decrease was observed when using causal window shapes that only rely on past and present T-F units, which is an important aspect for real-time applications.

TABLE II. Classification accuracy of individual T-F units measured in HIT-FA % as a function of the window shape for $\Delta t = 3$ and $\Delta f = 9$.

Spectro-temporal integration		AMS features	
Window shape	Number of T-F units	Linearly scaled AMS	Logarithmically scaled AMS
Rectangle	24.6	63.0	67.5
Rectangle causal	16.4	60.0	67.2
Plus	10.2	60.8	66.8
Plus causal	9.2	59.3	66.8

Figure 6 shows the effect of the spectro-temporal integration stage as a function of the SNR [Fig. 6(a)] and for the different background noises [Fig. 6(b)]. The integration stage was based on a causal plus-shaped integration window that spans three adjacent time frames and nine frequency channels. A performance increase compared to the results shown in Fig. 4 (without integration) for both the linear and logarithmic AMS feature representations was obtained. This improvement in terms of speech segregation performance was close to 13% and most prominent at low SNRs. Moreover, the integration stage seems particularly beneficial for non-stationary background noises, e.g., for the ICRA7 and the factory noise, as shown in Fig. 6(b).

D. Visualization of modulation-based speech segregation

An illustration of the modulation-based speech segregation is shown in Fig. 7. The IBM is presented in Fig. 7(a) for speech mixed with factory noise at 0 dB SNR. It can be seen that the distribution of speech-dominant T-F units in the IBM is quite compact. Figures 7(b) and 7(c) present the estimated IBMs using the linear and the logarithmically scaled AMS features, respectively, with the benefit of spectro-temporal integration represented in Figs. 7(d) and 7(e), respectively. In addition to the estimated IBM patterns, the average HIT-FA rates for each frequency channel are provided in the right part of each panel. When the IBM was estimated on the basis of individual T-F units [Figs. 7(b) and 7(c)], following Eq. (3), some speech-dominated T-F units, particularly at higher-frequency channels, were erroneously classified as background noise. In addition, several noise-dominated T-F units were classified as being speech dominated. In general, the logarithmic AMS features achieved a higher classification accuracy compared to the linear AMS features, especially at higher frequencies. The spectro-temporal integration stage in Figs. 7(d) and 7(e) reduced theses outliers and, more importantly, recovered many target-dominant T-F units at higher-frequency channels. Still, the linear AMS features missed many speech-dominant T-F units at higher-frequency channels, which cannot be recovered by the spectro-temporal integration stage.

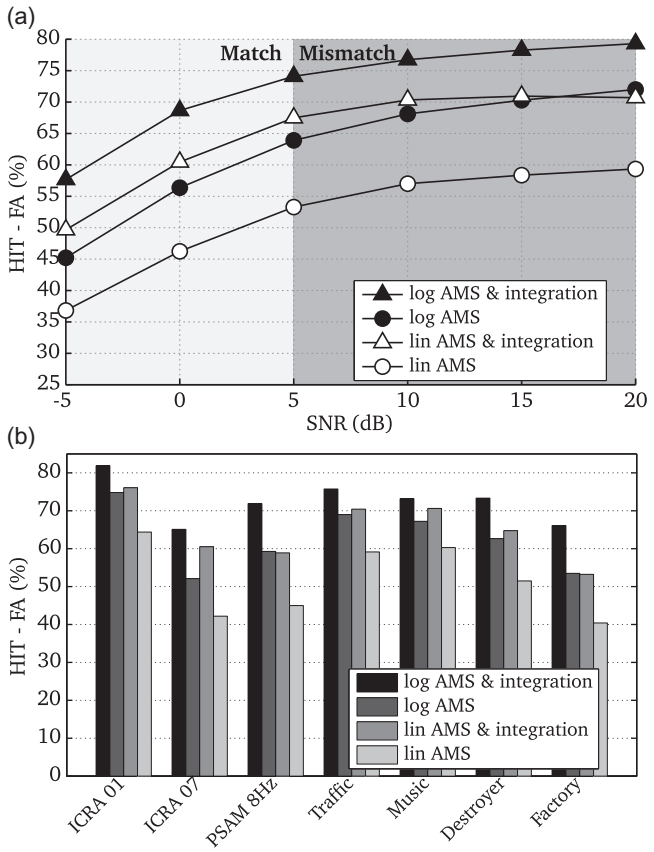

FIG. 6. Classification results of individual T-F units for various segregation systems as a function of (a) the SNR averaged across all background noises and frequency channels and (b) the background noise averaged across all SNRs and frequency channels.

V. DISCUSSION AND CONCLUSION

This study addressed the problem of segregating speech from a noisy mixture by the supervised learning of AMS features. It was shown that AMS features based on an auditory-inspired modulation filterbank with logarithmically scaled modulation filters were more robust in detecting speech-dominated T-F units than an AMS feature representation based on a linear frequency scale. The performance increase obtained with the logarithmic representation was mainly caused by the ability of the system to analyze low-frequency modulations and an increased robustness against interfering noise due to the increased bandwidth of the higher-frequency modulation filters. It was demonstrated that an accurate estimation of the IBM can be obtained by only exploiting a limited set of nine modulation features per frequency channel. This suggests that auditory-based features may be more informative and more robust than higher-dimensional variants of AMS features.

The classification results also illustrated that computational segregation systems can be quite sensitive to a mismatch between training and testing conditions. In the present study, the median-based normalization scheme allowed the segregation system to function over a wide range of SNRs, despite being only trained at low SNRs. This normalization scheme represented a pragmatic choice and may not be physiologically plausible. Conceptually, this normalization can be interpreted as an automatic gain control, reducing the effect of variations in the background noise level. However, auditory-inspired adaptation processes as, for example, reflected in auditory signal processing models (e.g., Meddis *et al.*, 1990; Dau *et al.*, 1996; Dau *et al.*, 1997a; Heinz *et al.*, 2001; Zilany *et al.*, 2009; Zilany *et al.*, 2014), might be more appropriate to provide a level-invariant AMS feature representation. In addition to the robustness to *unseen levels* of background noise, the ability of computational speech segregation to generalize to *unseen types* of interference represents another important property. This has not been considered in the present study, but should be evaluated in future investigations.

The spectro-temporal integration stage in the segregation system was shown to substantially improve the accuracy of the IBM estimation by ~12%, particularly for non-stationary background noises and low SNRs. The

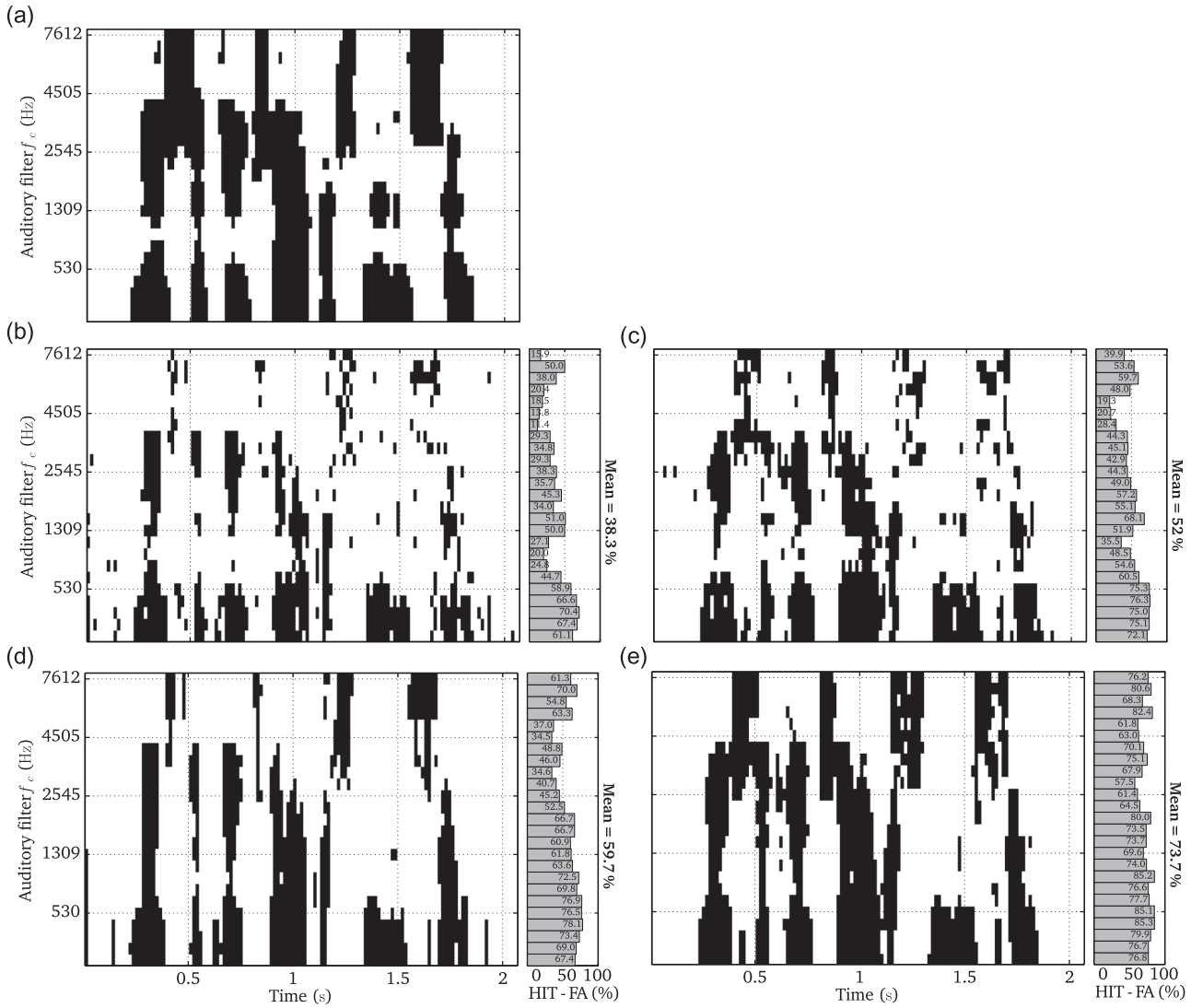

FIG. 7. IBM estimation and the frequency-dependent HIT-FA rates for an utterance mixed with factory noise at 0 dB SNR. (a) IBM, (b) estimated IBM using linear AMS features, (c) estimated IBM using logarithmic AMS features, (d) estimated IBM using linear AMS features and spectro-temporal integration, and (e) estimated IBM using logarithmic AMS features and spectro-temporal integration.

observation that across-frequency integration was more beneficial than the integration across adjacent time frames indicates that the decision about speech or noise activity has to be made with a relatively high temporal accuracy to avoid temporal smearing. This is consistent with other studies that reported that only IBMs with a high temporal resolution are capable of improving speech intelligibility (Anzalone *et al.*, 2006). It was shown in the present study that a plus-shaped, causal integration window achieved a similar performance to a rectangular, non-causal integration window that spans more than twice the T-F units. A causal integration stage implies that the method can be used in real-time applications, such as hearing aids. In addition, since the *a posteriori* speech presence probability distribution across the integration window is learned by a classifier, a lower-dimensional integration window simplifies the training procedure and will facilitate the combination of AMS features with other feature types that might further improve the distinction between speech and noise activity.

The approach presented here, indeed, is limited because only AMS features have been considered. In comparison to stationary interferences, the accuracy of estimating the IBM is reduced in the presence of noise with speech-like modulation characteristics. Thus, the combination of AMS features with other complementary features is likely to further improve the speech segregation performance. For example, the combination of AMS features with pitch (Han and Wang, 2012) or periodicity features (May and Dau, 2013) has been shown to improve the mask estimation accuracy as opposed to a system that solely relies on AMS features. Furthermore, the analysis of common onsets and offsets might be beneficial in order to organize and integrate T-F units originating from the same source across frequency (Hu and Wang, 2007). However, it seems important to study separately the robustness of the individual types of features and their ability to generalize to unseen acoustic conditions before combining them into a higher-dimensional feature vector.

Finally, the segregation system has been evaluated here by computing the difference between correctly classified speech-dominant T-F units and erroneously identified noise-dominant T-F units. Although this technical measure has been

shown to correlate with human speech intelligibility (Kim *et al.*, 2009), the next step would be to use the presented speech segregation system as a front-end for speech enhancement. Specifically, the knowledge about speech-dominant T-F units could be used to attenuate the T-F units that are dominated by the interfering background noise. Ultimately, the ability of such a speech enhancement system to improve the intelligibility of noisy speech for human listeners needs to be evaluated in corresponding behavioral listening tests.

ACKNOWLEDGMENTS

The authors would like to acknowledge the support by EU FET Grant TWO!EARS, 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.

Bacon, S. P., and Grantham, D. W. (**1989**). "Modulation masking: Effects of modulation frequency, depths, and phase," J. Acoust. Soc. Am. **85**, 2575–2580.

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

Büchler, M. C. (**2002**). "Algorithms for sound classification in hearing instruments," Ph.D. thesis, Swiss Federal Institute of Technology, Zurich, Switzerland.

Chang, C. C., and Lin, C. J. (**2001**). "LIBSVM: A library for support vector machines," Software is available at www.csie.ntu.edu.tw/∼cjlin/libsvm (Last viewed November 2014).

Christiansen, S. K., Jepsen, M. L., and Dau, T. (**2014**). "Effects of tonotopicity, adaptation, modulation tuning and temporal coherence in 'primitive' auditory stream segregation," J. Acoust. Soc. Am. **135**, 323–334.

Cooke, M. (**2005**). "Making sense of everyday speech: A glimpsing account," in *Speech Separation by Humans and Machines*, edited by P. Divenyi (Kluwer Academic, Dordrecht, The Netherlands), Chap. 21, pp. 305–314.

Cooke, M. (**2006**). "A glimpsing model of speech perception in noise," J. Acoust. Soc. Am. **119**, 1562–1573.

Cooke, M., Green, P., Josifovski, L., and Vizinho, A. (**2001**). "Robust automatic speech recognition with missing and unreliable acoustic data," Speech Commun. **34**, 267–285.

Dau, T., Kollmeier, B., and Kohlrausch, A. (**1997a**). "Modeling auditory processing of amplitude modulation. I. Detection and masking with narrow-band carriers," J. Acoust. Soc. Am. **102**, 2892–2905.

Dau, T., Kollmeier, B., and Kohlrausch, A. (**1997b**). "Modeling auditory processing of amplitude modulation. II. Spectral and temporal integration," J. Acoust. Soc. Am. **102**, 2906–2919.

Dau, T., Püschel, D., and Kohlrausch, A. (**1996**). "A quantitative model of the 'effective' signal processing in the auditory system. I. Model structure," J. Acoust. Soc. Am. **99**, 3615–3622.

Dreschler, W. A., Verschuure, H., Ludvigsen, C., and Westermann, S. (**2001**). "ICRA noises: Artificial noise signals with speech-like spectral and temporal properties for hearing instrument assessment," Audiology **40**, 148–157.

Drullman, R., Festen, J. M., and Plomp, R. (**1994**). "Effect of temporal envelope smearing on speech reception," J. Acoust. Soc. Am. **95**, 1053–1064.

Ewert, S. D., and Dau, T. (**2000**). "Characterizing frequency selectivity for envelope fluctuations," J. Acoust. Soc. Am. **108**, 1181–1196.

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

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.

Heinz, M. G., Colburn, H. S., and Carney, L. H. (**2001**). "Evaluating auditory performance limits: I. One-parameter discrimination using a computational model for the auditory nerve," Neural Comput. **13**, 2273–2316.

Houtgast, T. (**1989**). "Frequency selectivity in amplitude-modulation detection," J. Acoust. Soc. Am. **85**, 1676–1680.

Hu, G., and Wang, D. L. (**2007**). "Auditory segmentation based on onset and offset analysis," IEEE Trans. Audio, Speech, Lang. Process. **15**, 396–405.

Jørgensen, S., and Dau, T. (**2011**). "Predicting speech intelligibility based on the signal-tonoise envelope power ratio after modulation-frequency selective processing," J. Acoust. Soc. Am. **130**, 1475–1487.

Jørgensen, S., Ewert, S. D., and Dau, T. (**2013**). "A multi-resolution envelope-power based model for speech intelligibility," J. Acoust. Soc. Am. **134**, 1–11.

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. L. (**2009**). "Role of mask pattern in intelligibility of ideal binary-masked noisy speech," J. Acoust. Soc. Am. **126**, 1415–1426.

Kollmeier, B., and Koch, R. (**1994**). "Speech enhancement based on physiological and psychoacoustical models of modulation perception and binaural interaction," J. Acoust. Soc. Am. **95**, 1593–1602.

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. WASPAA* (New Paltz, NY).

May, T., and Dau, T. (**2014**). "Requirements for the evaluation of computational speech segregation systems," J. Acoust. Soc. Am. **136**, EL398–EL404.

May, T., and Gerkmann, T. (**2014**). "Generalization of supervised learning for binary mask estimation," in *Proceedings of IWAENC* (Juan les Pins, France).

May, T., van de Par, S., and Kohlrausch, A. (**2012a**). "A binaural scene analyzer for joint localization and recognition of speakers in the presence of interfering noise sources and reverberation," IEEE Trans. Audio, Speech, Lang. Process. **20**, 2016–2030.

May, T., van de Par, S., and Kohlrausch, A. (**2012b**). "Noise-robust speaker recognition combining missing data techniques and universal background modeling," IEEE Trans. Audio, Speech, Lang. Process. **20**, 108–121.

McDermott, J. H., and Simoncelli, E. P. (**2011**). "Sound texture perception via statistics of the auditory periphery: Evidence from sound synthesis," Neuron **71**, 926–940.

Meddis, R., Hewitt, M. J., and Shackleton, T. M. (**1990**). "Implementation details of a computation model of the inner hair-cell auditory-nerve synapse," J. Acoust. Soc. Am. **87**, 1813–1816.

Meyer, B. T., Brand, T., and Kollmeier, B. (**2011**). "Effect of speech-intrinsic variations on human and automatic recognition of spoken phonemes," J. Acoust. Soc. Am. **129**, 388–403.

Nielsen, J. B., and Dau, T. (**2011**). "The Danish hearing in noise test," Int. J. Audiol. **50**, 202–208.

Sroka, J. J., and Braida, L. D. (**2005**). "Human and machine consonant recognition," Speech Commun. **45**, 401–423.

Tchorz, J., and Kollmeier, B. (**2003**). "SNR estimation based on amplitude modulation analysis with applications to noise suppression," IEEE Trans. Audio, Speech, Lang. Process. **11**, 184–192.

Varga, A. P., and Steeneken, H. J. M. (**1993**). "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," Speech Commun. **12**, 247–251.

Wang, D. L. (**2005**). "On ideal binary mask as the computational goal of auditory scene analysis," in *Speech Separation by Humans and Machines*, edited by P. Divenyi (Kluwer Academic, Dordrecht, The Netherlands), Chap. 12, pp. 181–197.

Wang, Y., and Wang, D. L. (**2013**). "Towards scaling up classification-based speech separation," IEEE Trans. Audio, Speech, Lang. Process. **21**, 1381–1390.

Wang, Y., Han, K., and Wang, D. L. (**2013**). "Exploring monaural features for classificationbased speech segregation," IEEE Trans. Audio, Speech, Lang. Process. **21**, 270–279.

Wang, D. L., Kjems, U., Pedersen, M. S., and Boldt, J. B. (**2009**). "Speech intelligibility in background noise with ideal binary time-frequency masking," J. Acoust. Soc. Am. **125**, 2336–2347.

Zilany, M. S. A., Bruce, I. C., and Carney, L. H. (**2014**). "Updated parameters and expanded simulation options for a model of the auditory periphery," J. Acoust. Soc. Am. **135**, 283–286.

Zilany, M. S. A., Bruce, I. C., Nelson, P. C., and Carney, L. H. (**2009**). "A phenomenological model of the synapse between the inner hair cell and auditory nerve: Long-term adaptation with power-law dynamics," J. Acoust. Soc. Am. **126**, 2390–2412.