Improving Speech Intelligibility in Noise Using Environment-Optimized Algorithms

Gibak Kim and Philipos C. Loizou, *Senior Member, IEEE*

***Abstract*—While most speech enhancement algorithms improve speech quality, they may not improve speech intelligibility in noise. This paper focuses on the development of an algorithm that can be optimized for a specific acoustic environment and improve speech intelligibility. The proposed method 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 signal or the noise masker. Target-dominated T-F units are retained while masker-dominated T-F units are discarded. The Bayesian classifier is trained for each acoustic environment using an incremental approach that continuously updates the model parameters as more data become available. Listening experiments were conducted to assess the intelligibility of speech synthesized using the incrementally adapted models as a function of the number of training sentences. Results indicated substantial improvements in intelligibility (over 60% in babble at −5 dB SNR) with as few as ten training sentences in babble and at least 80 sentences in other noisy conditions.**

***Index Terms*—Environment-optimized algorithms, speech enhancement, speech intelligibility.**

I. Introduction

LARGE advances have been made in the development of enhancement algorithms that can suppress background noise and improve speech quality [1]. Considerably smaller progress has been made, however, in designing algorithms that can improve speech intelligibility. As demonstrated in [2], algorithms that improve speech quality do not necessarily improve speech intelligibility. This is most likely due to the distortions introduced to the speech signal. In contrast to speech quality, intelligibility relates to the understanding of the underlying message or content of the spoken words, and is often measured by counting the number of words identified correctly by human listeners. Intelligibility can potentially be improved only by suppressing the background noise without distorting the underlying target speech signal. 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 [2], [3], and perhaps this was due to the fact that algorithms were sought that would work for all types of maskers (noise) and for all signal-to-noise ratio (SNR) levels, clearly an ambitious goal. In some speech recognition applications (e.g., voice dictation) and hearing aid applications (e.g., [4]), however, the algorithm can be speaker and/or environment dependent.

Several environment-dependent algorithms have been proposed recently in [5]–[10]. The innovation of these algorithms lies in the derivation of spectral weighting functions (gain functions) that have been trained in a data-driven fashion based on various error criteria. Unlike the gain functions derived for minimum mean square error (MMSE) and maximum *a posteriori* (MAP) estimators [11]–[13], the gain functions derived in [7]–[10] make no assumptions about the probability density functions (pdf) of the complex clean and noise spectra. Fingscheidt *et al.* [10] have used a large corpus of clean speech and noise data to train frequency-specific gain functions for a specific noise environment. The gain functions were expressed as a function of the *a posteriori* and *a priori* SNRs (computed using a modified decision-directed approach [11]) and were derived by minimizing various perceptually motivated distance metrics [14]. The data-derived gain functions were stored in look-up tables indexed by the *a posteriori* and *a priori* SNRs, and used for enhancing speech in the trained acoustic environments. When tested in automotive environments, the data-driven approach [10] outperformed conventional algorithms (e.g., MMSE) both in terms of speech distortion and noise attenuation. The data-driven method proposed in [8] compared favorably to current state-of-the art noise suppression algorithms.

The above data-driven and/or environment-optimized algorithms performed well in terms of improving speech quality, but have not been evaluated in terms of speech intelligibility. Given our experience with MMSE-based speech enhancement algorithms [2], we do not expect significant improvements in intelligibility with these algorithms.

This paper takes a different approach that does not rely on the derivation of spectral weighting (gain) functions, but rather focuses on the reliable classification of the spectral SNR in two regions. The pursued approach is motivated by intelligibility studies of speech synthesized using the ideal binary mask (IdBM) [15]–[17], which in turn requires access to the SNR at each frequency bin. The ideal binary mask (originally known as *a priori* mask [18]) 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 (i.e., SNR > 0 dB) than the interfering noise (masker), and removes the regions that are weaker than the interfering

Manuscript received April 03, 2009; revised December 21, 2009. Date of publication January 26, 2010; date of current version September 08, 2010. This work was supported by the National Institute of Deafness and other Communication Disorders, National Institute of Health, under Grant R01 DC007527. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Malcolm Slaney.

The authors are with the Department of Electrical Engineering, University of Texas at Dallas, Richardson, TX 75083-0688 USA (e-mail: imkgb27@gmail.com; loizou@utdallas.edu).

Digital Object Identifier 10.1109/TASL.2010.2041116

noise (i.e., SNR $<$ 0 dB). Previous studies have shown that multiplying the ideal binary mask with the noise-masked signal can yield large gains in intelligibility, even at extremely low ($-5, -10$ dB) SNR levels [15], [16]. In these studies, prior knowledge of the true spectral SNR and subsequently the ideal binary mask was assumed. In practice, however, the binary mask needs to be estimated from the corrupted signal requiring an accurate estimate (and classification) of the spectral SNR. In our previous work [19], we have proposed a speech enhancement method which estimates the binary mask using a Bayesian classifier and synthesizes the enhanced signal by binary masking (i.e., multiplying by a binary gain function) the noisy spectra. This algorithm decomposes the input signal into T-F units with the use of a crude auditory-like filterbank and uses a simple binary Bayesian classifier to retain target-dominant[1] T-F units while removing masker-dominant units. Amplitude modulation spectrograms (AMS) [20] were used as features for training Gaussian mixture models (GMMs) to be used as classifiers. Unlike most speech enhancement algorithms [1], the proposed algorithm did not require speech/noise detection nor the estimation of noise statistics. This method was evaluated using listening tests and shown to achieve large gains in speech intelligibility at extremely low SNR levels. The listening tests were focused on extremely low SNR levels (e.g., -5 dB), such as those encountered in military applications, restaurants and manufacturing facilities, since speech intelligibility by normal-hearing listeners is known to suffer primarily at such low SNR levels.

The approach proposed in [19] required hundreds of sentences for training, and the batch training procedure used was burdensome in terms of computational requirements, thereby hampering rapid adaptation to new listening environments. In this paper, we investigate alternative training procedures for adapting/updating the model parameters for fast adaptation to new acoustic environments. More precisely, we consider an incremental training approach which starts from an initial model trained with a small amount of data and updates the model parameters as more data become available. Listening experiments were conducted to assess the performance of the incrementally adapted GMMs as a function of the number of sentences used for training. Speech was synthesized with the adapted GMMs and presented to normal-hearing listeners for identification. The aim of the listening experiments is to determine the minimum amount (in terms of duration) of training data required to obtain significant improvements in intelligibility relative to that of unprocessed (noise-masked) sentences.

II. Binary-Mask Based Speech Enhancement Algorithm

Fig. 1 shows the block diagram of the proposed algorithm [19], 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 Gaussian mixture models (GMMs) representing two feature classes: target speech dominating the masker and masker dominating target speech. In [21], harmonicity-based features were extracted directly from the speech signal, and used in a Bayesian classifier to estimate the binary mask. The reliability, however, of the harmonicity cues depends largely on the pitch estimation algorithm, which is often inaccurate in low SNR (e.g., SNR $<$ 0 dB) environments. In this paper, AMS are used as features, as they are neurophysiologically and psychoacoustically motivated [20], [22]. In the enhancement stage, a Bayesian classifier is used to classify the time–frequency 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.

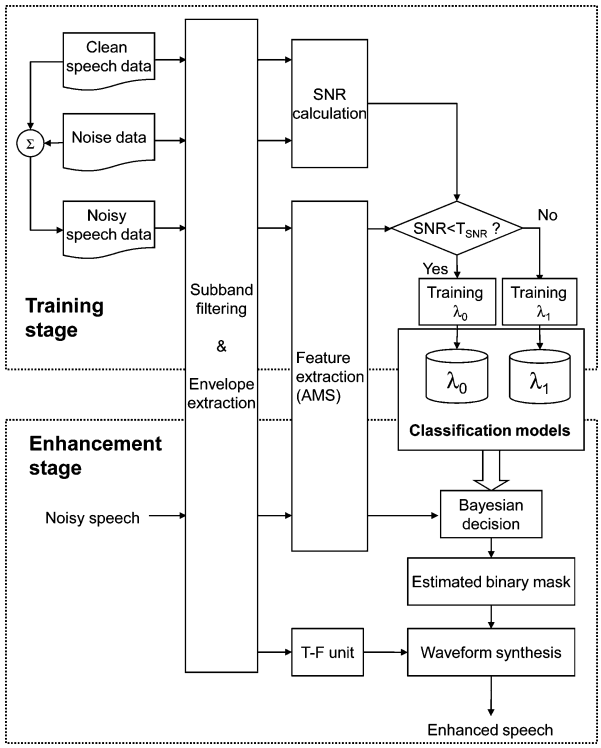

Fig. 1. Block diagram of the training and enhancement stages for the speech enhancement based on the binary masking of T-F units.

A. Feature Extraction

The noisy speech signal is first bandpass filtered into 25 channels according to a mel-frequency spacing, spanning a 6 kHz (68.5–6000 Hz) bandwidth. The sampling rate was 12 kHz. The envelopes in each subband are computed by full-wave rectification and then decimated by a factor of 3. The decimated envelopes are segmented into overlapping segments of 128 samples (32 ms) with 50% overlap. Each segment is Hann windowed and transformed using a 256-point fast Fourier transform (FFT) following zero-padding. The FFT computes the modulation spectrum of each subband, with a frequency resolution of 15.6 Hz. Within each subband, the FFT magnitudes are multiplied by 15 triangular windows spaced uniformly across the

[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 may occur when the chosen threshold is lower than 0 dB).

15.6–400 Hz range and summed up to produce 15 modulation spectrum amplitudes. The 15 modulation amplitudes represent the AMS feature vector [23], which we denote by $\mathbf{a}(t,\ l)$, where t indicates the time (frame) index and l 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

$$\mathbf{x}(t,\ l) = [\mathbf{a}(t,\ l),\ \Delta\mathbf{a}_T(t,\ l),\ \Delta\mathbf{a}_L(t,\ l)] \tag{1}$$

where

$$\begin{aligned}
\Delta\mathbf{a}_T(1,\ l) &= \mathbf{a}(2,\ l) - \mathbf{a}(1,\ l), \quad t = 1 \\
\Delta\mathbf{a}_T(t,\ l) &= \mathbf{a}(t,\ l) - \mathbf{a}(t-1,\ l), \quad t = 2, \ldots, T \qquad (2) \\
\Delta\mathbf{a}_L(t,\ 1) &= \mathbf{a}(t,\ 2) - \mathbf{a}(t,\ 1), \quad l = 1 \\
\Delta\mathbf{a}_L(t,\ l) &= \mathbf{a}(t,\ l) - \mathbf{a}(t,\ l-1), \quad l = 2, \ldots, L \qquad (3)
\end{aligned}$$

where $\Delta\mathbf{a}_T(t,\ l)$ and $\Delta\mathbf{a}_L(t,\ l)$ denote the delta feature vectors computed across time and frequency, respectively, and T is the total number of segments in each sentence. The number of subbands, L, was set to 25 in this work, and the total dimension of the feature vector $\mathbf{x}(t,\ l)$ was $45\ (= 3 \times 15)$.

B. Training Stage

We use a Bayesian classifier to estimate the binary mask of each T-F unit. The distribution of the feature vectors of each class was represented with a Gaussian Mixture Model (GMM) composed of the parameters $\{w_k,\ \mathbf{m}_k,\ \mathbf{\Sigma}_k\},\ k = 1, 2, \ldots, K$ where K is the number of mixture components, w_k is the mixture weight, $\mathbf{m}_k$ is the (D-dimensional) mean, and $\mathbf{\Sigma}_k$ is the covariance matrix. The two classes, denoted as C_0 for mask 0 (masker-dominated T-F units), and C_1 for mask 1 (target-dominated T-F units), were further subdivided into two smaller classes, i.e., $C_0 = \{C_0^0, C_0^1\}$, $C_1 = \{C_1^0, C_1^1\}$. This subclass division yielded faster convergence in GMM training and better classification. In the training stage, the noisy speech spectrum, $X(t,\ l)$, was classified into one of four subclasses as follows:

$$X(t,\ l) \in \begin{cases} C_0^0, & \text{if } \xi(t,\ l) < T_{\mathrm{SNR0}} \\ C_0^1, & \text{if } T_{\mathrm{SNR0}} \le \xi(t,\ l) < T_{\mathrm{SNR}} \\ C_1^0, & \text{if } T_{\mathrm{SNR}} \le \xi(t,\ l) < T_{\mathrm{SNR1}} \\ C_1^1, & \text{if } T_{\mathrm{SNR1}} \le \xi(t,\ l) \end{cases} \tag{4}$$

where $\xi(t,\ l)$ is the local (true) SNR computed as the ratio of envelope energies of the (clean) target speech and masker signals, and $T_{\mathrm{SNR0}},\ T_{\mathrm{SNR1}},\ T_{\mathrm{SNR}}$ are thresholds. The T_{SNR0} was chosen in the training stage so as to have equal amount of training data in the C_0^0 and C_0^1 classes. Classification performance was not found to be sensitive to this threshold value. In the training stage, each T-F unit of the noisy speech was classified into one of four subclasses $\left(C_0^0,\ C_0^1,\ C_1^0,\ C_1^1\right)$ according to their local SNR. The SNR threshold 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 nonuniform masking of speech across the speech spectrum. We utilized 256-mixture[2] Gaussian models for modeling the distributions of the feature vectors in each class. Full covariance matrices[3] were used for each mixture and the initial Gaussian model parameters were obtained by k-means clustering. The *a priori* probability for each subclass $\left(P(C_0^0),\ P(C_0^1),\ P(C_1^0),\ P(C_1^1)\right)$ was calculated by counting the number of feature vectors belonging to the corresponding class and dividing that by the total number of feature vectors.

C. Enhancement Stage

In the enhancement stage, the binary masks are estimated by a Bayesian classifier which compares the *a posteriori* probabilities of the classification models. 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 waveform.

1) Bayesian Decision: The T-F units are classified as C_0 or C_1 by comparing two *a posteriori* probabilities, $P\left(C_0|\mathbf{x}(t,\ l)\right)$ and $P\left(C_1|\mathbf{x}(t,\ l)\right)$. This comparison produces an estimate of the binary mask, $G(t,\ l)$, as follows:

$$G(t,\ l) = \begin{cases} 0, & \text{if } P\left(C_0|\mathbf{x}(t,\ l)\right) > P\left(C_1|\mathbf{x}(t,\ l)\right) \\ 1, & \text{otherwise} \end{cases} \tag{5}$$

where $P\left(C_0|\mathbf{x}(t,\ l)\right)$ is computed using Bayes' theorem as follows:

$$\begin{aligned} P\left(C_0|\mathbf{x}(t,\ l)\right) &= \frac{P\left(C_0, \mathbf{x}(t,\ l)\right)}{P\left(\mathbf{x}(t,\ l)\right)} \\ &= \frac{P(C_0^0)P\left(\mathbf{x}(t,\ l)|C_0^0\right) + P(C_0^1)P\left(\mathbf{x}(t,\ l)|C_0^1\right)}{P\left(\mathbf{x}(t,\ l)\right)}. \end{aligned} \tag{6}$$

The *a posteriori* probability $P\left(C_1|\mathbf{x}(t,\ l)\right)$ is computed similarly.

2) Waveform Synthesis: Fig. 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 again [24]. 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 (5). Finally, the estimated target signal is reconstructed by summing the weighted responses of the 25 filters. Fig. 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 (5)] and synthesized waveform are shown in Figs. 3(c) and 3(d), respectively.

[2]Singularities may occur in the estimation of the covariance matrix due to insufficient amount of training data, high dimensionality, and unrestricted form of covariance matrix. When a singularity was detected, we disabled the corresponding mixture by setting the mixture weight to zero. In this paper, the effective number of mixture varied across frequency bands and ranged from 138 to 256 mixtures.

[3]We attempted to use diagonal covariance matrices by de-correlating the features using transforms such as the discrete cosine transform (DCT) and Karhunen-Loéve transform (KLT), but no significant benefit was observed. We believe that this was due to the strong data-dependency of the AMS features.

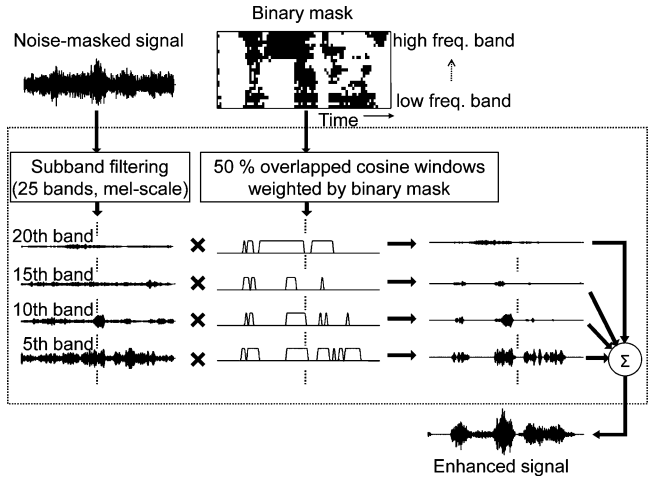

Fig. 2. Block diagram of the waveform synthesis stage of the proposed algorithm.

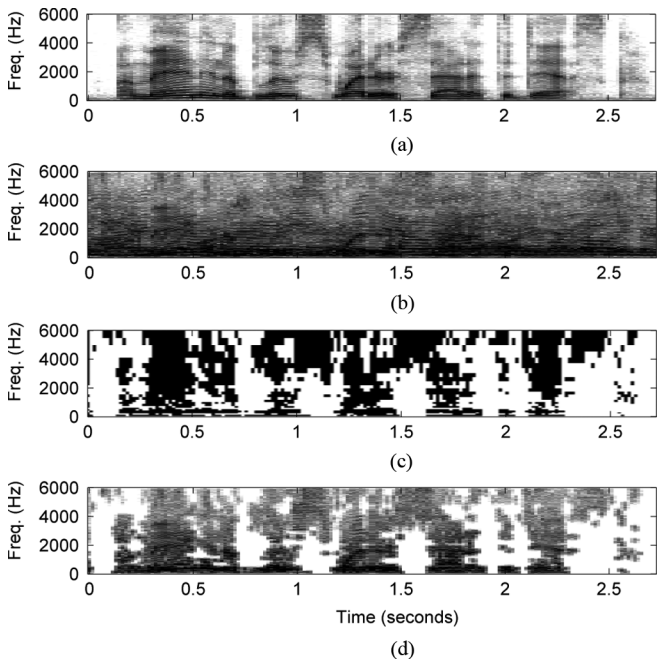

Fig. 3. (a) Wide-band spectrogram of an IEEE sentence in quiet. (b) Spectrogram of a sentence corrupted by multitalker babble at -5 dB SNR. (c) Binary mask estimated using (5), with black pixels indicating target-dominated T-F units and white pixels indicating masker-dominated T-F units. (d) Synthesized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

III. Adaptation to New Noise Environments

In the previous section, we described the enhancement of noise-masked speech based on the estimation of binary masks. In spite of the good performance obtained with GMMs trained in multiple listening environments [19], a user may encounter a new type of noise which is not included in the multiple-noise training set. There are several ways of handling a new noisy environment. One approach is to use a multi-style noise model trained on multiple types of noise. We tried such an approach, but the performance was less than satisfactory[4]. An alternative approach is to adapt the model parameters in the new environment. For rapid adaptation to a new noisy environment, we consider incrementally updating the GMM parameters to accommodate the new data, starting from an initial model trained with small amounts of data.[5] Next, we describe the incremental GMM adaptation technique used. Unlike the batch-training approach which requires access to the whole data set, the incremental training approach continuously adapts the model parameters as new data arrive. Consequently, the computational load of the incremental approach is smaller than the load of the batch-training approach.

A. Initial Model

Access to a small amount of speech data recorded in quiet is assumed for the training of the initial model. Such data can be stored in memory. In a new listening environment, noise-only data are collected and mixed with ten sentences of clean speech (stored in memory) at SNR levels of -5, 0, 5 dB. The distribution of each class can be modeled with only a small number of mixture components (e.g., 8), given the small number of sentences (e.g., ten sentences) in the training data. Although the method of splitting or adding Gaussian mixtures can be used to increase the number of mixture components as more data become available, we considered a simpler way of training the GMMs with 256-mixture components from the beginning. In the incremental training approach adopted, we only update the parameters of each Gaussian while fixing the number of mixtures used. The initial model was built using the following two steps. First, 32 different eight-mixture models were created, based on the same training data, by repeating the initial training procedure 32 times. At each training iteration, the initial centroids for the k-means clustering are randomly chosen, leading to 32 different models. In the second step, the initial model with 256 mixtures is created by merging the 32 models trained with eight mixtures. Since the same training data is used for the training of all the eight-mixture models, the initial 256-mixture model has much redundancy, suggesting that many Gaussian components are similar to each other. The redundancy of the models is analyzed and discussed in more detail in Section IV-A.

B. Incremental Training

We assume that an independent observation sequence is available at each updating step n for the $l\text{th}$ subband, and that the $n\text{th}$ observation sequence is given by

$$\mathbf{X}^{(n)}(l) = \left\{\mathbf{x}^{(n)}(1,l),\ \mathbf{x}^{(n)}(2,l), \ldots,\ \mathbf{x}^{(n)}(T_n,l)\right\} \quad (7)$$

[4]We built a multi-style noise trained model based on 34 types of noises and tested the model at -5 dB SNR on three types of noise environments (babble, factory, speech-shaped noise) not included in the training. The performance, measured in terms of detection rates (hit—false alarm rate), were 15.31% (babble), 20.87% (factory), 15.29% (speech-shaped noise). This performance was significantly lower than that attained by the incremental approach. It was even worse than that obtained by the initial models which were trained using only ten sentences.

[5]We also tried other adaptation techniques such as the MAP technique [25] and the maximum-likelihood linear regression (MLLR) [26] technique based on multi-style noise trained models, but performance was found to be poor compared to that obtained with the incremental training approach.

where T_n is the number of frames in the nth update, and $\mathbf{x}^{(n)}(t,\ l)$ is given by (1). For incremental training based on small amounts of data, we adopted the quasi-Bayes approach which provides an approximate solution by conducting recursive Bayes estimation of the mixture coefficients [27], [28]. Consider a mixture model $\lambda = \{w_k,\ \mathbf{m}_k,\ \mathbf{\Sigma}_k\}, k = 1, \ldots, K$ where w_k is the mixture weight, $\mathbf{m}_k$ is the $D \times 1$ mean vector, and $\mathbf{\Sigma}_k$ is the $D \times D$ covariance matrix of the kth mixture. Assuming independence between the parameters of the individual mixture components $(\mathbf{m}_k,\ \mathbf{\Sigma}_k)$ and the set of the mixture weights $(w_1, \ldots, w_K)$, the initial prior pdf of model λ, $g(\lambda)$, is assumed to be the product of the prior pdfs [25], [27], [28]

$$g(\lambda) = g(w_1, \ldots, w_K) \prod_{k=1}^{K} g(\mathbf{m}_k, \mathbf{\Sigma}_k) \tag{8}$$

$$g(w_1, \ldots, w_K) \propto \prod_{k=1}^{K} w_k^{\nu_k - 1} \tag{9}$$

$$\begin{aligned} g(\mathbf{m}_k, \mathbf{\Sigma}_k) \propto\ & |\mathbf{\Sigma}_k|^{-(\alpha_k - D)/2} \\ & \times \exp\left[-\frac{\tau_k}{2}(\mathbf{m}_k - \boldsymbol{\mu}_k)^T \mathbf{\Sigma}_k^{-1}(\mathbf{m}_k - \boldsymbol{\mu}_k)\right] \\ & \times \exp\left[-\frac{1}{2}\mathrm{tr}\left(\mathbf{U}_k \mathbf{\Sigma}_k^{-1}\right)\right] \end{aligned} \tag{10}$$

where ν_k, τ_k, $\boldsymbol{\mu}_k$, α_k, $\mathbf{U}_k$ are the hyperparameters of the prior density such that $\nu_k > 0$, $\alpha_k > D - 1$, $\tau_k > 0$, $\boldsymbol{\mu}_k$ is a vector of dimension D and $\mathbf{U}_k$ is a $D \times D$ positive definite matrix, and $\mathrm{tr}(\cdot)$ denotes the trace of a matrix. The approximate MAP estimation is obtained by finding model parameters that maximize the posterior density $g(\lambda|\varphi)$ at each step, where φ denotes the set of hyperparameters. For the training sequence $\mathbf{X}_n$, the hyperparameters are updated as follows (the subband index l is omitted for the sake of brevity)

$$\tau_k^{(n)} = \rho \tau_k^{(n-1)} + \chi_k^{(n)} \tag{11}$$

$$\alpha_k^{(n)} = \rho\left(\alpha_k^{(n-1)} - D\right) + D + \chi_k^{(n)} \tag{12}$$

$$\nu_k^{(n)} = m\rho\left(\nu_k^{(n-1)} - 1\right) + 1 + \chi_k^{(n)} \tag{13}$$

$$\boldsymbol{\mu}_k^{(n)} = \beta^{(n)} \bar{\mathbf{x}}_k^{(n)} + \left(1 - \beta^{(n)}\right) \boldsymbol{\mu}_k^{(n-1)} \tag{14}$$

$$\begin{aligned} \mathbf{U}_k^{(n)} = & \rho \mathbf{U}_k^{(n-1)} + \mathbf{S}_k^{(n)} + \frac{\rho \tau_k^{(n-1)} \chi_k^{(n)}}{\rho \tau_k^{(n-1)} + \chi_k^{(n)}} \\ & \cdot \left(\bar{\mathbf{x}}_k^{(n)} - \boldsymbol{\mu}_k^{(n-1)}\right)\left(\bar{\mathbf{x}}_k^{(n)} - \boldsymbol{\mu}_k^{(n-1)}\right)^T \end{aligned} \tag{15}$$

where

$$\gamma_k^{(n)}(t) = \frac{w_k \cdot b_k\left(\mathbf{x}^{(n)}(t)\right)}{\sum_{m=1}^{K} w_m \cdot b_m\left(\mathbf{x}^{(n)}(t)\right)} \tag{16}$$

$$\begin{aligned} b_k(\mathbf{x}(t)) = & \frac{1}{(2\pi)^{D/2} |\mathbf{\Sigma}_k|^{1/2}} \\ & \cdot \exp\left(-\frac{1}{2}(\mathbf{x}(t) - \boldsymbol{\mu}_k)^T \mathbf{\Sigma}_k^{-1}(\mathbf{x}(t) - \boldsymbol{\mu}_k)\right) \end{aligned} \tag{17}$$

$$\chi_k^{(n)} = \sum_{t=1}^{T_n} \gamma_k^{(n)}(t) \tag{18}$$

$$\bar{\mathbf{x}}_k^{(n)} = \frac{\sum_{t=1}^{T_n} \gamma_k^{(n)}(t) \mathbf{x}^{(n)}(t)}{\chi_k^{(n)}} \tag{19}$$

$$\mathbf{S}_k^{(n)} = \sum_{t=1}^{T_n} \gamma_k^{(n)}(t)\left(\mathbf{x}^{(n)}(t) - \bar{\mathbf{x}}_k^{(n)}\right)\left(\mathbf{x}^{(n)}(t) - \bar{\mathbf{x}}_k^{(n)}\right)^T \tag{20}$$

$$\beta^{(n)} = \frac{\chi_k^{(n)}}{\rho \tau_k^{(n-1)} + \chi_k^{(n)}} \quad n = 2,\ 3,\ \ldots \tag{21}$$

and ρ is a forgetting coefficient, which was set to $\rho = 0.9$ in this work. From the initial model, the hyperparameters are initialized as

$$\tau_k^{(1)} = \chi_k^{(1)} \tag{22}$$

$$\nu_k^{(1)} = 1 + \tau_k^{(1)} \tag{23}$$

$$\alpha_k^{(1)} = D + \tau_k^{(1)} \tag{24}$$

$$\boldsymbol{\mu}_k^{(1)} = \mathbf{m}_k^{(1)} \tag{25}$$

$$\mathbf{U}_k^{(1)} = \tau_k^{(1)} \cdot \mathbf{\Sigma}_k^{(1)} \tag{26}$$

where $\mathbf{m}_k^{(1)}$ and $\mathbf{\Sigma}_k^{(1)}$ are the mean vector and covariance matrix of the initial model, respectively. Finally, the parameters of the GMMs are updated using the hyperparameters as follows:

$$\hat{w}_k = \frac{\nu_k^{(n)} - 1}{\sum_{k=1}^{K}\left[\nu_k^{(n)} - 1\right]} \tag{27}$$

$$\hat{\mathbf{m}}_k = \boldsymbol{\mu}_k^{(n)} \tag{28}$$

$$\hat{\mathbf{\Sigma}}_k = \left(\alpha_k^{(n)} - D\right)^{-1} \mathbf{U}_k^{(n)}. \tag{29}$$

In this paper, three iterations of the above expectation–maximization (EM) procedure were performed.

IV. Experimental Results

Listening experiments were conducted to assess the performance of the incrementally adapted GMMs as a function of the number of sentences used for training. Speech was synthesized (see Fig. 2) with the adapted GMMs using the algorithm outlined in Section II, and presented to normal-hearing listeners for identification. The aim of the listening experiments is to determine the smallest number of sentences, or equivalently the minimum amount (in terms of duration) of training data required to obtain significant improvements in intelligibility relative to that of unprocessed (noise-masked) sentences.

A. Materials and Procedure

1) Speech and Noise Material: Sentences taken from the IEEE database [29] were used as test material for the listening experiments. The IEEE sentences are phonetically balanced with relatively low word-context predictability. The IEEE corpus was recorded in a soundproof booth and is available from [1]. The sentences were originally recorded at a sampling rate of 25 kHz and downsampled to 12 kHz. Three types of noise were used as maskers: factory noise, babble, and train noise. The factory noise was taken from the NOISEX database [30], and the babble from the Auditec CD (St. Louis, MO)

recorded by 20 talkers with equal number of male and female speakers. The train noise was recorded inside a train running at a speed of about 100 km/h. 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.

2) Model Training: A total of 200 sentences were used to train the incrementally updated models. The 200 sentences were grouped into 11 sets (S1-S11): the first two sets (S1,S2) contained ten sentences each, and the remaining nine sets (S3-S11) contained 20 sentences each. The ten sentences in the first set (S1) were used to train eight-mixture GMMs. As mentioned in Section III-A, a total of 32 different eight-mixture GMMs were modeled with S1 and merged to produce 256-mixture GMMs. These models served as the initial models for the incremental GMM training algorithm. After creating the initial model with S1, the model was incrementally updated with each data set starting from set S2 through S11 using the algorithm outlined in Section III.

3) Model Redundancy: As stated in Section III-A, the initial model has much redundancy, suggesting that many Gaussian components are similar to each other. The degree of redundancy, however, is expected to become smaller as the models are incrementally updated with more data. To examine the degree of redundancy in the GMMs, we used the symmetric Kullback–Leibler divergence. For two Gaussian distributions $\mathcal{N}_1(\mathbf{m}_1,\ \mathbf{\Sigma}_1)$ and $\mathcal{N}_2(\mathbf{m}_2, \mathbf{\Sigma}_2)$, the divergence is given by

$$\begin{aligned} D_{SKL}(\mathcal{N}_1,\ \mathcal{N}_2) &= D_{KL}(\mathcal{N}_1||\mathcal{N}_2) + D_{KL}(\mathcal{N}_2||\mathcal{N}_1) \\ &= \frac{1}{2}\left(\log\left(\frac{\det\mathbf{\Sigma}_2}{\det\mathbf{\Sigma}_1}\right) + \mathrm{tr}(\mathbf{\Sigma}_2^{-1}\mathbf{\Sigma}_1)\right. \\ &\quad \left.+(\mathbf{m}_2-\mathbf{m}_1)^T\mathbf{\Sigma}_2^{-1}(\mathbf{m}_2-\mathbf{m}_1) - D\right) \\ &\quad + \frac{1}{2}\left(\log\left(\frac{\det\mathbf{\Sigma}_1}{\det\mathbf{\Sigma}_2}\right) + \mathrm{tr}(\mathbf{\Sigma}_1^{-1}\mathbf{\Sigma}_2)\right. \\ &\quad \left.+(\mathbf{m}_1-\mathbf{m}_2)^T\mathbf{\Sigma}_1^{-1}(\mathbf{m}_1-\mathbf{m}_2) - D\right) \end{aligned} \tag{30}$$

where $\mathbf{m}_1$, $\mathbf{m}_2$, $\mathbf{\Sigma}_1$, and $\mathbf{\Sigma}_2$ are the $D \times 1$ mean vectors and $D \times D$ covariance matrices, respectively. Specifically, we calculated the divergence between two Gaussians (from all possible combinations) in each class and counted the number of Gaussian pairs which had a smaller divergence than a certain value (small divergence values indicate high similarity between two distributions). Fig. 4 shows the number of Gaussian pairs averaged over four classes and 25 subbands for the initial model and incrementally updated models. The initial model was trained with ten sentences, and then updated with another ten sentences. After that, 20 sentences were used for every update. As can be seen from this figure, as the models are incrementally updated, the number of Gaussian pairs with smaller divergence than a certain value decreases, which in turn implies a decrease in the model redundancy.

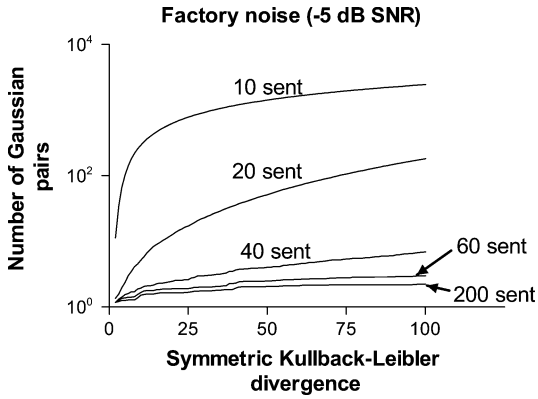

Fig. 4. Plot showing the number of Gaussian pairs with Kullback–Leibler divergence smaller than a certain value. This number was averaged over 25 subbands and four classes.

B. Listening Tests

Nine normal-hearing listeners participated in the listening experiments. The normal-hearing listeners were paid for their participation. The listeners participated in a total of 21 conditions ($=$ 6 training sets $\times 3$ maskers $+3$ unprocessed conditions). The six training sets included respectively: 10, 20, 40, 80, 140, and 200 sentences. The duration of each sentence was approximately 2.5 s. The experiments were performed in a soundproof 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. The listening level was controlled by each individual but was fixed throughout the test for a particular subject. Prior to the sentence test, each subject listened to a set of noise-masked sentences to become familiar with the testing procedure. Five-minute breaks were given to the subjects every 30 minutes. A total of 20 sentences were used per condition, and none of the sentences were repeated across conditions. The order of the conditions was randomized across subjects. Listeners were asked to write down the words they heard, and intelligibility performance was assessed by counting the number of words identified correctly.

C. Results: Intelligibility Evaluation

GMMs adapted using 10–200 sentences were used to synthesize (based on the algorithm presented in Section II) speech mixed with babble (-5 dB SNR), factory noise (-5 dB SNR) and train noise (-10 dB SNR). The synthesized speech was presented to human listeners for identification, and the results, scored in terms of percentage of words identified correctly, are plotted in Fig. 5 as a function of the amount of training data used (i.e., accumulated number of sentences used in each model). The word identification scores obtained with GMMs batch-trained with 390 sentences [19] are plotted for comparison. As expected, word identification accuracy improved as more training data were included. Performance in the babble conditions improved substantially from 15% correct obtained with unprocessed sentences to 55% correct with 20-sentence models, and to 80% correct with 140-sentence models. While the initial model trained with ten sentences provided improvement in the case of babble, a larger number of sentences was

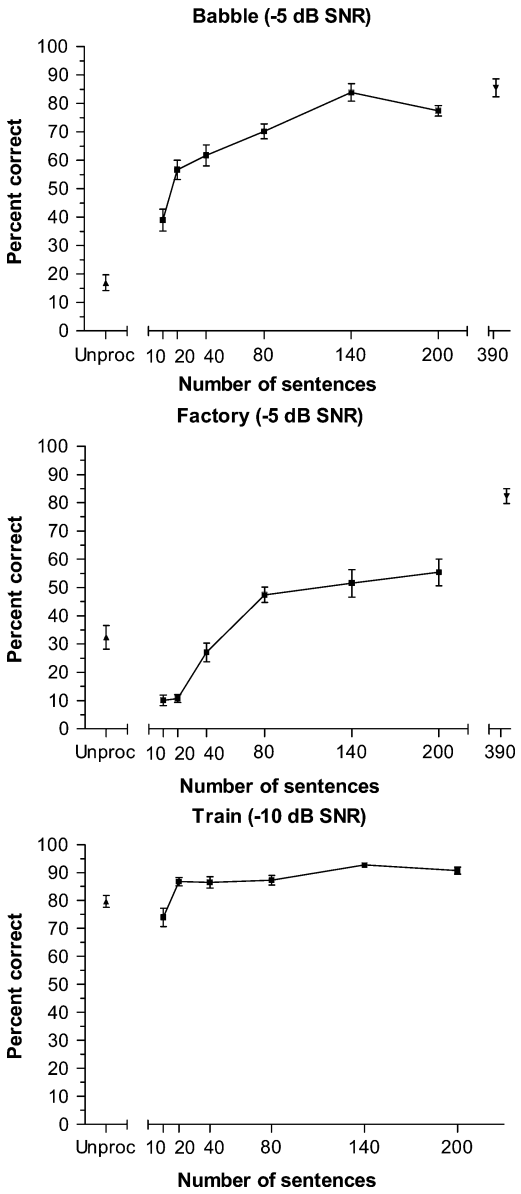

Fig. 5. Mean intelligibility scores (by normal-hearing listeners) for corrupted (unprocessed) sentences (denoted as Unproc) and sentences synthesized using incrementally updated models as a function of the number of accumulated sentences used in training. The intelligibility scores obtained with batch-trained models based on 390 sentences are also shown for comparison [19]. Error bars indicate standard errors of the mean.

required to yield a better score in the case of factory and train noise. Large improvement in intelligibility was also noted with factory noise, when 80 or more sentences were used to train the GMMs. The improvement in train noise conditions was smaller, partly because we were limited by ceiling effects (e.g., performance saturated near 90%–100%). The intelligibility of speech mixed with train noise is considerably higher than that obtained in babble (even at lower SNRs) because the train noise is modulated, and as such it provides to the listeners the opportunity to "glimpse" the target during the silent gaps or waveform "dips" of the train noise [1, Ch. 4]. In brief, the improvement in intelligibility was found to be statistically significant for all three maskers tested.

Analysis of variance (ANOVA) [31], with repeated measures, indicated a significant effect ($F_{6,48} = 205.1$, $p < 0.0005$ for babble; $F_{6,48} = 58.6$, $p < 0.0005$ for factory noise; $F_{6,48} = 24.9$, $p < 0.0005$ for train noise) of the amount of training data on sentence intelligibility. Post-hoc tests, according to Fisher's least significant difference (LSD) [31], were run to assess significant differences in intelligibility scores in the various conditions. For babble, the score obtained with ten sentences was significantly ($p < 0.0005$) higher than the score obtained with unprocessed (noise-masked) sentences. The score obtained with 80-sentence models did not differ significantly ($p = 0.116$) from the score obtained with the 200-sentence model. The score obtained with 140 sentences was not significantly ($p = 0.377$) higher than the score obtained with 200 sentences, but was higher than the score obtained with 80 sentences. For factory noise, asymptotic performance was attained with 80-sentence models. Scores obtained with 80 sentences did not differ significantly ($p = 0.304$) from scores obtained with 200 sentences. For the train noise, scores obtained with 80 sentences did not differ significantly ($p = 0.312$) from scores obtained with 200 sentences. Furthermore, the scores obtained with 20-sentence models was higher than that obtained with unprocessed (noise-masked) sentences, and the difference was marginally significant ($p = 0.05$). The score obtained with 80-sentence models was significantly ($p = 0.029$) higher than the score obtained with unprocessed sentences.

In summary, the above statistical analysis indicates that in most cases, large gains in intelligibility can be obtained with as few as ten sentences and as many as 80 sentences. For two of the three maskers tested, asymptotic performance was obtained with 80 sentences. For babble, performance obtained with batch training (390 sentences) was comparable to that attained with 140 sentences. That was not the case for factory noise, as the performance with 390 sentences was significantly higher than that obtained with 200 sentences. Overall, a minimum of 80 sentences was required to achieve substantial gains in intelligibility, relative to that obtained with unprocessed (noise-masked) sentences, for the three maskers tested. This amount of training data is considerably smaller than what has been used by others for estimating the binary mask (e.g., [21]) or estimating the gain function [8], [10]. In [21] for instance, approximately 1–2 hours of training data were required. Aside from the limited amount of training data required, another advantage of the proposed approach is that the GMM training used in this work does not require access to a labeled speech corpus, while other studies [21], [32], [33] required the use of accurate F0 detection algorithms or voiced/unvoiced segmentation algorithms.

As mentioned earlier, the GMMs were trained with data embedded in -5 to 5 dB SNR. This raises the question as to whether the performance of the GMMs would degrade if tested in an environment with SNR level outside the range of -5 to 5 dB. To test this SNR mismatch issue, we performed additional listening experiments wherein the GMMs were trained in -5 to 5 dB SNR levels, but tested at 10 dB SNR. Three additional listeners participated in the intelligibility listening tests, and the results are shown in Fig. 6. As can be seen, intelligibility scores were high and unaffected for the babble and train maskers. Consistent with the data in Fig. 5, 20–40 sentences seemed to be

TABLE I
HIT (HIT) AND FALSE ALARM RATES (FA) OBTAINED USING THE INCREMENTALLY UPDATED MODELS

		Accumulated number of sentences used in each model						
		10	20	40	80	140	200	390
Babble	HIT	70.81	68.46	77.95	83.48	85.06	85.84	89.56
	FA	23.07	15.52	17.84	17.42	14.93	16.02	19.60
	HIT-FA	47.74	52.94	60.11	66.06	70.13	69.82	69.96
	Error rate	25.08	20.95	19.24	17.11	14.93	15.43	16.60
Factory	HIT	46.28	43.16	61.46	70.40	72.33	71.12	82.10
	FA	16.13	14.51	18.30	14.31	17.03	12.37	16.26
	HIT-FA	30.15	29.10	43.16	56.09	55.30	58.75	65.84
	Error rate	25.57	25.33	23.17	18.00	19.52	16.22	16.66
Train	HIT	82.09	89.35	91.25	86.73	87.19	87.25	–
	FA	42.65	38.19	37.05	29.96	29.52	27.73	
	HIT-FA	39.44	51.16	54.20	56.77	57.67	59.52	
	Error rate	31.81	25.66	24.46	22.44	21.80	21.14	

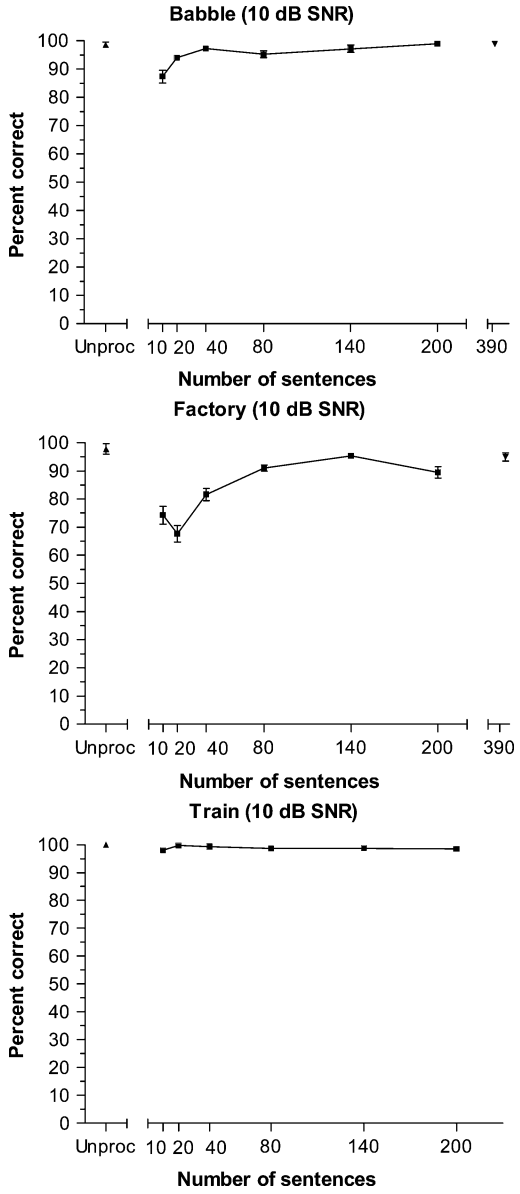

Fig. 6. Mean intelligibility scores (by normal-hearing listeners) for corrupted (unprocessed) sentences (denoted as Unproc) and sentences synthesized using incrementally updated models as a function of the number of accumulated sentences used in training. The GMMs were trained using data embedded in -5 to 5 dB SNR, but tested at 10-dB SNR. The intelligibility scores obtained with batch-trained models based on 390 sentences are also shown for comparison. Error bars indicate standard errors of the mean.

sufficient to reach high levels ($> 90\%$ correct) of performance. For the factory masker, 80 sentences were needed at least to reach high levels of performance. Overall, the data shown in Fig. 6 do not exhibit a sensitivity to the SNR level and for the most part are in agreement and consistent with the data obtained in Fig. 5.

D. Results: Objective Evaluation

We attribute the large gains in intelligibility obtained with the proposed algorithm (Section II) to the accurate classification of T-F units into target-dominant and masker dominant T-F units. To quantify the accuracy of the GMM-based SNR classifier, we computed the hit (HIT) and false alarms (FA) of the same test sets used in the listening experiments. The classification accuracy, expressed in terms of HIT and FA, of the trained GMM classifiers is tabulated in Table I as a function of the number of accumulated sentences used in the training. We also calculated the error rates assessing the classifier's performance without making a distinction between miss or false alarm errors. In terms of reduction in error rates (computed relative to ten-sentence models), substantial reduction was noted in all three maskers tested and ranged from a 34% error reduction (obtained with train noise with 200-sentence models) to 38% error reduction (obtained with babble with 200-sentence models).

In terms of detection rates, the hit rate improved as more training data were included and in most cases the false alarm rate decreased. Perceptually, the two types of errors that can be introduced, namely miss ($=$ 1-HIT) and false alarm, are not equivalent [16]. This is so, because the false alarm errors will possibly introduce more noise distortion, as T-F units that would otherwise be zeroed-out (presumably belonging to the masker) would now be retained. The miss errors will likely introduce speech distortion, as these errors are responsible for zeroing out T-F units that are dominated by the target signal and should therefore be retained. To account for the combined effect of both errors (miss and false alarm), we propose the use of the difference metric, $d =$ HIT-FA. Table I tabulates the difference metric d as a function of the number of accumulated sentences used in the training. As can be seen, the value of the difference metric d increases as more training data are included suggesting possibly a correlation between d and speech intelligibility scores. To examine this, we computed the correlation between the difference metric d and speech intelligibility scores

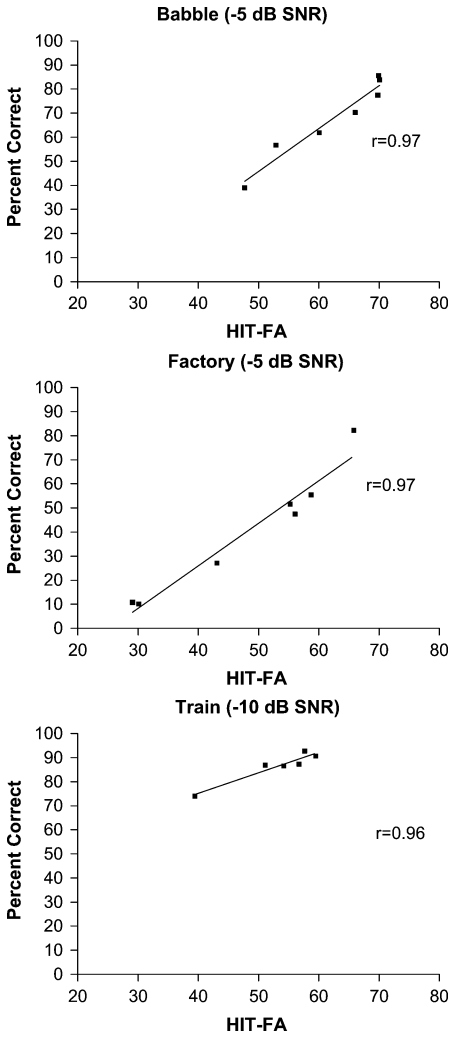

Fig. 7. Scatter plots showing the correlation between human listener's recognition scores and the difference metric HIT-FA.

using Pearson's product–moment coefficient. The resulting correlation coefficient was consistently high $(r = 0.96 - 0.97)$ for the three maskers tested. Scatter plots of d and speech intelligibility scores are shown in Fig. 7. As demonstrated by the data in Fig. 7, the difference metric d can serve as an effective objective measure for predicting speech intelligibility of algorithms that estimate the binary mask. Previous studies have used an SNR measure [32] computed based on the normalized difference between the signals synthesized (in the time domain) using the ideal and estimated binary masks. While the SNR measure is reasonable, it has not been validated with listening experiments; hence, it is not clear whether the SNR measure correlates with the subjective quality or intelligibility of speech synthesized using estimated binary masks. In contrast, the proposed difference metric d is relatively simple to compute and has been shown (Fig. 7) in the present study to predict reliably speech intelligibility.

Fig. 8 compares performance, in terms of the difference metric (HIT-FA), between the batch training and the incremental training methods. Performance (in terms of the difference metric) with the incremental training approach is

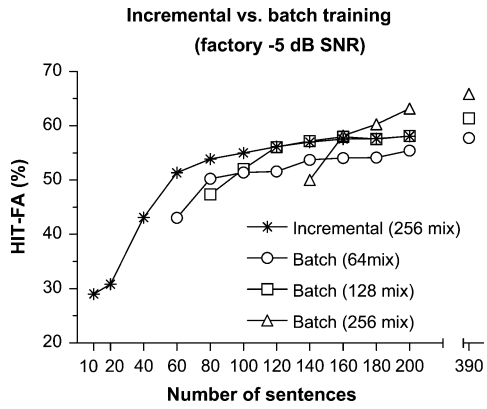

Fig. 8. Performance (expressed in terms of the difference between the hit rate and false alarm rates, i.e., HIT-FA) comparison between models which were either incrementally updated or batch trained. Performance with the batch-trained models is plotted only for a large number of sentences, as that was required to train GMMs with a large number of mixtures.

slightly lower than that obtained using the batch training approach when more than 160 sentences are used for training, but it is comparable to that obtained by the batch training approach when fewer than 160 sentences are used.

It should be noted that the hit rates obtained by the GMM binary classifiers (Table I) are substantially higher than those obtained with conventional noise-reduction algorithms [19], [34], which estimate the SNR in each T-F unit using the decision-directed approach. When tested in three different environments, the MMSE algorithm [11], for instance, yielded hit rates in the range of 57%-68%, and false alarm rates in the range of 52%-66% [19]. The false rate obtained with the MMSE algorithm is substantially higher than that attained by the GMM classifiers (Table I), and the hit rate is substantially lower than that obtained by the GMM classifiers. Similar performance was also obtained when the Wiener algorithm was used [19]. This outcome might explain, at least, partially why current noise reduction algorithms, even the most sophisticated ones, do not improve speech intelligibility [2].

V. Conclusion

Large gains in intelligibility were achieved with the proposed algorithm using a limited amount of training data. In most conditions, a minimum of 80 sentences was found to be sufficient to obtain significant improvements in intelligibility. The intelligibility of speech processed by the proposed algorithm was substantially higher than that achieved by human listeners listening to unprocessed (corrupted) speech. 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. 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 (MFCCs) [35] features commonly used in speech recognition, the AMS features capture information about amplitude and frequency modulations, known to be critically important for speech intelligibility [36]. An objective measure based on the classification accuracy (HIT-FA)

of the Bayesian classifier was also proposed for predicting intelligibility of speech synthesized by algorithms that estimate the binary mask. This measure was found to predict reliably $(r = 0.97)$ speech intelligibility. Overall, the findings from this study suggest that algorithms that can estimate (or classify) reliably the SNR in each T-F unit *can* improve speech intelligibility.

REFERENCES

[1] P. C. Loizou, *Speech Enhancement: Theory and Practice*. Boca Raton: CRC, 2007.

[2] Y. Hu and P. C. Loizou, "A comparative intelligibility study of single-microphone noise reduction algorithms," *J. Acoust. Soc. Amer.*, vol. 122, pp. 1777–1786, 2007.

[3] J. S. Lim, "Evaluation of a correlation subtraction method for enhancing speech degraded by additive white noise," *IEEE Trans. Acoust. Speech Signal Process.*, vol. ASSP-26, no. 5, pp. 471–472, Oct. 1978.

[4] J. A. Zakis, H. Dillon, and H. J. McDermott, "The design and evaluation of a hearing aid with trainable amplification parameters," *Ear Hear.*, vol. 28, no. 6, pp. 812–830, 2007.

[5] J. E. Porter and S. F. Boll, "Optimal estimators for spectral restoration of noisy speech," in *Proc. Int. Conf. Acoust. Speech Signal Process.*, 1984, pp. 18A.2.1–18A.2.4.

[6] S. Srinivasan, J. Samuelsson, and W. B. Kleijn, "Codebook driven short-term predictor parameter estimation for speech enhancement," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 1, pp. 163–176, Jan. 2006.

[7] J. Erkelens, J. Jensen, and R. Heusdens, "A general optimization procedure for spectral speech enhancement methods," in *Proc. Eur. Signal Proc. Conf.*, Florence, Italy, Sep. 2006.

[8] J. Erkelens, J. Jensen, and R. Heusdens, "A data-driven approach to optimizing spectral speech enhancement methods for various error criteria," *Speech Commun.*, vol. 49, pp. 530–541, 2007.

[9] T. Fingscheidt and S. Suhadi, "Data-driven speech enhancement," in *Proc. ITG-Fachtagung Sprachkommunikation*, Kiel, Germany, 2006.

[10] T. Fingscheidt, S. Suhadi, and S. Stan, "Environment-optimized speech enhancement," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 4, pp. 825–834, May 2008.

[11] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-32, no. 6, pp. 1109–1121, Dec. 1984.

[12] T. Lotter and P. Vary, "Speech enhancement by MAP spectral amplitude estimation using a super-Gaussian speech model," *EURASIP J. Appl. Signal Process.*, vol. 7, pp. 1110–1126, 2005.

[13] C. Bin and P. C. Loizou, "A Laplacian-based MMSE estimator for speech enhancement," *Speech Commun.*, pp. 134–143, 2007.

[14] P. C. Loizou, "Speech enhancement based on perceptually motivated Bayesian estimators of the magnitude spectrum," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 5, pp. 857–869, Sep. 2005.

[15] D. Brungart, P. Chang, B. Simpson, and D. Wang, "Isolating the energetic component of speech-on-speech masking with ideal time-frequency segregation," *J. Acoust. Soc. Amer.*, vol. 120, pp. 4007–4018, 2006.

[16] N. Li and P. C. Loizou, "Factors influencing intelligibility of ideal binary-masked speech: Implications for noise reduction," *J. Acoust. Soc. Amer.*, vol. 123, no. 3, pp. 1673–1682, 2008.

[17] N. Li and P. C. Loizou, "Effect of spectral resolution on the intelligibility of ideal binary masked speech," *J. Acoust. Soc. Amer.*, vol. 123, no. 4, pp. 59–64, 2008.

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

[19] G. Kim, Y. Lu, Y. Hu, and P. C. Loizou, "An algorithm that improves speech intelligibility in noise for normal-hearing listeners," *J. Acoust. Soc. Amer.*, vol. 126, no. 3, pp. 1486–1494, 2009.

[20] B. Kollmeier and R. Koch, "Speech enhancement based on physiological and psychoacoustical models of modulation perception and binaural interaction," *J. Acoust. Soc. Amer.*, vol. 95, no. 3, pp. 1593–1602, 1994.

[21] M. Seltzer, B. Raj, and R. Stern, "A Bayesian classifier for spectrographhic mask estimation for missing feature speech recognition," *Speech Commun.*, vol. 43, pp. 379–393, 2004.

[22] G. Langner and C. Schreiner, "Periodicity coding in the inferior colliculus of the cat. I: Neuronal mechanisms," *J. Neurophysiol.*, vol. 60, no. 6, pp. 1799–1822, 1988.

[23] J. Tchorz and B. Kollmeier, "SNR estimation based on amplitude modulation analysis with applications to noise suppression," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 3, pp. 184–192, May 2003.

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

[25] J.-L. Gauvain and C.-H. Lee, "Maximum *a posteriori* estimation for multivariate Gaussian mixture observations of Markov chains," *IEEE Trans. Speech Audio Process.*, vol. 2, no. 2, pp. 291–298, Apr. 1994.

[26] C. J. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Comput. Speech Lang.*, vol. 9, pp. 171–185, 1995.

[27] Q. Huo, C. Chan, and C.-H. Lee, "On-line adaptation of the SCHMM parameters based on the segmental quasi-Bayes learning for speech recognition," *IEEE Trans. Speech Audio Process.*, vol. 4, no. 2, pp. 141–144, 1996.

[28] Q. Huo and C.-H. Lee, "On-line adaptive learning of the continuous density hidden markov model based on approximate recursive Bayes estimate," *IEEE Trans. Speech Audio Process.*, vol. 5, no. 2, pp. 161–172, Mar. 1997.

[29] "IEEE recommended practice for speech quality measurements," *IEEE Trans. Audio Electroacoust.*, vol. 19, no. 3, pp. 225–246, Sep. 1969.

[30] A. Varga and H. J. M. Steeneken, "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.*, vol. 12, pp. 247–251, 1993.

[31] L. Ott, *An Introduction to Statistical Methods and Data Analysis*, 3rd ed. Boston, MA: PWS-Kent, 1988.

[32] G. Hu and D. L. Wang, "Monaural speech segregation based on pitch tracking and amplitude modulation," *IEEE Trans. Neural Netw.*, vol. 15, no. 5, pp. 1135–1150, Sep. 2004.

[33] G. Hu and D. L. Wang, "Segregation of unvoiced speech from nonspeech interference," *J. Acoust. Soc. Amer.*, vol. 124, pp. 1306–1319, 2008.

[34] Y. Hu and P. C. Loizou, "Techniques for estimating the ideal binary mask," in *Proc. 11th Int. Workshop Acoust. Echo Noise Control*, Sep. 2008.

[35] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. Acoust. Speech Signal Process.*, vol. ASSP-28, no. 4, pp. 357–336, Aug. 1980.

[36] F.-G. Zeng, K. Nie, G. S. Stickney, Y.-Y. Kong, M. Vongphoe, A. Bhargave, C. Wei, and K. Cao, "Speech recognition with amplitude and frequency modulations," *Proc. Nat. Acad. Sci. USA*, vol. 102, no. 7, pp. 2293–2298, 2005.

Gibak Kim received the B.S. and M.S. degrees in electronics engineering and the Ph.D. degree in electrical engineering from Seoul National University, Seoul, Korea, in 1994, 1996, and 2007, respectively.

From 1996 to 2000, he was with the Machine Intelligence Group, Department of the Information Technology, LG Electronics Inc., Seoul. He also worked at Voiceware, Ltd., from 2000 to 2003, as a Senior Research Engineer involved in the development of the automatic speech recognizer. Since 2007, he has been a Research Associate at the University of Texas at Dallas, Richardson, working on the development of noise-reduction algorithms that can improve speech intelligibility. His general research interests are in speech enhancement, speech recognition, and microphone array signal processing.

Philipos C. Loizou (S'90–M'91–SM'04) received the B.S., M.S., and Ph.D. degrees in electrical engineering from Arizona State University, Tempe, in 1989, 1991, and 1995, respectively.

From 1995 to 1996, he was a Postdoctoral Fellow in the Department of Speech and Hearing Science, Arizona State University, working on research related to cochlear implants. He was an Assistant Professor at the University of Arkansas, Little Rock, from 1996 to 1999. He is now a Professor and holder of the Cecil and Ida Green Chair in the Department of Electrical Engineering, University of Texas at Dallas, Richardson. His research interests are in the areas of signal processing, speech processing, and cochlear implants. He is the author of the textbook *Speech Enhancement: Theory and Practice* (CRC, 2007) and coauthor of the textbooks *An Interactive Approach to Signals and Systems Laboratory* (National Instruments, 2008) and *Advances in Modern Blind Signal Separation Algorithms: Theory and Applications* (Morgan & Claypool, 2010).

Dr. Loizou is a Fellow of the Acoustical Society of America. He is currently an Associate Editor of the IEEE TRANSACTIONS ON BIOMEDICAL ENGINEERING and the *International Journal of Audiology*. He was an Associate Editor of the IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING (1999-2002), IEEE SIGNAL PROCESSING LETTERS (2006-2009), and is currently a member of the Speech Technical Committee of the IEEE Signal Processing Society.