GMM AND KERNEL-BASED SPEAKER RECOGNITION WITH THE ISIP TOOLKIT

Tales Imbiriba[1], Aldebaro Klautau[1], Naveen Parihar[2], Sridhar Raghavan[2] and Joseph Picone[2]
[1]Signal. Processing Laboratory,
Universidade Federal do Para, Brazil
[2]Institute. for Signal. And Information Processing,
Mississippi State University, USA,
Web: www.laps.ufpa.br, and
www.isip.msstate.edu

ABSTRACT

This paper describes an open source framework for developing speaker recognition systems. Among other features, it supports kernel classifiers, such as the support and relevance vector machines. The paper also presents results for the IME corpus using Gaussian mixture models, which outperforms previously published ones, and discusses strategies for applying discriminative classifiers to speaker recognition.

1. INTRODUCTION

Speech research technologies, such as speech recognition and speaker verification/identification systems, require an integration of knowledge across several domains, which include signal processing and machine learning. With the constant evolution and ever increasing complexity of the speech research technology, the development of a state-of-the-art speech recognition system or a speaker verification/identification system becomes a time-consuming and infrastructure-intensive task. This problem is especially relevant for researchers in third-world countries, such as Brazil, where only few research groups have access to state-of-the-art systems. Besides, very few groups in such countries have access to an industry standard speech corpus. The few corpora that are freely available are often too small for serious research.

Hence, the adoption of public domain software and corpora is very important for advancing the state-of-art, and allowing a proper comparison among well-established and new techniques. In Brazil, the Military Institute of Engineering (IME) has recently released the IME 2002[3], a Brazilian Portuguese

[2] Work supported by the National Science Foundation under Grant No. IIS-0085940. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the NSF.
[3] Work supported by FAPERJ, Brazil, under grant number E-26/171.307/2001.

corpus for speaker recognition. This corpus was made available free of charge to several research groups. It is potentially the main resource for Brazilian speech researchers working in the area of speaker recognition. Together, the ISIP public domain technology and the IME corpus, compose an excellent framework that can serve as the test bed for comparing various speaker recognition technologies. Moreover, this framework also allows for the replication of the published results.

In this paper we present this freely available framework, and describe our results of a GMM-based speaker verification system on the IME corpus. We show that these results are better, in most cases, than the previously published ones. More importantly, all results published in this paper can be easily replicated on other sites. All the necessary experimental framework and software is available on our web sites. We also present a brief review of architectures for using kernel classifiers in speaker recognition, discuss associated issues, and present preliminary results for the support vector machine (SVM) approach.

This paper is organized as follows. Section 2 presents an overview of the ISIP public domain toolkit. Section 3 briefly discusses kernels classifiers and architectures of speaker recognition systems. Finally, in Section 4, we present simulation results using GMM and SVM-based speaker recognition systems on the IME corpus. Our conclusions are stated in Section 5.

2. ISIP SPEAKER VERIFICATION SYSTEM

Since 1994, ISIP has been developing free public domain software for the speech research community [1-8]. Since 1998, we have focused on the development of a modular, flexible, and extensible recognition research environment, which we refer to as the production system [5, 6]. The toolkit contains many common features found in modern speech to text systems: a GUI-based front end tool that converts the signal to a sequence of feature vectors, an HMM-based acoustic model trainer, and a time-synchronous hierarchical Viterbi decoder.

Recently, we have added a SVM trainer, a SVM classifier, and a set of supporting utilities to the production system that allows a user to build a hybrid HMM/SVM based speech recognition system [7]. We have also added a maximum likelihood linear regression (MLLR) technique for speaker adaptation, and extended the production system to allow for speaker verification experiments. We are also planning to add a hybrid HMM/RVM (relevance vector machine) based speech recognition system [8]. All these new features will be embedded in the next release (*r00_n12*) of the production system.

All the components introduced above are developed based on an extensive set of foundation classes (IFCs). IFCs are a set of C++ classes organized as libraries in a hierarchical structure. These classes are targeted to satisfy the needs of rapid prototyping and lightweight programming without sacrificing the efficiency. Some key features include: UNICODE support for multilingual applications; math classes that provide basic linear algebra and efficient matrix manipulations; memory management and tracking; system and I/O libraries that abstract users from details of the operating systems. The complete software toolkit and the associated documentation are freely available on-line at *http://www.isip.msstate.edu/projects/speech/index.html*.

Next, we illustrate the support for speaker recognition in the ISIP toolkit by briefly describing the general approach to speaker verification, as shown in Figure 1. The ISIP toolkit contains utilities for each of the following stages: feature extraction, pattern matching, and decision mechanism. Additionally, an enrollment is required to generate the models for each speaker.

During the speaker verification process, the sequences of feature vectors, extracted from the speech signal, are compared to a model representing the claimed speaker via pattern matching. The speaker models are obtained from a previous enrollment process. An utterance-based score is used to accept or reject a speaker through hypothesis testing.

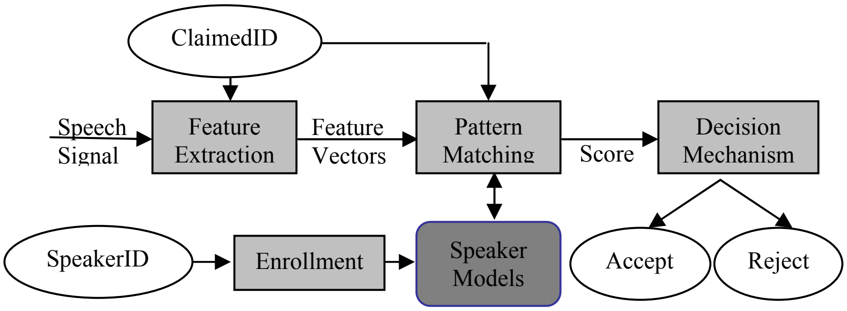

Figure 1 Architecture of a Typical Speaker Verification System

In the feature extraction stage, the perceptually relevant and meaningful information is extracted from the input speech signal. For example, an industry standard Mel-frequency cepstral coefficients (MFCC) front end is typically employed to extract 12 Mel-frequency cepstral coefficients (MFCC) plus the log energy at a frame rate of 100 frames per second. In order to model the spectral variation of the speech signal, the first and second order derivatives of the 13 coefficients are appended to yield a total of 39 coefficients per frame. Another popular front end that can be used for speaker recognition is based on the perceptual linear prediction (PLP) coefficients.

In the pattern matching stage, GMM is the most commonly adopted technique. More specifically, GMM is a special case of a Bayes classifier that obtains the likelihood of an observation using a mixture model λ of M multivariate Gaussians given by:

$$. p(\vec{x}/\lambda) = \sum_{i=1}^{M} p_i b_i(\vec{x}) \qquad (1)$$

where $\vec{x}$ is a D-dimensional feature vector, $i = 1,...,M$ mixture components, $b_i(\vec{x})$ are the mixture densities, and p_i are the mixture weights. Each mixture component is a D-variate Gaussian distribution given by:

$$b_i(\vec{x}) = \frac{1}{(2\Pi)^{D/2}\left|\sum_i\right|^{1/2}} \exp\{-\frac{1}{2}(\vec{x}-\vec{\mu})' \sum_i^{-1}(\vec{x}-\vec{\mu})\} \tag{2}$$

where μ_i is the mean vector and $\sum_i$ is the covariance of the i^{th} mixture component.

In GMM-based speaker verification, the binary decision to accept or reject a claimed identity is based on the likelihood score. If the null-hypothesis (H_0) represents the fact that the speaker is an imposter, and the alternative hypothesis (H_1) represents the fact that the speaker is whom he claims to be, then, the likelihood ratio $\lambda_A(z)$ of the claimed speaker A gives us the following decision criteria,

$$\lambda_A(z) = \frac{p_A(z|H_1)}{p_A(z|H_0)} \tag{3}$$

and if

$$\begin{array}{l} \lambda_A(z) \leq T \text{ , choose } \ldots H_0; \\ \lambda_A(z) > T \text{ , choose } \ldots H_1 \end{array} . \tag{4}$$

where $p_A(z \mid H_0)$ represents the conditional density of the likelihood score generated by the imposters' model (or "universal background model") and $p_A(z \mid H_1)$ represents the conditional density of the likelihood score generated by speaker A using his own model. The variable T is our acceptance or rejection threshold.

The next section details the pattern matching stage with an emphasis on speaker recognition based on kernel classifiers.

3. ARCHITECTURES FOR SPEAKER RECOGNITION

The speaker recognition problem is closely related to the conventional supervised classification. Hence, we start by providing few related definitions. In the classification scenario one is given a training set $\{(x_1, y_1),\ldots, .(x_N, y_N)\}$ containing N examples, which are considered to be independently and identically distributed (iid) samples from an unknown but fixed distribution P(x, y). Each example (x, y) consists of a vector $x \in \mathbf{X}$ of dimension L (called instance) and a label $y \in \{1,\ldots,Y\}$. A classifier is simply a mapping

F:**X**→{1,…,Y}. Of special interest are binary classifiers, for which Y=2, and for mathematical convenience, sometimes the labels are y ∈ {−1, 1} (e.g., SVM) or y ∈ {0, 1} (e.g., RVM).

Some classifiers are able to provide confidence-valued scores $f_i(x)$, for each class i = 1,…,Y. A special case occurs when the scores correspond to a distribution over the labels, i.e., $f_i(x) \geq 0$ and $\sum_i f_i(x) = 1$. Commonly, these classifiers use the max-wins rule $F(x)$ = $\arg_i$ max $f_i(x)$. When the classifier is binary, only a single score $f_i(x)$ ∈ R is needed. For example, if y ∈ {−1, 1}, the final decision $F(\mathrm{x}) = \mathrm{sign}(f_i(x))$ is simply the sign of the score.

In speaker verification systems, the input consists of a matrix **X** ∈ $\mathrm{R}^{T\times Q}$ that corresponds to a segment of speech. The number T of rows is the number of frames (or blocks) of speech, and Q (columns) represents the number of parameters representing each frame. If T is fixed (say, corresponding to 5 milliseconds), **X** could be turned into a vector of dimension $T \times Q$, and one would end up with a conventional classification problem. However, in unrestricted text speaker verification, any comparison between elements of two such vectors could fail if they represent different sounds. Hence, verification systems often adopt alternative architectures, which are similar, but do not exactly match the stated definition of a classifier. We now present the two most popular architectures:

3.1. Frame-based

Under this architecture, a frame-based system is trained using a generative or a discriminative learning. The input to confidence-valued classifiers is a frame X_t with dimension $L = Q$, and the total score of an utterance is computed by summing the score of each frame (possibly, in the log-domain). The Expectation-Maximization algorithm is often used for learning the generative classifiers for this architecture. This approach is adopted in the conventional GMM-based system for text independent speaker recognition, and the HMM-based system for text dependent speaker recognition. Alternatively, one can use discriminative learning, such as the SVM for training the classifiers. The main disadvantage of applying discriminative learning to this architecture is that the number of training examples can be too large for some discriminative techniques. For example, SVM training time scales with the square of the number of training examples, while GMM training scales linearly.

3.2. Hybrid

The hybrid architecture combines advantages of the generative and discriminative learning techniques through the use of the Fisher kernel [9]. It has been applied to the speaker verification tasks in [10] and [11]. In summary, first, the generative learning is used to obtain a generative model such as the GMM. Then, the discriminative learning is applied on new "features" obtained using the generative model. Under the hybrid architecture, the number of features for each input utterance is not given by $T \times Q$ as for the frame-based architecture, but it is equal to the number of parameters in the generative model (for e.g., the means and variances of a GMM).

4. EXPERIMENTAL RESULTS

In this section, first, we briefly describe the IME corpus, and then, present simulations results for a GMM-based speaker verification system, and preliminary results for a SVM-based speaker verification system. For the experiments, we mostly followed the approach adopted in [12].

4.1. IME corpus and experiments design

The IME corpus has been made available by the Signal Processing Group at IME (*http://www.ime.eb.br/~labvoz/*). The utterances in this database were collected from cellular and wired phone calls made by 75 speakers, through the D41ESC Dialogic board, which supports 8-bit PCM μ and A-laws. The speech files are stored in the Microsoft RIFF format as 8-bit PCM linear. Note that one would expect 12 or more bits per sample when expanding from the logarithmic to a linear scale.

We made two modifications to the corpus. The original 11-digit file names (e.g., *12151110051.wav*) were converted to names such as *id001.cel.train.man.RJ.cn.42.wav*, where a dot separates the information fields. These fields represent a unique speaker ID, cellular or wired phone, train or test, gender, speaker geographical origin, recording conditions, speaker's age and file extension (wav). Secondly, we eliminated silence from the utterances using a voice activity detector that is based on the signal energy.

4.2. GMM results

Model selection for GMM is often very simple. In this work, after few experiments using a validation set, we adopted 20 Gaussians per model. For the verification experiments, we used a universal background models (UBM) with 32 Gaussians. All the GMMs, representing the speakers in the database, were trained with 60 seconds of speech. For testing, we used segments of the following durations: 20, 10, and 5 seconds.

We compared the best results presented in [12] with the ones obtained with seven different front ends: a) the conventional 15 MFCCs, plus energy and two first derivatives (called here MFCCEDA48); b) MFCCEDAZ48, where Z indicates cepstral mean normalization (CMN); c) only 12 static coefficients with CMN (MFCCZ12); d) 12 PLPs, plus energy and two derivatives (PLPEDA39); e) PLPEDAZ39 and f) PLPZ7 with 7 static PLPs. Figure 2 and Figure *3* show

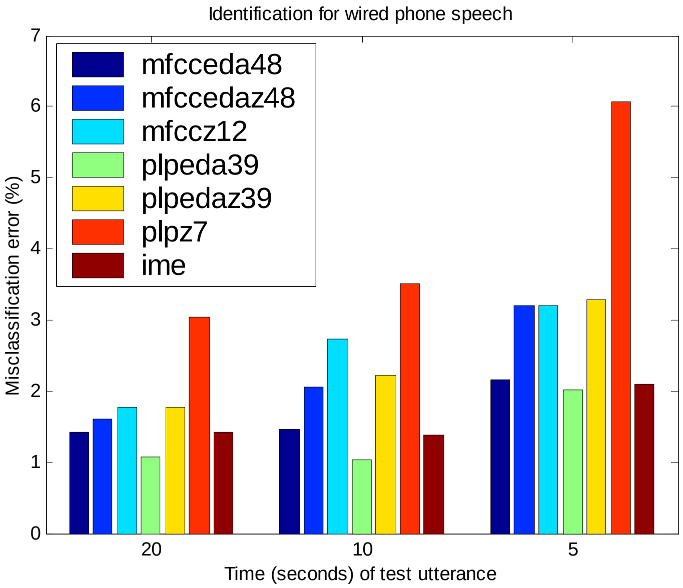

Figure 2- Identification error for wired phone speech for three durations of the test utterances: 5, 10, and 20 milliseconds.

the results obtained for wired and cellular phone speech, respectively. The results labeled as "*ime*" were obtained directly from [12], and correspond to a front end with 22 parameters (15 MFCCs plus 7 Hurst parameters), and the multi-dimensional fractional Brownian motion classifier (see [12] for details).

The standard front end plpeda39 outperformed the others as shown in Figure 2. It can be observed from Figure *3* that the plpedaz39 front end achieves the best result in all but one simulation: cellular phone speech with test utterances of 20 seconds. In this case, the error rate presented in [12] is 1.8%, which is very close to the 1.3% for wired phone speech, while, on average, the error rates in [12] for cellular are 5.8% higher than for wired phone speech. The results also show that though CMN is effective for cellular, it decreases the performance for wired phone speech.

In the next subsection, we present a comparison of the best GMM results with the preliminary results obtained with SVMs. The results presented in this section can be replicated by following the information on experimental framework provided in [13].

4.3. SVM results

Usually, model selection for kernel classifiers demands strategies more sophisticated than for a generative classifier. In our first experiments with SVM, all the model parameters were selected using ten-fold cross-validation (CV) on the training set, as commonly done in several domains. However, standard CV

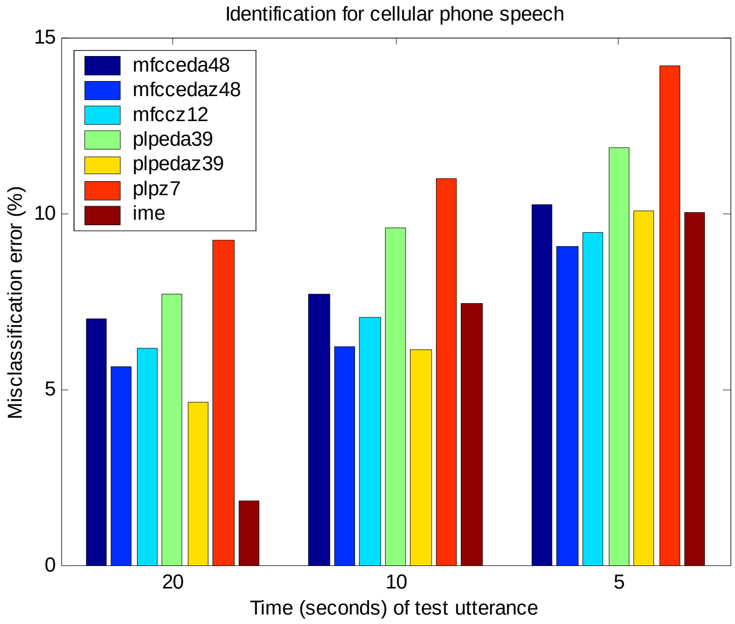

Figure 3- Identification error for cellular phone speech for three durations of the test utterances: 5, 10 and 20 milliseconds.

does not take in account that neighbor speech frames are correlated, and the random split of speech parameter vectors into folds for training and test sets can eventually make the test set too close to the training set. Better results were obtained using a validation set that does not overlap the training set.

The ISIP SVM trainer supports linear, polynomial, and radial-basis function (RBF) kernels. Among non-linear kernels, the Gaussian RBF kernel, when well-tuned, is competitive in terms of accuracy with any other kernel [14]. In this work, we exclusively adopted the Gaussian kernel, given by

$$K(x, x') = e^{-\gamma \|x - x'\|^2} . \tag{5}$$

The parameters γ and, C were selected using an iterative tuning process on a development test set.. First we selected C as 1, and then, chose γ by starting with γ = 1, then increasing (or decreasing) γ by a factor of 2 (or 0.5) until we did not get any further improvements. We iterated this process for three consecutive times. After choosing γ with C=1, we optimize C using the same iterative tuning procedure. Although this type of parameter tuning process is computationally intensive, it has an advantage of minimal manual intervention.

We also normalized the parameter vectors to have elements in the [0, 1] range. The normalization parameters were computed using the training set, and

	20 sec.	**10 sec.**	**5 sec.**
Previous [12]	0.67	0.85	1.13
MFCCEDA48-**GMM**	0.05	0.89	1.78
PLPEDA39-**GMM**	1.78	1.34	2.67
MFCCEDA48-**SVM**	1.10	1.40	2.30

*Table 1. C*omparison of Equal error rate (EER) for speaker verification systems using wired phone speech (%).

then, used to normalize both the training and the test sets.

In our experiments, we adopted a *frame-based* approach. Up to date, the *hybrid* is clearly the most successful architecture (see, e.g., [11]). However, given that the GMM-based systems follow a frame-based architecture, we chose to use it in our preliminary experiments with kernel classifiers in order to better understand the tradeoffs between the generative and discriminative approach.

More specifically, we consider the speaker verification problem an excellent domain to stress the differences between generative and discriminative learning. This is an important research topic in machine learning nowadays, and recent theoretical results have brought some light to this matter (see, e.g., [15]). For example, when training data is scarce, generative classifiers can outperform discriminative ones. Hence, our investigation aims to confirm these claims and elaborate new ones using speaker verification experiments. As shown in Table 1, our preliminary results show that the current performance achieved by SVM using the frame-based architecture is outperformed by GMM.

5. CONCLUSIONS

In this paper, we described the ISIP toolkit, an open source framework for developing speaker, and speech recognition systems. Among other features, the ISIP toolkit supports GMM and SVM-based systems. We also presented new results for the IME corpus using Gaussian mixture models, which outperformed the previously published results [12] in most cases.

The future work concentrates on using the speaker verification problem to investigate the tradeoffs between generative and discriminative learning. We are currently estimating the variation in the performance of GMM and SVM-based systems with respect to the amount of training data. These experiments will be useful to validate the conclusions in [15]. We are also planning to expand the experimentation using the SVM system on hybrid architectures with an aim to outperform the GMM based system.

6. REFERENCES

1. N. Deshmukh, A. Ganapathiraju, and J. Picone, “Hierarchical Search for Large Vocabulary Conversational Speech Recognition,” *IEEE Signal Processing Magazine*, vol. 16, no. 5, pp. 84-107, September 1999.

2. R. Sundaram, J. Hamaker, and J. Picone, “TWISTER: The ISIP 2001 Conversational Speech Evaluation System,” presented at the *Speech Transcription Workshop*, Linthicum Heights, Maryland, USA, May 2001.

3. F. Zheng, and J. Picone, “Robust Low Perplexity Voice Interfaces,”, MITRE Corporation, May 15, 2001.
(*http://www.isip.msstate.edu/publications/reports/index.htm.)*

4. N. Parihar, and J. Picone, “DSR Front End LVCSR Evaluation,” AU/384/02 Aurora Working Group, December 2002.
(*http://www.isip.msstate.edu/projects/aurora)*

5. B. Jelinek, F. Zheng, N. Parihar, J. Hamaker, and J. Picone, “Generalized Hierarchical Search in the ISIP ASR System,” *Proceedings of the Thirty-Fifth Asilomar Conference on Signals, Systems, and Computers*, vol. 2, pp. 1553-1556, Pacific Grove, California, USA, November 2001.

6. K. Huang, and J. Picone, “Internet-Accessible Speech Recognition Technology,” presented at the *IEEE Midwest Symposium on Circuits and Systems*, Tulsa, Oklahoma, USA, August 2002.

7. A. Ganapathiraju, J. Hamaker, and J. Picone, “Applications of Support Vector machines to Speech Recognition,” *IEEE Transactions on Signal Processing*, Fall 2004.

8. J. Hamaker, “Sparse Bayesian Methods for Continuous Speech Recognition,” Ph.D. Dissertation Proposal, Department of Electrical and Computer Engineering, Mississippi State University, March 2002.

9. T. Jaakkola, and D. Haussler, “Exploiting generative models in discriminative classifiers,” *Proc. of Tenth Conference on Advances in Neural Information Processing Systems*, 1998.

10. N. Smith, and M Gales, “Speech Recognition using SVMs”, Advances in Neural Information Processing Systems 14, MIT Press, 2002.

11. V. Wan, and S. Renals, “Speaker Verification using Sequence Discriminant Support Vector Machines”, to appear in *IEEE Transactions on Speech and Audio Processing*.

12. R. Sant’Ana, R. F. Coelho and A. Alcaim, “A New Text-Independent Automatic Speaker Recognition System Based on the Hurst Parameter” (in Portuguese), in XX SBT, Rio, Brazil, 2003.

13. T. Imbiriba, “Speaker Recognition using the IME Corpus”. Universidade Federal do Para. Graduation Project Report (in preparation).

14. R. Rifkin and A. Klautau, “In Defense of One-Vs-All Classification”, In Defense of One-Vs-All Classification, 5(1), 101-141, 2004. *http://www.jmlr.org/*

15. R. Raina, Y. Shen, A. Y. Ng, and A. McCallum, “Classification with Hybrid Generative/Discriminative Models”, In *Advances in Neural Information Processing Systems, 16*, 2004.