IN-SET / OUT-OF-SET SPEAKER RECOGNITION: LEVERAGING THE SPEAKER AND NOISE BALANCE*

Matthew R. Leonard and John H.L. Hansen

Speech and Speaker Modeling Group
Center for Robust Speech Systems (CRSS)
University of Texas at Dallas, Richardson, TX, 75083, U.S.A
{Matthew.Leonard@student. , John.Hansen@}utdallas.edu

ABSTRACT

This study addresses the problem of identifying in-set versus out-of-set speakers in noise for limited train/test durations in situations where rapid detection and tracking is required. The objective is to form a decision as to whether the current input speaker is accepted as a member of the enrolled in-set group or rejected as an outside speaker. A new scoring algorithm that combines scores across an energy-frequency grid is developed where high-energy speaker dependent frames are fused with weighted scores from low-energy noise dependent frames. By leveraging the balance between the speaker versus the background noise environment, it is possible to see an improvement in equal error rate performance. Using an initial form of the algorithm with speakers from the TIMIT database with 5 seconds of train and 2 seconds of test, the average relative EER performance improvement is 27.4%. The results confirm that for situations in which the background environment type remains constant between train and test, an in-set/out-of-set speaker recognition system that takes advantage of information gathered from the environmental noise can be formulated which realizes significant improvement.

Index Terms: Speaker recognition, in-set/out-of-set, environmental noise

I. INTRODUCTION

In-set/out-of-set speaker recognition systems are useful for situations in which it is important to detect and track the presence of a group of speakers. Examples of speech systems that benefit from in-set/out-of-set recognition include dialog systems, communications systems, spoken document retrieval, and security applications that allow access to private information only to people belonging to a specific group of authorized users [1]. The objective of an in-set/out-of-set speaker recognition system is to make a decision as to whether to accept the claim that the current input speaker is a legitimate member of the enrolled in-set group, or to reject the claim and classify the speaker as an outside speaker. Generally speaking, the number of speakers seen for in-set and out-of-set speaker recognition in our scenarios is in the range of 50-100.

The evaluation of in-set speaker recognition is based on two error measurements. The first, false rejection (FR), occurs when a member of the enrolled in-set group is rejected and classified as belonging to the out-of-set; the second, false acceptance (FA), occurs when an outside speaker is accepted as being part of the in-set group. One of the main challenges for this type of system is the effective rejection of outliers, while still allowing variability for the in-set speakers, such as interspeaker variations at the segmental level [2].

Basic in-set/out-of-set speaker recognition is performed as follows: a speaker-independent universal background model (UBM) is generated from an available set of non-target speakers. Models are then trained for all target in-set speakers. When a test speaker is submitted to the system, features for that speaker are generated and tested against each of the in-set speaker models as well as the UBM. Once these scores are generated, the classified model with the highest probability is selected. Thus, for *N* in-set speakers, an input speaker has *N+1* possible classifications; it is either one of the *N* in-set speakers, or it belongs to the UBM. If the classification is one of the in-set speakers, the input speaker's claim is accepted; otherwise, the claim is rejected. For this study, modeling is done with what has become the dominant approach in text-independent speaker recognition: Gaussian Mixture Models (GMMs) with a UBM and maximum *a posteriori* (MAP) speaker adaptation [3].

This paper is organized as follows. In Section II, the corpus and background noise types utilized are discussed. Section III covers noise and in-set/out-of-set speaker recognition. Section IV presents our new approach which incorporates models across an energy-frequency grid to obtain from both high-energy speaker dependent frames and low-energy noise dependent frames, and reports on a series of experimental results for an intial version of the approach. Section V presents the foundation of the full selective leveraging system. Finally, conclusions and discussion of future work is presented in Section VI.

*This work was sponsored by RADC under Grant A40104, and the University of Texas at Dallas under Project Emmitt.

1-4244-1484-9/08/$25.00 ©2008 IEEE

II. CORPUS AND NOISE TYPES

The system evaluation performed in this paper is based on the TIMIT database, down sampled to 8 kHz. In the present scenario, there are 60 total speakers, divided into 15 in-set speakers and 45 out-of-set speakers. Since the system emphasizes rapid detection and tracking, both the training and testing data are of short durations. Specifically, training data of 5 seconds duration and test data of 2 seconds duration are used.

Six different car noises were used for this study. These include (all models manufactured by General Motors Corporation) a Blazer (BLA) SUV, a Cavalier (CAV) car, an Express (EXP) van, a S10 (S10) pickup truck, a Silverado (SIL) full-size truck, and a Venture (VEN) minivan. Separate samples of each vehicle noise were used to degrade speakers for development, test, and train to ensure open segment time variability.

III. NOISE AND IN-SET/OUT-OF-SET SPEAKER RECOGNITION

While most in-set/out-of-set speaker recognition systems work reasonably well under clean conditions, the introduction of noise corruption causes a significant change in the equal error rate of the system. Numerous techniques have been suggested as a means to suppress noise from the speech signal in order to decrease the error; these include spectral subtraction and quantile-based noise reduction. Recently, some studies have considered using the noise context as an information source for which the system can adapt its decisions. Akbacak and Hansen [5] proposed the framework of Environmental Sniffing which can detect, classify, and track acoustical environmental structure in order to seek out detailed information that characterizes these conditions and use that knowledge to direct the processing of speech systems. In another example, Müller considered estimating the acoustic context in order to determine whether or not certain acoustic classifiers would be reliable for speaker classification [6].

For the purposes of this study, we assume that the background environment is the same for a particular speaker between train and test. This assumption allows for the use of the noise as an aid in the successful acceptance or rejection of an input speaker for in-set/out-of-set speaker recognition. While this assumption cannot be made for every scenario, there are many applications in which it applies. In the case where the rapid detection and tracking of speakers in a relatively short time frame is necessary, generally the speaker will be in the same environmental context.

There are many applications where rapid detection and tracking of speakers over audio streams is necessary, such as spoken document retrieval or monitoring pilots during air traffic control. For example, consider the tracking of various TV anchors and correspondents reporting the news. The main anchors will be, with a high degree of certainty, reporting from within the studio, and thus a noise model based on the background acoustics of the studio should be seen both during the training and test phases. Likewise, the traffic correspondent reporting from a helicopter will always have the helicopter as his or her environmental context. Another scenario in which this assumption can be made is for the monitoring and tracking of commercial communications at airports involving ground and air units. Commercial aircraft communications will have pilots in the same aircraft during take-off, travel, and landing. The pilot will have a distinctive noise environment when compared to the driver of a baggage transportation vehicle or an air traffic controller. Furthermore, it is highly unlikely that these speakers would switch environments during a restricted time period, as the probability of one individual such as a pilot to be trained and rated to fly an aircraft as well as also being an operator of a transport vehicle can safely be assumed to be very small.

It is important to note that the background noise information is not the main focus of the in-set/out-of-set speaker recognition system; rather, the noise context plays a role in that the knowledge can be used to augment the speaker-dependent information the system is already using as the basis for a large part of its decision.

IV. LEVERAGE APPROACH: SPKR+ENV

The new approach proposed in this study to increase performance of in-set/out-of-set speaker recognition is to take advantage of the assumption that a given speaker will remain in the same noise environment between train and test phases.

The 'standard' method used as our baseline is one in which the speech waveform is applied to an energy threshold; frames with an energy above the threshold are used to train a GMM for the enrolled speakers, while frames with energy lower than the threshold are set aside. The new method developed here employs an energy-frequency grid to tag input frames. In the simplest case, low energy frames (separated from high energy frames by an energy threshold λ) are used to train a separate GMM which has noise or silence content, with some low energy consonant information (see Fig. 1). When an input speech signal is submitted, the system evaluates the scores associated with in-set and out-of-set cases for both high energy and low energy GMMs. Next, a weight (β) is applied to the low energy scores and combined with those from the high energy GMM to create an overall leveraged (SPKR+ENV) score. The final decision is based on these scores and the EER performance is calculated. The system is then enhanced by generalizing the frame scheme to a grid for both energy and frequency.

Speaker recognition systems generally set aside low energy frames, since they normally contain low-energy consonants or silence which are prone to noise. Since

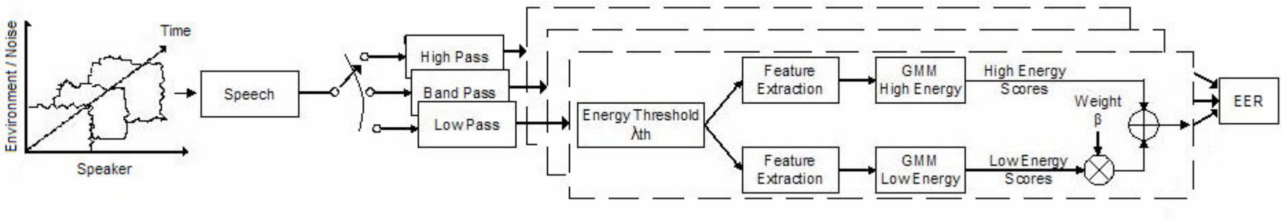


Figure 1 – Leveraged (SPKR+ENV) System Approach for In-Set/Out-of-Set Speaker Recognition

Our method for in-set/out-of-set speaker recognition considers very small amounts of train (5secs) and test (2secs) data, setting aside any data could lower performance in noise free scenarios. Therefore, the proposed method assumes that high energy frames have speaker dependent phonemic content with some background environmental structure, and low energy frames have primarily environmental content with some speaker dependent consonant information. We employ an in-set framework similar to our earlier work [1], where the speaker size is 60, with a 15/45 in-set/out-of-set size. By varying the weight value β (see Fig. 1), it is possible to control the emphasis placed on low energy environmental centric scores.

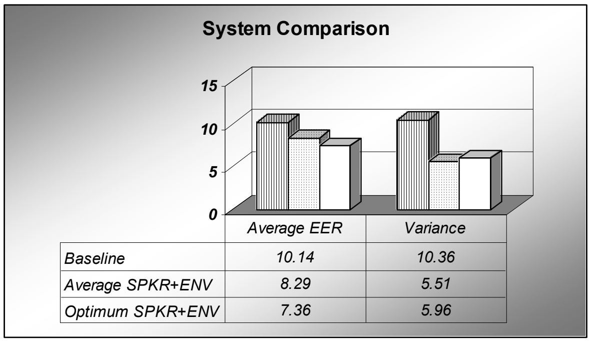
In order to provide a comparative analysis of the benefits of the new (SPKR+ENV) leverage approach, several test sets were randomly created with an equal number of each of the 6 noise types; of the 60 speakers, each of the 6 noise types were used for a random set of 10 speakers, and the speakers that comprised a certain noise type were kept constant between train and test. The same 15 speakers were chosen as the in-set speakers for all of the tests. Eight random noise combinations were used, labeled Sets A through H.

Once the test sets were constructed, the baseline method of setting aside the low energy frames was implemented for in-set/out-of-set speaker recognition, yielding an average EER of 10.14.

Next, the simplest case of the leveraged (SPKR+ENV) approach (no frequency partitioning) was employed, with λ = 0.3 and with β values ranging from 0.05 to 1.00, where the scores from the GMM trained on low energy frames was counted at most as equal to the scores from the high energy frames (β = 1.0), with a minimum weight of 5% (β = 0.05). The average results from this method resulted in an EER of 8.29, with an average relative performance gain of 18.3%.

For each of the evaluation sets, the optimum performance occurred when β = 0.70, with an average EER of 7.36. These results indicate that for the optimum tested value of β, the absolute improvement was 2.78%, with a relative improvement (decrease in EER) of 27.4%.

The evaluation process was also run using an energy threshold of λ = 0.1. While still yielding improved performance over the baseline method, the improvement was less than that of the system when λ = 0.3. The results of the λ = 0.3 experiments can be seen in Table 1.



	Average EER	Variance
Baseline	10.14	10.36
Average SPKR+ENV	8.29	5.51
Optimum SPKR+ENV	7.36	5.96

Table 1 – System Comparison for EER and Variance

Fig. 2 shows the detection error tradeoff (DET) curves for the baseline and optimum SPKR+ENV methods. This curve demonstrates how the SPKR+ENV algorithm improves overall EER performance.

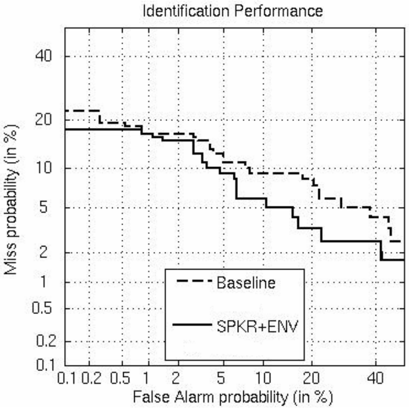


Figure 2 – DET Curve for Baseline and Optimum SPKR+ENV Methods

V. SELECTIVE LEVERAGING FRAMEWORK

The next step in enhancing the SPKR+ENV algorithm was to formulate a method in which the noise environments could be evaluated for a decision on how to apply the leveraging process. A framework for this type of selectively leveraged SPKR+ENV system has been developed that evaluates the speakers using a grid with both energy and frequency thresholds, partitioning the dimensions as can be seen in Fig. 3. The goal of the partitioning is that some of the partitions will contain speaker dependent traits while other partitions will contain noise dependent traits.

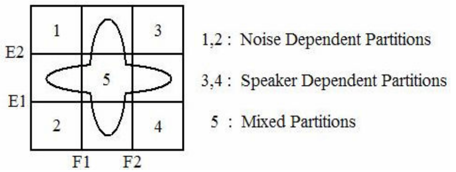

1
3
E2
5
E1
2
4
F1
F2
1,2 : Noise Dependent Partitions
3,4 : Speaker Dependent Partitions
5 : Mixed Partitions

Figure 3 – Decision Grid Consisting of Energy Versus Frequency Partitioning

An examination was performed for BLA, CAV, and EXP data files using partitioning variables of F1 = 300 Hz, F2 = 600 Hz, E1 = 0.1 (normalized), and E2 = 0.3 (normalized). The total number of frames distributed to each of the partitions was tabulated, and the percent of the total frames calculated. Fig. 4 shows the percent of total frames results for the Low Frequency / High Energy partition (labeled '1' in Fig. 3) after being passed through a five-point median filter. This figure clearly demonstrates three distinct bands for the BLA, CAV, and EXP degraded files, confirming that environmental background would be a useful discriminatory trait with frequency dependency.

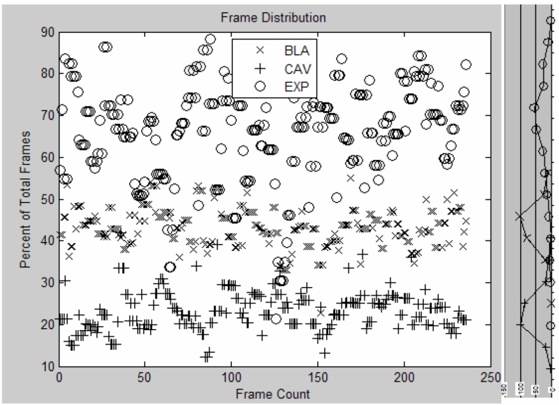


Figure 4 – Frame Distribution for Low Frequency / High Energy Partition for BLA, CAV, and EXP files

Therefore, it is possible to develop a SPKR+ENV system that front-end analyzes the noise content using the grid partitioning method before leveraging the background noise. An application of this analysis is to verify that the speaker stays in the same environment between train and test, and upon verification apply the SPKR+ENV process.

VI. DISCUSSION AND CONCLUSIONS

Speaker recognition systems must overcome a range of issues in order to achieve and maintain performance in diverse environments. For in-set/out-of-set speaker recognition, where limited amounts of train/test data are available, leveraging knowledge of the acoustic environment offers an additional dimension to improve system performance. The proposed (SPKR+ENV) system shows significant performance improvement by leveraging environmental structure, which virtually all other algorithms intentionally ignore.

By taking into account low energy noise-dependent frames we significantly increase the performance for in-set/out-of-set speaker recognition. For cases in which the speaker distribution was evenly distributed across 6 car noise types, the average relative performance improvement was 27.4%. This indicates that for situations in which one can assume that the background speaker environment of a speaker remains constant (but randomly distributed across the in-set speakers) between train and test phases, the leveraged (SPKR+ENV) approach will eliminate 1 out of every 4 decision errors. Since we are focused on a scenario in which both train and test data are of short durations (5 seconds and 2 seconds respectively), this reduction in error is particularly beneficial because of the limitation a lack of data places on other methods such as spectral subtraction.

One important aspect of error reduction is whether the errors fall in false accept (FA) or false reject (FR) categories. If the models for two speakers are similar, but their acoustic backgrounds are different, knowledge of the background noise should help drive the models farther apart. An analysis of errors from both the baseline and the SPKR+ENV algorithm shows that the leveraged method eliminates many of the false accept errors.

Several further enhancements are possible given the framework of the SPKR+ENV algorithm, in particular optimization of the frequency partitioning threshold, and the formulation of an algorithmic approach to adaptively optimize the frequency and energy thresholds for an unknown background environment.

While the leveraged (SPKR+ENV) approach presented here cannot be applied to every in-set/out-of-set speaker recognition scenario, it does significantly improve performance when utilized in systems looking at the rapid detection and tracking of speakers that remain in the same noise environment between train and test phases. It also allows us to achieve an upper bound on performance improvement when noise is present.

REFERENCES

[1] P. Angkititrakul, J.H.L. Hansen, "Discriminative In-Set/Out-of-Set Speaker Recognition", *IEEE Trans. Audio, Speech, and Language Proc.*, vol. 15, no. 2, pp. 498-508, Feb. 2007.

[2] G. Doddington, "Speaker recognition-identifying people by their voices," *Proc. IEEE*, vol. 73, no. 11, pp. 1651–1664, 1985.

[3] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Process.*, vol. 10, no. 1–3, pp. 19–41, 2001.

[4] J.H.L. Hansen, L.M. Arslan, "Robust Feature-Estimation and Objective Quality Assessment for Noisy Speech Recognition using the Credit Card Corpus", *IEEE Trans. Speech & Audio Proc.*, vol. 4, no. 3, pp. 169-184, May 1995.

[5] M. Akbacak, J.H.L. Hansen, "Environmental Sniffing: Noise Knowledge Estimation for Robust Speech Systems", *IEEE Trans. Audio,Speech,and Lang. Proc.*,vol.15,no.2, pp.465-477, Feb. 2007.

[6] C. Müller, "Estimating the Acoustic Context to Improve Speaker Classification", German Research Center for Artificial Intelligence, 2005.