An algorithm that improves speech intelligibility in noise for normal-hearing listeners

Gibak Kim, Yang Lu, Yi Hu, and Philipos C. Loizou[a)]
Department of Electrical Engineering, University of Texas at Dallas, Richardson, Texas 75080

(Received 30 October 2008; revised 27 March 2009; accepted 1 July 2009)

Traditional noise-suppression algorithms have been shown to improve speech quality, but not speech intelligibility. Motivated by prior intelligibility studies of speech synthesized using the ideal binary mask, an algorithm is proposed that decomposes the input signal into time-frequency (T-F) units and makes binary decisions, based on a Bayesian classifier, as to whether each T-F unit is dominated by the target or the masker. Speech corrupted at low signal-to-noise ratio (SNR) levels (−5 and 0 dB) using different types of maskers is synthesized by this algorithm and presented to normal-hearing listeners for identification. Results indicated substantial improvements in intelligibility (over 60% points in −5 dB babble) over that attained by human listeners with unprocessed stimuli. The findings from this study suggest that algorithms that can estimate reliably the SNR in each T-F unit can improve speech intelligibility.
 [DOI: 10.1121/1.3184603]

PACS number(s): 43.72.Ar, 43.72.Dv [MSS] Pages: 1486–1494

I. INTRODUCTION

Dramatic advances have been made in automatic speech recognition (ASR) technology (Rabiner, 2003). Despite these advances, human listener's word error rates are often more than an order of magnitude lower than those of state-of-the art recognizers in both quiet and degraded environments (Lippmann, 1997, Sroka and Braida, 2005; Scharenborg, 2007). Large advances have also been made on the development of algorithms that suppress noise without introducing much distortion to the speech signal (Loizou, 2007). These algorithms, however, have been shown to improve primarily the subjective quality of speech rather than speech intelligibility (Hu and Loizou, 2007a, 2007b). Speech quality is highly subjective in nature and can be easily improved, at least to some degree, by suppressing the background noise. In contrast, intelligibility is related to the underlying message or content of the spoken words and can be improved only by suppressing the background noise without distorting the underlying target speech signal. Designing such algorithms has been extremely challenging, partly because of inaccurate and often unreliable estimates of the background noise (masker) signal from the corrupted signal (often acquired using a single microphone). Algorithms that would improve intelligibility of speech in noisy environments would be extremely useful not only in cellphone applications but also in hearing aids/cochlear implant devices. The development of such algorithms has remained elusive for several decades (Lim, 1978; Hu and Loizou, 2007a), and perhaps this was due to the fact that algorithms were sought that would work for all types of maskers and for all signal-to-noise ratio (SNR) levels, clearly an ambitious goal. In some ASR applications (e.g., voice dictation) and hearing aid applications (e.g., Zakis *et al.*, 2007), however, the algorithm can be speaker and/or masker dependent. Such an approach was taken in this study.

The approach that is being pursued in the present study was motivated by intelligibility studies of speech synthesized using the ideal binary mask (IdBM) (Brungart *et al.*, 2006; Li and Loizou, 2008b, 2008a). The IdBM is a technique explored in computational auditory scene analysis (CASA) that retains the time-frequency (T-F) regions of the target signal that are stronger than the interfering noise (masker), and removes the regions that are weaker than the interfering noise. Previous studies have shown that multiplying the IdBM with the noise-masked signal can yield large gains in intelligibility, even at extremely low (−5, −10 dB) SNR levels (Brungart *et al.*, 2006; Li and Loizou, 2008b). In these studies, prior knowledge of the true IdBM was assumed. In practice, however, the binary mask needs to be estimated from the corrupted signal. Motivated by the successful application of the IdBM technique for improvement of speech intelligibility, we focused on developing a classifier that would identify T-F units as either target-dominated or masker-dominated.[1] This is a conceptually and computationally simpler task than attempting to mimic the human auditory scene analysis using grouping and segmentation principles (Hu and Wang, 2004, 2008; Wang and Brown, 2006), such as common periodicity across frequency, common offsets and onsets, and common amplitude and frequency modulations. Such techniques would require the reliable detection of F0 and onset/offset segments in noise, a formidable task. The challenge faced in the present work is in the design of an accurate classifier capable of operating at negative SNR levels, wherein performance of normal-hearing (NH) listeners is known to degrade. While many techniques have been proposed to estimate the IdBM (Wang and Brown, 2006; Hu and Loizou, 2008), none of the techniques were evaluated with human listeners at extremely low (negative) SNR levels. Most of the proposed algorithms have been

[a)]Author to whom correspondence should be addressed. Electronic mail: loizou@utdallas.edu

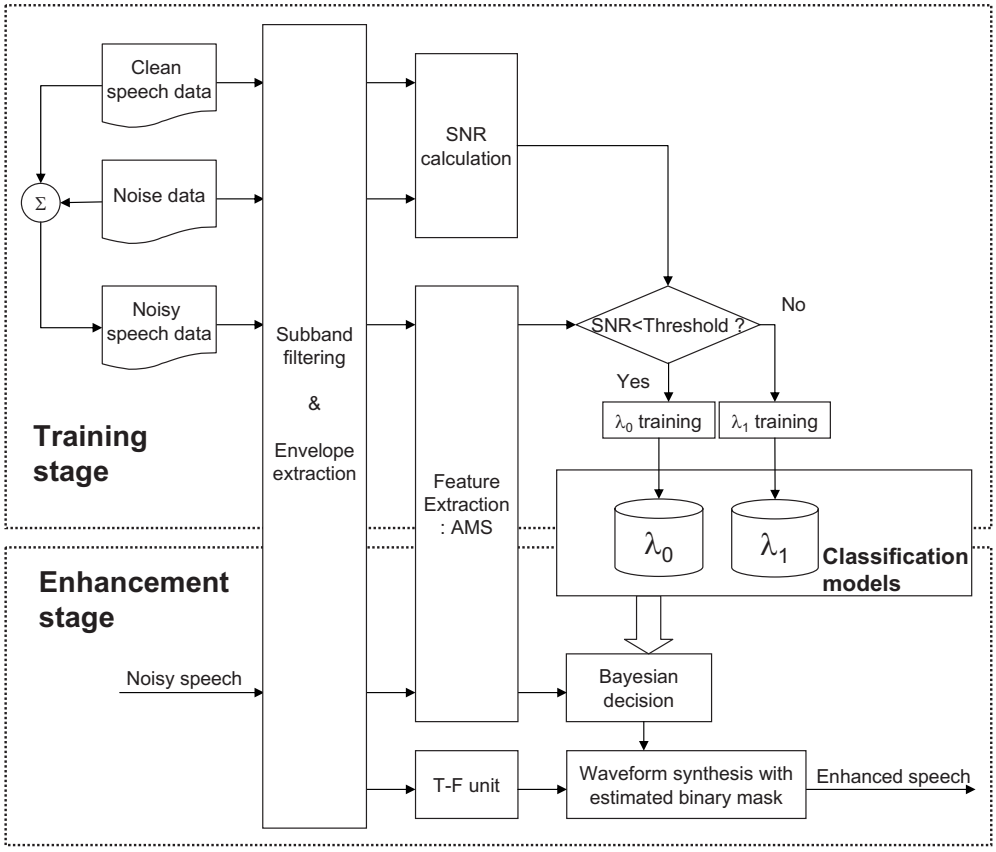

FIG. 1. Block diagram of the training and enhancement stages of the proposed algorithm.

evaluated using objective measures (Hu and Wang, 2004, 2008) and in terms of ASR error rates (Seltzer *et al.*, 2004) rather than in terms of speech intelligibility scores.

The goal of this study is to evaluate the intelligibility of speech synthesized via an algorithm that decomposes the input signal into T-F regions, with the use of a crude auditory-like filterbank, and uses a simple binary Bayesian classifier to retain target-dominated spectro-temporal regions while removing masker-dominated spectro-temporal regions. Amplitude modulation spectrograms (AMSs) (Kollmeier and Koch, 1994) were used as features for training Gaussian mixture models (GMMs) to be used as classifiers. Unlike most noise-suppression algorithms (Loizou, 2007), the proposed algorithm requires no speech/noise detection nor the estimation of noise statistics. Speech corrupted at low SNR levels by different types of maskers is synthesized using this algorithm and presented to human listeners for identification. The present work tests the hypothesis that algorithms that make use of knowledge of when the target is stronger than the masker (at each T-F unit) *can* improve speech intelligibility in noisy conditions.

II. PROPOSED NOISE-SUPPRESSION ALGORITHM

Figure 1 shows the block diagram of the proposed algorithm, consisting of a training stage (top panel) and an intelligibility enhancement stage (bottom panel). In the training stage, features are extracted, typically from a large speech corpus, and then used to train two GMMs representing two feature classes: target speech dominating the masker and masker dominating target speech. AMS are used in this work as features, as they are neurophysiologically and psychoacoustically motivated (Kollmeier and Koch, 1994; Langner and Schreiner, 1988). In the enhancement stage, a Bayesian classifier is used to classify the T-F units of the noise-masked signal into two classes: target-dominated and masker-dominated. Individual T-F units of the noise-masked signal are retained if classified as target-dominated or eliminated if classified as masker-dominated, and subsequently used to reconstruct the enhanced speech waveform.

A. Feature extraction

The noisy speech signal is first bandpass filtered into 25 channels according to a mel-frequency spacing (shown in the subband filtering block in Fig. 1). The envelopes in each band are computed by full-wave rectification and then decimated by a factor of 3 (shown in the envelope extraction block in Fig. 1). The decimated envelope signals are subsequently segmented into overlapping segments of 128 samples (32 ms) with an overlap of 64 samples. Each segment is Hanning windowed and following zero-padding, a 256-point fast Fourier transform (FFT) is computed. The FFT computes the modulation spectrum in each channel, with a frequency resolution of 15.6 Hz. Within each band, the FFT magnitudes are multiplied by 15 triangular-shaped windows spaced uniformly across the 15.6–400 Hz range and summed up to produce 15 modulation spectrum amplitudes. The 15 modulation amplitudes represent the AMS feature vector (Tchorz and Kollmeier, 2003), which we denote

by $\mathbf{a}(\tau,k)$, where τ indicates the time index and k indicates the subband. In addition to the AMS feature vector, we also include delta features to capture feature variations across time and frequency. The overall feature vector is given by

$$\mathbf{A}(\tau,k)=[\mathbf{a}(\tau,k),\Delta\mathbf{a}_T(\tau,k),\Delta\mathbf{a}_K(\tau,k)], \tag{1}$$

where

$$\Delta\mathbf{a}_T(1,k)=\mathbf{a}(2,k)-\mathbf{a}(1,k),\quad \tau=1,$$

$$\Delta\mathbf{a}_T(\tau,k)=\mathbf{a}(\tau,k)-\mathbf{a}(\tau-1,k),\quad \tau=2,\ldots,T, \tag{2}$$

$$\Delta\mathbf{a}_K(\tau,1)=\mathbf{a}(\tau,2)-\mathbf{a}(\tau,1),\quad k=1,$$

$$\Delta\mathbf{a}_K(\tau,k)=\mathbf{a}(\tau,k)-\mathbf{a}(\tau,k-1),\quad k=2,\ldots,K, \tag{3}$$

where $\Delta\mathbf{a}_T(\tau,k)$ and $\Delta\mathbf{a}_K(\tau,k)$ denote the delta feature vectors computed across time and frequency, respectively, and T is the total number of segments. The number of subbands, K, was set to 25 in this work, and the total dimension of the feature vector $\mathbf{A}(\tau,k)$ was 45 $(=3\times15)$.

B. Training stage

A two-class Bayesian classifier was used to estimate the binary mask for each T-F unit. The distribution of the feature vectors of each class was represented with a GMM. The two classes, denoted as λ_0 for mask 0 (masker-dominated T-F units) and λ_1 for mask 1 (target-dominated T-F units), were further subdivided into two smaller classes, i.e., $\lambda_0=\{\lambda_0^0,\lambda_0^1\}$, $\lambda_1=\{\lambda_1^0,\lambda_1^1\}$. This sub-class division yielded faster convergence in GMM training and better classification. In the training stage, the noisy speech spectrum, $Y(\tau,k)$, at time slot τ and k-th subband, was classified into one of four subclasses as follows:

$$Y(\tau,k)\in\begin{cases}\lambda_0^0 & \text{if } \xi(\tau,k)<T_{\mathrm{SNR0}}\\ \lambda_0^1 & \text{if } T_{\mathrm{SNR0}}\le\xi(\tau,k)<T_{\mathrm{SNR}}\\ \lambda_1^0 & \text{if } T_{\mathrm{SNR}}\le\xi(\tau,k)<T_{\mathrm{SNR1}}\\ \lambda_1^1 & \text{if } T_{\mathrm{SNR1}}\le\xi(\tau,k),\end{cases} \tag{4}$$

where $\xi(\tau,k)$ is the local (true) SNR computed as the ratio of envelope energies of the (clean) target speech and masker signals, and T_{SNR0}, T_{SNR1}, and T_{SNR} are thresholds. The T_{SNR0} was chosen in the training stage so as to have equal amount of training data in the λ_0^0 and λ_0^1 classes. Classification performance was not found to be sensitive to this threshold value. The SNR threshold, T_{SNR}, was set to −8 dB for the first 15 frequency bands (spanning 68–2186 Hz) and to −16 dB for the higher frequency bands. This was done to account for the non-uniform masking of speech by the maskers across the spectrum. We utilized 256-mixture Gaussian models for modeling the distributions of the feature vectors in each class. The initial Gaussian model parameters (mixture weights, mean vectors, and covariance matrices) were obtained by running 15 iterations of the k-means clustering algorithm. Full covariance matrices were used for each mixture. If a particular covariance matrix was found to be singular during training, the corresponding mixture weight was set to zero. The final GMM parameters were obtained using the expectation-maximization training algorithm (Dempster *et al.*, 1977). The *a priori* probability for each sub-class $[P(\lambda_0^0),P(\lambda_0^1),P(\lambda_1^0),P(\lambda_1^1)]$ was calculated by counting the number of feature vectors belonging to the corresponding class and dividing that by the total number of feature vectors.

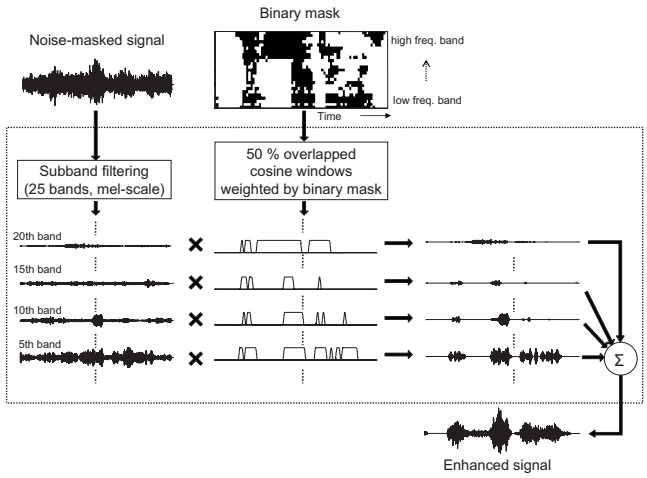

FIG. 2. Block diagram of the waveform synthesis stage of the proposed algorithm.

C. Enhancement stage

In the enhancement stage, the binary masks of each T-F unit are first estimated using a Bayesian classifier. Each T-F unit of noisy speech signal is subsequently retained or eliminated by the estimated binary mask and synthesized to produce the enhanced speech waveforms.

1. Bayesian classification

The T-F units are classified as λ_0 or λ_1 by comparing two *a posteriori* probabilities, $P(\lambda_0|\mathbf{A}_Y(\tau,k))$ and $P(\lambda_1|\mathbf{A}_Y(\tau,k))$. This comparison produces an estimate of the binary mask, $G(\tau,k)$, as follows:

$$G(\tau,k)=\begin{cases}0 & \text{if } P(\lambda_0|\mathbf{A}_Y(\tau,k))>P(\lambda_1|\mathbf{A}_Y(\tau,k))\\ 1 & \text{otherwise},\end{cases} \tag{5}$$

where $P(\lambda_0|\mathbf{A}_Y(\tau,k))$ is computed using Bayes' theorem as follows:

$$P(\lambda_0|\mathbf{A}_Y(\tau,k))=\frac{P(\lambda_0,\mathbf{A}_Y(\tau,k))}{P(\mathbf{A}_Y(\tau,k))}=\frac{P(\lambda_0^0)P(\mathbf{A}_Y(\tau,k)|\lambda_0^0)+P(\lambda_0^1)P(\mathbf{A}_Y(\tau,k)|\lambda_0^1)}{P(\mathbf{A}_Y(\tau,k))}. \tag{6}$$

The *a posteriori* probability $P(\lambda_1|\mathbf{A}_Y(\tau,k))$ is computed similarly.

2. Waveform synthesis

Figure 2 shows the block diagram of the waveform synthesis stage. The corrupted speech signal is first filtered into 25 bands (same bands used in the feature-extraction stage). To remove across-channel differences, the output of each filter is time reversed, passed through the filter, and reversed

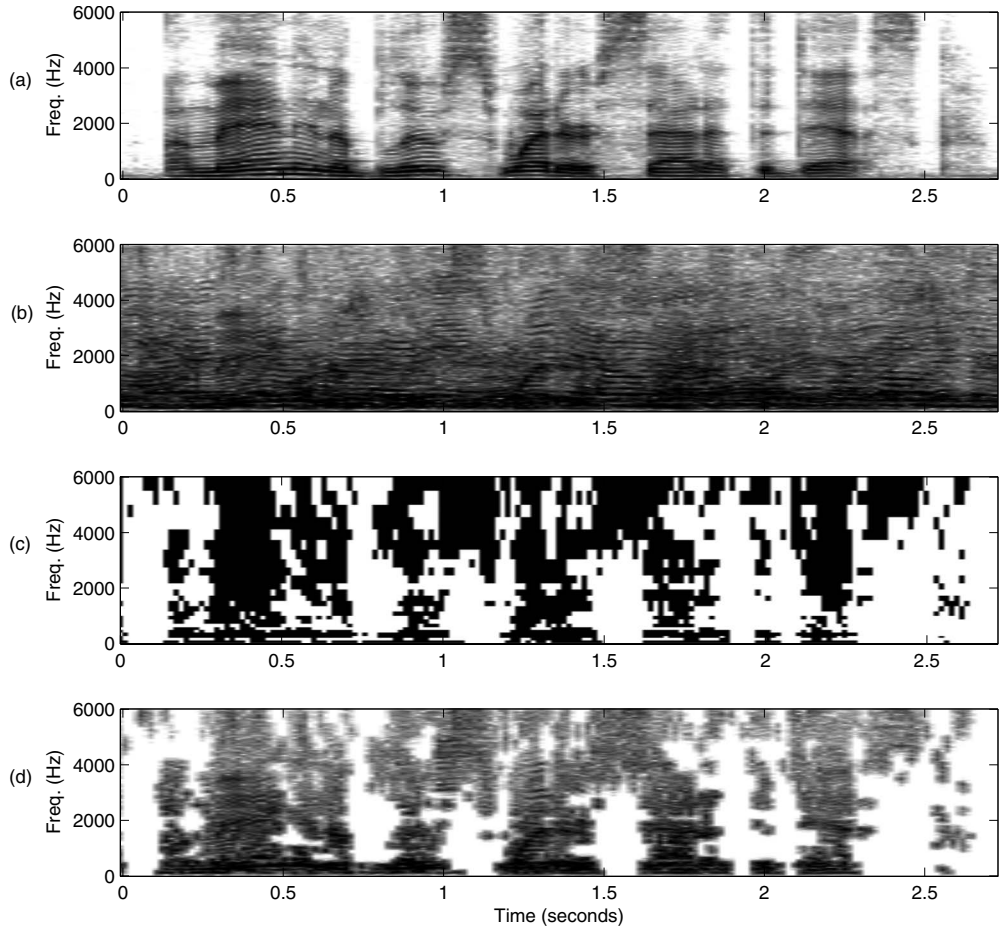

FIG. 3. (a) Wide-band spectrogram of an IEEE sentence in quiet. (b) Spectrogram of corrupted sentence by multitalker babble at −5 dB SNR. (c) Binary mask estimated using Eq. (5), with black pixels indicating target-dominated T-F units and white pixels indicating masker-dominated T-F units. (d) Synthesized signal obtained by multiplying the binary mask shown in panel (c) with the corrupted signal shown in panel (b).

again (Wang and Brown, 2006). The filtered waveforms are windowed with a raised cosine every 32 ms with 50% overlap between segments, and then weighted by the estimated binary mask [Eq. (5)]. Finally, the estimated target signal is reconstructed by summing the weighted responses of the 25 filters. Figure 3 shows an example spectrogram of a synthesized signal using the proposed algorithm. In this example, the clean speech signal [Fig. 3(a)] is mixed with multitalker babble at −5 dB SNR [Fig. 3(b)]. The estimated binary mask [as per Eq. (5)] and synthesized waveform are shown in Figs. 3(c) and 3(d), respectively.

III. LISTENING EXPERIMENTS

A. Stimuli

Sentences taken from the IEEE database (IEEE, 1969) were used as test material. The sentences in the IEEE database are phonetically balanced with relatively low word-context predictability. The sentences were produced by one male and one female speaker in a sound-proof booth using Tucker-Davis Technologies (TDT) recording equipment. The sentences were originally recorded at a sampling rate of 25 kHz and downsampled to 12 kHz. Three types of noise (20-talker babble, factory, speech-shaped noise) were used as maskers. The (steady) speech-shaped noise was stationary having the same long-term spectrum as the sentences in the IEEE corpus. The factory noise was taken from the NOISEX database (Varga and Steeneken, 1993), and the babble (20 talkers with equal number of female and male talkers) was taken from the Auditec CD (St. Louis, MO).

A total of 390 IEEE sentences were used to train the GMM models. These sentences were corrupted by three types of noise at −5, 0, and 5 dB SNR. The maskers were randomly cut from the noise recordings and mixed with the target sentences at the prescribed SNRs. Each corrupted sentence had thus a different segment of the masker, and this was done to evaluate the robustness of the Bayesian classifier in terms of generalizing to different segments of the masker having possibly different temporal/spectral characteristics. Three different training sets were prepared to train three GMM models and three test sets were used for the evaluation of the GMM models. Two types of GMM models were trained: (1) a single-noise GMM model (denoted as sGMM) trained only on a single type of noise (tested with the same type of noise) and (2) a multi-noise GMM model (denoted as mGMM) trained on all three types of noise (tested with one of the three types of noise). The latter models (mGMM) were used to assess the performance and robustness of a single GMM model in multiple noisy environments. As we were limited by the total number of sentences available in the IEEE corpus, we used different sets of training sentences

randomly assigned to the various conditions. This was necessary to avoid testing NH listeners with the same sentences used in the training stage. More specifically, three sets of training data were created with each set having 390 sentences and two training sets having an overlap of 150 sentences. There was no overlap between the training and testing sets in any condition.

B. Procedure

A total of 17 NH listeners (all native speakers of English) were recruited for the listening tests. All subjects were paid for their participation. The listeners were randomly assigned to the conditions involving processed IEEE sentences produced by the male and female speakers (eight listened to the IEEE sentences produced by the male speaker and nine listened to the IEEE sentences produced by the female speaker). Subjects participated in a total of 24 conditions [=2 SNR levels (−5 dB, 0 dB) × 4 processing conditions × 3 types of maskers]. The processing conditions included speech processed using (1) sGMM models, (2) mGMM models, (3) the idBM, and (4) the unprocessed (noise-masked) stimuli. The IdBM condition was included as a control condition to assess the performance of the proposed algorithms relative to the ideal condition in which we have *a priori* knowledge of the local SNR and IdBM. The IdBM was obtained by comparing the local (true) SNR against a predefined threshold. The SNR at each T-F unit was computed as the ratio of the envelope energies of the (clean) target speech and masker signals in each unit. The IdBM takes a value of 1 if the local SNR is greater than the threshold and takes the value of 0 otherwise.

The experiments were performed in a sound-proof room (Acoustic Systems, Inc.) using a PC connected to a Tucker-Davis system 3. Stimuli were played to the listeners monaurally through Sennheiser HD 250 Linear II circumaural headphones at a comfortable listening level. Prior to the sentence test, each subject listened to a set of noise-corrupted sentences to be familiarized with the testing procedure. During the test, subjects were asked to write down the words they heard. The whole listening test lasted for about 2–3 h, which was split into two sessions each lasting 1–1.5 h. 5 min breaks were given to the subjects every 30 min. Two lists of sentences (i.e., 20 sentences) were used per condition, and none of the lists were repeated across conditions. The sentence lists were counterbalanced across subjects. Sentences were presented to the listeners in blocks, and 20 sentences were presented in each block for each condition. The order of the conditions was randomized across subjects.

IV. RESULTS

The mean performance, computed in terms of percentage of words identified correctly by the NH listeners, are plotted in Fig. 4 for sentences produced by male (top panel) and female speakers (bottom panel). A substantial improvement in intelligibility was obtained with the proposed algorithm using both sGMM and mGMM models, compared to that attained by human listeners with unprocessed (corrupted) speech. The improvement (over 60% points in some

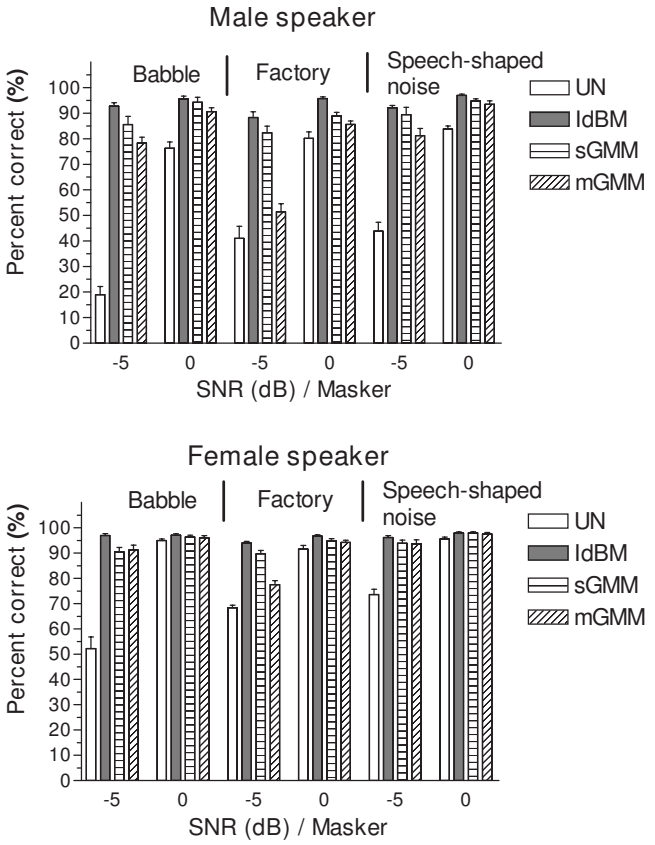

FIG. 4. Mean speech recognition scores obtained by 17 NH listeners for corrupted (unprocessed) sentences (denoted as UN), sentences processed using the sGMM (single-noise trained GMMs) and mGMM models (multiple-noise trained GMMs), and sentences processed using the IdBM in the various SNR/masker conditions. Error bars indicate standard errors of the mean.

cases) was more evident at −5 dB SNR levels for all three maskers tested. Performance at 0 dB SNR in the female-speaker conditions (bottom panel of Fig. 4) was limited in most cases by ceiling effects. Analysis of variance (with repeated measures) indicated significant effect of masker type [$F(2,14)=41.2$, $p<0.0005$], significant effect of SNR level [$F(1,7)=583.3$, $p<0.0005$], and significant effect of processing algorithm [$F(3,21)=314.1$, $p<0.0005$]. All interactions were found significant ($p<0.05$). *Post-hoc* analysis (Schéffe), corrected for multiple comparisons, was done to assess significant differences between conditions. For the male-speaker data (top panel of Fig. 4), performance at −5 dB SNR with mGMM models was significantly ($p<0.0005$) higher than that attained by the listeners in all baseline masker conditions (unprocessed sentences) except the factory condition. Performance at −5 and 0 dB SNR with sGMM models was significantly ($p<0.005$) higher than that attained in all baseline masker conditions. For the female-speaker data, performance at −5 dB SNR with sGMM and mGMM models was significantly ($p<0.0005$) higher than performance obtained with unprocessed speech in all masker conditions. There was no significant ($p>0.05$) difference in scores between the various algorithms in the 0 dB SNR masker conditions, as performance was limited by ceiling effects. Consistent with prior studies (Brungart *et al.*, 2006; Li and Loizou, 2008b), highest performance was obtained

TABLE I. Hit (HIT) and false alarm (FA) rates obtained using the sGMM and mGMM models for the male-speaker and female-speaker data in the various masker conditions.

			Babble		Factory		Speech-shaped	
Speaker	Model	Performance	−5 dB	0 dB	−5 dB	0 dB	−5 dB	0 dB
Male	sGMM	HIT	86.96%	82.49%	80.17%	75.18%	88.30%	84.99%
		FA	14.54%	10.43%	15.27%	11.44%	12.20%	9.48%
		HIT-FA	72.42%	72.06%	64.90%	63.74%	76.10%	75.51%
	mGMM	HIT	78.24%	75.94%	75.36%	75.79%	76.90%	76.61%
		FA	18.83%	17.69%	24.12%	22.91%	12.75%	13.24%
		HIT-FA	59.41%	58.25%	51.24%	52.88%	64.15%	63.37%
Female	sGMM	HIT	89.95%	86.21%	83.28%	79.52%	89.28%	85.82%
		FA	15.23%	11.36%	17.26%	12.12%	12.65%	10.26%
		HIT-FA	74.72%	74.85%	66.02%	67.40%	76.63%	75.56%
	mGMM	HIT	82.28%	79.46%	82.58%	81.93%	81.76%	80.34%
		FA	18.03%	16.63%	25.84%	21.60%	13.49%	13.86%
		HIT-FA	64.25%	62.83%	56.74%	60.33%	68.27%	66.48%

with the IdBM. Performance with sGMM models was comparable to that obtained with the IdBM in nearly all conditions. Note that with the exception of one condition (factory noise at −5 dB SNR), performance with mGMM models did not differ significantly ($p>0.05$) from that obtained with sGMM models, an outcome demonstrating the potential of the proposed approach in training a single GMM model that would be effective in multiple listening environments.

To quantify the accuracy of the binary Bayesian classifier, we computed the average hit (HIT) and false alarm (FA) rates for three test sets not included in the training. Each test set comprised of 60 sentences, for a total of 180 sentences corresponding to 893,950 T-F units (35,758 frames × 25 frequency bands) for the male-speaker sentences and 811,750 T-F units (32 470 × 25 frequency bands) for the female-speaker sentences. HIT and FA rates were computed by comparing the estimated binary mask against the (oracle) IdBM. Table I shows the results obtained using sGMM and mGMM models in the various masker conditions. High hit rates (lowest with factory noise at 0 dB, male speaker; 75.18%) and low false-alarm rates (highest with factory noise at −5 dB, female speaker; 17.26%) were obtained with sGMM models. The hit rate obtained with mGMM models was about 10% lower than that of sGMM models for the male speaker. The difference was much smaller for the female speaker (about 5%). As demonstrated in Li and Loizou, (2008b), low false alarm rates (<20% assuming high hit rates) are required to achieve high levels of speech intelligibility.

To predict the intelligibility of speech synthesized using the estimated binary masks (based on the Bayesian classifier), we propose a simple metric based on the difference between the hit rate and false alarm rates, i.e., HIT-FA. This metric bears resemblance to the sensitivity index, d', used in psychoacoustics (Macmillan and Creelman, 2005). The index d' is derived assuming a Gaussian distribution of responses. No such assumptions are made with the use of the HIT-FA difference metric. A modestly high correlation ($r=0.80$) was obtained between this simple difference metric and speech intelligibility scores based on data from the same three test sets used in the listening experiments (see Fig. 5). More generally, the difference metric, $d_\alpha=\alpha\cdot H-(1-\alpha)\cdot FA$, can be used to obtain higher correlation by optimizing the value of α for different speech materials. A value of $\alpha=0.3$ yielded a maximum correlation of $r=0.84$ for our test materials (IEEE sentences), suggesting that more weight needs to be placed on FA values, an outcome consistent with intelligibility studies (Li and Loizou, 2008b). Table II shows the performance (in terms of HIT and FA rates) of two conventional noise reduction algorithms, the Wiener algorithm (Scalart and Filho, 1996) and the MMSE algorithm (Ephraim and Malah, 1984). The binary mask was estimated by comparing the SNR in each frequency bin against the same threshold T_{SNR} used in the proposed algorithm (see Sec II B). The SNR was estimated from the corrupted signal using the decision-directed approach (Ephraim and Malah, 1984). As can be seen, the hit rates obtained by the GMM binary classifiers (Table I) are substantially higher than those obtained with conventional noise reduction algorithms. This outcome might explain, at least, partially why current noise reduction algorithms, even the most sophisticated ones, do not improve speech intelligibility (Hu and Loizou, 2008).

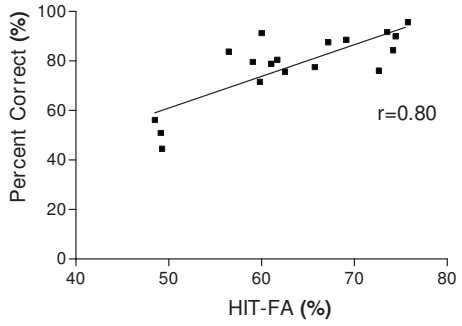

FIG. 5. Scatter plot showing the correlation between human listener's recognition scores, obtained for the male-speaker data at −5 dB in the three masker conditions, and a metric based on the difference between the resulting hit-rate (HIT) and false alarm (FA) values of the binary Bayesian classifier.

TABLE II. Binary mask accuracy obtained by two conventional noise reduction algorithms for the male-speaker data at −5 dB SNR.

	Babble		Factory		Speech-shaped	
	Wiener	MMSE	Wiener	MMSE	Wiener	MMSE
HIT	54.60%	68.44%	53.14%	57.52%	58.59%	58.89%
FA	55.62%	66.94%	54.48%	60.38%	50.44%	52.24%
HIT-FA	−1.02%	1.50%	−1.34%	−2.86%	8.15%	6.65%

To assess the robustness of the binary classifier in terms of handling speakers not included in the training, we performed a cross-gender test wherein we used the male-trained models to classify AMS features extracted from the female-speaker data, and vice versa (see Table III). The performance obtained with the cross-gender models was comparable to that obtained with the same-speaker models (Table I) showing differences ranging from 2.25% (factory noise, female speaker) up to 9.77% (speech-shaped noise, male speaker).

Finally, to quantify the gain in classification accuracy obtained with the proposed delta features [Eqs. (2) and (3)], we compared the hit and false alarm rates obtained with and without the use of delta features (see Table IV). The same test set used in the evaluation of the cross-gender models (Table III) was used in the evaluation of the delta features. As can be seen, delta features improved the hit rate considerably (by as much as 20% in some cases), without increasing the false alarm rate.

V. DISCUSSION AND CONCLUSIONS

Large gains in intelligibility were achieved with the proposed algorithm (Fig. 4). The intelligibility of speech processed by the proposed algorithm was substantially higher than that achieved by human listeners listening to unprocessed (corrupted) speech, particularly at extremely low SNR levels (−5 dB). We attribute this to the accurate classification of T-F units into target- and masker-dominated T-F units, and subsequently reliable estimation of the binary mask. As demonstrated by several intelligibility studies with NH listeners (Brungart *et al.*, 2006; Li and Loizou, 2008b) access to reliable estimates of the binary mask can yield substantial gains in intelligibility. The accurate classification of T-F units into target- and masker-dominated T-F units was accomplished with the use of neurophysiologically-motivated features (AMS) and carefully designed Bayesian classifiers (GMMs). Unlike the mel-frequency cepstrum coefficients (Davis and Mermelstein, 1980) features commonly used in ASR, the AMS features capture information about amplitude and frequency modulations, known to be critically important for speech recognition (Zeng *et al.*, 2005). Furthermore, the proposed delta features [Eqs. (2) and (3)] are designed to capture to some extent temporal and spectral correlations. Unlike the delta features commonly used in ASR (Furui, 1986), the simplified delta features proposed in Eq. (2) use only past information and are therefore amenable to real-time implementation with low latency.

GMMs are known to accurately represent a large class of feature distributions, and as classifiers, GMMs have been used successfully in several applications and, in particular speaker recognition (e.g., Reynolds and Rose, 1995). Other classifiers (e.g., neural networks, and support vector machines) could alternatively be used (Tchorz and Kollmeier, 2003). Our attempt, however, to use neural networks[2] as classifiers was not very successful as poorer performance was observed, particularly when different segments (randomly cut) of the masker were mixed with each test sentence (as done in the present study).

There exist a number of differences in our approach that distinguishes it from previous attempts to estimate the binary mask. First, their approach is simple as it is based on the design of an accurate (binary) Bayesian classifier. Others (Wang and Brown, 2006) focused on developing sophisticated grouping and segmentation algorithms that were motivated largely by existing knowledge in auditory scene analysis (Bregman, 1990). Second, the resolution of the auditory filters used in the present work is crude compared to that used by humans. A total of 128 Gammatone filters have been used by others (Brungart *et al.*, 2006; Hu and Wang, 2004)

TABLE III. Classification of male-speaker data using the female-speaker model (sGMM) and classification of the female-speaker data using the male-speaker model (sGMM) at −5 dB SNR.

		Male-speaker model			Female-speaker model		
		Babble	Factory	Speech-shaped	Babble	Factory	Speech-shaped
Male-speaker data	HIT	87.82%	82.88%	89.04%	79.82%	75.89%	78.68%
	FA	16.06%	16.97%	12.20%	15.34%	16.41%	11.61%
	HIT-FA	71.76%	65.91%	76.84%	64.48%	59.48%	67.07%
Female-speaker data	HIT	88.22%	82.68%	88.62%	89.52%	82.42%	88.81%
	FA	18.78%	17.34%	16.11%	13.72%	14.83%	10.83%
	HIT-FA	69.44%	65.34%	72.51%	75.80%	67.59%	77.98%

TABLE IV. Performance comparison, in terms of hits and false alarm rates, between the AMS feature vectors and AMS+Delta feature vectors for the male-speaker data at −5 dB SNR.

	Babble		Factory		Speech-shaped	
	AMS only	AMS+Delta	AMS only	AMS+Delta	AMS only	AMS+Delta
HIT	79.46%	87.82%	60.58%	82.88%	76.12%	89.04%
FA	18.19%	16.06%	19.32%	16.97%	15.84%	12.20%
HIT-FA	61.27%	71.76%	41.26%	65.91%	60.28%	76.84%

for modeling the auditory periphery. A smaller number (25) of channels was used in this work for two reasons: (a) to keep the feature dimensionality small and (b) to make it appropriate for hearing aid and cochlear implant applications, wherein the signal is typically processed through a small number of channels. Previous work in our laboratory (Li and Loizou, 2008a) demonstrated that spectral resolution has a significant effect on the intelligibility of IdBM speech, but the use of 25 channels seemed to be sufficient for accurate speech recognition. Third, our approach required limited amount of training data. Fewer than 400 sentences (~20 min) were used for training compared to thousands of sentences (~1–2 h) used by others (Seltzer *et al.*, 2004). Finally, the GMM training used in this work does not require access to a labeled speech corpus, while the methods proposed by others required the use of accurate F0 values or voiced/unvoiced segmentation (Hu and Wang, 2004, 2008; Seltzer *et al.*, 2004).

The proposed algorithm can be used not only for robust ASR (e.g., Cooke *et al.*, 2001) or cellphone applications but also for hearing aids or cochlear implant devices. Modern hearing aids use sound classification algorithms (e.g., Nordqvist and Leijon, 2004) to identify different listening situations and adjust accordingly hearing aid processing parameters. All advantages cited above make the proposed approach suitable for trainable hearing aids (Zakis *et al.*, 2007) and cochlear implant devices. As these devices are powered by a digital signal processor chip, the training can take place at the command of the user whenever in a new listening environment. Following the training stage, the user can initiate the proposed algorithm to enhance speech intelligibility in extremely noisy environments (e.g., restaurants). As shown in Sec. III, a single GMM trained on multiple types of noise (mGMM) can yield high performance; however, a user might encounter a new type of noise not included in the training set. In such circumstances, either new training needs to be initiated or perhaps adaptation techniques can be used to adapt the parameters of existing GMM models to the new data (Reynolds *et al.*, 2000).

Humans outperform ASR and CASA systems on various recognition tasks and are far better at dealing with accents, noisy environments, and differences in speaking style/rate (Lippmann, 1997; Scharenborg, 2007). Neither ASR nor CASA algorithms, however, have yet reached the level of performance obtained by human listeners, despite the level of sophistication built in these algorithms (Lippmann, 1997). The present study demonstrated that if the goal of CASA is to design algorithms that would perform as well or better than humans, it is not always necessary to mimic all aspects of the human auditory processing. Knowledge of when the target is stronger than the masker in each T-F unit is all that is required to achieve high levels of speech understanding (Li and Loizou, 2008b). This reduces the problem to that of designing an accurate binary classifier [see Eq. (5)]. Computers can generally be trained to classify accurately not only binary datasets (as in the present work) but also complex data patterns. The humans' ability, however, to detect the target signal in the presence of a masker within a critical band is limited by simultaneous (and temporal) masking and is dependent on several factors including the masker frequency (in relation to the target's), the masker level and the type of masker (e.g., tonal or noise-like) (Moore, 2003). The present study demonstrated that computer algorithms can be designed to overcome these shortcomings and subsequently improve speech intelligibility in noisy conditions.

ACKNOWLEDGMENTS

This research was supported by Grant No. R01 DC007527 from National Institute of Deafness and other Communication Disorders, NIH.

[1]A T-F unit is said to be target-dominated if its local SNR is greater than 0 dB and is said to be masker-dominated otherwise. These definitions can be extended by using a threshold other than 0 dB. In this paper, we define a target-dominated unit as a T-F unit for which the SNR is greater than a predefined threshold even if the power of the target signal is smaller than that of the masker (this occurs when the chosen threshold is <0 dB).

[2]A standard feed-forward neural network was trained with the same AMS feature vectors using the back-propagation algorithm. The network consisted of an input layer of 375 neurons (15×25), a hidden layer with 225 neurons, and an output layer with 25 output neurons, one for each channel. The output neuron activities indicated the respective SNR in each channel. The predicted SNR values from the output layer were compared against a SNR threshold of −8 dB to estimate the binary mask.

Bregman, A. S. (**1990**). *Auditory Scene Analysis* (MIT, Cambridge, MA).
Brungart, D., Chang, P., Simpson, B., and Wang, D. (**2006**). "Isolating the energetic component of speech-on-speech masking with ideal time-frequency segregation," J. Acoust. Soc. Am. **120**, 4007–4018.
Cooke, M., Green, P., Josifovski, L., and Vizinho, A. (**2001**). "Robust automatic speech recognition with missing and unreliable acoustic data," Speech Commun. **34**, 267–285.
Davis, S. B., and Mermelstein, P. (**1980**). "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," IEEE Trans. Acoust., Speech, Signal Process. **ASSP-28**, 357–336.
Dempster, A. P., Laird, N. M., and Rubin, D. B. (**1977**). "Maximum likelihood from incomplete data via the EM algorithm," J. R. Stat. Soc. Ser. B (Methodol.) **39**, 1–38.
Ephraim, Y., and Malah, D. (**1984**). "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," IEEE Trans. Acoust., Speech, Signal Process. **ASSP-32**, 1109–1121.

Furui, S. (**1986**). "Speaker independent isolated word recognition using dynamic features of speech spectrum," IEEE Trans. Acoust., Speech, Signal Process. **ASSP-34**, 52–59.
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. (**2008**). "Segregation of unvoiced speech from nonspeech interference," J. Acoust. Soc. Am. **124**, 1306–1319.
Hu, Y., and Loizou, P. C. (**2007a**). "A comparative intelligibility study of single-microphone noise reduction algorithms," J. Acoust. Soc. Am. **122**, 1777–1786.
Hu, Y., and Loizou, P. C. (**2007b**). "Subjective evaluation and comparison of speech enhancement algorithms," Speech Commun. **49**, 588–601.
Hu, Y., and Loizou, P. C. (**2008**). "Techniques for estimating the ideal binary mask," in The 11th International Workshop on Acoustic Echo and Noise Control, Seattle, WA
IEEE (**1969**). "IEEE recommended practice for speech quality measurements," IEEE Trans. Audio Electroacoust. **17**, 225–246.
Kollmeier, B., and Koch, R. (**1994**). "Speech enhancement based on physiological and psychoacoustical models of modulation perception and binaural interaction," J. Acoust. Soc. Am. **95**, 1593–1602.
Langner, G., and Schreiner, C. (**1988**). "Periodicity coding in the inferior colliculus of the cat. I. Neuronal mechanisms," J. Neurophysiol. **60**, 1799–1822.
Li, N., and Loizou, P. C. (**2008a**). "Effect of spectral resolution on the intelligibility of ideal binary masked speech," J. Acoust. Soc. Am. **123**, EL59–EL64.
Li, N., and Loizou, P. C. (**2008b**). "Factors influencing intelligibility of ideal binary-masked speech: Implications for noise reduction," J. Acoust. Soc. Am. **123**, 1673–1682.
Lim, J. S. (**1978**). "Evaluation of a correlation subtraction method for enhancing speech degraded by additive white noise," IEEE Trans. Acoust., Speech, Signal Process. **26**, 471–472.
Lippmann, R. P. (**1997**). "Speech recognition by machines and humans," Speech Commun. **22**, 1–15.
Loizou, P. C. (**2007**). *Speech Enhancement: Theory and Practice* (CRC, Boca Raton, FL).
Macmillan, N., and Creelman, D. (**2005**). *Detection Theory: A User's Guide* (Lawrence Erlbaum Associates, New York).
Moore, B. (**2003**). *An Introduction to the Psychology of Hearing* (Academic, London).
Nordqvist, P., and Leijon, A. (**2004**). "An efficient robust sound classification algorithm for hearing aids," J. Acoust. Soc. Am. **115**, 3033–3041.
Rabiner, L. (**2003**). "The power of speech," Science **301**, 1494–1495.
Reynolds, D., and Rose, R. (**1995**). "Robust text-independent speaker identification using Gaussian mixture speaker models," IEEE Trans. Speech Audio Process. **3**, 72–83.
Reynolds, D., Quatieri, T., and Dunn, R. (**2000**). "Speaker verification using adapted Gaussian mixture models," Digit. Signal Process. **10**, 19–41.
Scalart, P., and Filho, J. (**1996**). "Speech enhancement based on a priori signal to noise estimation," in *Proceedings of the IEEE International Conference on Acoustics, Speech, Signal Processing*, pp. 629–632.
Scharenborg, O. (**2007**). "Reaching over the gap: A review of efforts to link human and automatic speech recognition research," Speech Commun. **49**, 336–347.
Seltzer, M., Raj, B., and Stern, R. (**2004**). "A Bayesian classifier for spectrographic mask estimation for missing feature speech recognition," Speech Commun. **43**, 379–393.
Sroka, J. J., and Braida, L. D. (**2005**). "Human and machine consonant recognition," Speech Commun. **45**, 401–423.
Tchorz, J., and Kollmeier, B. (**2003**). "SNR estimation based on amplitude modulation analysis with applications to noise suppression," IEEE Trans. Speech Audio Process. **11**, 184–192.
Varga, A., and Steeneken, H. J. M. (**1993**). "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," Speech Commun. **12**, 247–251.
Wang, D. L., and Brown, G. J. (**2006**). *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications* (Wiley, Hoboken, NJ).
Zakis, J. A., Dillon, H., and McDermott, H. J. (**2007**). "The design and evaluation of a hearing aid with trainable amplification parameters," Ear Hear. **28**, 812–830.
Zeng, F.-G., Nie, K., Stickney, G. S., Kong, Y.-Y., Vongphoe, M., Bhargave, A., Wei, C., and Cao, K. (**2005**). "Speech recognition with amplitude and frequency modulations," Proc. Natl. Acad. Sci. U.S.A. **102**, 2293–2298.