# SPEAKER IDENTIFICATION IN THE PRESENCE OF PACKET LOSSES

*Deva K. Borah and Phillip DeLeon*

New Mexico State University
Klipsch School of Electrical and Computer Engineering
Box 30001 / Dept.3-0
Las Cruces, New Mexico USA 88003
{dborah, pdeleon}@nmsu.edu

## ABSTRACT

Gaussian mixture model (GMM)-based speaker identification systems have proved remarkably accurate for large populations using reasonable lengths of high-quality test utterances. Test utterances, however, aquired from cellular telephones or over the Internet (VoIP) may have dropouts due to packet loss. In our research, we have demonstrated that for small packet sizes, these losses can result in degraded accuracy of the speaker identification system. It is shown that by training the GMM model with lossy speech packets, corresponding to the loss rate experienced by the speaker to be identified, significant performance improvement is obtained. In order to avoid the prior estimation of the packet loss rate experienced by the test subject, we propose an algorithm to identify the user based on maximizing the *a posteriori* probability over the GMM models of the users trained with several packet loss rates. It is shown that the proposed algorithm provides excellent identification performance.

## 1. INTRODUCTION

The objective of a speaker identification algorithm is to determine which voice sample from a set of known voice samples best matches the characteristics of an unknown input voice sample. This involves extraction of speaker-dependent features from the known voice samples, model building for each known sample, and eventual matching of the features extracted from the unknown voice sample. Of various speaker identification techniques [1], the Gaussian mixture model (GMM)-based speaker identification algorithm has shown to be remarkably successful in identifying speakers from a large population. The GMM approach provides a probabilistic model where an implicit segmentation of the speech into phonetic sound classes prior to speaker model training takes place. It has been found [2] that the performance of the GMM-based method is near $100\%$ up to a population size of 630 speakers using the TIMIT speech database (clean speech) with about 24 seconds of training and 6 seconds of test utterances. The performance degraded significantly for telephone-quality speech and is near $60\%$ for a similar size population.

Recently there has been an interest in studying the performance of speaker identification algorithms in the context of mobile wireless channels. It is well known that in order to achieve high transmission efficiency, speech signals in such systems undergo speech coders and decoders which modify the original voice signal. In addition, the uncertain wireless channel can cause data packet loss during deep fading periods. The effect of GSM (Global System for Mobile Communication) coders on speaker recognition has been investigated in [3]. It has been shown that the usage of GSM coding significantly degrades performance. By extracting features directly from the encoded bit stream, the work in [3] is able to improve the performance of the system. However, to our knowledge the effects of packet loss due to the mobile wireless channel have not been investigated.

In this paper, we consider the problem of speaker identification in the presence of packet losses. This study is directly relevant for wireless channels and the VoIP environments. Since our focus is to identify the effects of packet loss due to fading in the wireless channel or due to a delay/congestion problem in the VoIP network we do not incorporate speech coders and decoders in our model. Each data packet contains a fixed number of speech samples and the loss of a packet results in the loss of the speech samples contained in the packet. Our study shows that for small packet sizes, these losses can result in degraded accuracy of the speaker identification system. We next show that by training the GMM model with lossy packets, corresponding to the loss rate experienced by the speaker to be identified, significant performance improvement is obtained. In order to avoid the estimation of the packet loss rate, we propose an algorithm to identify the user based on maximizing the *a posteriori* probability over the GMM models of the users trained with several packet loss rates. It is shown that the proposed algorithm provides excellent identification perfor-

mance.

## 2. OVERVIEW OF THE GMM-BASED SPEAKER IDENTIFICATION SYSTEM

Briefly, a speaker identification system works as follows. Prior to speaker identification, the system must first be trained, i.e. create a table associating each individual speaker with a distinguishing set of parameters based on the individual's speech signal. Afterward, a new speech signal from an unknown user is acquired and a parameter set is determined. A comparison is made with the unknown individual's parameter set and the entries in the table in order to determine a closest "match" and subsequent identification of the speaker. In the following subsections, we provide more details regarding the GMM-based speaker identification system as reported in [2].

### 2.1. Speech Analysis and Feature Extraction

The first stage in either the training or identification stage is to perform an analysis of the speech signal and extract distinguishing features. Fig. 1 illustrates the steps involved in the feature extraction [4]. First, silence must be removed from the utterance, $u(n)$ (samples assumed to be normalized). In our implementation, we measured the signal energy in 3ms non-overlapping windows and compared to a threshold set to 0.012 (found through experiment). If the energy was below the threshold, we removed the 3ms segment (which is assumed to be silence) from the utterance. Next the short-time Fourier transform (STFT), $X(m,k)$ is computed from the silence-removed, utterance $x(n)$. The STFTs (1024-point) are computed using 20ms Hamming-windowed segments with 50% overlap. Magnitude-squared data is computed from the STFT, i.e. spectrogram and a pre-emphasis is optionally applied in order to boost the higher frequencies. Next a 20-channel, mel-scale filterbank, shown in Fig. 2 is applied in order to weight the spectrogram [4]. The filterbank is designed with triangular responses and the first ten center frequencies are uniformly spaced over 1kHz while the second ten center frequencies are logarithmically spaced over the remaining 3kHz. The filters, $F_l$ are normalized according to their bandwidth. The log-energy, $y(m,l)$ of each channel is calculated and the DCT of the vector is computed. The resulting feature vector is the $20 \times 1$ mel-cepstrum, $\mathbf{Y}\,(m)$ computed every 10ms. In our experiments, we used 90s of speech for training and 15s for the identification. The speech data is acquired from the YOHO Speaker Verification corpus by concatenating available files for each individual speaker [5].

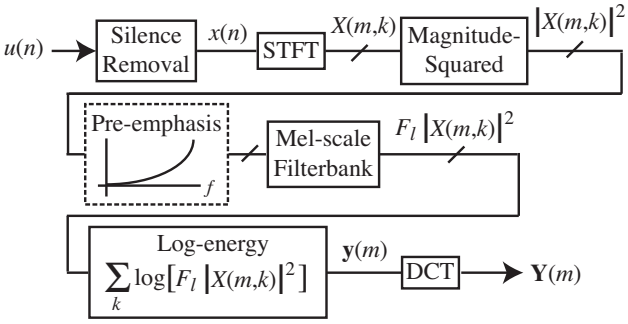


**Fig. 1**. Mel-scale cepstral feature analysis

### 2.2. GMM Description, ML Parameter Estimation and Speaker Identification

The probability density function of the feature vector $\mathbf{Y}$ of a given speaker is modeled as a Gaussian mixture given by

$$p(\mathbf{Y}|\lambda_s) = \sum_{i=1}^{W} \left\{ \frac{w_i}{\sqrt{(2\pi)^L \sigma_{i,1}\sigma_{i,2}\ldots\sigma_{i,L}}} \times \exp\left( -\frac{1}{2} \sum_{k=1}^{L} \frac{|Y_k - m_{i,k}|^2}{\sigma_{i,k}^2} \right) \right\} \quad (1)$$

where $W$ is the number of mixture components, $L$ is the feature vector length, $w_i$ is the weight of the $i$-th mixture component, and $m_{i,k}$ and $\sigma_{i,k}$ denote the mean and the variance respectively of the $k$-th component of the feature vector corresponding to the $i$-th mixture component. The weights, means and the variances are collectively represented by the parameter $\lambda_s$ for the $s$-th speaker. Thus each speaker $s$ is represented by a GMM and is referred to by his/her model $\lambda_s$.

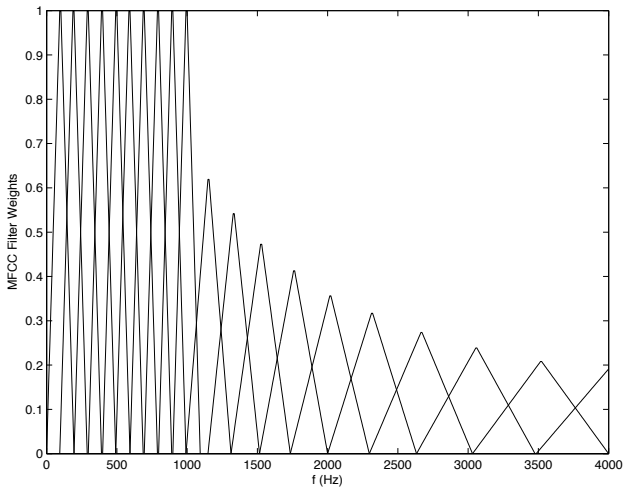


**Fig. 2**. Mel-scale filter bank

Once the feature vectors have been extracted from speech training data, the first step in the development of the speaker identification algorithm is to extract the model parameter $\lambda$ for that speaker. It is known that a maximum likelihood (ML) parameter estimation approach results in a difficult nonlinear optimization problem. Therefore, iterative techniques, such as the expectation maximization (EM) algorithm, have been employed that guarantee convergence to local minima [2]. The EM algorithm begins with an initial parameter estimate, and then iteratively improves upon the previous estimates with new updated estimates. The iterations continue until some convergence threshold is reached.

Once the GMM parameters of all the speakers in the training set are obtained, the next step of identification begins. In identification, it is typically assumed that all the $S$ speakers in the training set are equally likely. In that case, it is well-known that the maximum a posteriori (MAP) detection becomes the ML detection for the user estimate $\hat{S}$ given by

$$\hat{S} = \arg \max_{1 \leq s \leq S} \prod_{i=1}^{T} p(\mathbf{Y}_i | \lambda_s) \quad (2)$$

under the assumption that the observations are independent. $T$ is the number of training vectors.

## 3. SPEAKER IDENTIFICATION OVER CHANNELS WITH PACKET LOSS

We first assume that during training, the speech utterances are complete, i.e. no interruptions due to packet loss. Therefore, only the test data are incomplete due to packet loss. In order to simplify, we apply the packet loss model (described below), to fixed-sized packets each assumed to have a fixed number of speech samples. However, packets usually contain *coded* speech which would imply that the loss of the packet would represent the loss of numerous speech samples depending on the coding scheme and compression ratio.

### 3.1. Packet Loss Model

The packet loss model used in our study is the well-known Gilbert-Elliot channel [6]. This channel has two states: good and bad. When the state is 'good', the transmitted packet is received without any error, and during a 'bad' state the packet is considered lost. This type of packet loss model has been widely used both in the wireless communications literature and in the internet traffic modeling area. In our study, we simulate the samples of a Rayleigh fading channel using the Jakes model with the Doppler spectrum given by

$$S(f) = \frac{1}{\pi f_D \sqrt{1 - f^2/f_D^2}} \quad (3)$$

for $|f| \leq f_D$, where $f_D$ is called the Doppler frequency. The power of the samples over the packet duration is calculated and the channel is considered 'good' if the power is above a certain threshold.

### 3.2. Results

Using the GMM-based speaker identification system described in Section 2 together with test data subjected to packet loss, we simulated speaker identification for various packet sizes and different packet loss rates. The number of Gaussian mixture components $W = 10$, and the feature vector length is $L = 19$. Figure 3 illustrates consistently good speaker identification rates ($\approx 95\%$) for packet sizes above 32 samples/packet even with packet loss rates of 40%. However, with smaller packets (8 and 16 samples/packet), the performance noticeably degrades. In the case of 8 samples/packet, the performance is 68% correct identification with 20% packet loss and only 33% correct identification with 40% packet loss. The small size packet losses directly affect the components of the feature vectors changing their statistics.

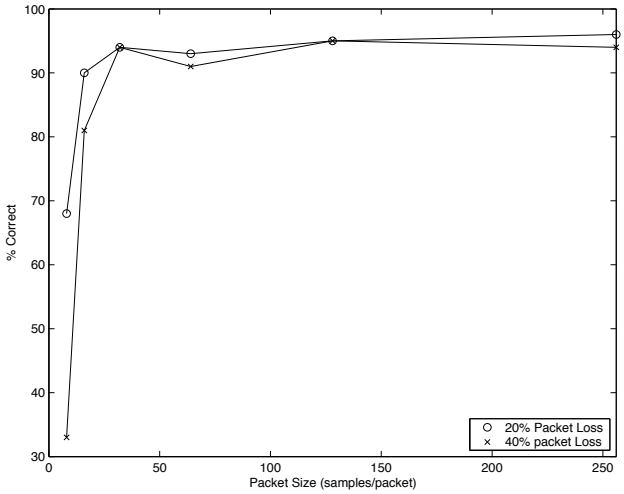


**Fig. 3**. Speaker identification performance as a function of packet size. Packet loss rates of 20% and 40% are used.

## 4. IMPROVING SPEAKER IDENTIFICATION OVER LOSSY, PACKET CHANNELS

In this section, we propose to use a lossy packet training approach for improving the speaker identification performance in lossy channels. When the packet loss rate of the unknown speaker is known or can be accurately estimated, the same losses can be applied to training data for all $S$ users prior to identification thereby providing a better match between training and test data. As shown in Fig. 4, with lossy

test data ($30\%$ packet loss) but lossless training data ($0\%$ packet loss), the recognition rate is $35\%$. However, with the proposed method which instead uses lossy training data ($30\%$ packet loss) the recognition rate improves to above $90\%$. However, when a large mismatch occurs between the actual packet loss rate for test data and that applied to the training data, performance will be degraded. In the figure, the identification rate has decreased to $89\%$ when a $50\%$ loss rate is used in the training data but a $30\%$ loss rate actually occurs in the test data. It is observed from the figure that the performance is relatively insensitive to *small* errors in the loss rate estimation.

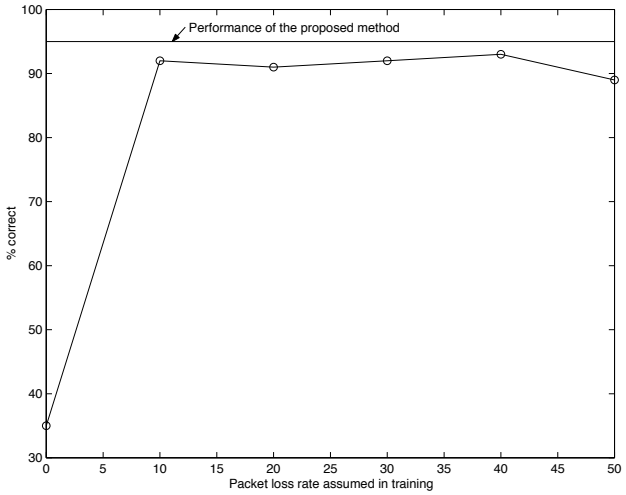


**Fig. 4**. Performance of GMM-based speaker identification when lossy packets (with different assumed loss rates) are used for training. The unknown speaker has a packet loss rate of $30\%$.

In order to avoid the estimation of the loss rate altogether, we propose to use a set of GMM parameters for each speaker's training data with different packet loss rates applied. We exploit the relative insensitivity of small errors in loss rate estimation, and choose $M = 6$ loss rates of $0\%$, $10\%$, $20\%$, $30\%$, $40\%$, and $50\%$. The algorithm then obtains the MAP estimate over the *set* of loss models as

$$\hat{S} = \arg \max_{1 \le k \le S, 1 \le l \le M} \prod_{i=1}^{T} p(\mathbf{Y}_i | \lambda_{s,l}) \qquad (4)$$

where $\lambda_{s,l}$ denotes the GMM parameters for speaker $s$ under the loss rate model $l$, $1 \le l \le M$. It is observed from Fig. 4 that the identification performance has improved to $95\%$ without any explicit loss rate estimation of the channel.

In Figs. 5 and 6, we study the performance of the proposed method for identification data of various lengths. Both the figures show that identification performance increases with the length of the data. Packets containing more samples show better performance with small identification data. As more data are used, the performance behavior becomes nearly similar. Figure 6 shows that in more lossy channels, identification performance is poorer for smaller identification data. The figure also shows results for packets with no losses. In our study, we have used a hard measure of speaker identification unlike a soft measure as in [2]. Therefore, ignoring minor variations, the $20\%$ loss rate case behaves similar to the no loss case. For larger identification data, behavior for different loss rates becomes similar to the no loss case.

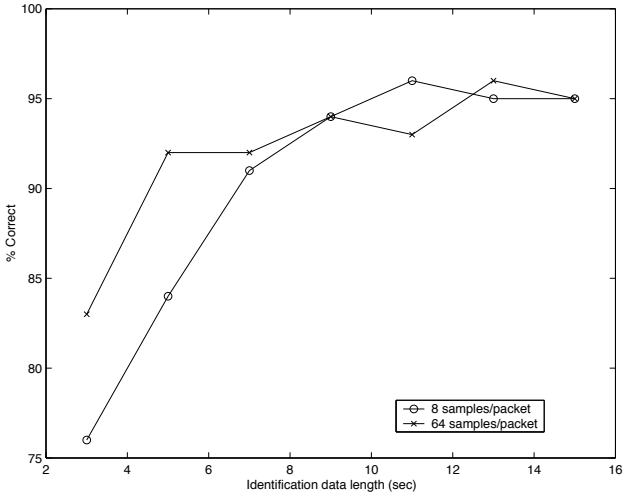


**Fig. 5**. Speaker identification performance as a function of identification data length with packet length as the parameter. The packet loss rate is $40\%$.

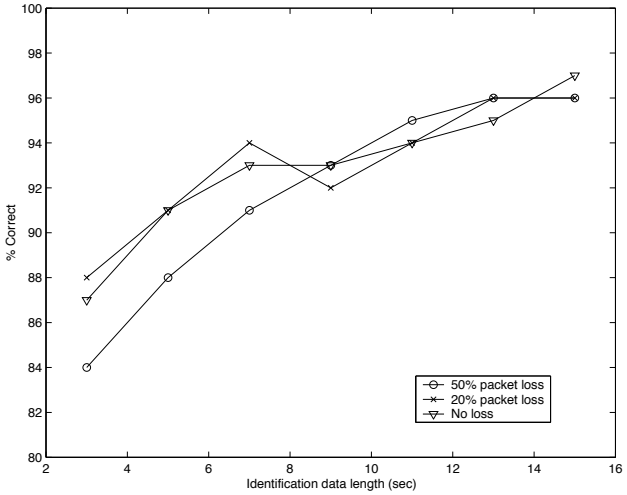


**Fig. 6**. Speaker identification performance as a function of identification data length with packet loss rate as the parameter. Each packet contains 16 samples.

## 5. CONCLUSIONS

In this paper we have demonstrated that when test utterances are acquired over lossy, packet channels, speaker identification rates quickly decrease as the packet size gets smaller (approximately 8-16 speech samples/packet). An algorithm for improving speaker identification in lossy channels is proposed. The algorithm uses a set of GMM models for several packet loss rate models for each known spkeaker, and the best speaker match is identified over all the loss model sets. It has been found that the proposed method results in excellent identification performance.

## 6. REFERENCES


[1] T. F. Quatieri, *Discrete-time Speech Signal Processing Principles and Practice*, Prentice-Hall, Inc., New Jersey, 2002.

[2] D. Reynolds and R. Rose, "Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models," *IEEE Trans. Signal Processing*, vol. 3, no. 1, pp. 72–83, Jan. 1995.

[3] L. Besacier, S. Grassi, A. Dufaux, M. Ansorge, and F. Pellandini, "GSM speech coding and speaker recognition," in *Proc. IEEE ICASSP'00*, June 2000.

[4] S. Davis and P. Mermelstein, "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-28, no. 4, pp. 357–366, Aug. 1980.

[5] J. Campbell, "Testing with the YOHO CD-ROM Voice Verification Corpus," *Proc. Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, 1995.

[6] A. J. Goldsmith and P. P. Varaiya, "Capacity, Mutual Information, and Coding for Finite-State Markov Channels," *IEEE Trans. Signal Processing*, vol. 42, pp. 868–886, May 1996.