AN INTEGRATED SPEECH–BACKGROUND MODEL FOR ROBUST SPEAKER IDENTIFICATION*

D. A. Reynolds and R. C. Rose
Lincoln Laboratory, MIT
244 Wood St.
Lexington, MA 02173

ABSTRACT

This paper examines a procedure for text independent speaker identification in noisy environments where the interfering background signals cannot be characterized using traditional broadband or impulsive noise models. In the procedure, both the speaker and the background processes are modeled using mixtures of Gaussians. Speaker and background models are integrated into a unified statistical framework allowing the decoupling of the underlying speech process from the noise corrupted observations via the expectation-maximization algorithm. Using this formalism, speaker model parameters are estimated in the presence of the background process, and a scoring procedure is implemented for computing the speaker likelihood in the noise corrupted environment. Performance is evaluated using a 16 speaker conversational speech database with both "speech babble" and white noise background processes.

1. INTRODUCTION

In this paper we examine the use of signal decomposition for robust text-independent speaker identification in the presence of an interfering background signal. There are many speaker identification applications that must be carried out in situations involving background signals that are difficult to model using existing noise suppression techniques. Of particular interest are background signals, such as competing speech, which do not fit into traditional broadband or impulsive noise models. To accommodate these background signals, we use a statistical formulation which allows us to estimate speaker model parameters from speech recorded in these difficult environments.

In previous work, good text-independent speaker identification performance was obtained on a wideband conversational speech task [1]. The parameters of speaker dependent Gaussian mixture models (GMM) were trained using the iterative EM algorithm, and the GMM based speaker classifier provided high recognition rates for input utterances as short as 1 second. A more recent study [2] examined techniques for robust speaker identification in a noisy telephone channel environment. The focus of the study was to evaluate a technique for the combined modeling of speech and noise similar to [3] with respect to a particular implementation of a spectral subtraction noise pre-processor. Here, the background was assumed to be broadband Gaussian noise. It was found that, while the front-end processing provided good performance when exactly "tuned" to the noise characteristics, the integrated noise model gave similar improvements in identification performance without requiring *ad-hoc* tuning of environment dependent thresholds. Based on the above results, the aim of this present study is to extend the integrated speech/background speaker dependent GMM to handle a more general class of background signals.

We present a robust text-independent speaker identification system which uses separate background and speech models combined in a unified statistical framework. Under this framework, the system provides a means for decomposing a corrupted signal into speech and background processes. We have developed a procedure for estimating the parameters of a speaker's GMM in the presence of a non-Gaussian background and a scoring procedure for computing the likelihood of a corrupted observation sequence as it traverses through the speech–background state space.

The paper is organized as follows. The next section introduces the combined speech-background model and the mechanism for speaker model decoding. Section 3. describes the procedure for estimating the speaker model parameters from noise corrupted observations. Finally, in Section 4., experimental results are presented, and discussion and summary are provided in Section 5..

2. THE SPEECH-BACKGROUND MODEL

This section introduces the model that is used to address two problems related to speaker representation in the presence of a non–Gaussian background environment. The first problem, maximum likelihood speaker decoding, is to find the speaker that is most likely to have generated a set of noisy observations. In addressing the first problem, we begin with parametric model representations for a population of S speakers λ_{s_k}, $k = 1, \ldots, S$ and a parametric model representation for background λ_b. In this case, all models are estimated from independent measurements so the interaction between the speaker and background processes is not considered an issue in training [4]. The second problem, speaker model parameter estimation, is to estimate speaker model parameters from noise corrupted observations where the interaction between speaker and background is considered an important issue. In this case, background model parameters are estimated from independent observations, so that prior knowledge of the background process is assumed to exist.

2.1. Model Description

Viewed as a generative process, the combined speech-background model in figure 1 generates two concurrent and independent sequences of hidden states; $I = \{i_1, \ldots, i_T\}$ for the speech model and $J = \{j_1, \ldots, j_T\}$ for the background model. These state sequences are then turned into a sequence of speech or signal vectors $X = \{x_1, \ldots, x_T\}$ and background vectors $Y = \{y_1, \ldots, y_T\}$ through two sets of state dependent probability density functions. To accommodate non-Gaussian signal and background processes, the

*THIS WORK WAS SPONSORED BY THE DEPARTMENT OF THE AIR FORCE. THE VIEWS EXPRESSED ARE THOSE OF THE AUTHORS AND DO NOT REFLECT THE OFFICIAL POLICY OR POSITION OF THE US GOVERNMENT.

0-7803-0532-9/92 $3.00 © 1992 IEEE

Gaussian mixture densities

$$p(x_t \mid \lambda_s) = \sum_{i=1}^{M} p_i \, b_i(x_t) \qquad (1)$$

$$p(y_t \mid \lambda_b) = \sum_{j=1}^{N} q_j \, a_j(y_t) \qquad (2)$$

are used to model the speech and background processes respectively. In equations 1 and 2, $b_i(x_t)$ and $a_j(y_t)$ represent Gaussian densities defined over D dimensional speech and background vectors respectively, and p_i and q_i represent the mixture weights. It is assumed that the parameters of the background pdf are obtained independently (i.e. from non-speech regions in the input utterance). Therefore, as a notational shorthand, reference to λ_b will be dropped, and the speaker model will be referred to simply as λ, where $\lambda = \{p_i, \mu_i, \sigma_i\}$, $i = 1, \ldots, M$.

At each time step, the speech and background vectors are combined through a general function of speech and background $f(x_t, y_t)$ to produce the corrupted observation sequence $Z = \{z_1, \ldots, z_T\}$. The observed signal density for state (i_t, j_t) is determined by both the speech and background models and the form of $f()$ as follows,

$$p(z_t \mid i_t, j_t, \lambda) = \oint_{C_t} b_{i_t}(x_t) \, a_{j_t}(y_t) \, dx_t dy_t \qquad (3)$$

where C_t denotes the contour defined by $z_t = f(x_t, y_t)$. Equation (3) is the main link needed to tie the observed process Z to the hidden processes X and Y. From this we are able to derive integrated scoring and training procedures.

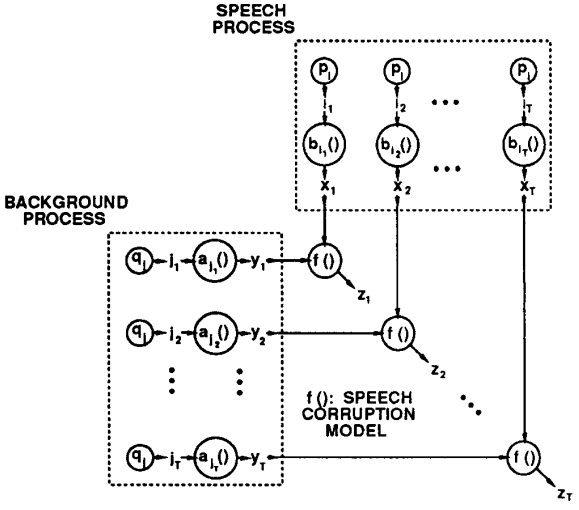

Figure 1: Integrated speech-background speaker model.

2.2. Model Scoring

As with the basic GMM speaker ID system, the score for a test utterance is the likelihood of the observation sequence given a speaker model. However, with the speech-background model we now must compute the likelihood of a corrupted observation over a $M \times N$ "sheet" formed by the Cartesian product of the signal and background states. This 3-D state space lattice is depicted in Figure 2. The likelihood of an observation vector z_t is computed as

$$p(z_t|\lambda) = \sum_{i=1}^{M} \sum_{j=1}^{N} p_i \, q_j \, p(z_t \mid i, j, \lambda). \qquad (4)$$

Assuming independent observations, the product of these likelihoods for an input utterance is computed for each speaker model and the most likely speaker is identified. All speaker models use the same background model for computing observation likelihoods.

It is evident from equation (4) that the likelihood computations can be quite burdensome for large products of M and N. However, it should be possible to greatly reduce computations by using some form of "Viterbi scoring" only over small regions of the state-space similar to the method in [4].

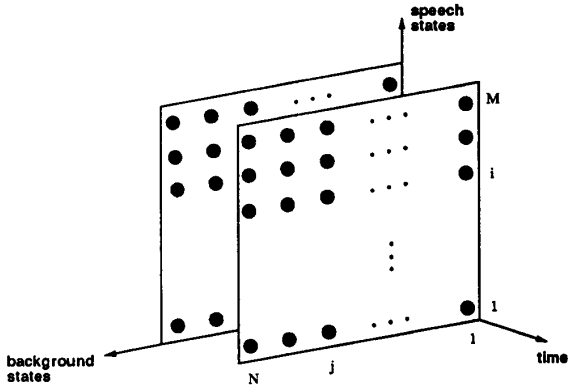

Figure 2: Speech-background state space.

3. PARAMETER ESTIMATION

Using the above model formulation, this section addresses the problem of estimating speaker model parameters from corrupted observations. Since no closed form maximum likelihood parameter estimation procedure exists for this problem, we will follow the development in [2] and derive iterative re-estimation equations via Baum's auxiliary function. The parameter estimation procedure is summarized for a general corruption function $f()$, then the pdfs are specified for a particular choice of $f()$, and finally the properties of the estimated models are considered in terms of familiar noise masking concepts.

3.1. Equation Derivation

The likelihood function of the observation sequence Z in terms of the complete data (X, Y, I, J) is,

$$P(Z \mid \lambda) = \sum_{I} \sum_{J} \oint_{C} P(X, Y, I, J \mid \lambda) \, dX dY \qquad (5)$$

where,

$$P(X, Y, I, J \mid \lambda) = \prod_{t=1}^{T} p_{i_t} \, b_{i_t}(x_t) \, q_{j_t} \, a_{j_t}(y_t). \qquad (6)$$

The term $\sum_I$ denotes the sum over all length T state sequences and $\oint_C$ represents a T-fold iterated integral over the contours C_t defined by $z_t = f(x_t, y_t)$.

The auxiliary function for the above likelihood function can then be shown to be,

$$Q(\lambda,\bar{\lambda}) = \sum_I \sum_J \oint_C P(X,Y,I,J \mid \lambda) \cdot \log P(X,Y,I,J \mid \bar{\lambda})\, dX dY \quad (7)$$

Given an initial set of model parameters, we can obtain new model parameters that will increase (or at least not decrease) the likelihood of the observations by maximizing $Q(\lambda,\bar{\lambda})$ with respect to each of the model parameters.

Substituting in (6) and assuming independence over time and between the hidden variables, we can rewrite (7) as follows,

$$Q(\lambda,\bar{\lambda}) = \sum_{t=1}^{T} \sum_{i=1}^{M} \sum_{j=1}^{N} \oint_{C_t} \gamma_t(i,j) \cdot \log(\bar{p}_i\, \bar{b}_i(x_t)\, q_j\, a_j(y_t))\, dx_t dy_t \quad (8)$$

where,

$$\gamma_t(i,j) = \sum_I \sum_J \eta_t(i,j,I,J)\, P(X,Y,I,J \mid \lambda)\, dX dY \quad (9)$$

and

$$\eta_t(i,j,I,J) = \begin{cases} 1 & \text{if } i_t = i \text{ and } j_t = j \\ 0 & \text{otherwise} \end{cases} \quad (10)$$

It is clear from Equation 8 that using the the auxiliary function allows us to isolate terms that are pertinent to the desired speaker model parameters and solve for them under the appropriate constraints.

Maximizing equation (8) with respect to $\bar{p}_i$ under the constraint $\sum_{i=1}^{M} \bar{p}_i = 1$, yields the mixture weights re-estimation equation,

$$\bar{p}_i = \frac{1}{T} \sum_{t=1}^{T} \sum_{j=1}^{N} p(i,j \mid z_t, \lambda). \quad (11)$$

Assuming each density has a diagonal covariance matrix, each vector element of the speaker model's means and variances can be estimated individually and we will use the notation $\bar{\mu}_i$ and $\bar{\sigma}_i^2$ to refer to an arbitrary vector element. Maximizing $Q(\lambda,\bar{\lambda})$ in (8) with respect to $\bar{\mu}_i$ and $\bar{\sigma}_i^2$ separatly gives the following re-estimation equations.

$$\bar{\mu}_i = \frac{\sum_{t=1}^{T} \sum_{j=1}^{N} p(i,j \mid z_t, \lambda)\, E\{x_t \mid z_t, i, j, \lambda\}}{\sum_{t=1}^{T} \sum_{j=1}^{N} p(i,j \mid z_t, \lambda)} \quad (12)$$

and

$$\bar{\sigma}_i^2 = \frac{\sum_{t=1}^{T} \sum_{j=1}^{N} p(i,j \mid z_t, \lambda)\, E\{x_t^2 \mid z_t, i, j, \lambda\}}{\sum_{t=1}^{T} \sum_{j=1}^{N} p(i,j \mid z_t, \lambda)} - \bar{\mu}_i^2 \quad (13)$$

where $E\{\}$ is the expectation operator.

These new model parameters can be used to estimate the required probabilities and expectations in (11), (12), and (13) to again refine the model. This process of iteratively computing conditional expectations and using them to obtain maximum likelihood estimates of the model parameters is known as the expectation maximization (EM) algorithm [5]. It has been shown in [5], that the EM algorithm is guaranteed to converge to a local maximum likelihood model.

To compute the conditional expectations and probabilities needed for the parameter estimations, we need to define the noise corruption function $f()$ which determines the form of the observation pdf (equation (3)). Note that the above derivation is general and applies for all functions $f()$ for which $z = f(x,y)$ defines a one-dimensional contour in the x-y plane. The choice of $f()$ will depend on how the corruption signal is combined with the speech and the features used in the observation vectors. Several examples of $f()$ and the corresponding conditional expectations are given in [2]. Assuming that the speech and background signals are added together and using log-energy filterbank features, the noise corruption function is well modeled by the max() function [3]. With this choice, the observation pdf in (3) has the following form [6],

$$p(z_t \mid i_t, j_t, \lambda) = b_{i_t}(z_t)\, A_{j_t}(z_t) + B_{i_t}(z_t)\, a_{j_t}(z_t) \quad (14)$$

with $A(z)$ and $B(z)$ denoting cumulative densities. For this observation pdf, the equations for $E\{x_t \mid z_t, i, j, \lambda\}$ and $E\{x_t^2 \mid z_t, i, j, \lambda\}$ needed in (12) and (13) are given in [3].

3.2. Noise Masking

It is interesting to examine how the training procedure performs a form of noise masking. In many cases of low SNR, several of the frequency bands (elements of an observation vector) will be completely buried in noise. Thus the training data contains no speech information for that band. In practice the speaker model must be initialized from the corrupted data and this noise masking manifests itself by producing speaker model components which correspond only to background observations. Figure 3b shows such as initial model for a single frequency band. The true signal pdf (solid curve) and background pdf (dashed curve) are shown in figure 3a.

The mechanisms of the re-estimation equations work such that at each iteration those signal components which overlap with the background model are moved in the negative direction (toward $-\infty$). As seen in figure 3c, after training (10 iterations in this case) the final model appears to still retain components which correspond to the background process. However, these components have mean levels which are below the background mean level. Thus when these models are used for recognition, in a presumably lower or equal SNR environment, the background model for scoring will mask these errant signal components (via the pdf in (14)) and only the components corresponding to the actual speech will be used to discriminate among the speakers. This masking, as related to the Klatt noise masking procedure, is discussed in [4] and [7].

4. EXPERIMENTS

4.1. Databases

The speech data is from the wideband portion of the KING conversational speech corpus. We used a subset of 16 speakers (all male) with five sessions per speaker (each session is approximately 40 seconds in duration). For the experiments two sessions were used for training models and the remaining three were used for testing.

The background signals came from the NATO RSG-10 noise database. A white noise signal and a "speech babble" signal were chosen for background signals. The "speech babble" is a recording of 100 people speaking in a canteen and was chosen since it represents a contrast to traditional noise (e.g. white) and thus should be a good test candidate for the integrated model. The speech and background signals were digitally summed to create the corrupted signals used in the experiments, allowing the use of a variety of different background signals on the same speech data over a range of signal-to-noise ratios.

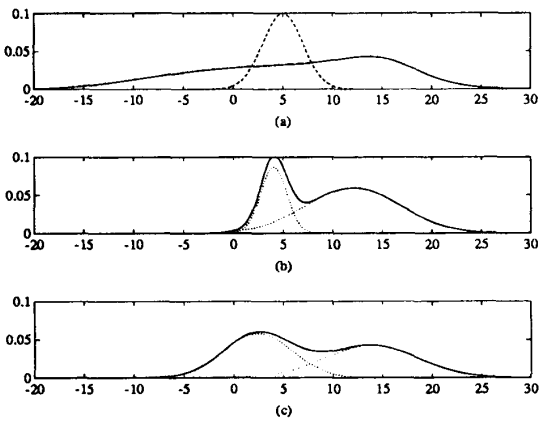

Figure 3: Example of noise masking in integrated training. (a) True signal pdf (solid) and bkg pdf (dashed). (b) Initial signal model from corrupted data. (c) Final signal model.

4.2. Results

Our initial experiments focused on two basic paradigms. The first is a cross channel paradigm (clean train / noisy test) and the second paradigm corresponds to both model estimation and classification in noisy environments (noisy train / noisy test). Tables 1 and 2 show the results for these experimental paradigms using the speech babble and white noise background signals. In the tables we compare the speaker identification performance of the basic GMM system (GMM) to the integrated background system (GMM-IB) for 5 and 10 second test utterance lengths. All noisy environments are at an average SNR of 10dB. Speaker models had 25 component Gaussians, the speech babble model had 5 components and the white noise model had a single component.

	GMM		GMM-IB	
	5 sec	10 sec	5 sec	10 sec
Clean/Noisy	26.4	25.8	73.8	78.8
Noisy/Noisy	75.9	83.3	79.3	86.5

Table 1: Speaker identification results (percent correct) with speech babble at SNR=10dB.

	GMM		GMM-IB	
	5 sec	10 sec	5 sec	10 sec
Clean/Noisy	10.4	10.6	63.6	68.6
Noisy/Noisy	72.0	78.4	73.4	79.8

Table 2: Speaker identification results (percent correct) with white noise at SNR=10dB.

5. DISCUSSION

There a several things we can note from the results. First, it is clear that under clean train / noisy test situations there is a drastic degradation in recognition performance in the basic GMM system. However, using the integrated model we are able to greatly improve performance , demonstrating the effectiveness of using the background model in recognition under mis-matched train/test conditions. Thus, given we have previously trained speaker models , we can use these for recognition in a different environment without retraining. All that is required is a sample of the new background signal.

Also, note that for the matched noisy/noisy case, the intergated model actually performs better than the straight GMM model. While the optimal situation is to train and test in the same noise conditions, the integrated model has the advantage of using the same background model for all speaker models in training and testing. This avoids producing different speaker scores for a test vector which is mostly due to the background process. Only the pdf truly indicative of a speaker should be used to discriminate among the speakers.

Finally, it is evident that the integrated model is equally effective for traditional white noise and the more complex speech babble. Indeed this is one of the major advantages of the integrated model. No environmental tuning or thresholding is required, only a sample of the interfering process is needed.

6. SUMMARY

This paper has presented a robust speaker modeling technique capable of operating on speech corrupted with non-Gaussian interfering signals. An iterative maximum-likelihood EM algorithm was derived for estimating speaker model parameters from corrupted speech along with a scoring procedure for corrupted input speech.

This new technique is very powerful and applicable to a wide variety of interfering signals. The consistent good performance obtained by the integrated model of speech and background demonstrate the importance of integrating prior knowledge of the background process directly in the speaker classifier. Experimental results demonstrated a large performance improvement over a GMM classifier that did not incorporate knowledge of the background under both mis-matched and optimal matched training testing conditions.

REFERENCES

[1] R. C. Rose and D. A. Reynolds, "Text-independent speaker identification using automatic acoustic segmentation," in *International Conference on Acoustics, Speech, and Signal Processing*, IEEE, 1990.

[2] R. C. Rose, J. A. Fitzmaurice, E. M. Hofstetter, and D. A. Reynolds, "Robust speaker identification in noisy environments using noise adaptive speaker models," in *International Conference on Acoustics, Speech, and Signal Processing*, IEEE, 1991.

[3] A. Nadas, D. Nahamoo, and M. A. Picheny, "Speech recognition using noise-adaptive prototypes," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. ASSP-37, pp. 1495–1503, 1989.

[4] A. P. Varga and R. E. Moore, "Hidden markov model decomposition of speech and noise," in *International Conference on Acoustics, Speech, and Signal Processing*, IEEE, 1990.

[5] A. Dempster, N. Laird, and D. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *Journal of the Royal Statistical Society*, vol. 39, pp. 1–38, 1977.

[6] A. Papoulis, *Probability, Random Variables, and Stochastic Processes*. New York, NY: McGraw-Hill, 2 ed., 1984.

[7] J. Holmes and N. Sedgwick, "Noise compensation for speech recognition using probabilistic models," in *International Conference on Acoustics, Speech, and Signal Processing*, IEEE, 1986.