The impact of exploiting spectro-temporal context in computational speech segregation

Bentsen, Thomas; Kressner, Abigail Anne; Dau, Torsten; May, Tobias

Published in:
Journal of the Acoustical Society of America

Link to article, DOI:
10.1121/1.5020273

Publication date:
2018

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

Citation (APA):
Bentsen, T., Kressner, A. A., Dau, T., & May, T. (2018). The impact of exploiting spectro-temporal context in computational speech segregation. Journal of the Acoustical Society of America, 143(1), 248-259. DOI: 10.1121/1.5020273

The impact of exploiting spectro-temporal context in computational speech segregation

Thomas Bentsen,[a] Abigail A. Kressner, Torsten Dau, and Tobias May
Hearing Systems Group, Department of Electrical Engineering, Technical University of Denmark, 2800 Kongens Lyngby, Denmark

(Received 14 July 2017; revised 5 October 2017; accepted 12 December 2017; published online 19 January 2018)

Computational speech segregation aims to automatically segregate speech from interfering noise, often by employing ideal binary mask estimation. Several studies have tried to exploit contextual information in speech to improve mask estimation accuracy by using two frequently-used strategies that (1) incorporate delta features and (2) employ support vector machine (SVM) based integration. In this study, two experiments were conducted. In Experiment I, the impact of exploiting spectro-temporal context using these strategies was investigated in stationary and six-talker noise. In Experiment II, the delta features were explored in detail and tested in a setup that considered novel noise segments of the six-talker noise. Computing delta features led to higher intelligibility than employing SVM based integration and intelligibility increased with the amount of spectral information exploited via the delta features. The system did not, however, generalize well to novel segments of this noise type. Measured intelligibility was subsequently compared to extended short-term objective intelligibility, hit–false alarm rate, and the amount of mask clustering. None of these objective measures alone could account for measured intelligibility. The findings may have implications for the design of speech segregation systems, and for the selection of a cost function that correlates with intelligibility.

https://doi.org/10.1121/1.5020273

[MAH] Pages: 248–259

I. INTRODUCTION

The overall goal of computational speech segregation systems is to automatically segregate a target speech signal from interfering noise. These systems are relevant for many practical applications, e.g., as pre-processors in communication devices such as hearing aids or cochlear implants (Brungart *et al.*, 2006; Li and Loizou, 2008; Wang *et al.*, 2008) or front-ends in speech and speaker recognition systems for human-computer interfaces (Cooke *et al.*, 2001; May *et al.*, 2012a,b). One frequently-used single-channel approach, termed the ideal binary mask (IBM) technique (Wang, 2005), separates a time–frequency (T–F) representation of noisy speech into target-dominated and interference-dominated T–F units. Given *a priori* knowledge about the target and the interfering signal, the IBM is constructed by comparing the signal-to-noise ratio (SNR) in individual T–F units to a local criterion (LC). The resulting IBM is a binary matrix where T–F units with SNRs exceeding the LC are considered target-dominated and labeled one, and zero otherwise. Many studies have used IBMs to segregate a target speech signal from a noisy mixture and demonstrated large intelligibility improvements (Brungart *et al.*, 2006; Wang *et al.*, 2008; Kjems *et al.*, 2009). However, *a priori* knowledge about the target and the interfering noise is rarely available in realistic conditions, and therefore, the goal of segregation systems is to estimate the IBM based on the noisy speech signal. This challenge of obtaining an estimated IBM is typically approached by employing supervised learning strategies (Wang, 2005), which generally consist of a feature extraction front-end and a classification back-end. The front-end extracts a set of acoustic features which attempt to capture speech- and interference-specific properties. The distributions of speech and interference-dominated T–F units are then learned by a classification back-end, through an initial training stage (Kim *et al.*, 2009; Han and Wang, 2012; Healy *et al.*, 2013; May and Dau, 2014a).

When analyzing binary mask patterns, speech-dominated T–F units tend to cluster in spectro-temporal regions, forming so-called *glimpses*, and the size of these glimpses, denoted the glimpse proportion in the model by Cooke (2006), has been shown to correlate with speech intelligibility scores from normal-hearing (NH) listeners (Cooke, 2006; Barker and Cooke, 2007). Consequently, several studies have tried to exploit spectro-temporal contextual information in speech to improve the performance of computational speech segregation systems by predominantly using two strategies. One strategy is to exploit the context in the front-end by calculating so-called *delta features* (Kim *et al.*, 2009; Hu and Loizou, 2010; May and Dau, 2014b), which capture feature variations across time and frequency. Alternatively, the context can be exploited in the back-end, where the posterior probability of speech presence obtained from a first classifier can be learned by a second classifier across a spectro-temporal window of T–F units, where the amount of spectro-temporal context can be controlled by the

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

size of the window function (Han and Wang, 2012; Healy *et al.*, 2013; May and Dau, 2014a). Some studies have combined both strategies in the front-end and in the back-end (Healy *et al.*, 2013; May and Dau, 2013).

The performance of computational speech segregation systems and the effectiveness of different system configurations have primarily been evaluated based on the hit–false alarm (H–FA) rate, which calculates 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)] (Kim *et al.*, 2009; Han and Wang, 2012; Healy *et al.*, 2013; May and Dau, 2013, 2014a,b). However, it has recently been shown that speech intelligibility scores strongly depend on both the *distribution* of mask errors and the H–FA rate (Kressner and Rozell, 2015, 2016; Kressner *et al.*, 2016). Specifically, Kressner and Rozell (2015) developed a graphical model to systematically measure the influence of clustering of T–F units on the intelligibility of binary-masked speech and showed that the intelligibility was *reduced* when masks contained an increased amount of clustering among T–F units, but the same mask error rates. Thus, the applicability of the H–FA rate as the sole objective measure to optimize or evaluate computational segregation systems has come into question. However, the impact of the different spectro-temporal context-exploring strategies on the amount of clustering of T–F units, or on speech intelligibility, has not yet been analyzed.

Kim *et al.* (2009) were the first to report speech intelligibility improvements for a computational speech segregation system based on Gaussian mixture models (GMMs). They considered a high complexity GMM classifier with 256 components in the back-end for modeling the distribution of the feature vectors in a restricted setup in which the same short noise recording was used during training and testing. By using such a setup, it was possible to achieve high H–FA rates and improve speech intelligibility scores by up to 60% compared to unprocessed noisy speech for NH subjects (Kim *et al.*, 2009). A high complexity classifier is able to learn all spectro-temporal characteristics of the noise, if the same short noise recording is used during training and testing, resulting in high H–FA rates (May and Dau, 2014b) and, most likely, also the high intelligibility scores observed in Kim *et al.* (2009). The restricted setup therefore has a high potential to improve speech intelligibility and can be used to investigate the behavior of the segregation system by comparing different system configurations. The ability of segregation systems to generalize to unseen acoustic conditions, such as novel segments of the same noise and novel noise types, is, however, an important and active research field (Healy *et al.*, 2015; Chen *et al.*, 2016b) and needs to be addressed at the same time.

In the present study, two experiments were conducted by measuring word recognition scores (WRSs) in NH listeners. In Experiment I, the impact of exploiting spectro-temporal context in the front-end and the back-end of a segregation system, based on GMMs, was systematically investigated to identify the best performing strategy for the system. Specifically, the extraction of the delta features (Kim *et al.*, 2009) was considered in the front-end, and the two-layer classification stage from May and Dau (2014a) was employed in the back-end. Different system configurations were compared here, which either incorporated spectro-temporal context only in the front-end, only in the back-end or in both. These configurations were compared to a baseline configuration that did not include any of the strategies in the front-end and the back-end. This experiment was conducted in a restricted setup, similar to Kim *et al.* (2009), with high potential to improve speech intelligibility. Furthermore, the effect of the GMM classifier complexity in a segregation system was also investigated by comparing the results obtained with 16 GMM components and 64 GMM components. In Experiment II, the best performing strategy from Experiment I was explored in detail, and the generalization ability was subsequently evaluated in a less restricted setup that considered a mismatch in noise segments during training and testing. Finally, the intelligibility scores from both experiments were related to predictions from objective measures[1] from the extended short-term objective intelligibility (ESTOI) (Jensen and Taal, 2016), the H–FA rate (Kim *et al.*, 2009), and the amount of clustering among T–F units in binary masks (Kressner and Rozell, 2015). The primary focus of the later analysis was to guide the selection of a cost-function that correlates with speech intelligibility for future applications in computational speech segregation systems.

II. THE SEGREGATION SYSTEM

The segregation system consisted of a feature extraction front-end and a classification back-end (May *et al.*, 2015). Figure 1 illustrates the processing stages of the system. Each of these stages is described in more detail below.

A. Front-end

The noisy speech was sampled at a rate of 16 kHz and decomposed into $K = 31$ frequency channels by employing an all-pole version of the gammatone filterbank (Lyon, 1996), whose center frequencies were equally spaced on the equivalent rectangular bandwidth (ERB) scale between 80 and 7642 Hz. Previous studies (Kim *et al.*, 2009; May and Dau, 2014a; May *et al.*, 2015) have successfully exploited modulations in the speech and the interferer by extracting amplitude modulation spectrogram (AMS) features (Kollmeier and Koch, 1994; Tchorz and Kollmeier, 2003). To derive the AMS features in each subband, the envelope 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 computed over the entire envelope signal. 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 cutoff frequency of the modulation low-pass filter was calculated as the inverse of the window duration to ensure that at least one full period of the modulation frequency was included in the window, and subsequently adjusted to the nearest power of 2 integer (May *et al.*, 2015). Using a time

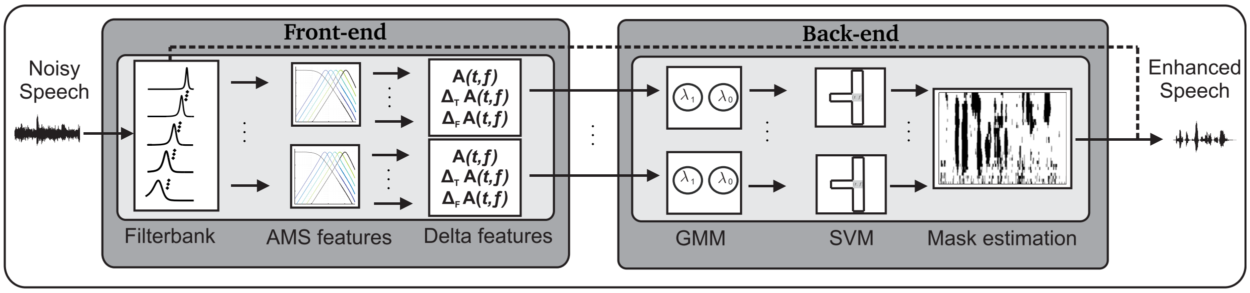

FIG. 1. (Color online) Block diagram of the speech segregation system. The system consists of a feature extraction front-end and a classification back-end. In the front-end, the noisy speech is first decomposed by a gammatone filterbank. Then, AMS features are extracted and delta features are computed. The back-end consists of two layers with a GMM classifier in the first layer and a SVM classifier in the second layer. Finally, the estimated ideal binary mask is applied to the subband signals of the noisy speech, as illustrated by the dashed line, in order to reconstruct the target signal.

frame duration of 32 ms then resulted in a cutoff frequency of 32 Hz. The root-mean-square value of each modulation filter was then calculated across each time frame with a 75% overlap. The extraction of the AMS features resulted in a six-dimensional feature vector for each T–F unit $\mathbf{A}(t,f) = \{M_1(t,f), \ldots, M_6(t,f)\}^T$. The delta features across time (Δ_T) and frequency (Δ_F) can be appended to the feature vector $\mathbf{A}(t,f)$ according to previous studies (Kim *et al.*, 2009; Han and Wang, 2012; May and Dau, 2013), resulting in a feature vector $\mathbf{X}(t,f)$ for each individual T–F unit at time frame t and subband f that consists of

$$\begin{aligned}
\mathbf{X}(t,f) &= [\mathbf{A}(t,f), \Delta_T\mathbf{A}(t,f), \Delta_F\mathbf{A}(t,f)]\\
\Delta_T\mathbf{A}(t,f) &= \begin{cases} \mathbf{A}(2,f) - \mathbf{A}(1,f) & \text{if } t = 1\\ \mathbf{A}(t,f) - \mathbf{A}(t-1,f) & \text{otherwise}\end{cases}\\
\Delta_F\mathbf{A}(t,f) &= \begin{cases} \mathbf{A}(t,2) - \mathbf{A}(t,1) & \text{if } f = 1\\ \mathbf{A}(t,f) - \mathbf{A}(t,f-1) & \text{otherwise.}\end{cases}
\end{aligned} \tag{1}$$

Instead of the calculation in Eq. (1), delta features that only operate across frequency can be considered and appended symmetrically to the AMS features for a resulting feature vector $\mathbf{X}(t,f)$,

$$\begin{aligned}
\mathbf{X}(t,f) &= \left[\mathbf{A}(t,f), \Delta_{f-k}\mathbf{A}(t,f), \Delta_{f+k}\mathbf{A}(t,f)\right]\\
\Delta_{f-k}\mathbf{A}(t,f) &= \mathbf{A}(t,f) - \mathbf{A}(t,f-k),\\
&\qquad \forall k \in \{n \in [1;K] | f - n \geq 1\}\\
\Delta_{f+k}\mathbf{A}(t,f) &= \mathbf{A}(t,f) - \mathbf{A}(t,f+k),\\
&\qquad \forall k \in \{n \in [1;K] | f + n \leq K\}.
\end{aligned} \tag{2}$$

In Eq. (2), k indicates the considered number of subbands in the calculation, and K the number of gammatone filters. Appending the delta features to the feature vector in Eqs. (1) and (2) increased the amount of exploited spectro-temporal context, but also the size of the feature vector; e.g., appending $\Delta_T\mathbf{A}(t,f)$ and $\Delta_F\mathbf{A}(t,f)$ from Eq. (1) to $\mathbf{A}(t,f)$ would increase the feature vector from 6 to 18 dimensions.

B. Back-end

Similar to previous studies, the classification back-end consisted of a two-layer segregation stage (Healy *et al.*, 2013; May and Dau, 2014a; May *et al.*, 2015). In the first layer, a 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, the LC was applied to the *a priori* SNR, and the *a priori* probabilities $P(\lambda_{1,f})$ and $P(\lambda_{0,f})$ were computed by counting the number of feature vectors for each of the classes $\lambda_{1,f}$ and $\lambda_{0,f}$ during training. The GMM classifier output was given as the posterior probability of speech and noise presence $P(\lambda_{1,f}|\mathbf{X}(t,f))$ and $P(\lambda_{0,f}|\mathbf{X}(t,f))$, respectively,

$$P\left(\lambda_{1,f}|\mathbf{X}(t,f)\right) = \frac{P\left(\lambda_{1,f}\right)P\left(\mathbf{X}(t,f)|\lambda_{1,f}\right)}{P(\mathbf{X}(t,f))}, \tag{3}$$

$$P\left(\lambda_{0,f}|\mathbf{X}(t,f)\right) = \frac{P\left(\lambda_{0,f}\right)P\left(\mathbf{X}(t,f)|\lambda_{0,f}\right)}{P(\mathbf{X}(t,f))}. \tag{4}$$

For each subband, the computed posterior probabilities of speech $P(\lambda_{1,f}|\mathbf{X}(t,f))$ were processed by a linear support vector machine (SVM) classifier (Chang and Lin, 2011) across a spectro-temporal window $\mathcal{W}$ (May and Dau, 2014a),

$$\bar{\mathbf{X}}(t,f) = \{P(\lambda_{1,u}|\mathbf{X}(u,v)) : (u,v) \in \mathcal{W}(t,f)\}. \tag{5}$$

The size of the window $\mathcal{W}$ determined the amount of spectro-temporal context exploited around the considered T–F unit. A causal and plus-shaped window function $\mathcal{W}$ was used here, where the window size with respect to time and frequency was controlled by Δt and Δf, respectively. Further details regarding the choice of the second-layer classifier and the size and shape of the window function $\mathcal{W}$ can be found in May and Dau (2014a).

III. METHODS

A. Configurations

To systemically analyze the impact of spectro-temporal context strategies in the front-end and the back-end, four system configurations were tested in Experiment I (see Table I). The "No context" configuration denotes the baseline configuration with no delta features in the front-end and no spectro-temporal integration in the back-end, corresponding to setting the window size $\mathcal{W}$ to unity ($\Delta t = 1, \Delta f = 1$). The

TABLE I. Configurations in Experiment I.

Configurations	Front-end		Back-end	
			$\mathcal{W}$ size	
	Feature vector $\mathbf{X}(t,f) =$	Feature dimension	Δt	Δf
No context	$[\mathbf{A}(t,f)]$	6	1	1
Front-end	$[\mathbf{A}(t,f), \Delta_T\mathbf{A}(t,f), \Delta_F\mathbf{A}(t,f)]$	18	1	1
Back-end	$[\mathbf{A}(t,f)]$	6	3	9
Front- & back-end	$[\mathbf{A}(t,f), \Delta_T\mathbf{A}(t,f), \Delta_F\mathbf{A}(t,f)]$	18	3	9

"Front-end" configuration includes the delta features, while the "Back-end" configuration includes the second-layer classification stage in the back-end ($\Delta t = 3, \Delta f = 9$). The "Front- & back-end" configuration employs both the front-end and the back-end spectro-temporal context strategies.

In Experiment II, the delta features were explored in details in order to investigate the potential of this strategy in the segregation system. Four configurations were selected (see Table II). The system configuration "Front-end" is the baseline configuration for the analysis across frequency and appends only $\Delta_F\mathbf{A}(t,f)$ to $\mathbf{A}(t,f)$. The configurations "3 subbands," "7 subbands," and "11 subbands" include $k = 1$, $k = 3$, and $k = 5$ lower and upper subbands to $\mathbf{A}(t,f)$.

B. Stimuli

The speech material came from the Danish Conversational Language Understanding Evaluation (CLUE) database (Nielsen and Dau, 2009). It consists of 70 sentences in seven lists for training and 180 sentences in 18 balanced lists for testing, and is spoken by a male Danish talker. Noisy speech mixtures were created by mixing individual sentences with a stationary (ICRA1) and a fluctuating six-talker (ICRA7) noise (Dreschler *et al.*, 2001). A Long Term Average Spectrum (LTAS) template was computed based on the CLUE corpus and the LTAS of each noise masker was adjusted to the template LTAS. A randomly-selected noise segment was used for each sentence. In order to avoid onset effects in the intelligibility test (Nielsen and Dau, 2009), the noise segment started 1000 ms before the speech onset and ended 600 ms after the speech offset. However, the objective measures were computed only for the regions between speech onset and offset.

C. System training and evaluation

In Experiment I, the segregation system was trained separately for the two noise types limited to 10 s in duration. Originally, the ICRA1 consists of a 60 s noise recording and ICRA7 of a 600 s recording (Dreschler *et al.*, 2001). The first layer of the classification back-end consisted of a subband GMM classifier with either 16 or 64 components and full covariance matrices. The classifiers were first initialized by 15 iterations of the K-means clustering algorithm, followed by five (for 16 GMMs) or 50 (for 64 GMMs) iterations of the expectation-maximization algorithm. The classifiers were trained with the 70 training sentences that were each mixed three times with a randomly-selected noise segment from 10 s noise recordings at -5, 0, and 5 dB SNR. The subsequent linear SVM classifier was trained for each subband with only ten sentences mixed at -5, 0, and 5 dB SNR. Afterwards, a re-thresholding procedure was applied (Han and Wang, 2012; May and Dau, 2014a) using a validation set of ten sentences, where new SVM decision thresholds were obtained which maximized the H–FA rates. Both the first and second-layer classifiers employed an LC of -5 dB in a similar manner as previous findings (Han and Wang, 2012; May and Dau, 2014b). The segregation system was evaluated with the 180 CLUE sentences. Each sentence was mixed with the noises at -5 dB SNR using the same limited noise recordings from the training session.

TABLE II. Configurations in Experiment II.

Configurations	Front-end	
	Feature vector $\mathbf{X}(t,f) =$	Feature dimension
Front-end	$[\mathbf{A}(t,f), \Delta_F\mathbf{A}(t,f)]$	12
3 subbands	$[\mathbf{A}(t,f), \Delta_{F-1}\mathbf{A}(t,f), \Delta_{F+1}\mathbf{A}(t,f)]$	18
7 subbands	$[\mathbf{A}(t,f), \Delta_{F-1}\mathbf{A}(t,f), \Delta_{F+1}\mathbf{A}(t,f), ..., \Delta_{F+3}\mathbf{A}(t,f)]$	42
11 subbands	$[\mathbf{A}(t,f), \Delta_{F-1}\mathbf{A}(t,f), \Delta_{F+1}\mathbf{A}(t,f), ..., \Delta_{F+5}\mathbf{A}(t,f)]$	66

Experiment II only tested the highly non-stationary ICRA7 noise type in a less restricted setup. This noise type is more likely to challenge a speech segregation system than the stationary ICRA1. The full noise recording of 600 s was divided into one half recording for training and one half recording for testing. The training and evaluation was similar to Experiment I. The first layer of the classification back-end had a complexity of 16 GMMs with full covariance matrix. The complexity choice is discussed in Sec. V B.

D. Test procedure and subjects

In Experiment I, the following 24 conditions were tested: (Noisy speech, No integration, Front-end, Back-end, Front- & back-end, IBM) × (ICRA1, ICRA7) × (16 GMMs, 64 GMMs). The total number of conditions (24) exceeded the number of available CLUE test lists (18). Therefore, to be able to randomly assign one condition to one test list, the experiment was conducted with two subject groups, each with $n = 15$ NH listeners. The first subject group was tested with the 12 conditions corresponding to the classifier complexity of 16 GMMs, and the second group was tested with the 12 conditions with only 64 GMMs. The following five conditions were tested in Experiment II: Noisy speech, Front-end, 3 subbands, 7 subbands, and 11 subbands. The experiment was conducted with one subject group with $n = 20$ NH listeners that differed from the subject groups used in Experiment I. In this experiment, 13 other conditions were also tested that were not relevant to this study.

The listener age was between 20 and 32 yr with a mean of 24.5 yr in Experiment I and a mean of 26.7 yr in Experiment II. Requirements for participation were: (1) age between 18 and 40 years, (2) audiometric thresholds of less than or equal to 20 dB hearing level (HL) in both ears (0.125 to 8 kHz), (3) Danish as native language, and (4) no previous experience with the Hearing In Noise Test (HINT) (Nielsen and Dau, 2011) or CLUE (Nielsen and Dau, 2009). The total experimental time was about 2 h in Experiment I and about 1.5 h in Experiment II, including the screening process. The subjects were paid for their participation.

The experiments consisted of a training and testing session. During the training session, five randomly selected sentences from the training set were presented for each of the 12 conditions to familiarize the subject to the task. Subsequently, each subject heard one list per condition, and conditions and lists were randomized across subjects. The sentences were presented diotically to the listener via headphones (Sennheiser HD650) in an acoustically and electrically shielded booth. Prior to the actual experiments, the headphones were calibrated by first adjusting to a reference sound pressure level (SPL) value and then performing a headphone frequency response equalization. During the experiment, the sentences were adjusted to the desired presentation level, and the equalization filters were applied. The SPL was set to a comfortable level of 65 dB. The presentation level was only increased after the training session if the subject reported back that the level was too low. The level never exceeded 70 dB SPL for any subject. For each sentence, the subjects were instructed to repeat the words they heard, and an operator scored the correctly understood words via a MATLAB interface. The subjects were told that guessing was allowed. They could listen to each sentence only once, and breaks were allowed according to the subject's preference.

E. Statistical analysis

Intelligibility scores were reported as a percentage of correctly scored words, i.e., the WRS, at −5 dB SNR. The WRSs were computed per sentence and averaged across sentences per list. The averaged WRSs were used to construct a linear mixed effect model for each experiment. In Experiment I, the three fixed factors of the mixed model were the system configuration (four levels), the noise type (two levels), and the classifier complexity (two levels). The subjects were treated as a random factor, as is standard in a repeated measure design. The intelligibility scores in Experiment I followed a normal distribution. All fixed effects, all interactions between fixed effects, and the random effect were initially included in the model. The model was then reduced by performing a backward elimination of all random and fixed interactions that were non-significant. This included all of the interaction terms between the random effect (subjects) and the fixed factors (configuration, noise type, and classifier complexity) and the interaction term between all three fixed factors. In Experiment II, the only fixed factor was system configuration (four levels) and subjects were treated as a random factor. The intelligibility scores in Experiment II also followed a normal distribution.

All levels were tested at a 5% significance level. To visualize the data, the least-squares means and 95% confidence intervals were extracted from the model. To assess any difference between conditions, the differences of the least-squares means were computed and the p values were adjusted following the Tukey multiple comparison testing. To evaluate potential speech intelligibility improvements, Paired Students t-tests between the noisy speech and each of the system configurations were constructed and tested at a 5% significance level.

F. Objective measures

Three different objective measures were compared to the intelligibility scores in each experiment: ESTOI (Jensen and Taal, 2016), H–FA rate (Kim *et al.*, 2009), and the clustering parameter γ (Kressner and Rozell, 2015). The ESTOI (Jensen and Taal, 2016) is a modified version of the short-term objective intelligibility (STOI) index (Taal *et al.*, 2011) to better account for modulated noise maskers. The STOI metric is based on a short-term correlation analysis between the clean and the degraded speech (Taal *et al.*, 2011), mapped to a value between 0 and 1. The ESTOI improvements (Δ ESTOI) were reported here as the relative difference between the predicted ESTOI values for the processed and the unprocessed noisy speech baselines. To compute the H–FA rate, the correctly classified speech-dominated T–F units and incorrectly classified noise-dominated T–F units were derived by comparing the estimated IBM with the IBM. The H–FA rates and the ESTOI improvements were averaged across all 180 test sentences. The clustering parameter γ was learned across all 180 test sentences by the graphical model described in Kressner and Rozell (2015). Given a set of binary masks, the graphical model estimates the amount of clustering γ between T–F units within the masks as a single number. γ quantifies how much more likely neighboring T–F units are to have the same label (speech-dominated or noise-dominated) as opposed to different labels. Therefore, binary masks with T–F units that are twice as likely to have the same label than a different label as their neighboring units would be described by $\gamma = 2.0$. Binary masks with T–F units that are equally likely to be in the same state as their neighbors would have a $\gamma = 1.0$, indicating that the labels of the T–F units would be uniformly and randomly distributed. Therefore, a mask with $\gamma = 2.0$ will contain more clustering among the T–F units than a mask with $\gamma = 1.0$ (Kressner and Rozell, 2015). To illustrate the γ parameter, Fig. 2 shows binary masks for one particular CLUE sentence mixed with ICRA7 noise at −5 dB SNR with the respective γ values, shown in parenthesis. Figure 2(a) shows the IBM and Figs. 2(b)–2(e) present the estimated IBMs for the four tested system configurations listed in Table I. The two mask error types, misses and false alarms, are shown on top of the binary masks for a visualization of the error distributions. Comparing the masks for the four tested system configurations, the masks from Fig. 2(d) and Fig. 2(e) contain a larger amount of clustering than the masks in Fig. 2(b) and Fig. 2(c).

IV. RESULTS

A. Experiment I: Impact of exploiting spectro-temporal context

Figure 3 shows intelligibility scores obtained with the four system configurations ("No Context," "Front-end,"

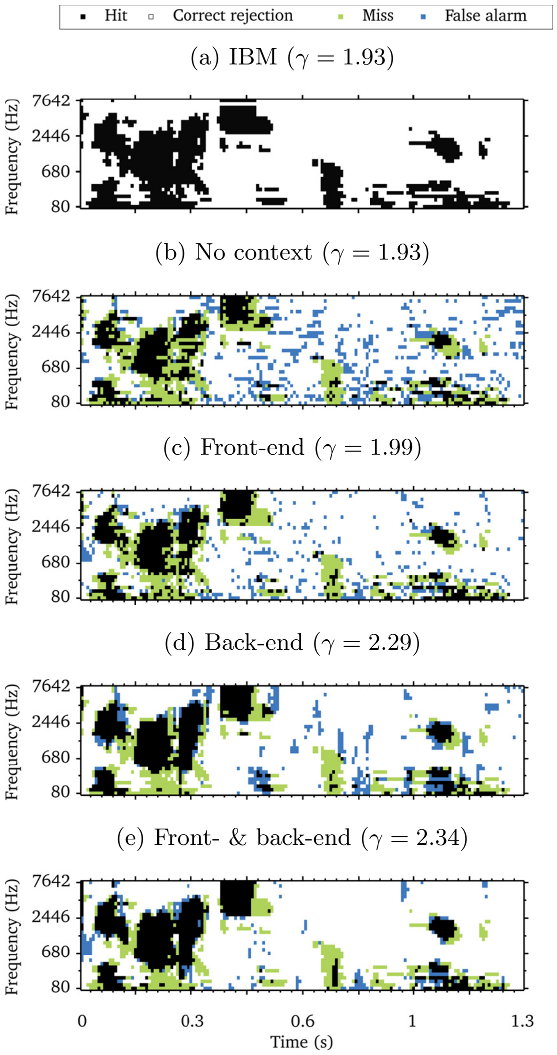

FIG. 2. (Color online) Binary masks for a CLUE sentence mixed with ICRA7 noise at -5 dB SNR. Misses (target-dominated T–F units erroneously labeled as masker-dominated) and false alarms (masker-dominated T–F units erroneously labeled as target-dominated) are shown on top of the masks.

"Back-end," and "Front- & back-end") in the two noise types (ICRA1 and ICRA7) considered in Experiment I. Results are shown for the two classifier complexities, namely 16 GMMs in Fig. 3(a) and 64 GMMs in Fig. 3(b). The condition with the unprocessed noisy speech (diamonds) represented the baseline, and the IBM condition (stars) was considered as the ideal reference. For the baseline and the ideal reference, sample means across subjects and 95% Students t-based confidence intervals of the mean were computed. For the system configurations, the least square means and 95% confidence intervals from the fitted linear mixed effect model were considered.

The baseline in Fig. 3 differed across noise types, with WRS of about 50%–55% for the stationary ICRA1 and 65% for the fluctuating ICRA7, presumably because the participants were able to listen in the dips in the six-talker noise. For the IBM conditions, WRS of close to 100% was achieved for both noise types. This was expected as the IBM exploited the *a priori* information about the speech and the noise signals.

There was an effect of system configuration depending on the classifier complexity and on the noise type. Most importantly, the "Front-end" configuration led to significantly higher intelligibility scores than the "Back-end" configuration for both noise types and both classifier complexities ($p < 0.0001$). Specifically, the WRS increased by 18.0% in ICRA1 and 23.1% in ICRA7 with 16 GMMs [Fig. 3(a)], and 28.8% in ICRA1 and 34.0% in ICRA7 with 64 GMMs [Fig. 3(b)]. This particular finding suggests that extracting and appending the delta features to the AMS features in the front-end is a more effective way of exploiting spectro-temporal contextual information than using the SVM-based integration strategy in the back-end. In all four combinations, except with 16 GMMs in the case of the ICRA1 noise, the "Front-end" configuration led to significantly larger scores than the "No context" configuration, which emphasizes that it is more effective to exploit contextual information in the front-end of the system than not considering any strategy at all. Finally, the "Front- & back-end" configuration also led to significantly higher scores than the "Back-end" configuration in all four combinations of noise type and classifier complexity. However, the mean scores for the "Front- & back-end" were generally lower than for the "Front-end." This suggests that employing both strategies is more effective to exploit spectro-temporal context than just employing the SVM-based integration strategy in the back-end alone, but the combination of the two strategies does not lead to better results than the front-end strategy alone.

There was also an effect of the classifier complexity that depended on the system configuration and the noise type. By comparing the results in Figs. 3(a) and 3(b), significantly higher scores were obtained for the "Front-end" configuration with 64 GMMs than with 16 GMMs for both noise types. Specifically, the WRS increased by 12.6% in ICRA1 ($p < 0.05$) and 19.5% in ICRA7 ($p < 0.0001$). However, the scores for the "Back-end" configuration did not change significantly across classifier complexity for either noise type. Most importantly, the ranking of the system configurations remained unchanged across classifier complexity.

The measured intelligibility scores from Fig. 3 were converted into WRS improvements relative to the unprocessed noisy speech, ΔWRS. Figures 4(a) and 4(b) show ΔWRS as a function of the system configuration, noise type, and classifier complexity. Significant improvements, based on the Paired Students t-tests, are indicated by an asterisk (*). Significant improvements of about 50% for ICRA1 and 35% for ICRA7 over noisy speech were obtained with the IBM. For 64 GMMs in Fig. 4(b), the configurations "No Context" ($t[14] = -2.16$, $p = 0.02$), "Front-end" ($t[14] = -4.29, p =< 0.001$), and "Front- & back-end" ($t[14] = -2.82, p = 0.007$) for ICRA1 led to significant improvements and for the ICRA7, only the "Front-end" ($t[14] = -7.44, p =< 0.001$) led to a significant improvement. To evaluate the potential of the objective measures, the measured intelligibility scores were related to predictions from each of the objective measures described in Sec. III F. Figure 4 also shows the objective measures ΔESTOI [Figs. 4(c) and 4(d)], H–FA rates [Figs. 4(e) and

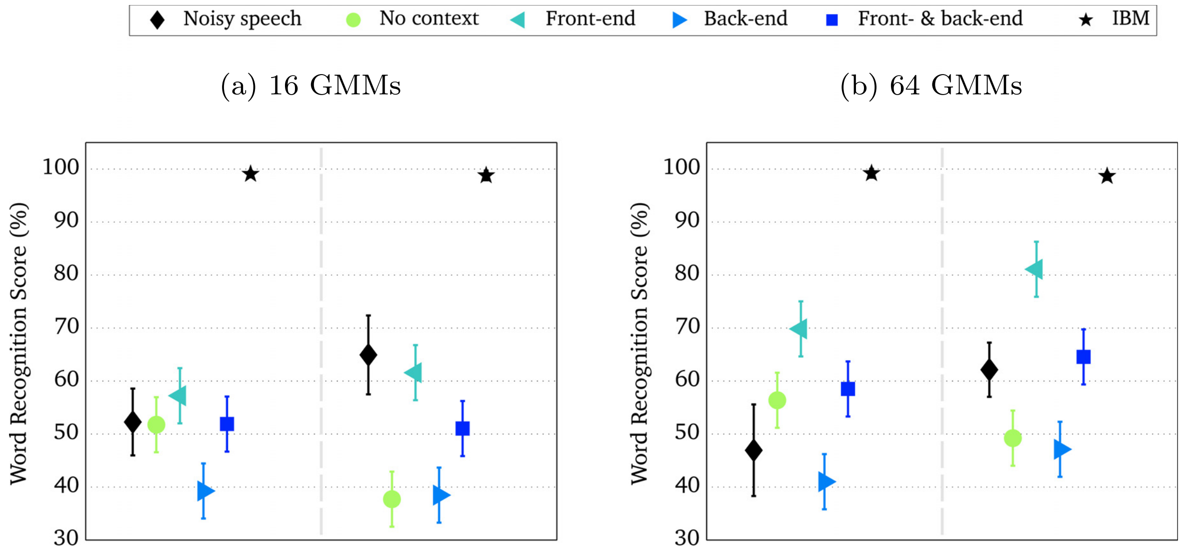

FIG. 3. (Color online) Experiment I's WRSs at -5 dB SNR of the four different system configurations ("No Context," "Front-end," "Back-end," and "Front- & back-end") for the two noise types (ICRA1 and ICRA7) and for the two classifier complexities plotted in panel (a) (16 GMMs) and panel (b) (64 GMMs). The condition with the unprocessed noisy speech represented the baseline and the IBM condition was considered as the ideal reference. For the baseline and the ideal reference, sample means across subjects and 95% Students t-based confidence intervals of the mean were computed. For all system configurations in all combinations of noise type and classifier complexity, the least square means and 95% confidence intervals from the fitted linear mixed effect model were plotted.

4(f)] and γ [Figs. 4(g) and 4(h)] in Experiment I. ΔESTOI indicates the increase in ESTOI relative to the unprocessed noisy speech. The largest predicted improvement was observed for the configuration "Front- & back-end," and the lowest predicted improvement was found for the "No context" configuration in all combinations of noise type and classifier complexity level. This is in conflict with the measured ΔWRS in Figs. 4(a) and 4(b) where the "Front-end" configurations led to the largest improvements. By comparing Figs. 4(c) and 4(d), it can be seen that larger ESTOI improvements were generally observed with 64 GMMs compared to 16 GMMs. This is consistent with the measured WRS improvements in Figs. 4(a) and 4(b).

Figures 4(e) and 4(f) show the H–FA rates. The segregation system generally produced higher H–FA rates in the presence of the stationary noise than in the presence of the non-stationary six-talker noise. The six-talker noise contains spectro-temporal modulations, similar to modulations in the target speech signal, and it will be more difficult for the classifier to separate the speech modulations from the six-talker noise modulations. In all combinations of noise type and classifier complexity, the lowest H–FA rates were observed for the "No context" configuration and the highest H–FA rates were found for the "Front- & back-end" configuration. Also, larger H–FA rates were obtained for the "Back-end" than for the "Front-end" configuration, which is not consistent with Figs. 4(a) and 4(b). Furthermore, higher H–FA rates were obtained with 64 GMMs in Fig. 4(f) than with 16 GMMs in Fig. 4(e). A comparison with the measured WRS improvements in Figs. 4(a) and 4(b) indicated a conflict with this prediction, since the "Front-end" configuration led to the highest intelligibility scores, but not the highest H–FA rates. Finally, it is observed that a small increase of H–FA [from Fig. 4(e) to Fig. 4(f)] corresponds to a large increase of WRS [from Fig. 4(a) to Fig. 4(b)] from 16 GMMs classifier to the 64 GMMs classifier. This was found for both noise types.

Figures 4(g) and 4(h) show the γ values learned by the graphical model. The IBM itself contains a certain level of clustering, due to the compact representation of speech-dominated T–F units forming glimpses of the target signal. The γ values from system configurations that exploited spectro-temporal context through the SVM based integration strategy in the back-end ("Back-end" and "Front- & back-end") were consistently larger than the γ values learned over masks from the "Front-end" and the "No context" configurations. Furthermore, the "Front-end" did not lead to larger γ values than the "No context." This suggests that computing delta features in the front-end does not increase the amount of clustering in contrast to employing a spectro-temporal SVM based integration strategy in the back-end. The effect of exploiting spectro-temporal context in binary masks was visualized in Fig. 2 in Sec. III. Figures 2(d) and 2(e) showed masks with a larger amount of T–F clustering than the masks in Figs. 2(b) and 2(c), and a visual inspection of the example utterance indicated that the erroneous T–F units became more clustered in Figs. 2(d)–2(e). Finally, a comparison of Figs. 4(g) and 4(h) suggests that the amount of clustering in the mask is not affected by the classifier complexity in the segregation system, as γ remains unchanged.

B. Experiment II: Exploring delta features and the system generalization ability

Figure 5 shows intelligibility scores obtained in Experiment II with the four system configurations ("Front-end," "3 subbands," "7 subbands," and "11 subbands") tested in the less restricted setup in ICRA7 noise. For all four configurations, the $\Delta_T\mathbf{A}(t,f)$ from Eq. (1) was not

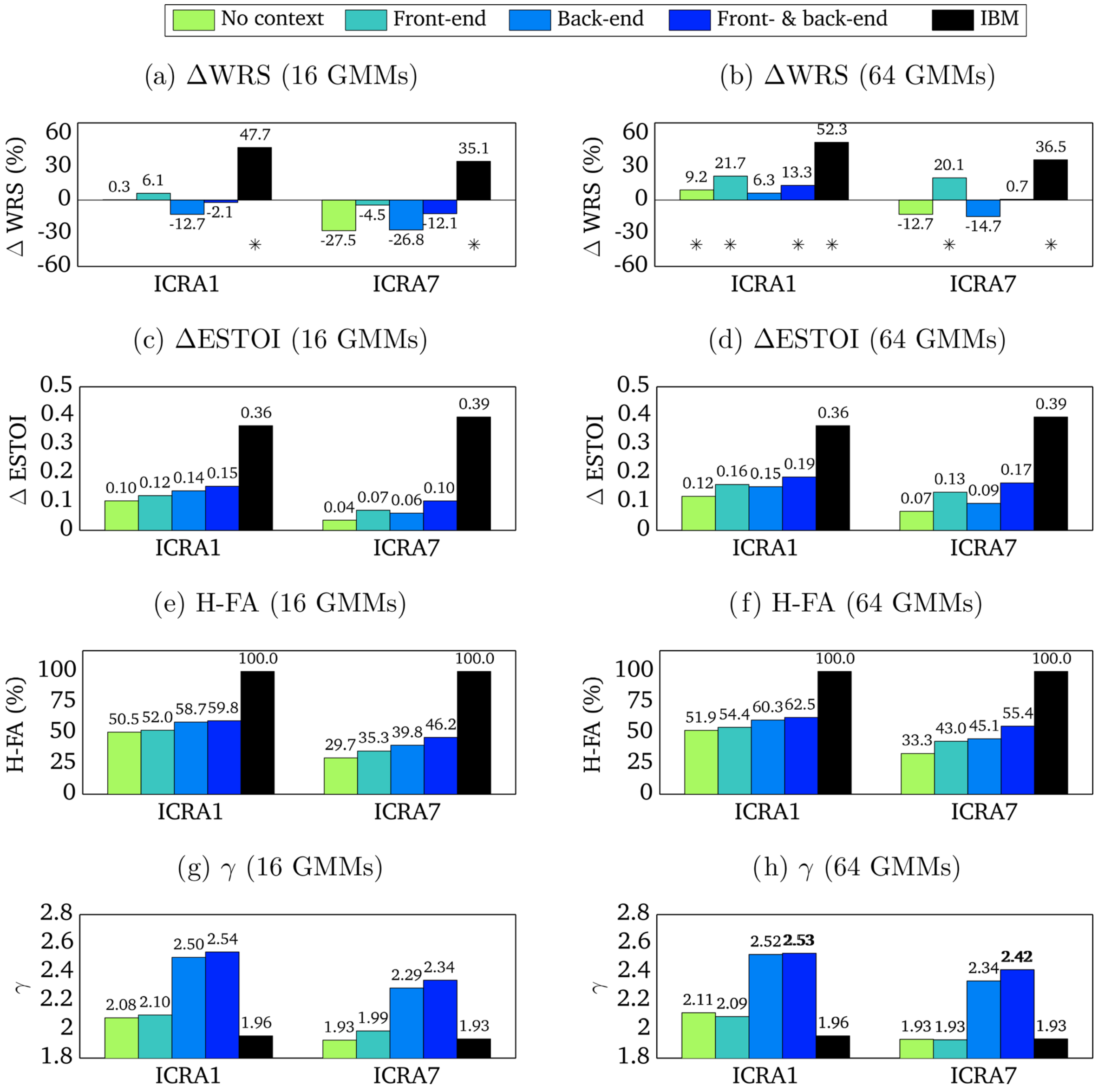

FIG. 4. (Color online) Experiment I's ΔWRS relative to noisy speech (first row of panels), ΔESTOI relative to noisy speech (second row of panels), H–FA rates (third row of panels), and γ values (fourth row of panels) for the four different system configurations with the two noise types (ICRA1 and ICRA7) and with the two classifier complexities in (a) and in (b). The IBM has been included as the ideal reference. WRS improvements are derived from the Paired Students t-tests and significant improvements (on a 5% significance level) are marked with an asterisk (*). All objective measures are evaluated at -5 dB SNR.

appended to the feature vector in Eq. (2). This decision was based on an analysis of the objective measures prior to Experiment II, which showed no change in the objective measures when $\Delta_T\mathbf{A}(t,f)$ was left out. In Fig. 5, the level of the noisy speech was consistent with the level in Experiment I for ICRA7 (see Fig. 3). In this experiment, there was an effect of system configuration. The intelligibility scores were significantly higher in the "3 subbands" configuration than the "Front-end" configuration by 10.7% ($p < 0.01$) and from the "3 subbands" to the "7 subbands" configuration by 8.2% ($p < 0.05$). The "7 subbands" and the "11 subbands" configurations did not differ significantly. This finding indicated that appending more subbands, as proposed in Eq. (2), can lead to significantly higher intelligibility until a plateau at $k = 5$ with "11 subbands." Figure 6 presents the intelligibility improvements and objective measure predictions for Experiment II. In Fig. 6(a), the Paired Students t-tests showed that all system configurations led to significantly smaller intelligibility scores than the noisy speech, despite an increase in intelligibility over appended subbands. Therefore, none of the system configurations were able to improve speech intelligibility in the less restricted setup. Since this setup included novel noise segments in testing not seen during training, this suggested that the segregation system did not generalize well to unseen noise segments of the six-talker noise.

In Fig. 6(b), all predicted ΔESTOI values were positive, and the largest predicted improvements were observed for the configurations "7 subbands" and "11 subbands." This was not consistent with results from the listener study in Fig. 6(a), where no WRS improvements were observed, which highlights the discrepancy between predicted and measured intelligibility improvements in this study. The H–FA rate in

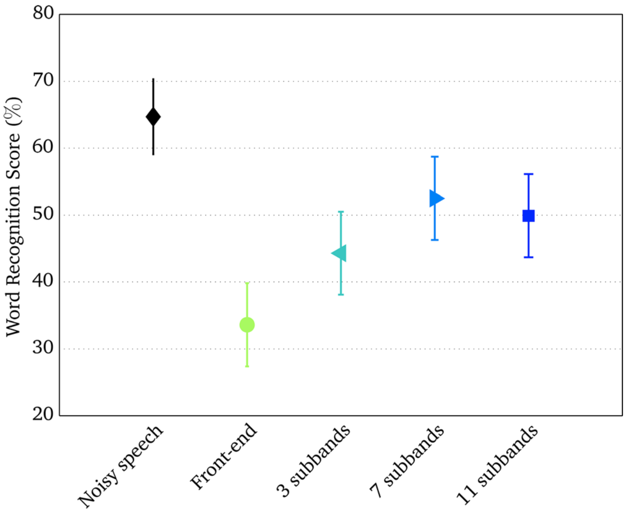

FIG. 5. (Color online) Experiment II's WRSs at −5 dB SNR with the four different system configurations ("Front-end," "3 subbands," "7 subbands," and "11 subbands") in ICRA7. The condition with the unprocessed noisy speech represented the baseline. For the baseline, sample means across subjects and 95% Students *t*-based confidence intervals of the mean were computed. For all system configurations, the least square means and 95% confidence intervals from the fitted linear mixed effect model were plotted.

Fig. 6(c) increased with the number of appended subbands, whereas the rates were comparable for "7 subbands" and "11 subbands." As observed in Experiment I, a small change in H–FA had a large impact on the measured intelligibility scores. This was illustrated by comparing Fig. 4(e) for the ICRA7 noise and Fig. 6(c). A H–FA rate of 35.3% in Fig. 4(e) corresponded to a 4.5% decrease in WRS for the "Front-end" configuration, whereas a H–FA of 33.6% in Fig. 6(c) corresponded to a 31.1% decrease in WRS over noisy speech. With respect to clustering [Fig. 6(d)], γ did not change with the system configuration, suggesting that the amount of clustering in the mask is not affected by appending more subbands to the AMS features. This is in contrast to the Experiment I where the SVM integration stage in the back-end increased both H–FA and γ.

V. DISCUSSION

A. The impact of exploiting spectro-temporal context

The measured intelligibility scores in Experiment I (Sec. IV A) showed that the front-end strategy, where the system was given access to both the AMS features and the delta features, led to significantly higher intelligibility scores than employing the back-end strategy, which incorporated the SVM-based spectro-temporal integration. The scores were consistently higher for the front-end strategy than the back-end strategy, regardless of the noise type and classifier complexity. Moreover, compared to the unprocessed noisy speech, the back-end strategy actually had a detrimental effect on the intelligibility scores. The comparison of the objective measures in Fig. 4 (Sec. IV A) indicated that the back-end strategy increased the H–FA rates over the front-end strategy but, at the same time, increased the amount of clustering of individual T–F units. The visual inspection of the illustrated mask examples in Fig. 2 (Sec. III F) furthermore suggested that the increased amount of clustering implied an increased clustering of the misses and false alarms. Previously, it was shown that clustering of the two error types results in reduced intelligibility scores despite having the same classification accuracy (Kressner and Rozell, 2015), which may explain the detrimental effect of the back-end strategy on the present intelligibility scores. Furthermore, computing delta features in the front-end had a positive effect on speech intelligibility. The intelligibility scores were significantly higher than the scores with the

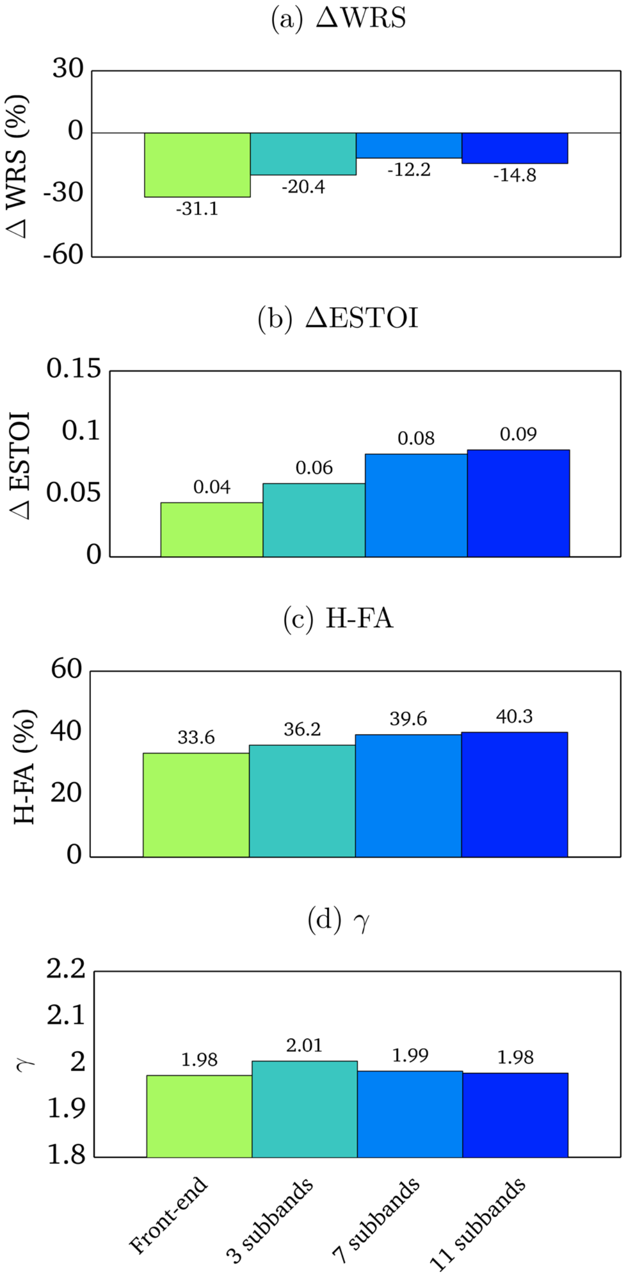

FIG. 6. (Color online) Experiment II's ΔWRS relative to noisy speech (first row of panels), ΔESTOI relative to noisy speech (second row of panels), H–FA rates (third row of panels), and γ values (fourth row of panels) with the four different system configurations in ICRA7. WRS improvements are derived from the Paired Students *t*-tests and significant improvements (on a 5% significance level) are marked with an asterisk (*). All objective measures are evaluated at −5 dB SNR.

configuration that did not employ any of the strategies, and improvements over noisy speech were significant for the higher complexity classifier of 64 GMMs. Because of the detrimental effect of the back-end strategy on intelligibility, combining both strategies simultaneously in the front-end and in the back-end did not lead to the largest measured intelligibility scores in Sec. IV A. This contradicted the findings in Fig. 4(e) and Fig. 4(f) (Sec. IV A) where a higher H–FA rate was found when combining the strategies than employing only one of the strategies, consistent with the literature (Healy *et al.*, 2013; May and Dau, 2013). The results from Experiment I therefore suggest that, in the considered segregation system, a better spectro-temporal strategy is to compute delta features of the AMS features in the front-end rather than employing the selected SVM-based integration strategy in the back-end. This study, however, did not consider the effects of changing the shape and the size of the window in the back-end on measured intelligibility. Also, the effect of employing a different second-layer classifier is currently unknown. Healy *et al.* (2013) considered a similar two-layer classification stage, but they employed deep neural networks (DNNs) in a DNN–DNN layer with an integration window of size five time frames and 17 subbands of the 64 channels. They reported significant improvements in intelligibility scores with this system, but did not quantify the impact of the back-end strategy alone.

In Experiment II, the front-end strategy was explored in detail by appending delta features computed from symmetrical subbands. Results in Sec. IV B showed that the intelligibility scores increased with the number of appended subbands up to $k = 5$ bands where the improvement reached a plateau. This indicated that intelligibility increased with the amount of spectral information in the speech that was exploited up to $k = 5$ subbands. The same trend was observed for the H–FA rate in Fig. 6. Appending the delta features across frequency increased the size of the feature vector, and the larger amount of training data led to improvements in H–FA rate for the higher complexity classifier of 64 GMMs compared to the 16 GMMs classifier. Moreover, the amount of clustering among the T–F units in Experiment II was equal to the amount of clustering for the front-end strategy in Experiment I and remained constant with the number of appended subbands. This is in line with the notion from Experiment I that increased accuracy without increased clustering among the T–F units can lead to higher intelligibility scores.

Other strategies exists that exploit the contextual information in speech. In contrast to the delta features, which work on a subband level, temporal context can also be exploited by stacking feature frames as input to broadband DNNs for classification (Wang *et al.*, 2014; Chen *et al.*, 2016b). However, the impact of this particular strategy on intelligibility scores, or any of the objective measures, has not been quantified, which makes a comparison to the strategies in the present study challenging.

B. The generalization ability of the segregation system

In Experiment I, a restricted setup from Kim *et al.* (2009), with matched noises during training and testing, was used in order to facilitate a comparison of the system configurations, and for a comparison across GMM classifier complexity. May and Dau (2014b) compared H–FA rates for matched and mismatched noise segments of the same noise type in training and testing as a function of the number of GMMs in the classification stage. A high complexity classifier of 256 GMMs employed in Kim *et al.* (2009) was able to learn all spectro-temporal characteristics of the noise, when the same short noise segment was used in training and testing. This was due to an over-fitting of the segregation system which resulted in high H–FA rates (May and Dau, 2014b) and potentially explains the high intelligibility scores obtained in the study. In Experiment I, these observations from May and Dau (2014b) were verified. The measured intelligibility scores of the front-end strategy were higher with 64 GMMs in Fig. 3(b) compared to the lower complexity classifier of 16 GMMs in Fig. 3(a). Employing the same amount of components as in Kim *et al.* (2009) would likely result in intelligibility scores at ceiling and close to the IBM.

The ability of segregation systems to generalize to acoustic conditions not seen during training is a very important aspect. In Experiment II, novel noise segments in testing not seen during training were considered. Despite the fact that intelligibility increased with appended subbands in Fig. 6(a), none of the configurations were able to improve speech intelligibility over noisy speech, suggesting that the system did not generalize well to unseen noise segments of the six-talker noise. This noise type contains spectro-temporal modulations very similar to modulations in the target speech signal. Therefore, the task of improving intelligibility in a realistic setup is non-trivial. According to May and Dau (2014b), the H–FA rates were generally lower when the considered segregation system was tested with unseen noise segments of the same noise recording, and the rates decrease with increasing GMM classifier complexity. Therefore, in a more realistic setup like in Experiment II, choosing a lower complexity classifier will reduce the risk of over-fitting the system (May and Dau, 2014b), however at the expense of lower H–FA rates and lower intelligibility outcomes.

Other studies have successfully demonstrated a generalization ability to acoustical mismatches by employing DNNs because of their predictive power and the ability to benefit from large-scale training for feature learning (Healy *et al.*, 2015; Chen *et al.*, 2016a, 2016b). In Healy *et al.* (2015), a four-hidden layer DNN was applied and tested on novel segments of the same noise type, which led to a 25% improvement in WRS in 20-talker babble at −5 dB SNR in NH listeners, but no improvement in cafeteria noise. In Chen *et al.* (2016b), a multi-conditional training set was introduced, and a classifier was trained using a five-hidden layer DNN and tested for a range of novel noise types. For the same 20-talker noise at −5 dB SNR, they were able to improve the WRS by approximately 10% in NH listeners. The amount of training employed in these two studies, however, differs from the current study. In Healy *et al.* (2015) $560 \times 50 = 28\,000$ utterances were used for each noise type and SNR, and in Chen *et al.* (2016b) 640 000 utterances were used in the multi-conditional training set. In the current study, only 210 utterances were used for training of the GMM classification stage. The capability of the DNNs to handle large-scale training data is most likely key to an increased ability to generalize to the unseen acoustical conditions.

C. Implications for cost function design

Kressner *et al.* (2016) highlighted potential limitations of STOI in predicting the intelligibility of binary-masked speech. In the present study, ESTOI was employed instead of STOI, but several observations indicated that ESTOI has similar limitations as STOI. First of all, in Experiment I, the ranking of the system configurations for the ESTOI improvements conflicted with the ranking of the configurations for the measured intelligibility improvements, as was observed in Fig. 4. Second, in Experiments I and II, ESTOI predicted improvements of the system configurations when no intelligibility improvements were actually present. In Experiment I, the listener study only revealed improvements for configurations with the 64 GMMs classifier, and in Experiment II, no improvements were observed at all. Therefore, ESTOI alone is not able to account for the observations in this study. Furthermore, the H–FA metric was also not able to correctly predict the ranking of the system configurations in Experiment I. Specifically, the H–FA rate was consistently higher for the back-end strategy than the front-end strategy, despite the fact that the intelligibility study revealed an opposite effect. Therefore, it is possible to construct a segregation system that is able to improve H–FA and ESTOI, but at the same time fails to improve speech intelligibility scores in noisy conditions. This reveals the limitations of the two measures and emphasizes the need of a single objective measure that comprehensively predicts segregation performance and correlates well with intelligibility for speech segregation systems.

The findings from Experiment I and II have important implications for the design of cost functions in computational speech segregation systems. Monitoring the amount of mask clustering γ in the estimated IBMs seems critical as the clustering among erroneously-labeled T–F units should be minimized. The IBM itself inherently contains clustering, and the obtained γ value can be regarded as the accepted amount of clustering among the correctly-labeled T–F units. Therefore, an appropriate cost function should maximize the H–FA rate and approximate γ as close as possible to γ of the IBM.

VI. CONCLUSION

In this study, two experiments were conducted with NH listeners. In Experiment I, the impact of spectro-temporal context in a computational speech segregation system was investigated by considering two strategies in the system front-end and back-end, respectively. The experiment showed that computing delta features in the front-end led to higher speech intelligibility than employing an SVM-based integration strategy in the back-end. The results were consistent across different noise types and for different classifier complexities. In Experiment II, the delta features were explored in detail and tested in a setup that considered novel noise segments of the same six-talker noise. Intelligibility scores increased with the amount of spectral information exploited, but the segregation system did not generalize well to novel noise segments of this particular noise type. The intelligibility scores were subsequently compared to predictions from several objective measures. The comparison showed that no single measure could account for all intelligibility scores, and therefore emphasizes the need of a single objective measure that comprehensively predicts segregation performance and correlates well with intelligibility. The findings from the present study may have implications for the design of computational speech segregation systems, in which spectro-temporal context should be incorporated without increasing the amount of clustering among erroneous labeled T–F units. Furthermore, the findings can help select a cost function that correlates with intelligibility. According to the results in the present study, the cost function should maximize the H–FA rate and approximate the γ value as close as possible to the γ of the IBM.

ACKNOWLEDGMENTS

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 No. DFF-5054-00072.

[1] Predictions in Experiment I were based on simulations with the objective measures and these predictions have been presented at the 17th Annual Conference of the International Speech Communication Association, San Francisco, USA and published as part of the conference proceedings in Bentsen *et al.* (2016)

Barker, J., and Cooke, M. (**2007**). "Modelling speaker intelligibility in noise," Speech Commun. **49**(5), 402–417.

Bentsen, T., May, T., Kressner, A., and Dau, T. (**2016**). "Comparing the influence of spectro-temporal integration in computational speech segregation," in *Proceedings of Interspeech 2016*, September 8–12, San Francisco, CA, pp. 170–174.

Brungart, D. S., Chang, P. S., Simpson, B. D., and Wang, D. (**2006**). "Isolating the energetic component of speech-on-speech masking with ideal time-frequency segregation," J. Acoust. Soc. Am. **120**(6), 4007–4018.

Chang, C. C., and Lin, C. J. (**2011**). "LIBSVM: A library for support vector machines," ACM Trans. Intell. Syst. Technol. **2**(3), 27:1–27:27. Software is available at http://www.csie.ntu.edu.tw/∼cjlin/libsvm.

Chen, J., Wang, Y., and Wang, D. (**2016a**). "Noise perturbation for supervised speech separation," Speech Commun. **78**, 1–10.

Chen, J., Wang, Y., Yoho, S. E., Wang, D., and Healy, E. W. (**2016b**). "Large-scale training to increase speech intelligibility for hearing-impaired listeners in novel noises," J. Acoust. Soc. Am. **139**(5), 2604–2612.

Cooke, M. (**2006**). "A glimpsing model of speech perception in noise," J. Acoust. Soc. Am. **119**(3), 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**(3), 267–285.

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**(3), 148–157.

Han, K., and Wang, D. L. (**2012**). "A classification based approach to speech segregation," J. Acoust. Soc. Am. **132**(5), 3475–3483.

Healy, E. W., Yoho, S. E., Chen, J., Wang, Y., and Wang, D. (**2015**). "An algorithm to increase speech intelligibility for hearing-impaired listeners in novel segments of the same noise type," J. Acoust. Soc. Am. **138**(3), 1660–1669.

Healy, E. W., Yoho, S. E., Wang, Y., and Wang, D. L. (**2013**). "An algorithm to improve speech recognition in noise for hearing-impaired listeners," J. Acoust. Soc. Am. **134**(6), 3029–3038.

Hu, Y., and Loizou, P. C. (**2010**). "Environment-specific noise suppression for improved speech intelligibility by cochlear implant users," J. Acoust. Soc. Am. **127**(6), 3689–3695.

Jensen, J., and Taal, C. H. (**2016**). "An algorithm for predicting the intelligibility of speech masked by modulated noise maskers," IEEE Trans. Audio, Speech, Lang. Process. **24**(11), 2009–2022.

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**(3), 1486–1494.

Kjems, U., Boldt, J. B., Pedersen, M. S., Lunner, T., and Wang, D. (**2009**). "Role of mask pattern in intelligibility of ideal binary-masked noisy speech," J. Acoust. Soc. Am. **126**(3), 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**(3), 1593–1602.

Kressner, A. A., May, T., and Rozell, C. J. (**2016**). "Outcome measures based on classification performance fail to predict the intelligibility of binary-masked speech," J. Acoust. Soc. Am. **139**(6), 3033–3036.

Kressner, A. A., and Rozell, C. J. (**2015**). "Structure in time-frequency binary masking errors and its impact on speech intelligibility," J. Acoust. Soc. Am. **137**(4), 2025–2035.

Kressner, A. A., and Rozell, C. J. (**2016**). "Cochlear implant speech intelligibility outcomes with structured and unstructured binary mask errors," J. Acoust. Soc. Am. **139**(2), 800–810.

Li, N., and Loizou, P. C. (**2008**). "Factors influencing intelligibility of ideal binary-masked speech: Implications for noise reduction," J. Acoust. Soc. Am. **123**(3), 1673–1682.

Lyon, R. (**1996**). "All-pole models of auditory filtering," in *Proceedings of the International Symposium on Diversity in Auditory Mechanics*, June 24–28, Berkeley, CA, pp. 205–211.

May, T., Bentsen, T., and Dau, T. (**2015**). "The role of temporal resolution in modulation-based speech segregation," in *Proceedings of Interspeech*, September 6–10, Dresden, Germany, pp. 170–174.

May, T., and Dau, T. (**2013**). "Environment-aware ideal binary mask estimation using monaural cues," in *Proceedings of the 2013 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, October 20–23, New Paltz, NY.

May, T., and Dau, T. (**2014a**). "Computational speech segregation based on an auditory-inspired modulation analysis," J. Acoust. Soc. Am. **136**(6), 3350–3359.

May, T., and Dau, T. (**2014b**). "Requirements for the evaluation of computational speech segregation systems," J. Acoust. Soc. Am. **136**(6), EL398–EL404.

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**(7), 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**(1), 108–121.

Nielsen, J. B., and Dau, T. (**2009**). "Development of a Danish speech intelligibility test," Int. J. Audiol. **48**(10), 729–741.

Nielsen, J. B., and Dau, T. (**2011**). "The Danish hearing in noise test," Int. J. Audiol. **50**(3), 202–208.

Taal, C. H., Hendriks, R. C., Heusdens, R., and Jensen, J. (**2011**). "An algorithm for intelligibility prediction of timefrequency weighted noisy speech," IEEE Trans. Audio, Speech, Lang. Process. **19**(7), 2125–2136.

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**(3), 184–192.

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

Wang, D., Kjems, U., Pedersen, M. S., Boldt, J. B., and Lunner, T. (**2008**). "Speech perception of noise with binary gains," J. Acoust. Soc. Am. **124**(4), 2303–2307.

Wang, Y., Narayanan, A., and Wang, D. (**2014**). "On training targets for supervised speech separation," IEEE/ACM Trans. Audio, Speech, Lang. Process. **22**(12), 1849–1858.