A model for multitalker speech perception

Soundararajan Srinivasan[a)]
Biomedical Engineering Department, The Ohio State University, Columbus, Ohio 43210

DeLiang Wang[b)]
Department of Computer Science and Engineering and Center for Cognitive Science, The Ohio State University, Columbus, Ohio 43210

(Received 5 November 2007; revised 6 August 2008; accepted 18 August 2008)

A listener's ability to understand a target speaker in the presence of one or more simultaneous competing speakers is subject to two types of masking: energetic and informational. Energetic masking takes place when target and interfering signals overlap in time and frequency resulting in portions of target becoming inaudible. Informational masking occurs when the listener is unable to distinguish target and interference, while both are audible. A computational model of multitalker speech perception is presented to account for both types of masking. Human perception in the presence of energetic masking is modeled using a speech recognizer that treats the masked time-frequency units of target as missing data. The effects of informational masking are modeled as errors in target segregation by a speech separation system. On a systematic evaluation, the performance of the proposed model is in broad agreement with the results of a recent perceptual study. © *2008 Acoustical Society of America.* [DOI: 10.1121/1.2982413]

PACS number(s): 43.71.An, 43.72.Dv, 43.72.Ne [DOS] Pages: 3213–3224

I. INTRODUCTION

In everyday listening conditions, the acoustic input reaching our ears is often a mixture of multiple sound sources. In such situations, the human ability to perceive a target source is susceptible to the effects of masking, which is defined as the increase in the audibility threshold of the target (Mayer, 1876). In particular, our ability to attend to and understand a target speaker in the presence of other competing speakers is affected by energetic and informational masking (Tanner, 1958; Brungart *et al.*, 2001). Energetic masking refers to the phenomenon in which a stronger signal masks a weaker one within a critical band (Fletcher, 1940). Recently, the term informational masking has been used to refer to the perceptual degradation caused by the listener's inability to segregate a target signal from interference (Carhart *et al.*, 1969; Pollack, 1975) (for a review, see Watson, 2005). Informational masking, therefore, depends on the similarity of segregation cues such as voice characteristics (Brungart *et al.*, 2001) and spatial locations (Freyman *et al.*, 1999) associated with individual signals. In this paper, we propose a model for recognizing target speech in the presence of both energetic and informational masking under monaural conditions.

Spectrotemporal overlap between target and interference is a prime cause of energetic masking. Portions of a target signal subject to energetic masking become inaudible at the periphery of the auditory system and are unavailable for subsequent processing. A missing-data speech recognizer (Cooke *et al.*, 2001) is therefore used to model speech perception under energetic masking conditions. When target speech is contaminated by additive interference, some time-frequency (T-F) regions are dominated by target energy while some of the rest are dominated by interference. The missing-data method treats the former T-F units as reliable and the latter T-F regions as missing or unreliable during recognition.

The missing-data recognizer requires a T-F mask (typically binary) that provides information about which T-F regions of the mixture signal are reliable and unreliable. The task of generating such a mask is akin to the task of segregating the target from the mixture. The process by which the human auditory system is able to organize the acoustic input into components that correspond to individual sources in the input is known as auditory scene analysis (ASA) (Bregman, 1990). Therefore, informational masking is closely related to ASA. Hence, we adapt a monaural computational ASA (CASA) system to estimate a binary mask that selects those T-F regions of the mixture where the target is stronger than the interference (Hu and Wang, 2004). The system of Hu and Wang (2004) is a voiced-speech segregation system that utilizes differences in periodicity between target and interference. The similarities between target and interference characteristics degrade the performance of the CASA system and therefore contribute to informational masking in our model.

Lippmann and Carlson (1997) used *a priori* reliabilities of T-F regions to examine the effects of noise, low- and high-pass filtering, and interruptions by periodic silent gaps on the performance of a missing-data recognizer on a digit recognition task. This is then compared to human performance on a consonant-vowel-consonant syllable recognition task. Cooke (2006) also used a missing-data recognizer to model the effects of energetic masking on listeners' perception in babble noise. His model uses *a priori* knowledge of target-dominant T-F regions or "glimpses." In addition to

[a)]Present address: Robert Bosch LLC, Research and Technology Center North America, Pittsburgh, PA 15212. Electronic mails: srinivasan.36@osu.edu and soundar.srinivasan@us.bosch.com
[b)]Electronic mail: dwang@cse.ohio-state.edu

modeling energetic masking, Barker and Cooke (2004) modeled informational masking by performing grouping using trained speech models in a top-down manner (see Barker *et al.*, 2005). The present study addresses the problem of modeling the effects of both energetic and informational masking on multitalker speech perception by combining speech segregation based on *a priori* pitch and missing-data recognition (see Srinivasan and Wang, 2005a for an earlier version).

The model proposed here could also serve as an architecture for robust speech recognition in the presence of interfering speech sources. It is well known that the performance of automatic speech recognizers (ASRs) degrades rapidly in the presence of other sound sources (Huang *et al.*, 2001; Srinivasan, 2006). Speech recognizers are typically trained in an environment containing a single speech source and face a problem of mismatch when used in conditions where target speech occurs simultaneously with other sources. To mitigate the effect of this mismatch on recognition, "noisy" speech is typically preprocessed by speech separation systems. However, in many realistic applications, the output of typical speech segregation algorithms contains distortions in segregated speech not seen during ASR training. These distortions cause substantial degradation in recognition performance (Cooke *et al.*, 2001). Researchers have previously shown that combining binaural speech segregation with missing-data methods can improve target speech recognition under multitalker conditions (Roman *et al.*, 2003; Palomaki *et al.*, 2004). Unlike these binaural systems, the model presented here combines monaural target segregation and missing-data recognition. Our evaluations show that the proposed model can improve robust speech recognition under monaural multitalker conditions.

Models such as the articulation index (Fletcher, 1953) and the speech-transmission index (Steeneken and Houtgast, 1980) predict speech intelligibility in the presence of noise and other certain distortions. As pointed out by Cooke (2006), the use of an ASR system as a component in a model of human speech recognition additionally enables it to predict listener responses on a per-utterance basis. The purpose of the present study is twofold. First, similar to Cooke (2006), the proposed system is used to model listener performance on a per-utterance basis using a common multitalker speech corpus. Second, the proposed model is intended to serve as a potential paradigm for automatic speech recognition in multitalker situations.

The rest of the paper is organized as follows. In the next section, we briefly review a recent study that systematically examined the degradation in speech perception caused by energetic and informational masking using a binary masking procedure (Chang, 2004; Brungart *et al.*, 2006). Section III contains a detailed presentation of our proposed model. The model has been systematically evaluated on the same task used in the perceptual study presented in Sec. II. The evaluation results and a comparison with listener performance are presented in Sec. IV. Finally, conclusions are given in Sec. V.

II. ENERGETIC AND INFORMATIONAL MASKING EFFECTS IN MULTITALKER SPEECH PERCEPTION

A recent study (Chang, 2004; Brungart *et al.*, 2006) uses binary T-F masks to isolate the effects of energetic and informational masking on the intelligibility of a target speech signal in the presence of one or more competing speech signals. Specifically, Brungart *et al.* (2006) utilized an ideal binary mask that is obtained from premixed target and interference as follows. A unit in the ideal binary mask is assigned a value 1 if the signal-to-noise ratio (SNR) within the corresponding T-F unit exceeds a predefined local SNR criterion (LC) value; it is labeled 0 otherwise. To generate this mask, target and interfering signals are first analyzed using a 128-channel gammatone filterbank whose center frequencies are quasilogarithmically spaced from 80 Hz to 5 kHz (Patterson *et al.*, 1988). The energy at the output of each filter is calculated every frame for both target and interference. Each rectangular frame is 20 ms long with a 10 ms frame shift (a frame rate of 100 Hz). The ideal binary mask is used to resynthesize a signal by retaining only those T-F units in the mixture where the local SNR exceeds the specified LC value. Specifically, the gammatone filter responses of a mixture are weighted by the binary mask and summed across frequencies (after accounting for across-filter phase shifts) to yield a resynthesized signal. The resynthesized signal is then used in a series of experiments to study the effects of the number of interfering talkers and their sex on the two types of masking.

The speech corpus used by Chang (2004) and Brungart *et al.* (2006) is the coordinate response measure (CRM) corpus (Bolia *et al.*, 2000). This corpus consists of utterances from four male and four female speakers produced according to the grammar, "READY ⟨$call-sign⟩ GO TO ⟨$color⟩ ⟨$digit⟩ NOW." There are eight call signs, four colors, and eight numbers (Bolia *et al.*, 2000) and the target utterance always contains the call sign. "BARON;" e.g., "READY BARON GO TO RED ONE NOW." The interference utterance consists of a call sign, a color, and a number different from that of the target. Figure 1 shows the effect of applying the ideal binary mask to a mixture of two speech utterances from this corpus. Figures 1(a) and 1(b) show the cochleagrams of a target speech utterance and an interference utterance, respectively. A cochleagram is a T-F representation of a signal analogous to a spectrogram and is generated using the gammatone filterbank decomposition of a signal as described before (see Wang and Brown, 2006). The target signal corresponds to a male speech utterance, "READY BARON GO TO BLUE ONE NOW." The interference corresponds to a female utterance, "READY ARROW GO TO RED THREE NOW." Figure 1(d) shows the cochleagram of a mixture of target and interference at 0 dB SNR. The ideal binary mask for this mixture corresponding to a LC value of 0 dB is shown in Fig. 1(c). T-F units labeled 1 in the mask are shown in black and white represents the T-F units labeled 0. This mask is applied to the mixture in Fig. 1(d) and the results are presented in Fig. 1(e). Note that the application of the binary mask results in the removal of the interference-dominant T-F units from the mixture.

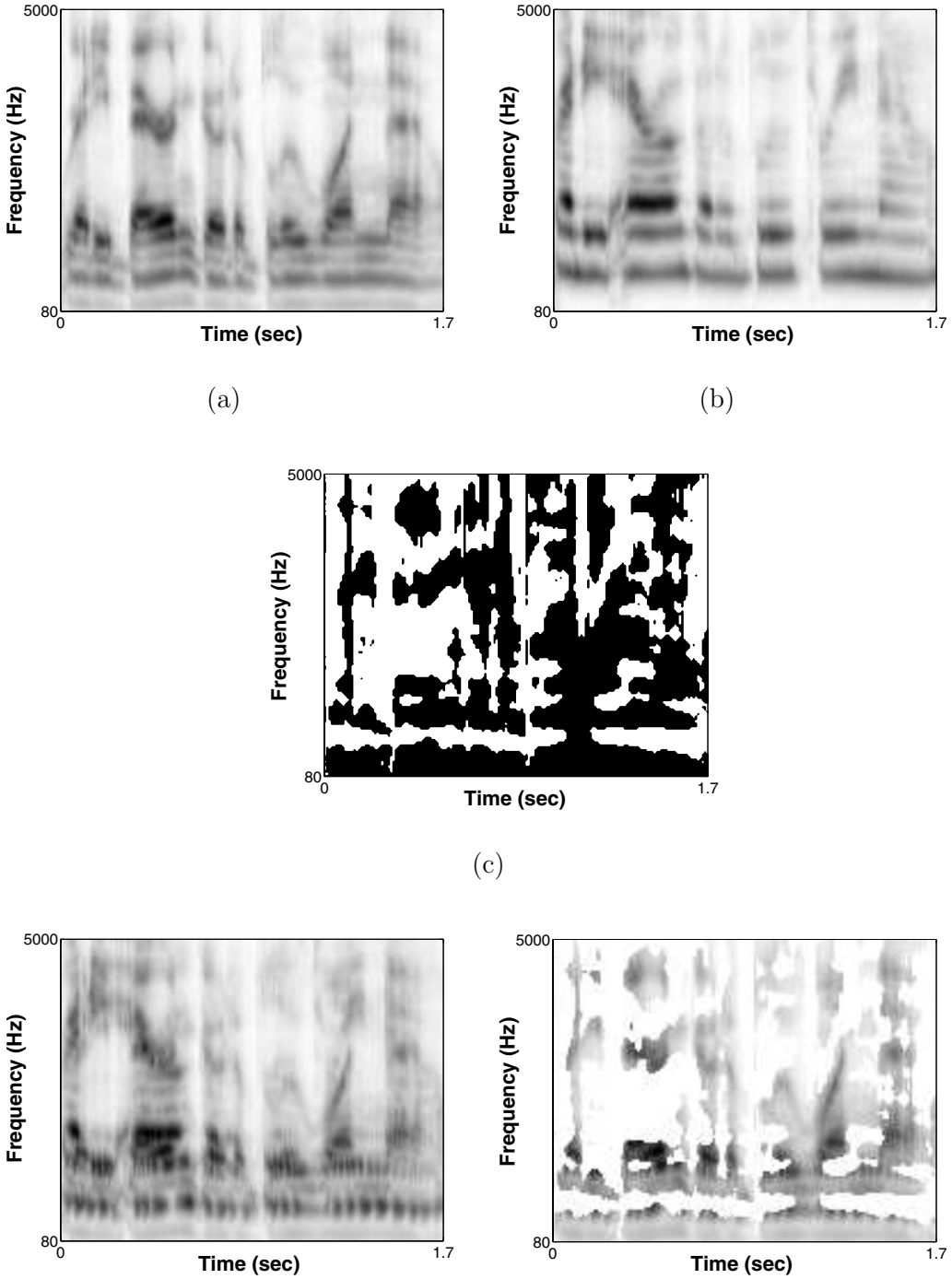


FIG. 1. An illustration of the ideal binary mask for a mixture of a male target utterance and a female interference. (a) The cochleagram of the male target utterance. (b) The cochleagram of the female interference utterance. (c) An ideal binary mask at 0 dB LC value. The target-dominant T-F units are marked black and the interference-dominant T-F units are marked white. (d) The cochleagram of the mixture. (e) The cochleagram obtained from (d) by applying the ideal mask in (c).

Using signals resynthesized by applying such ideal binary masks to multitalker speech mixtures, Brungart *et al.* (2006) asked listeners to identify the keywords (the color and the number) in the target phrase. The target-to-masker ratio (TMR) in the mixtures was fixed at 0 dB. While SNR refers to the ratio of target energy and combined interference energy, TMR refers to the ratio of target energy and energy of one interference in the mixture (interference signals have equal energy) (Brungart *et al.*, 2001). Figure 2 shows the effects of varying the number of competing talkers on the correct identification of keywords in the target phrase as a function of the LC value. The results are shown in terms of accuracy of recognizing both the color and the number in the target phrase at different LC values, ranging from −60 to +30 dB in steps of 3 dB. A control, "no mask" condition, is also included to assess listener performance directly on the

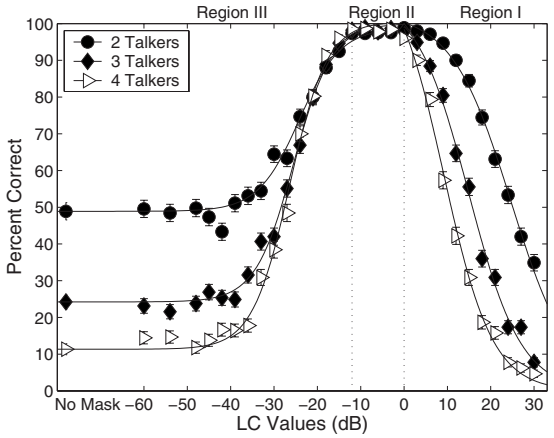


FIG. 2. Percentage of trials in which the listeners identified the keywords in the target phrase correctly (from Brungart *et al.*, 2006). The error bars represent 95% confidence intervals. The figure shows the effect of varying the number of competing talkers on listener performance as the LC value is varied in steps of 3 dB.

mixture in each multitalker condition. In other words, the signals generated in this control condition did not undergo any binary mask processing.

Based on the results in Fig. 2, Brungart *et al.* (2006) reached the following main conclusion. The information lost in the target signal due to energetic masking is argued to be approximately the same as that resulting from the resynthesis using an ideal binary mask with a LC value of 0 dB. Note that this mask removes all interference-dominant T-F units and, despite distortions caused by the removal of certain T-F regions of the mixture, the resulting intelligibility is quite high. Brungart *et al.* (2006) also suggested that the signal resynthesized from an ideal binary mask with a positive LC value of δ dB (region I) causes energetic masking equivalent to that by a mixture with SNR of $-\delta$ dB [see also Fig. 9(a)]. Therefore, the decrease in listener performance in region I as LC increases shows the increase in the deleterious effects of energetic masking. On the other hand, decreasing the LC value below 0 dB allows for the progressive introduction of interference energy in the resynthesized signal. This leads to an increase in the level of informational masking. Hence, energetic masking effects dominate listener perception in region I, while informational masking effects dominate in region III of Fig. 2. The signals employed in region II include some T-F units dominated by interference. However, listener performance seems to be largely unaffected by the interference contained in those units. Incidentally, region II, where near perfect performance is obtained, is centered at a LC value of −6 dB.

III. A COMPUTATIONAL MODEL

Listener performance gradually degrades as the LC value decreases in region III of Fig. 2. We propose to attribute this degradation to performance limitations of ASA. In other words, listeners use differences in the voice characteristics of target and interfering talkers to segregate the target speaker with varying degrees of success. Additionally, region II demonstrates the robust performance of listeners in the presence of energetic masking as isolated by the use of ideal binary masks. These observations motivate our model for multitalker speech perception shown in Fig. 3. The proposed model is a two-stage system that combines a CASA system with a missing-data ASR. The input to the model in Fig. 3 is a mixture of target and interference, sampled at 20 kHz. We use the same auditory filterbank decomposition of the input signal as used by Brungart *et al.* (2006) (see Sec. II). The output is used to generate feature vectors for recognition and as input to a monaural CASA system (Sec. III A). In our model, the monaural CASA system of Hu and Wang (2004) is adapted to segregate target from interference. The computational goal of the CASA system is an ideal binary mask (Wang, 2005). Human perception in the presence of energetic masking is modeled using a missing-data ASR that treats the masked T-F regions as missing. The similarities between target and interference cause, however, deviations in the estimation of the ideal binary mask by the CASA system. This corresponds to simulated informational masking in our model.

Figure 4 illustrates the effect of errors in the estimation of the ideal binary mask. Figure 4(a) shows the cochleagram of the same mixture shown in Fig. 1(d). Figure 4(b) shows the ideal binary T-F mask generated at the LC value of 0 dB. As mentioned before, this mask removes all interference-dominant T-F units. Figure 4(c) shows an estimated binary mask produced by the CASA system. Due to errors in target segregation, this mask contains many interference-dominant T-F units. Application of this mask to the cochleagram in Fig. 4(a), therefore, retains some interference energy in the masked mixture as shown in Fig. 4(e). Figure 4(d) shows the cochleagram obtained from (a) by applying the ideal mask in

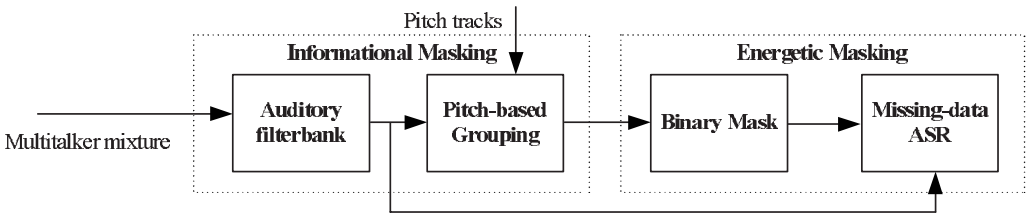


FIG. 3. Schematic of the proposed model. The input mixture signal is analyzed by an auditory filterbank in successive time frames. The output is fed to a monaural CASA system that uses pitch tracks of the individual sources in the mixture to produce a binary mask that selects the T-F regions in the mixture where target dominates interference. This mask is used by the missing-data recognizer to decode the input.

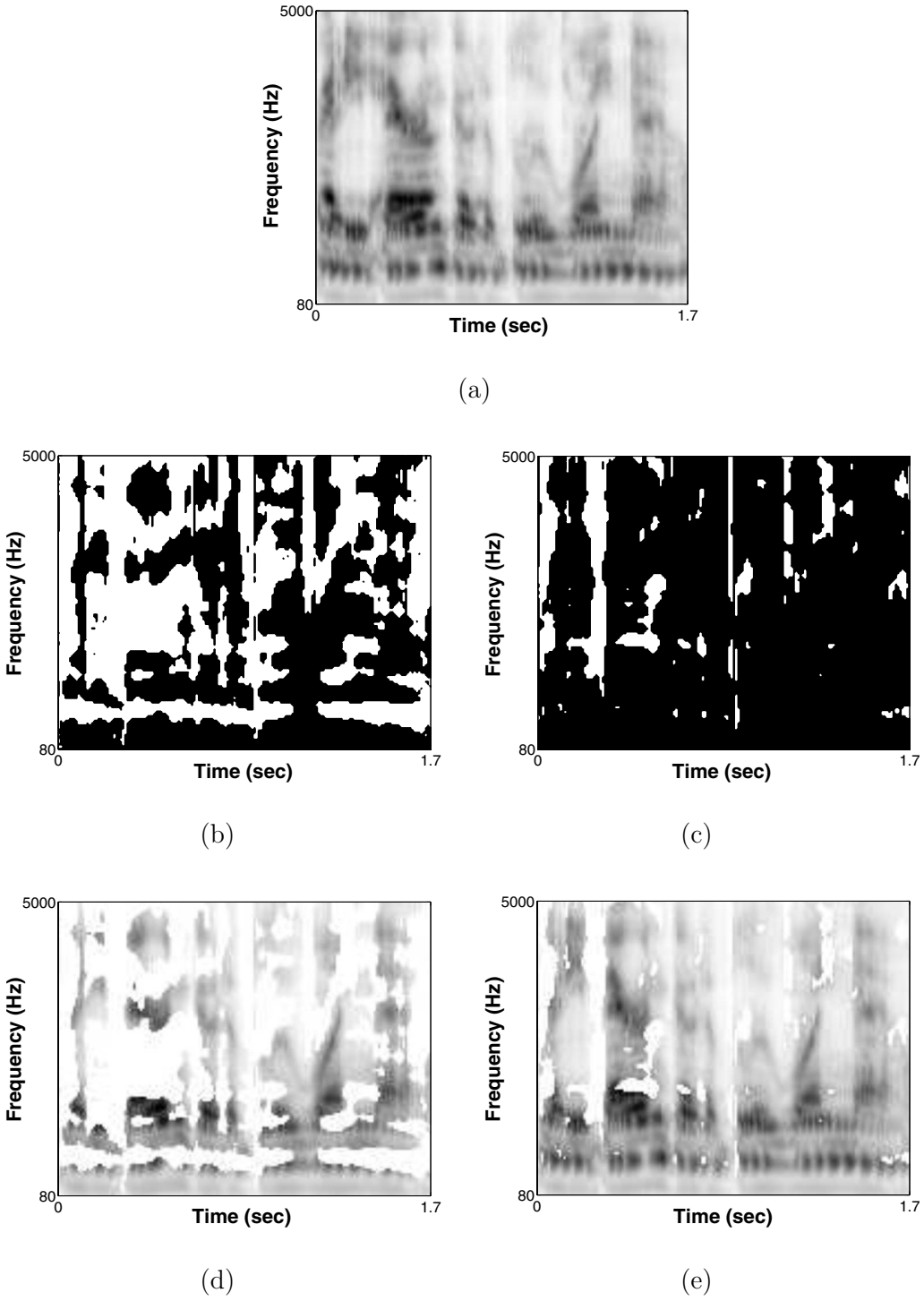


FIG. 4. An illustration of informational masking caused by deviations in target segregation via binary mask estimation. (a) The cochleagram of the mixture. (b) The ideal binary mask at 0 dB LC value. (c) An estimated binary mask. The target-dominant T-F units in (b) and (c) are marked black and the interference-dominant T-F units are marked white. (d) The cochleagram obtained from (a) by applying the ideal mask in (b). (e) The cochleagram obtained from (a) by applying the estimated mask in (c).

(b). The degradation in the model performance when using the cochleagram in Fig. 4(e) compared to the one in Fig. 4(d) is attributed to informational masking.

A. Pitch-based segregation of target

Under monaural conditions, the human auditory system is believed to segregate a target speech source from various interferences using several primitive cues, including differences in pitch (Bird and Darwin, 1997; Bregman, 1990; Brokx and Nooteboom, 1982) and onset, as well as prior knowledge (Bregman, 1990). Pitch has been successfully used for segregation of voiced speech in many CASA systems (Brown and Cooke, 1994; Hu and Wang, 2004). Here, we adapt the speech separation system by Hu and Wang (2004) to segregate target from interference. This system is chosen as it shows robust performance when tested with a

variety of intrusions. This system has two main stages: segmentation and grouping. In segmentation, the input signal is decomposed into a collection of contiguous T-F regions that are dominated by only one sound source. During grouping, the segments that likely belong to the same source are grouped together.

Consistent with psychophysical studies (Bird and Darwin, 1997), the system of Hu and Wang (2004) applies different processing strategies in the low- and high-frequency ranges. It has been shown that when neighboring channels respond to the same source, their autocorrelation responses are similar (Brown and Cooke, 1994; Wang and Brown, 1999). Therefore, cross correlation between adjacent channels indicates whether the filter channels respond to the same source. Hence, in the low-frequency range ($<$800 Hz), the system generates segments based on temporal continuity and cross-channel correlation. At higher frequencies, due to its wider bandwidth relative to harmonic spacing, a gammatone filter responds to multiple harmonics. Hence, the system uses the envelope characteristics of gammatone filter responses in the high-frequency range. Specifically, the cross-channel correlation of envelopes is used for segmentation along with temporal continuity.

For grouping segments, the system uses similarity in periodicity. The autocorrelation of a filter response in a frame encodes the periodicity within the corresponding T-F unit. Hence, if in a T-F unit the maximum autocorrelation value within the plausible pitch range has a lag consistent with the pitch lag of the target source in that frame, it is labeled as target-dominant (or 1); it is labeled interference dominant (or 0) otherwise (Hu and Wang, 2004). Furthermore, at high frequencies, a response envelope fluctuates at a rate consistent with the frequency of the dominant pitch, and amplitude modulation rates are therefore used for grouping (see also Cooke, 1993). To illustrate the potential of the proposed approach for segregation, pitch tracks and pitch strengths are derived *a priori* from premixed target and interfering signals using PRAAT (Boersma and Weenink, 2002). Note that robust multipitch tracking of two or more sources is currently a challenging problem (Wang and Brown, 2006).

Hu and Wang (2004) utilized only target pitch contours for grouping. However, listeners appear to utilize the pitch information of interfering sources too (de Cheveigne, 1997). In particular, the results of Culling *et al.* (2005) suggest that the auditory system can "perceptually remove" the harmonic components of at least one interference. Therefore, we adapt the system of Hu and Wang (2004) by expanding the grouping procedure as follows. First, we use the system of Hu and Wang (2004) to group segments based on the dominant pitch in a time frame. The dominant pitch in a frame is the one that has the highest associated pitch strength, which is measured as the height of the normalized autocorrelation value at the pitch lag. Second, the identity of the dominant pitch is then used to derive the final binary T-F mask in the following fashion. If the dominant pitch at a particular frame belongs to the interferer, we discard the grouped segments in that frame. Specifically, the grouped segments in that frame are set to 0 in the mask and the rest of the T-F units are set to 1 (see also de Cheveigne, 2005 for alternate approaches). On the other hand, if the dominant pitch belongs to the target, the grouped T-F units are retained. In this case, the grouped T-F units are labeled 1 in the mask, and other T-F units in that frame are labeled 0. We find that the proposed system performs better than the original system of Hu and Wang (2004). The performance improvement is especially significant when the original mixture SNR is high. In this case, canceling harmonic portions of interference helps retain more target-dominant T-F units than grouping harmonic portions of target alone. Figure 5 shows a comparison between the proposed system and the system of Hu and Wang (2004). The results are generated using the same target and interference in Figs. 1(a) and 1(b), mixed at a SNR of 20 dB. Figure 5(a) shows the binary mask containing the segregated voiced portions by primarily canceling the interference-dominant T-F units. The segregated T-F units are shown in black. All other T-F units are shown in white. Figure 5(b) shows the mask generated by the system of Hu and Wang (2004). For comparison, the ideal binary mask corresponding to this mixture, generated at 0 dB LC value, is shown in Fig. 5(c). Notice that the system of Hu and Wang (2004) wrongly labels some target-dominant T-F units as 0.

Figure 6 shows how the binary mask corresponding to the target is generated by using pitch strengths of both target and interference. Figure 6(a) shows a binary mask containing the segregated voiced portions corresponding to the dominant pitch in each frame. The segregated T-F units are labeled 1 and shown in black. All other T-F units are labeled 0 and shown in white. This mask is generated by using the mixture in Fig. 1(d) as input to the proposed CASA system. The dominant pitch in a frame may correspond to either target or interference. Figure 6(b) shows how pitch strengths of target and interference vary in the mixture. If the pitch strength of the interference is higher in a frame, mask labels are flipped; they are retained otherwise. The resulting mask is shown in Fig. 6(c). Currently, we do not process unvoiced frames. For such frames, all frequency units are labeled 0. The output of the CASA system in the form of a binary mask is then used by the missing-data ASR to recognize target speech.

B. Missing-data recognition

The feature vectors for the missing-data recognizer consist of the instantaneous Hilbert envelope at the output of each gammatone filter, smoothed using a first-order filter with 8 ms time constant and log compressed as suggested by Cooke *et al.* (2001). The missing-data recognizer is an HMM-based ASR that makes use of spectrotemporal redundancy in speech to recognize a noisy signal based on its target-dominant T-F units. Given a speech observation sequence X, the problem of word recognition is to maximize the posterior $P(W|X)$, where W is a valid word sequence. When parts of X are masked by noise or other distortions, a binary T-F mask can be used to partition X into its reliable and unreliable constituents as X_r and X_u, where $X=X_r \cup X_u$. The missing-data ASR treats the T-F regions labeled 0 as unreliable data during recognition. Specifically, it modifies the computation of the observation probability in a state of

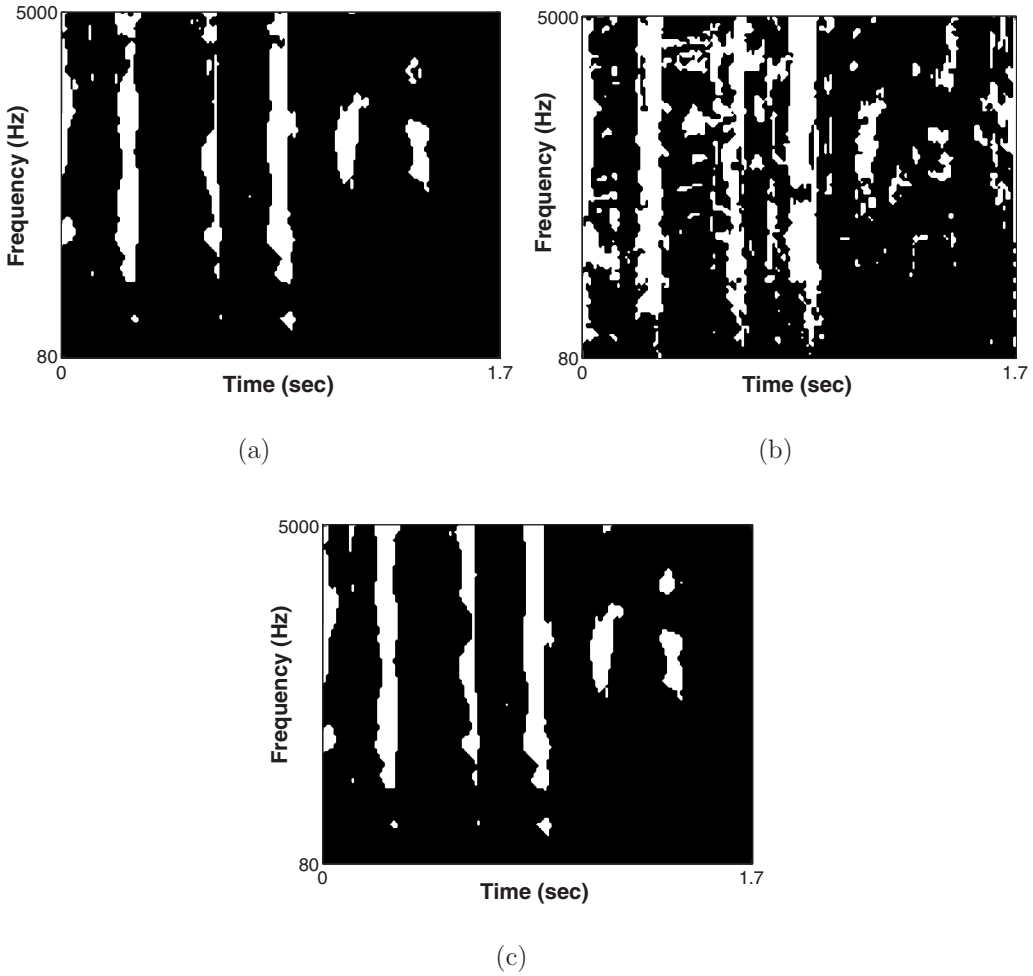


FIG. 5. An illustration of improved segregation using the proposed approach. (a). The binary mask generated by the proposed system. T-F units determined to be target-dominant are shown in black. T-F units consistent with the interference pitch are canceled and shown in white. (b) The binary mask produced by the system of Hu and Wang (2004). (c) The ideal binary mask.

an HMM (hidden Markov model) based ASR to handle missing or unreliable data (Cooke *et al.*, 2001). The observation density in an ASR is typically modeled using a mixture of Gaussians as follows:

$$p(x|q) = \sum_{k=1}^{M} p(k|q)p(x|k,q), \tag{1}$$

where x is the spectral energy feature vector in a frame, M is the number of mixture components, k is the mixture index, q is an HMM state, $p(k|q)$ is the mixture weight, and $p(x|k,q)=\mathcal{N}(x;\mu_{k,q},\Sigma_{k,q})$. Note that $\mathcal{N}(x;\mu_{k,q},\Sigma_{k,q})$ denotes that x follows a normal (Gaussian) distribution with mean $\mu_{k,q}$ and variance $\Sigma_{k,q}$. When parts of x are corrupted by interference, the missing-data ASR marginalizes unreliable feature dimensions in the computation of the likelihood in Eq. (1). Typically, the various dimensions of the feature vectors are modeled as independent given a mixture. Theoretically, this is a good approximation if an adequate number of mixtures are used (McLachlan and Basford, 1988). Hence, in the presence of unreliable data, the computation of the observation density is modified as

$$p(x|q) = \sum_{k=1}^{M} p(k|q) \prod_{j} p(x_{r,j}|k,q) \prod_{i} \int p(x_{u,i}|k,q)dx_{u,i}, \tag{2}$$

where $x_{r,j}$ and $x_{u,i}$ correspond to the spectral energies in a reliable (j) and an unreliable (i) feature dimension, respectively. Note that $p(x_{r,j}|k,q)=\mathcal{N}(x_{r,j};\mu_{k,q,j},\sigma^2_{k,q,j})$.

Furthermore, if the range for the true value of the unreliable feature is known, it provides bounds (limits) over which the unreliable feature is integrated. Under additive interference conditions, the true speech value $\tilde{x}_{u,i}$, in the unreliable part, may be constrained as $0 \leq \tilde{x}_{u,i} \leq y_{u,i}$ (Cooke *et al.*, 2001; Srinivasan and Wang, 2007), where $y_{u,i}$ is the observed (mixture) spectral energy. This constraint is then used as bounds on the integral in Eq. (2) as

$$p(x|q) = \sum_{k=1}^{M} p(k|q) \prod_{j} p(x_{r,j}|k,q) \prod_{i} \int_{0}^{y_{u,i}} p(x_{u,i}|k,q)dx_{u,i}. \tag{3}$$

This bounded marginalization method is shown in Cooke *et al.* (2001) to have a better recognition score than the

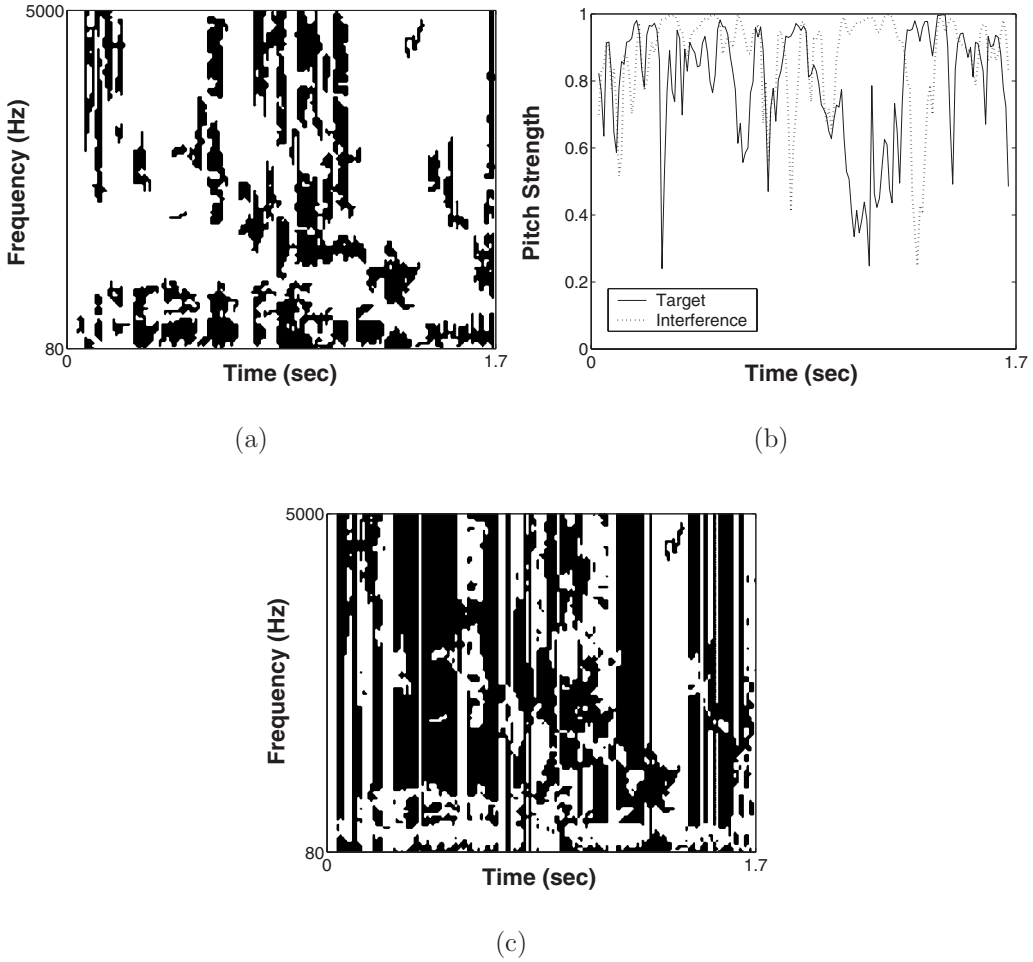


FIG. 6. An illustration of target segregation based on the dominant pitch. (a) The binary mask generated by the CASA system using the pitch cue. The input to the system is the mixture shown in Fig. 1(d). If the periodicity of a gammatone filter response is consistent with the dominant pitch in that frame, it is shown in black. All other T-F units are shown in white. (b) The variation in the strengths of target and interference pitch tracks across time. (c) The final mask is obtained from (a) by flipping the frame labels generated using the interference pitch.

simple marginalization method, and is hence used in all our experiments.

IV. EVALUATION RESULTS

To facilitate comparison with the behavioral data from Chang (2004) and Brungart *et al.* (2006), we have also evaluated our model using the CRM corpus (Bolia *et al.*, 2000). Twenty-three ("Ready," "Baron," "Charlie," "Arrow," "Laker," "Hopper," "Ringo," "Tiger," "Eagle," "Goto," "Blue," "Green," "Red," "White," "Now," and the numbers 1–8) speaker-independent, word-level HMMs are trained. All models have eight states, whose output distribution is modeled as a mixture of ten Gaussians with diagonal covariance. The models are trained using 1792 utterances from 3 male talkers (talkers 1–3) and 4 female talkers (talkers 4–7) in this database chosen arbitrarily. The testing data consist of utterances from a male talker (talker 0) not utilized during training and containing the call sign "BARON." Similar to the experiments by Brungart *et al.* (2006), one, two, and three utterances from the same talker are added to target as interference at 0 dB TMR. Thus the SNRs corresponding to two, three, and four talker mixture conditions are 0, −3, and −4.8 dB, respectively. Interference utterances contain call signs, numbers, and colors different from the target one. Each testing condition comprises 256 mixture utterances. An HMM toolkit, HTK (Young *et al.*, 2000), is used for training. For testing, a decoder incorporating the missing-data method is used. The task is to recognize both color and number in the

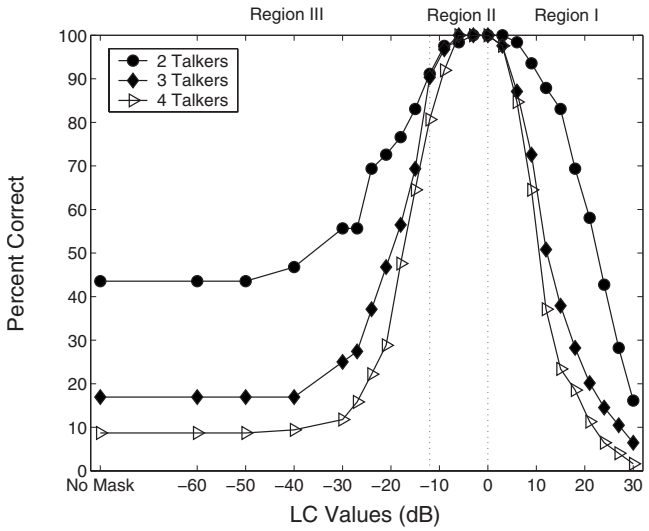


FIG. 7. Percentage of utterances in which the model identified the keywords in the target utterance correctly. The figure shows the model's performance with respect to the LC value under various multitalker conditions.

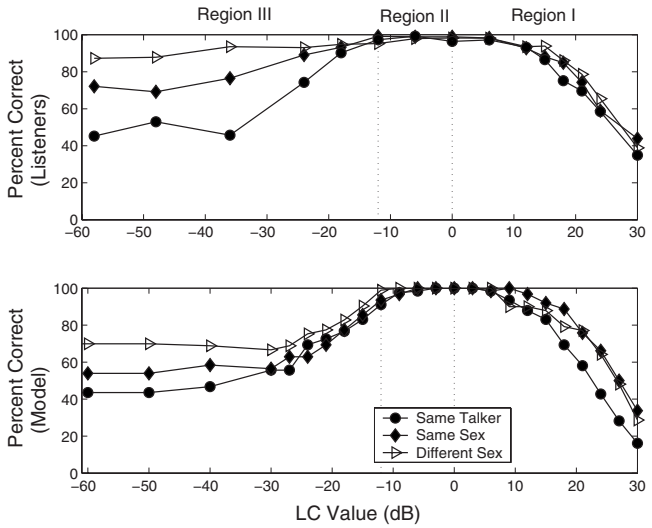


FIG. 8. The effect of voice characteristics on listener and model performance with two talkers as a function of the LC value. The top panel shows the performance of listeners in correctly identifying the keywords in the target utterance (from Chang, 2004). The bottom panel shows the performance of the proposed model on the same task.

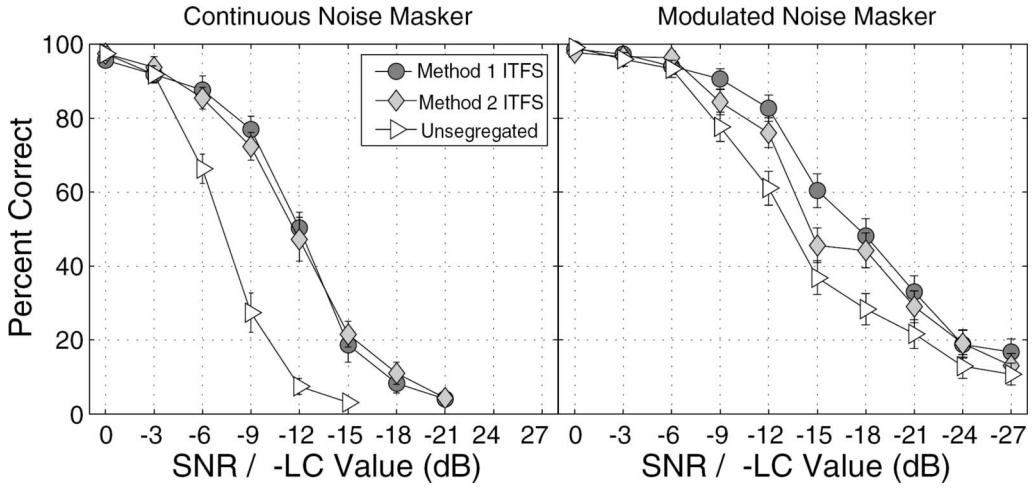
target utterance as described in Sec. II. We recognize the contents of the target utterance in the mixture by using the following grammar: "READY BARON GOTO ⟨$color⟩ ⟨$digit⟩ NOW." Recognition performance on the target-only test-data input is 100%. Note that the chance performance on this task is 3.125%.

Similar to the procedure employed by Brungart *et al.* (2006), the multitalker mixture is resynthesized using ideal binary masks with varying LC values (see also Sec. II). The resynthesized signal is used as input to the proposed model. Figure 7 shows the performance of the proposed model as a function of the LC value under different multitalker conditions. Similar to the evaluation used by Brungart *et al.* (2006), performance in Fig. 7 is shown in terms of percentage of utterances in which the model correctly identified all the keywords in the target phrase.

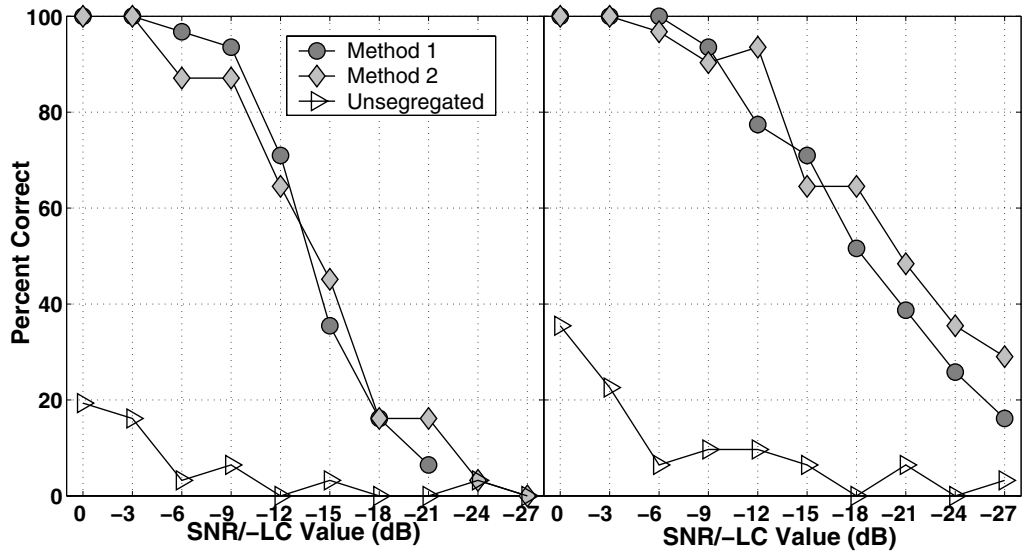
While the absolute recognition rates at some LC values differ from human performance, the model is able to simulate the general pattern of systematic listener performance seen in Fig. 2. As in Fig. 2, informational masking can be seen to have a larger effect on the recognition performance. The model saturates to the ceiling performance in region II. The improvement compared to the no mask condition shows that the removal of interference-dominant T-F units from the input contributes to substantial performance gains. As in Fig. 2, the peak performance is obtained in an interval of LC values approximately around −6 dB. As the number of talkers in the mixture increases, both model and human performance significantly degrade in region III. In comparison, smaller differences are seen in region I across the different multitalker conditions. Overall, the comparison with Fig. 2 shows that the proposed model is able to replicate listener perception under various multitalker conditions in the presence of both energetic (region I and parts of region II) and informational masking (region III).

Figure 8 shows the results of a second experiment in which the effects of voice characteristics of an interfering talker are examined. The top panel in Fig. 8 shows listener performance under three different interference conditions: Same talker, a different talker of the same sex, and a different sex talker (Chang, 2004). The bottom panel shows the performance of our model. In evaluating the model performance under the "same sex" condition, interference utterances are chosen randomly from talkers 1 to 3. For the "different sex" condition, interference utterances are chosen from talkers 4 to 7. The results of evaluation of our model and the human data show that the differences in voice characteristics between target and interference have only minor effects on the performance in regions I and II. The peak performance in both cases exhibits a plateau in region II. Recall that the SNR is the same across the three interference conditions. However, as with human listeners, the proposed model is able to utilize the larger differences in voice characteristics to obtain improved performance under the same sex and the different sex conditions in region III. Also, the model performance in the different sex condition is better that in the same sex condition. These indicate that the model is able to simulate the dependency of informational masking on the similarities in voice characteristics of target and interference.

Brungart *et al.* (2006) additionally examined the perception of a target speech utterance in the presence of nonspeech interferences. In this experiment, they considered two interfering sources: a continuous noise and a modulated noise. The continuous noise was generated by filtering a Gaussian noise source with the average long-term spectrum of all utterances in the CRM corpus. The modulated noise source was generated by further modulating the continuous noise masker with the envelope of a randomly chosen utterance in the CRM corpus (Brungart *et al.*, 2006). The resulting listener performance at various SNR and LC values is shown in Fig. 9(a). In "method 1," the SNR is fixed at 0 dB and the LC value is varied in steps of 3 dB. In "method 2," the LC value is kept constant at 0 dB and the SNR is varied from 0 to 27 dB in steps of 3 dB. Additionally, Brungart *et al.* (2006) performed a control experiment in which the listener performance was assessed without any binary mask processing at the aforementioned SNRs (equivalent to the no mask condition described in Sec. II), referred to as "unsegregated" in Fig. 9(a). The results in Fig. 9(a) were also used to validate the assumption that a δ dB increase in the LC value at a fixed SNR in region I in Fig. 2 is equivalent to a decrease in mixture SNR of $-\delta$ dB at a fixed LC value. Figure 9(b) shows the model performance in the continuous and the modulated noise conditions. Note that for both model and listeners, the performances in method 1 and method 2 conditions are quite similar. While the model performance in the unsegregated condition is well below the listener performance, the model is able to simulate the increased intelligibility in the modulated noise condition compared to that in the continuous noise condition. Additionally, the model performance is still better than a direct recognition of mixture speech. For example, at 0 dB SNR, the model achieves 16% and 17% absolute improvements in accuracy over the baseline recognition performance. The mask estimated by the proposed CASA system incorrectly labels many interference-



(a)



(b)

FIG. 9. The effect of noise characteristics on listener and model performance. For method 1, the performance is shown as a function of the LC value, while for method 2 and the unsegregated conditions, the performance is shown with respect to SNR. (a) The figure shows the performance of listeners in correctly identifying the keywords in the continuous and modulated noise conditions (from Brungart *et al.*, 2006). (b) The corresponding performance of the proposed model.

dominant T-F units as reliable, causing insufficient target segregation. This is the cause of the performance gap (compared to human listeners) in the unsegregated condition.

As described in Sec. I, the second objective of the present study is to examine the efficacy of the proposed model for robust speech recognition under multitalker conditions. In Fig. 7 and in the bottom panel of Fig. 8, the no mask condition represents a control condition in which the mixture signal is used directly as input to the model. The model performance with this input, therefore, shows the improvement over the baseline recognition performance. Table I shows the performance of the proposed model in the no mask condition. Performance is again reported in terms of the percentage of utterances in which the model correctly identified both keywords spoken by the target speaker. As mentioned before, the interference consists of one, two, and three utterances from the same talker as target at a TMR of 0 dB. Table I also shows the baseline performance obtained by using the mixture directly as input to the ASR. Note that, by performing segregation based on *a priori* pitch, the model is able to improve significantly over the baseline performance across all multitalker conditions. The performance

TABLE I. Keyword recognition accuracy (%) of the proposed model with respect to the number of talkers in the mixture. For comparison, the baseline accuracy is also shown.

No. of talkers	Baseline performance	Model performance
2	15.6	43.6
3	6.2	16.9
4	4.3	8.7

TABLE II. Keyword recognition accuracy (%) of the proposed model across various interference conditions. For comparion, the baseline accuracy is also shown.

Interference type	Baseline performance	Model performance
Same talker	15.6	43.6
Same sex	17.9	52.3
Different sex	21.2	69.9

improvement is especially substantial in the two talker condition. This is because the SNR in the two talker condition is higher. Also, with additional talkers, the mixture becomes more similar to babble and segregation based on dominant pitch deteriorates.

Similarly, Table II compares the model and the baseline performance under same talker, same sex and different sex conditions. There are two talkers in the mixture at an SNR of 0 dB. The pitch-based grouping component of the model is able to utilize the larger differences between target and interference pitch contours under same sex and different sex conditions to produce better segregation results. This leads to a substantial improvement in the performance of the missing-data recognizer. In contrast, the baseline performance only changes slightly under those conditions.

V. DISCUSSION

We have presented a model for monaural multitalker speech perception that is able to account for the effects of both energetic and informational masking in multitalker conditions. We have conducted a systematic comparison between the model performance and listener results on a common speech corpus and shown that the performance of the proposed model is in broad quantitative agreement with behavioral data. We have also shown that differences between target and interference pitch ranges contribute to a reduction in informational masking (see also Oh and Lutfi, 2000) by improving target speech segregation. Hence, the performance of the model is better in the same sex and the different sex conditions compared to that in the same talker condition. Notice that the best performance, for both human listeners and the proposed model, is obtained under the different sex condition. This is expected due to the difference in voice characteristics between target and interference being largest under this condition.

Our study also demonstrates that combining target speech segregation via CASA and missing-data recognition can provide significant improvement over baseline recognition performance. Hence, the proposed model shows potential for robust speech recognition under multitalker conditions. The present study, therefore, confirms previous findings that the harmonicity of voiced speech can be successfully exploited to estimate masks for missing-data recognition of monaural noisy speech mixtures (Seltzer *et al.*, 2000; Brown *et al.*, 2001; van Hamme, 2004). Additionally, our study extends previous studies to multitalker conditions.

The results of the present study could be used in conjunction with the findings of Brungart *et al.* (2006) and Chang (2004) to design appropriate binary masking thresholds for CASA systems. Note that CASA systems must estimate the ideal T-F mask directly from the mixture. As in the present study, the computational goal of most current CASA systems is an ideal binary mask generated an LC value of 0 dB. The results of Brungart *et al.* (2006) and Chang (2004) suggest that speech segregated by using a −6 dB LC value may be a better choice for improving speech intelligibility for human listeners. However, our results indicate that for missing-data recognition, this choice of LC value may only be appropriate in the two talker condition. For more than two talkers, the peak performance is centered at a value higher than −6 dB.

Our model utilizes only pitch information for grouping and obtains large improvements in recognition accuracy. The use of other ASA cues including common onset should help further enhance segregation, especially for unvoiced speech (Hu and Wang, 2005, 2007). These primitive grouping mechanisms may also be supplemented by schema-based, top-down segregation (Barker *et al.*, 2005; Srinivasan and Wang, 2005b, 2005c). For example, top-down processing may play an important role in segregation of unvoiced consonants. Top-down processing may also help in segregation of voiced speech when SNR is low. In general, top-down processing provides prior information for grouping, and in a complete system, top-down and bottom-up CASA should probably interact for maximal performance gains.

The model also needs to address sequential grouping of sources across time (Bregman, 1990). In our current work, we have avoided this problem by utilizing the *a priori* pitch information and the target grammar. While solutions based on speech or speaker models have been proposed (Barker *et al.*, 2005; Shao and Wang, 2006), such solutions are currently limited to mixtures of one talker and nonspeech interference or mixtures of two talkers. Future work needs to develop a general solution for arbitrary speech mixtures.

Note that for use as a robust speech recognition system, the main limitation is the assumption of *a priori* pitch tracks. While algorithms for forming multiple continuous pitch contours exist (Wu *et al.*, 2003), the problem of sequential grouping of pitch contours into pitch tracks has been little addressed and presents a major challenge for CASA systems. Recall that the performance improvements reported in this study come despite not segregating unvoiced speech. Segregation of unvoiced speech, as mentioned above, should further improve recognition results. Additional improvements can also be obtained by using binary T-F masks in an uncertainty decoding approach to robust speech recognition (Srinivasan and Wang, 2007).

ACKNOWLEDGMENTS

This research was supported in part by an AFOSR grant (FA9550-04-1-0117), an NSF grant (IIS-0534707), and an AFRL grant via Veridian. We thank P. S. Chang, D. S. Brungart, and L. L. Feth, for helpful discussions. A preliminary version of this work was presented in 2005 Interspeech.

Barker, J., and Cooke, M. P. (**2004**). "Modelling the intelligibility of multitalker speech in the CRM task," presented on the *International Conference on Auditory Scene Analysis and Speech Perception by Human and Ma-*

chine, Hanse Institute for Advanced Studies, Delmenhorst, Germany.
Barker, J. P., Cooke, M. P., and Ellis, D. P. W. (**2005**). "Decoding speech in the presence of other sources," Speech Commun. **45**, 5–25.
Bird, J., and Darwin, C. J. (**1997**). "Effects of a difference in fundamental frequency in separating two sentences," in *Psychophysical and Physiological Advances in Hearing*, edited by A. R. Palmer, A. Rees, A. Q. Summerfield, and R. Meddis (Whurr, London, UK), pp. 263–269.
Boersma, P., and Weenink, D. (**2002**). "PRAAT: Doing phonetics by computer, version 4.0.26," http://www.fon.hum.uva.nl/praat (last viewed October, **2007**).
Bolia, R. S., Nelson, W. T., and Ericson, M. A. (**2000**). "A speech corpus for multitalker communications research," J. Acoust. Soc. Am. **107**, 1065–1066.
Bregman, A. S. (**1990**). *Auditory Scene Analysis* (MIT, Cambridge, MA).
Brokx, J. P. L., and Nooteboom, S. G. (**1982**). "Intonation and the perceptual separation of simultaneous voices," J. Phonetics **10**, 23–36.
Brown, G. J., Barker, J., and Wang, D. L. (**2001**). "A neural oscillator sound separator for missing data speech recognition," in Proceedings of the International Joint Conference on Neural Networks'01, pp. 2907–2912.
Brown, G. J., and Cooke, M. P. (**1994**). "Computational auditory scene analysis," Comput. Speech Lang. **8**, 297–336.
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.
Brungart, D. S., Simpson, B. D., Ericson, M. A., and Scott, K. R. (**2001**). "Informational and energetic masking effects in the perception of multiple simultaneous talkers," J. Acoust. Soc. Am. **110**, 2527–2538.
Carhart, R., Tillman, T. W., and Greetis, E. S. (**1969**). "Perceptual masking in multiple sound backgrounds," J. Acoust. Soc. Am. **45**, 694–703.
Chang, P. S. (**2004**). "Exploration of behavioral, physiological, and computational approaches to auditory scene analysis," Master's thesis, Department of Computer Science & Engineering, The Ohio State University; http://www.cse.ohio-state.edu/pnl/theses.html (last viewed October, 2007).
Cooke, M. P. (**1993**). *Modeling Auditory Processing and Organization* (Cambridge University Press, Cambridge, UK).
Cooke, M. P. (**2006**). "A glimpsing model of speech perception in noise," J. Acoust. Soc. Am. **119**, 1562–1573.
Cooke, M. P., Green, P., Josifovski, L., and Vizinho, A. (**2001**). "Robust automatic speech recognition with missing and unreliable acoustic data," Speech Commun. **34**, 267–285.
Culling, J. F., Linsmith, G. M., and Caller, T. L. (**2005**). "Evidence for a cancellation mechanism in perceptual segregation by differences in fundamental frequency," J. Acoust. Soc. Am. **117**, 2600.
de Cheveigne, A. (**1997**). "Concurrent vowel identification. III. A neural model of harmonic interference cancellation," J. Acoust. Soc. Am. **101**, 2857–2865.
de Cheveigne, A. (**2005**). "The cancellation principle in acoustic scene analysis," in *Speech Separation by Humans and Machines*, edited by P. Divenyi (Kluwer Academic, Norwell, MA), pp. 245–259.
Fletcher, H. (**1940**). "Auditory patterns," Rev. Mod. Phys. **12**, 47–65.
Fletcher, H. (**1953**). *Speech and Hearing in Communication* (Van Nostrand, Princeton, NJ).
Freyman, R. L., Helfer, K. S., McCall, D. D., and Clifton, R. K. (**1999**). "The role of perceived spatial separation in the unmasking of speech," J. Acoust. Soc. Am. **106**, 3578–3588.
Hu, G., and Wang, D. L. (**2004**). "Monaural speech segregation based on pitch tracking and amplitude modulation," IEEE Trans. Neural Netw. **15**, 1135–1150.
Hu, G., and Wang, D. L. (**2005**). "Separation of fricatives and affricates," in Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing'05, pp. 1101–1104.
Hu, G., and Wang, D. L. (**2007**). "Auditory segmentation based on onset and offset analysis," IEEE Trans. Audio, Speech, Lang. Process. **15**, 396–405.
Huang, X., Acero, A., and Hon, H. (**2001**). *Spoken Language Processing* (Prentice-Hall, Upper Saddle River, NJ).
Lippmann, R. P., and Carlson, B. A. (**1997**). "Using missing feature theory to actively select features for robust speech recognition with interruptions, filtering, and noise," in Proceedings of the European Conference on Speech Communication and Technology'97, pp. 37–40.
Mayer, A. M. (**1876**). "Research in acoustics," Philos. Mag. **2**, 500–507.
McLachlan, G. J., and Basford, K. E. (**1988**). *Mixture Models: Inference and Applications to Clustering* (Dekker, New York).
Oh, E. L., and Lutfi, R. A. (**2000**). "Effect of masker harmonicity on informational masking," J. Acoust. Soc. Am. **108**, 706–709.
Palomaki, K. J., Brown, G. J., and Wang, D. L. (**2004**). "A binaural processor for missing data speech recognition in the presence of noise and small-room reverberation," Speech Commun. **43**, 361–378.
Patterson, R. D., Nimmo-Smith, I., Holdsworth, J., and Rice, P. (**1988**). "An efficient auditory filterbank based on the gammatone function," MRC Applied Psychology Unit (APU) Report No. 2341, Cambridge, UK.
Pollack, I. (**1975**). "Auditory informational masking," J. Acoust. Soc. Am. **57**, Supplement 1, p. 55.
Roman, N., Wang, D. L., and Brown, G. J. (**2003**). "Speech segregation based on sound localization," J. Acoust. Soc. Am. **114**, 2236–2252.
Seltzer, M. L., Raj, B., and Stern, R. M. (**2000**). "Classifier-based mask estimation for missing feature methods of robust speech recognition," in Proceedings of the International Conference on Spoken Language Processing '00, pp. 538–541.
Shao, Y., and Wang, D. L. (**2006**). "Model-based sequential organization in cochannel speech," IEEE Trans. Audio, Speech, Lang. Process. **14**, 289–298.
Srinivasan, S. (**2006**). "Integrating computational auditory scene analysis and automatic speech recognition," Ph.D. thesis, Biomedical Engineering Department, The Ohio State University, Columbus, OH.
Srinivasan, S., and Wang, D. L. (**2005a**). "Modeling the perception of multitalker speech," in Proceedings of the Interspeech '05, pp. 1265–1268.
Srinivasan, S., and Wang, D. L. (**2005b**). "Robust speech recognition by integrating speech separation and hypothesis testing," in Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing '05, Vol. **1**, pp. 89–92.
Srinivasan, S., and Wang, D. L. (**2005c**). "A schema-based model for phonemic restoration," Speech Commun. **45**, 63–87.
Srinivasan, S., and Wang, D. L. (**2007**). "Transforming binary uncertainties for robust speech recognition," IEEE Trans. Audio, Speech, Lang. Process. **15**, 2130–2140.
Steeneken, H. J. M., and Houtgast, T. (**1980**). "A physical method for measuring speech-transmission quality," J. Acoust. Soc. Am. **67**, 318–326.
Tanner, W. P., Jr. (**1958**). "What is masking?" J. Acoust. Soc. Am. **30**, 919–921.
van Hamme, H. (**2004**). "Robust speech recognition using cepstral domain missing data techniques and noisy masks," in Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing '04, Vol. **1**, pp. 213–216.
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, Norwell, MA), pp. 181–197.
Wang, D. L., and Brown, G. J. (**1999**). "Separation of speech from interfering sounds based on oscillatory correlation," IEEE Trans. Neural Netw. **10**, 684–697.
Wang, D. L., and Brown, G. J. (**2006**). *Computational Auditory Scene Analysis: Principles, Algorithms and Applications* (Wiley, New York, IEEE, Hoboken, NJ).
Watson, C. S. (**2005**). "Some comments on informational masking," Acta. Acust. Acust. **91**, 502–512.
Wu, M., Wang, D. L., and Brown, G. J. (**2003**). "A multipitch tracking algorithm for noisy speech," IEEE Trans. Speech Audio Process. **11**, 229–241.
Young, S., Kershaw, D., Odell, J., Valtchev, V., and Woodland, P. (**2000**). *The HTK Book (for HTK Version 3.0)* (Microsoft Corporation, Redmond, WA).