ICSI'S 2005 SPEAKER RECOGNITION SYSTEM

Nikki Mirghafori, Andrew O. Hatch, Steven Stafford, Kofi Boakye, Daniel Gillick, and Barbara Peskin

International Computer Science Institute
1947 Center Street, Suite 600
Berkeley, CA 94704
{nikki,ahatch,sjs,kaboakye,dgillick,barbara}@icsi.berkeley.edu

ABSTRACT

This paper describes ICSI's 2005 speaker recognition system, which was one of the top performing systems in the NIST 2005 speaker recognition evaluation. The system is a combination of four sub-systems: 1) a keyword conditional HMM system, 2) an SVM-based lattice phone n-gram system, 3) a sequential non-parametric system, and 4) a traditional cepstral GMM System, developed by SRI. The first three systems are designed to take advantage of higher-level and long-term information. We observe that their performance is significantly improved when there is more training data. In this paper, we describe these sub-systems and present results for each system alone and in combination on the Speaker Recognition Evaluation (SRE) 2005 development and evaluation data sets.

1. INTRODUCTION

Because of their effectiveness and relative simplicity, cepstral-based Gaussian Mixture Models (GMMs) [16] have become the mainstream approach for text independent speaker recognition systems. However, the cepstral bag-of-frames modeling, although powerful, does not take advantage of long-term information present in the speech signal. There have been myriad attempts, especially since the Johns Hopkins 2002 Workshop [18], to harness the power of such features. The systems developed at ICSI, and presented in this paper, are contributions to this endeavor.

ICSI's 2005 speaker recognition system is a weighted combination of four systems. Three of the systems were developed at ICSI and aim at employing higher-level information. These systems are:

- Keyword Conditional HMM System (WordHMM)
- SVM-based Lattice Phone N-gram System (PhoneNgram)
- Sequential Non-Parametric System (SNP)

Each system relies on either phone- or word-level recognition transcription. The fourth system, which has been developed by SRI, is a traditional cepstral-based GMM system. The four systems are combined at the score level using a neural network.

The nature of the data often determines the parameters of the research. Starting in year 2001, NIST added the "extended data task" to its yearly speaker recognition evaluation (SRE), with the intent "to foster new research ... through the discovery and exploitation of higher-level and more complex characteristics of a speaker's speech" [14]. The extended data task soon became a main focus of the evaluations. Our three systems are designed to use such long-term information and to model idiosyncratic speaker behaviors.

This paper is organized as follows: In Section 2, we describe the NIST Speaker Recognition Evaluation 2005 (SRE05) task. In section 3, the resources that are common to the systems, such as the development data and combination strategy, are discussed. In Section 4, the constituent system descriptions and experimental results on the development set are presented. Experimental results on the SRE05 evaluation set and an analysis of system combinations are presented in 5. Conclusions are presented in Section 6.

2. THE NIST SPEAKER RECOGNITION EVALUATION

To help the reader place this work in context, we briefly describe NIST SRE05. For complete details, please see the Evaluation Plan [15].

The NIST SRE05 is a part of yearly ongoing evaluations in speaker recognition. For this year's evaluation, data was selected from the Mixer corpus [13]. Five mintues of each conversation between unfamiliar speaking partners were selected, with the assumption (or hope) that each party spoke for roughly half the time. Data from different channel conditions (e.g., landline, cellular, hands-free) as well as languages other than English (e.g., Spanish, Mandarin, Arabic) were collected. The evaluation paradigm includes a 5x4 matrix of optional training (such as 10 seconds, 1 side, or 8 sides of a conversation) and testing (such as 10 seconds, or an entire conversation side) conditions. The *required* condition was chosen to be one in which the length of both the training and test segments were one conversation side each (referred to as "1 side training" in the rest of the paper), and included both English and non-English trials. The *common* subset included only English trials collected from a hand-held telephone.

Given our interest in long-term features, we also participated in the "8 side training" condition, using 8 conversation sides for training. In both cases, we evaluated the systems on test segment durations of 1 conversation side only.

Note that these conditions also include some non-English trials, where either the training, testing, or both segments are in a language other than English. Due to both the linguistic dependencies of some of our systems, as well as the availability of speech recognition output in English, we have chosen to concentrate on the English trials. Although for the SRE05 evaluation we did test a subset of our systems on the non-English trials, in this paper we focus on and only report results on the English trials.

This material is based upon work supported by the National Science Foundation under grant number 0329258.

In SRE, performance is measured primarily using the Decision Cost Function (DCF), although the traditional measure of equal error rate or EER (the point at which the rate of false alarms and misses are equal) is also reported. DCF is the weighted sum of miss and false alarm error probabilities, defined by:

$$C_{Det} = C_{Miss} * P_{Miss|Target} * P_{Target} + C_{FalseAlarm} * P_{FalseAlarm|NonTarget} * (1 - P_{Target})$$

where $(C_{Miss}, C_{FalseAlarm}, P_{Target})$ are defined to be (10, 1, 0.01). C_{Det} is normalized by dividing by the best cost that could be obtained without processing the data, i.e., by $(C_{Miss} * P_{Target})$, which effectively multiplies C_{Det} by 10.

3. COMMON RESOURCES

In this section, we describe resources that the systems share, such as the datasets, the combination strategy, and the ASR system.

3.1. Datasets

Five databases, all of which contain conversational telephone (landline or cellular) speech, were used for development and testing of the systems:

1. NIST SRE 2005 (Mixer)
2. NIST SRE 2004 (Mixer)
3. NIST SRE 2003 extended data (Switchboard-II phases 2 and 3)
4. NIST SRE 2002 (Switchboard Celluar)
5. Fisher

The NIST SRE 2005 was only used for the final evaluation. Subsets of the other four databases were used for training the background models, estimating the TNORM statistics, training the combination weights, and development testing . For details on the selection of the development subsets, see Section 1 of [9].

Background Models Training: Background data for the WordHMM, SNP, and PhoneNgram systems was selected from Fisher and SRE 2003 data sets. The SRI GMM system additionally used NIST SRE 2002 for background training. The amount of data used for background training by each system is detailed in Section 4.

TNORM Models Training: TNORM models [2] were trained using utterances from Fisher. The models were constructed from the speech in one conversation side of the middle 5-minute segment of the conversations. These TNORM models were from unique speakers and roughly gender-balanced. The data comprised a similar number of electret and cellphone channels, and a handful of carbon-button channels.

Combination Weights Training: We experimented with SRE04 eval data and subsets of SRE03 and Fisher to train the weights of the neural network combiner (discussed in Section 3.2). However, in the final system (which is reported in this paper), SRE04 eval data was used to estimate the neural network combination weights and the optimal operating point corresponding to the minimum DCF.

Development Testing: The SRE04 evaluation set, as well as subsets of SRE03 and Fisher were used for development testing. For brevity, we only report on the development testing on SRE04 eval data.

Evaluation Testing: The SRE05 evaluation set served as the held-out evaluation data. Table 1 shows the number of models and trials for SRE04 and SRE05 evaluation sets for all English trials. The "common condition" subset has fewer trials than the "all English" subset; hence, we report results on the latter for more statistically meaningful comparisons.

Eval Set	Train Cond	Models	Trials
SRE04	1side	479	15,317
SRE04	8side	225	7,336
SRE05	1side	598	20,907
SRE05	8side	464	16,053

Table 1. The number of all English models and trials for SRE04 and SRE05 evaluation sets.

3.2. Combination Strategy

The systems were combined using LNKNet software [12]. A neural network with no hidden layer and sigmoid output non-linearity was used. The combination of all four systems was used to obtain scores for the English trials.

As mentioned in the previous section, SRE04 data was used to estimate the optimal combination weights. SRE04 1side-1side and 8side-1side train/test conditions were used for SRE05 1side-1side and 8side-1side evaluation conditions, respectively. Furthermore, only the English trials in SRE04 were used to estimate weights and operating threshold for English trials in SRE05.

3.3. ASR System

The WordHMM and SNP systems rely on word recognition, using the word-level word and phone alignments, respectively. The PhoneNgram system uses lattices produced from open phone loop recognition. We used the output produced by SRI's DECIPHER 3xRT conversational telephone speech (CTS) recognition system. For the word-level recognition, we used models developed for the NIST RT-03F evaluation. The system was trained on Switchboard-I, some Switchboard-II, and CallHome English data, as well as Broadcast News and web data for the language model (no Fisher data was used in training the ASR system). For the details of the ASR system, see [11].

4. THE CONSTITUENT SYSTEMS

In this section, we describe the four sub-systems that made up our recognition system and present development results on the SRE04 dataset. Three of the systems, which aim to take advantage of higher-level information, were developed at ICSI. The fourth one, the cepstral GMM system, was developed at SRI and its output was shared with us. It is a common approach in the field to combine cepstral GMM systems with those designed for longer term features in order to take advantage of both short- and long-term information inherent in the speech signal and to evaluate the added information provided by the long-term features.

4.1. Keyword Conditional HMM System (WordHMM)

The main idea of the Keyword Conditional HMM system [3] is to capitalize on advantages of text-dependent systems in a text-

independent domain. Whole-word HMM models are trained on only a small subset of words (and word-pairs). We use a set of 19 keywords which we believe are rich with speaker characteristic cues. These keywords are drawn from the following classes:

- **Discourse markers**: actually, anyway, like, see, well, now, you_know, you_see, i_think, i_mean
- **Filled pauses**: um, uh
- **Backchannels**: yeah, yep, okay, uhhuh, right, i_see, i_know

The keyword HMMs were left-to-right state sequences with self-loops and no skips. Each state model consisted of a mixture of eight Gaussians and the number of states for each keyword was defined to be the smaller of the average number of phones in pronunciations of the word, multiplied by 3, and the median duration in frames, divided by four; that is $min(mean(NumPhones * 3), median(NumFrames/4))$. All modeling and scoring was performed using the HMM Toolkit, HTK [5].

The HMM feature vectors consisted of MFCCs C_0 through C_{19} and their deltas, producing a 40 element feature vector. Cepstral Mean Subtraction was performed over the speech regions of each conversation side, as determined by the ASR word alignments.

Universal Background Models (UBMs) were trained on 1,128 Fisher and 425 SRE03 conversation sides from unique speakers. Speaker models were obtained by MAP adapting the Gaussian means of the UBMs. In the cases where there were no speaker training data for a particular word, the UBM was simply copied as the speaker-specific model. This resulted in removing the influence of the keyword, as the contribution to the overall score was zero, due to the cancellation of target and background. The location of the keywords within the conversation was determined using word-level ASR alignment.

Each keyword appearing in the test segment was scored by taking the difference between the log probabilities obtained from scoring the speaker-specific and UBM models on the test tokens. The final score was obtained by adding these keyword scores and normalizing by the total number of frames. The final scores were T-normalized with 249 models constructed from unique conversation sides of the Fisher database. Figure 1 is a schematic diagram of the WordHMM system.

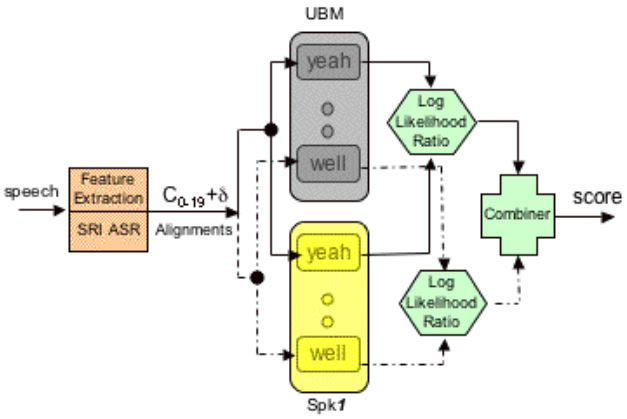


Fig. 1. Schematic diagram of the WordHMM system.

Table 2 shows the improvement in performance of the SRE05 WordHMM system compared to SRE04 evaluation and post-evaluation systems. The SRE04 evaluation system UBMs were trained entirely on SRE03 data, whereas those of the SRE04 post-evaluation system were trained entirely on Fisher data, which is a better match to SRE04 eval data. Therefore, the improvement between the first and second rows of the table is entirely due to the difference in data used to train the UBMs. This year's WordHMM system (the third row in the table) differed with the previous two systems in a number of ways. The most significant was major infrastructure changes which resulted in the ability to more accurately utilize the word alignment boundaries. Similarly, speed enhancements allowed the addition of TNORM as well as enabled us to better search the system parameter space, such as the number of states per model and number of Gaussians per state. This system utilized eight, versus the previously used four, Gaussians per state. Finally, the UBM models were trained on a combination of SRE03 and Fisher data.

WordHMM	**1side Train**		**8side Train**	
	EER	**DCF**	**EER**	**DCF**
SRE04 system	13.06%	0.526	8.85%	0.382
SRE04 post-eval	12.98%	0.445	7.06%	0.306
SRE05 system	11.38%	0.399	6.27%	0.224

Table 2. The improvement in the WordHMM system from SRE04 to SRE05. Results are shown on all English trials of SRE04 evaluation set. DCF is short for Min DCF in tables throughout this paper.

It is of interest to examine the performance of the WordHMM system alone, but particularly in combination with the traditional cepstral GMM system (discussed in Section 4.4). Table 3 shows the results of the WordHMM system on all English trials of SRE04. As mentioned in Section 3.2, the systems were combined at the score level using a neural network. We see that as the amount of training data increases, the relative performance gap between the WordHMM and GMM narrows. Also, the WordHMM contributes much more significantly in the reduction of error in the 8-side training condition, as expected.

WordHMM	**1side Train**		**8side Train**	
	EER	**DCF**	**EER**	**DCF**
WordHMM	11.38%	0.3990	6.27%	0.2244
GMM	7.73%	0.3113	4.96%	0.2115
WordHMM+ GMM	7.59% (2%)	0.2721 (13%)	4.08% (18%)	0.1672 (21%)

Table 3. The performance of the WordHMM system alone and in combination with the cepstral GMM. Results are on all English trials of SRE04 evaluation set. Values in ( ) are percent improvements relative to the GMM system alone.

4.2. SVM-based Lattice Phone N-gram System

The Phone N-gram (PhoneNgram) system uses an open-loop phone recognizer to generate phone lattices, which are then used to compute expected counts of phone n-grams. These expected counts are converted into relative frequencies, which form the feature vectors for training SVM-based speaker models [7]. This system builds on [1, 4] and dramatically improves the results by using lattices rather than 1-best phone recognition hypotheses.

As mentioned in Section 3.3, we used SRI's speech recognizer [11] to generate phone lattices for every conversation side. Our particular realization of the ASR system used gender-dependent, monophone acoustic models, where each monophone was modeled by a 3-state HMM. The acoustic models were trained on the Switchboard I corpus using MFCC features. Phone decoding was performed in open-loop mode (i.e. we used a unigram phone language model with uniform probabilities) with a vocabulary of 46 phone units.

For the SVM, one feature vector for every conversation side was used, where the features represent relative frequencies of the 8500 most frequent phone bigrams and trigrams. We used a linear kernel [4] to train an SVM-based model for every target speaker. The SVMs were trained using a one-versus-all approach, where the conversation sides from the target speaker's training data were used as positive training examples, and the conversation sides in a set of background data were used as negative training examples. For this system, we used the same background dataset as the WordHMM system. SVM training and scoring was done using the SVMlite package [8].

To score a given test-target pair, we simply applied the feature vector of the test conversation to the SVM output function of the target model. We then used TNORM to normalize the scores for every test conversation.

Table 4 shows the results of the PhoneNgram system on all English trials of SRE04 alone and in combination with the cepstral GMM system. For the 1-side training case, the performance of the PhoneNgram system is significantly worse than the GMM system, but with the addition of data in the 8-side training condition, the PhoneNgram system's performance dramatically improves to the point of matching the GMM's. Yet, the behaviors of the system are complementary, as their combination yields further improvement (˜30% for the 8-side training case).

PhoneNgram	**1side Train**		**8side Train**	
	EER	**DCF**	**EER**	**DCF**
PhoneNgram	12.09%	0.5408	4.96%	0.2358
GMM	7.73%	0.3113	4.96%	0.2115
PhoneNgram+ GMM	6.47% (16%)	0.2767 (11%)	3.64% (27%)	0.1443 (32%)

Table 4. The performance of the Phone N-gram system alone and in combination with cepstral GMM. Results are reported on all English trials of SRE04 evaluation set.

4.3. Sequential Non-Parametric (SNP) system

The SNP system uses a non-parametric technique, where no explicit speaker models are built. The system performs speaker detection by comparing a test segment directly to similar instances of that segment in the training data [6]. Figure 2 shows the system schematically.

The cepstral features employed by this system were identical to the ones used in the WordHMM system (MFCC C0-C19 plus deltas, with CMS). The system utilized ASR word alignments to compare phone-trigram regions in the test conversation with every occurrence of the phone-trigram in the training conversation(s). Dynamic Time Warping (DTW) was used to align frame sequences of different lengths. Frame normalized Euclidean distances were calculated between warped phone-trigram regions.

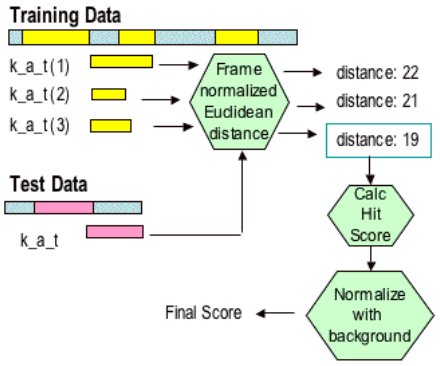


Fig. 2. Schematic diagram of the SNP system.

Using the best (smallest) Euclidean distance, the "Hit Score" was calculated, where:

$$\text{Hit Score} = \sum_{i \in \text{test tokens}} \frac{\text{number of matched frames in } i}{k^{\text{distance}[i]}} \quad (1)$$

This scoring method primarily captures positive evidence by placing exponentially higher weight on small distance values. The value of the constant k was empirically estimated to be 1.5.

Background normalization was implemented by scoring the test conversation against a background set in a similar procedure as the test-target scoring. The background set consisted of 60 SRE03 and 40 Fisher conversations from unique speakers. The test-target Hit Score was then divided by the test-background Hit Score.

Finally, ZNORM was applied to normalize the scores. ZNORM was scored in the typical method using 35 SRE03 conversations. TNORM was not performed due to time and computational constraints.

Table 5 shows the results of the SNP system on all English trials of SRE04 alone and in combination with the cepstral GMM system. As expected, the SNP system performs relatively better in the 8-side conversation training, where there is more data.

SNP	**1side Train**		**8side Train**	
	EER	**DCF**	**EER**	**DCF**
SNP	12.65%	0.5177	6.12%	0.3169
GMM	7.73%	0.3113	4.96%	0.2115
SNP+GMM	7.10% (8%)	0.2943 (6%)	4.37% (12%)	0.1777 (16%)

Table 5. The performance of the SNP system alone and in combination with cepstral GMM. Results are on all English trials of SRE04 evaluation set.

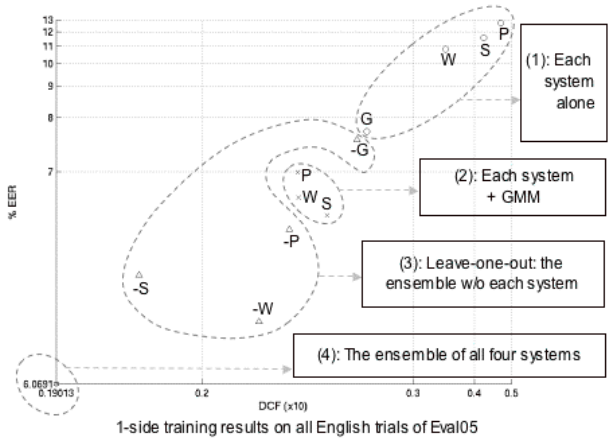
4.4. Cepstral GMM System

The cepstral Gaussian mixture model (GMM) was developed by our collaborators at SRI. The data was bandlimited to between 300-3300 Hz. 19 mel filters were used to compute 13 cepstral coefficients (C1-C13), and their delta, double delta, and triple-delta coefficients, producing a 52 dimensional feature vector. CMS was applied. The number of components of the GMM was chosen to be 2048. The background GMM was trained using data from

Fisher, SRE02, and SRE03. For channel normalization, feature mapping [17] was applied using gender- and handset-dependent models that were adapted from the background model. The resulting features were mean and variance normalized over the utterance. Speaker models were adapted from the background GMM using MAP adaptation of the means of the Gaussian components. Verification was performed using the 5-best Gaussian components per frame selected with respect to the background model scores. T-normalization, where the models were constructed from the Fisher database, was applied to the final scores. Performance of this system on all English trials of SRE04 is reported in Tables 3-5.

5. SRE05 RESULTS & ANALYSIS

The constellation plots [10] in Figures 3 and 4 show the performance (EER vs. DCF) of various system combinations on all English trials of SRE05 for the 1-side and 8-side training conditions, respectively. The first group on the top right (group 1) denotes the performance of each system alone. The second group denotes the performance of each system combined with the GMM. The third group is the combination of all systems minus the noted system (e.g., "-G" is the system which excludes GMM, and is the combination of WordHMM, PhoneNgram, and SNP). This group of performance points indicate the significance of the contribution of a paraticular sub-system to the ensemble: the further a point is from the origin, the more significant the contribution of the sub-system. Finally, the point at the origin (group 4) is the combination of all four systems.

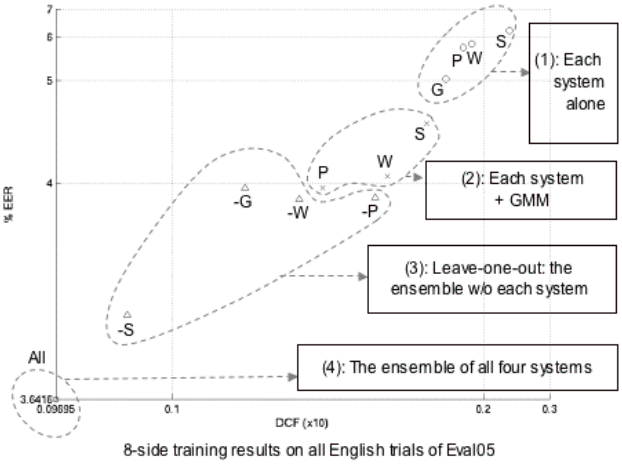


1-side training results on all English trials of Eval05

Fig. 3. Constellation plot with results of all the systems for 1-side Training condition on all English trials of SRE05. The labels on the graph are: (W: WordHMM), (P: PhoneNgram), (S: SNP), and (G: GMM).

Many observations can be made from these results. Common to the 1-side and 8-side training case, we see that the cepstral GMM is the best stand-alone system. Examining groups 2, we see that the combination of any system with the GMM produces a better system than the GMM (or any other system) alone. Also, it is reassuring to note that every system contributes to improving the final four-way combination results.

There are some notable differences in the performance of the 1-side and 8-side training cases. One is that leaving the GMM out of the combination in the 1-side training case degrades the results drastically, whereas, this is not the case in the 8-side training case.



8-side training results on all English trials of Eval05

Fig. 4. Constellation plot with results of all the systems for 8-side Training condition on all English trials of SRE05.

In the latter case, the composite is hurt at least as much by removing the WordHMM or the PhoneNgram systems. It appears that in the condition with more training data, the other systems perform well in the GMM's absence by effectively utilizing long-term information.

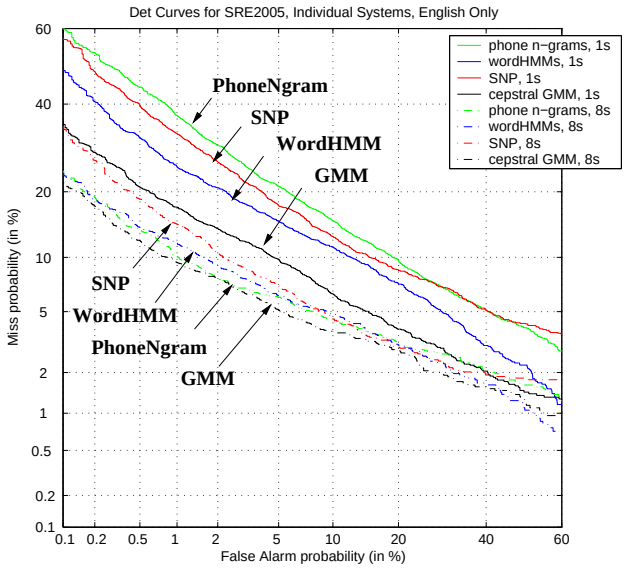


Fig. 5. The DET plot with results of each system alone.

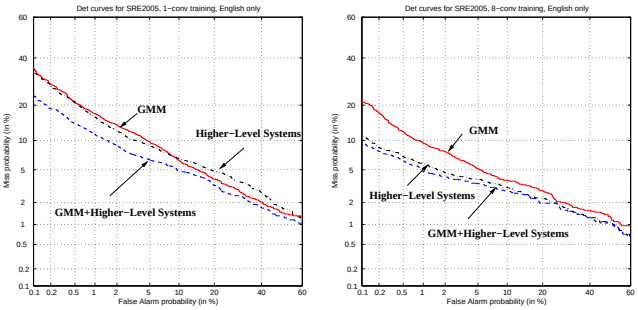
Figure 5 shows the DET curves for each system in groups 1. As we noted in Section 4.2, the performance of the PhoneNgram system improves more dramatically with increase of training data, from being ranked the worst system in the 1-side training condition to performing almost as well as the GMM system in the 8-side condition.

Table 6 shows the EER and DCF and Figure 6 shows the DET curves of the GMM system, the fusion of the three "higher-level" systems, and all four systems combined. We note that for the 1-side training condition, the combination of higher-level systems performs as well as the bag-of-frames cepstral GMM system

SRE05	1side Train		8side Train	
Systems	EER	DCF	EER	DCF
GMM	7.68%	0.2572	5.03%	0.1668
HighLev Sys	7.51%	0.2503	3.98%	0.1049
GMM+HighLev Sys	6.07%	0.1901	3.64%	0.0970

Table 6. Comparing the contribution of the cepstral GMM system with the combination of the three systems which model higher-level features ("HighLev Sys"). Results are reported on all English trials of SRE05 evaluation set.

alone. For the 8-side training condition, the combination of the three higher-level systems significantly outperforms the cepstral GMM stand-alone system. In both cases, the addition of the GMM system to the combination improves the results.



(a) 1-side Train (b) 8-side Train

Fig. 6. DET plots comparing the performance of GMM, the fusion of the "higher-level" systems, as well as the combination of all four systems.

6. CONCLUSIONS

In this paper we described our speaker recognition system developed for the NIST Speaker Recognition Evaluation 2005. This system was made up of four sub-systems, three of which were aiming to take advantage of long-term features, and the fourth was a traditional cepstral GMM system. Results were presented on NIST SRE04 and SRE05 evaluation sets.

The strength of our systems could be characterized as follows: the WordHMM capitalized on the finer-grained modeling of a text-dependent strategy in a text-independent domain. The PhoneNgram system took advantage of improved statistics in a lattice phone decoding approach to model what may resemble the phonetic pronunciation variations of the speaker. The SNP system used a non-parametric approach to compare test and training segments directly, while heavily biased towards positive evidence. Finally, the SRI GMM system is a state-of-the-art cepstral system. We observed that, as expected, the three "higher-level" systems perform particularly well in the training condition with more data. These relatively complementary strategies and systems combined well together to create one of the top performing systems at the 2005 NIST Speaker Recognition evaluation.

7. ACKNOWLEDGMENTS

We would like to thank our collaborators at SRI, especially Sachin Kajarekar and Andreas Stolcke, for providing us with the cepstral GMM system scores, ASR system outputs, assistance in building the phoneNgram system, and countless useful discussions. This work has also benefited greatly from the advice of George Doddington.

8. REFERENCES

[1] W. Andrews, M. Kohler, J. Campbell, J. Godfrey, and J. Hernandez-Cordero. Gender-dependent phonetic refraction for speaker recognition. In *ICASSP*, volume 1, pages 149–152, 2002.

[2] R. Auckenthaler, M. Carey, and H. Lloyd-Thomas. Score normalization for text-independent speaker verification systems. In *Digital Signal Processing*, volume 10, pages 42–54, 2000.

[3] K. Boakye and B. Peskin. Text-constrained speaker recognition on a text-independent task. In *Proc. Odyssey Speaker Recognition Workshop*, pages 129–34, Toledo, Spain, 2004.

[4] W.M. Campbell, J.P. Campbell, D.A. Reynolds, D.A. Jones, and T.R. Leek. Phonetic speaker recognition with support vector machines. In *NIPS*, volume 15, 2003.

[5] Cambridge University Engineering Department. The Hidden Markov Model Toolkit (HTK). http://htk.eng.cam.ac.uk/.

[6] D. Gillick, S. Stafford, and B. Peskin. Speaker detection without models. In *ICASSP*, pages I–757, 2005.

[7] A. Hatch, B. Peskin, and A. Stolcke. Improved phonetic speaker recognition using lattice decoding. In *ICASSP*, 2005.

[8] T. Joachims. Text categorization with support vector machines: Learning with many relevant features. In *European Conference on Machine Learning*, 1998.

[9] S. Kajarekar. Four weightings and a fusion: a cepstral-SVM system for speaker recognition. In *ASRU*, Cancun, Mexico, 2005.

[10] S. Kajarekar, L. Ferrer, E. Shriberg, K. Sonmez, A. Stolcke, and A. Venkataraman. SRI NIST 2005 speaker recognition evaluation system. Presented at the NIST Speaker Recognition Evaluation Workshop, Montreal, Canada, 2005.

[11] S. Kajarekar, L. Ferrer, E. Shriberg, K. Sonmez, A. Stolcke, A. Venkataraman, and J. Zheng. SRI's 2004 NIST speaker recognition evaluation system. In *ICASSP*, volume 1, pages 173–176, 2005.

[12] MIT Lincoln Labs. LNKNet. http://www.ll.mit.edu/IST/lnknet.

[13] A.F. Martin, D. Miller, M.A. Przybocki, J.P. Campbell, and H. Nakasone. Conversational telephone speech corpus collection for the NIST Speaker Recognition Evaluation 2004. In *4th International Conference on Language Resources and Evaluation*, pages 587–590, Lisbon, Portugal, 2004.

[14] National Institute of Standards and Technology. The NIST year 2003 speaker recognition evaluation plan. http://www.nist.gov/speech/tests/spk/2003/doc/2003-spkrec-evalplan-v2.2.pdf, 2003.

[15] National Institute of Standards and Technology. The NIST year 2005 speaker recognition evaluation plan. http://www.nist.gov/speech/tests/spk/2005/sre-05_evalplan-v6.pdf, 2005.

[16] D.A. Reynolds. Speaker identification and verification using Gaussian mixture speaker models. *Speech Communication*, 17:91–108, 1995.

[17] D.A. Reynolds. Channel robust speaker verification via feature mapping. In *ICASSP*, 2003.

[18] D.A. Reynolds et al. The SuperSID project: Exploiting high-level information for high-accuracy speaker recognition. In *Proc. IEEE ICASSP*, Hong Kong, 2003.