

THE CONTRIBUTION OF CEPSTRAL AND STYLISTIC FEATURES TO SRI'S 2005 NIST SPEAKER RECOGNITION EVALUATION SYSTEM

Luciana Ferrer^{1,2} Elizabeth Shriberg^{1,3} Sachin S. Kajarekar¹ Andreas Stolcke^{1,3}
Kemal Sönmez¹ Anand Venkataraman¹ Harry Bratt¹

¹ SRI International, Menlo Park, CA, USA

² Department of Electrical Engineering, Stanford University, Stanford, CA, USA

³ International Computer Science Institute, Berkeley, CA, USA

ABSTRACT

Recent work in speaker recognition has demonstrated the advantage of modeling stylistic features in addition to traditional cepstral features, but to date there has been little study of the relative contributions of these different feature types to a state-of-the-art system. In this paper we provide such an analysis, based on SRI's submission to the NIST 2005 Speaker Recognition Evaluation. The system consists of 7 subsystems (3 cepstral, 4 stylistic). By running independent N -way subsystem combinations for increasing values of N , we find that (1) a monotonic pattern in the choice of the best N systems allows for the inference of subsystem importance; (2) the ordering of subsystems alternates between cepstral and stylistic; (3) syllable-based prosodic features are the strongest stylistic features, and (4) overall subsystem ordering depends crucially on the amount of training data (1 versus 8 conversation sides). Improvements over the baseline cepstral system, when all systems are combined, range from 47% to 67%, with larger improvements for the 8-side condition. These results provide direct evidence of the complementary contributions of cepstral and stylistic features to speaker discrimination.

1. INTRODUCTION

Automatic speaker recognition is the task of identifying a speaker based on his or her voice. Conventional systems for this task use features extracted from very short time segments of speech, and model spectral information using Gaussian mixture models (GMMs) [1]. This approach, while successful in matched acoustic conditions, suffers significant performance degradation in the presence of handset mismatch or ambient noise. Furthermore, since spectral information is not modeled as a sequence, short-term cepstral modeling fails to capture longer-range stylistic aspects of a person's speaking behavior, such as lexical, rhythmic, and intonational patterns. Recently, it has been shown that systems based on longer-range stylistic features provide significant complementary information to the conventional system [2, 3]. In addition, modeling of spectral information by GMMs can be improved or complemented by the use of other modeling techniques like support vector machines (SVMs) [4,5], or by transformations of the cepstral space [6].

The National Institute of Standards in Technology (NIST) conducts annual speaker recognition evaluations (SREs) to allow

for meaningful comparisons of different approaches and to assess their performance relative to state-of-the-art systems. In this paper, we describe SRI's submission to the 2005 SRE. The system uses a number of novel long-range features, as well as new approaches to short-term cepstral modeling, and has achieved outstanding results in the evaluation. The main focus of this paper, besides describing the submitted system, is on the analysis of the relative importance of the cepstral and stylistic subsystems we have developed. This is essential for the understanding of the source of the achieved improvements in performance with respect to the baseline cepstral GMM system and for guiding future research.

The remainder of the paper is organized as follows. Section 2 briefly describes the evaluation setup, the development datasets and the speech recognition system used. Sections 3 and 4 summarize the subsystems included in our submission and the methods used to combine them. Section 5 presents results and an analysis of subsystem contributions. Final conclusions are given in Section 6.

2. BASIC SETUP

The 2005 NIST SRE dataset (referred to as SRE05) is part of the conversational speech data recorded in the Mixer project. The data contains mostly English speech and was recorded over telephone (landline and cellular) channels. The evaluation consists of twenty main conditions differing in the amount of available training and test data, and in the recording conditions [7]. The core condition, for which all evaluation participants are required to submit results, allows one side of a telephone conversation for training and another side for testing. The common condition is defined as the subset of trials for any of the main conditions for which all train and test conversations were spoken in English using handheld phones. We submitted results for the (1-side train, 1-side test) and (8-side train, 1-side test) conditions. The common condition subset for these conditions consisted of 20,907 and 15,947 trials respectively.

The main performance metric in the NIST SRE is the detection cost function (DCF), defined as the Bayes risk with $P_{target} = 0.01$, $C_{fa}=1$, and $C_{miss}=10$ [7]. In this paper, results are presented in terms of the minimum value of the DCF measure over all possible score thresholds and the equal error rate (EER), for the trials corresponding to the common condition.

The component subsystems and combiners described in this paper were developed using three different data sets:

Switchboard, Fisher, and SRE04 (last year’s evaluation set). The Fisher database consists of three disjoint sets: a background set and two test splits, Fisher1 and Fisher2. The set of Fisher trials was designed to be similar to the SRE04 set in the proportion of impostor to true-speaker trials and other characteristics. For a description of these databases, see [8]. The background data from Switchboard and Fisher (or, in some cases, a subset of it) was used for creation of the background models for all component systems described in Section 3. A set of 249 1-side speaker models balanced by gender and handset type, extracted from Fisher2 data, was used to compute TNORM scores [9]. SRE04, Fisher and SRE05 data were processed exactly the same way, using the same background models and TNORM speakers. The scores from Fisher and SRE04 were used for development.

Transcriptions were generated with SRI’s 3xRT CTS recognition system. The final pass of the recognition, with 21% word error rate (WER) on Fisher data, was used for all the component systems that needed transcriptions, except for the state duration system for which we used an earlier stage of the recognition (with 29% WER). This is because we found that for state duration features the less constrained automatic speech recognition (ASR) output leads to better speaker recognition performance. See [8] for more details on the ASR system used.

3. COMPONENT SUBSYSTEMS

Our submission consisted of the combined scores from seven subsystems. Three of those systems are based on cepstral features, while the other four aim to model longer-term stylistic features. Table 1 shows the performance of each of these systems for SRE05. All results shown are after TNORM.

Cepstral GMM system: This is a conventional cepstral Gaussian mixture model system adapted from a universal background model, using a 2048-component GMM. Its details are described in [8].

Cepstral SVM system: This system uses multiple projections of PCA-transformations of mean polynomial vectors over cepstral features (two of which are variance normalized). These features are modeled using SVMs, generating four separate scores that are combined with equal weight to produce the final score [5].

MLLR transform SVM system: This system uses as features the components of the maximum likelihood linear regression (MLLR) transforms used in SRI’s speech recognition system for speaker adaptation. The transform coefficients are modeled by SVMs [6].

Word N-gram SVM system: This system uses an SVM with a linear kernel with first-, second-, and third-order word N-gram frequencies as features [8].

SNERF system: This system uses a set of prosodic features extracted over automatically estimated syllables. The modeling is done by SVMs [10]. Apart from the features described in [10], a set of word-specific SNERFs were added to the vector of features. The addition of these new features produced an improvement of around 10% relative to the original SNERF system on the development sets, even though the new features are much more sparse (comprise only 24 unique words) and are already modeled indirectly in the overall SNERF system.

Table 1: Performance of components systems for SRE05 (DCF refers to minimum DCF).

System	Short name	1-side training		8-side training	
		DCFx100	%EER	DCFx100	%EER
Cepstral GMM	CepGm	2.48	7.17	1.69	4.91
MLLR SVM	CepMI	2.52	10.34	1.20	5.50
Cepstral SVM	CepSv	2.68	7.26	1.03	3.05
SNERF	StySn	5.22	14.06	2.75	6.52
State Dur	StySd	6.03	15.36	3.19	8.02
Word Dur	StyWd	7.83	19.23	3.74	8.62
Word N-gram	StyWn	8.60	24.58	4.84	11.25

Duration systems: Two sets of duration features – state- and word-level – modeled by GMMs are used in this system [11]. The phone-level durations were not used in this evaluation since they were found to be redundant in most cases given the other available systems.

With respect to last year’s evaluation, the cepstral SVM and the MLLR transform SVM systems are new additions. Also, the SNERF system has been improved significantly by the inclusion of the word-dependent features. Several of the systems we used last year were not used in this year’s system because we found that they did not significantly improve the performance on the SRE04 data.

4. SYSTEM COMBINATION METHODS

Three different combination methods were used for merging the scores produced by the subsystems into a single score. In all three cases, the scores from the subsystems are first normalized to have zero mean and unit variance with respect to the statistics in the set used for training the combiner.

Neural network (NN) combiner: The baseline combiner is a single-layer feed-forward network that uses a sigmoid output node during training and a linear output for the final predictions. The linear output allows better combination of these predictions with the ones from the other two combiners. The perceptron is trained to achieve minimum squared error with output labels 0 (impostor) and 1 (target). Target and impostor priors are set to 0.09 and 0.91 during training in order to optimize the DCF.

SVM combiner: Three SVMs with polynomial kernels of orders 1, 2, and 3 are trained with equal penalty for false acceptance and false rejection. The three scores obtained are averaged to produce the final score.

Class-dependent (CD) combiner: This combiner relies on clustering both the target models and the test utterances in a vector space defined by the MLLR features computed for the speaker during ASR [12]. For each class in the product set, that is, (target, test) pair of clusters, we allocate a separate combiner trained to fit the data in that class. During testing, a weighted average of the scores given by each of those combiners is used as the final score. The weights for averaging are given by the probability of the trial of belonging to each of the classes.

The combiner for the 2005 system was trained using the scores obtained for SRE04 data, which we believed to be a reasonable match to the SRE05 data. This proved to be a good choice: the EER for the NN combiner on SRE05 is around 5% better when the combiner is trained on SRE04 data than when it is trained on Fisher1 data.

The final score submitted for the evaluation was computed as the average of the scores given by the three combiners described above. Scores from the three combiners are previously normalized to zero mean and unit variance using the training set statistics. This allows for the usage of equal weights in the final sum. A few different weights for the three combiners were tried during development, but a simple average proved to be the more robust choice. This final combiner will be called NSC (NN+SVM+CD) combiner in the following sections.

5. RESULTS AND ANALYSIS

With all these systems available for combination and various different ways of combining them, several questions arise: Which systems are more important for the combination? Can we ignore some of them without losing accuracy? Does the importance of the systems depend on the amount of training data or on the combiner approach? In this section we will try to give answers to these questions.

Table 2 shows combination results for some meaningful subsets of systems. The first line corresponds to the cepstral GMM system alone. This system is the conventional speaker recognition system and is commonly used as the baseline against which new systems are compared. The second line shows the combination results of that system with the two novel cepstral systems, the cepstral SVM and the MLLR SVM. The combined system achieves an improvement in the DCF of 33% for the 1-side condition and 53% for the 8-side condition. Similar improvements are obtained when combining the baseline with the four stylistic systems: word N-gram, SNERF and both duration systems. Finally, when all systems are combined, the relative improvement over the baseline alone is 47% in the 1-side condition and 67% in the 8-side condition. Clearly, the benefit of the new systems, both cepstral and stylistic, increases as more data is available for training.

Table 2: Performance for the cepstral GMM (baseline) and the combination of that system with the rest of the cepstral systems, the stylistic systems and all system together.

Systems being combined	1-side training		8-side training	
	DCF _{x100}	%EER	DCF _{x100}	%EER
Baseline	2.48	7.17	1.69	4.91
Baseline + new cepstral	1.66	4.61	0.80	2.45
Baseline + stylistic	1.77	4.89	0.83	2.45
All systems combined	1.31	4.10	0.56	2.03

Tables 3 and 4 show the best combination results when we allow a fixed number of systems for the NN and the NSC combiners for both training conditions. Each pair of lines in these figures shows which systems lead to the best performance

for each combiner when N systems are allowed. There is only one line for the best 1-way because that choice is obviously independent of the combiner. Note first that the DCF obtained when we use the NSC combiner is always better than that with the NN combiner, for any number of allowed systems. Furthermore, except for the 7-way case, adding a new system always improves the performance for the NSC combiner (this is not always true for the NN combiner, as can be seen by comparing the 5-way with the 6-way results in Table 3). After the sixth system is added, though, we observe no improvement by adding the final system. In fact, in the 8-side condition the performance is significantly hurt by adding the state duration system to the combination. This indicates two things: first, the state duration system is most probably redundant once the other systems are being used, and second, our combiners are not able to handle redundant features well. Ideally, we should be able to detect these cases and ignore those systems that are not needed. To this end, further research on system selection and more robust combiners is needed.

Tables 3 and 4 also offer a great opportunity for analyzing the importance of the systems. Even though the best N systems for each value of N are chosen independently so as to optimize the performance for that number of systems, for the NSC combiner the subsets of systems chosen for a certain N always includes the subset chosen for $N-1$ systems. This was a remarkable finding. There is nothing forcing, say, the best 2-way combination to include the single best system, rather than two other systems that, when combined, give better performance than the best system alone. But given that the results turned out this way, we can very easily rank the importance of the seven systems by looking at the order in which they are being added as we allow more systems in the combination.

From the tables we see that the order in which systems are chosen is highly dependent on the amount of training data. In Table 1 we can see that the performance of the subsystems is, without exception, relatively closer to the baseline in the 8-side case than in the 1-side case. For example, the EER of the SNERF system is twice that of the baseline for the 1-side case, while it is only 30% worse for the 8-side case. This explains the bigger relative improvement obtained from combining the baseline with the other systems for the 8-side condition than the 1-side condition (Table 2), and it also explains the difference in the order in which the systems are added for those two conditions in Tables 3 and 4. Both the SNERF and the Word-Ngram systems are added earlier in the 8-side condition than in the 1-side condition where they have a worse performance relative to the baseline. Overall we see that both factors, the performance of the system with respect to the baseline and the amount of new information the system conveys about the speakers, affect which system is chosen next. This qualitative observation accounts for the alternated way stylistic and cepstral systems are added to the combination.

We are also interested in knowing how much improvement we have achieved since last year. For this, we compare the performance of last year's system and this year's system when run on the English-only SRE04 subset. (Note that these results do not agree with those in [8] because the latter corresponded to the common condition trials. We are presenting all-English results here because the number of trials is larger, allowing for more significant comparisons.) The two systems differ in many aspects: the background data this year includes both Fisher and

Switchboard data, while last year's included only Fisher data, new systems have been added, some old ones have been discarded, and some have been improved. Table 5 compares the results when the combiners are trained on Fisher1 data; we can see overall improvements between 25% and 43% in both DCF and EER.

Table 3: Best possible N-way combinations for the NN and the NSC combiners for the 1-side training condition. System names refer to those defined in Table 1.

N	Combiner	Cep Gm	Cep MI	Sty Sn	Sty Wd	Cep Sv	Sty Wn	Sty Sd	DCF x100
1	-								2.47
2	NN								1.98
	NSC								1.77
3	NN								1.67
	NSC								1.58
4	NN								1.58
	NSC								1.44
5	NN								1.49
	NSC								1.33
6	NN								1.60
	NSC								1.31
7	NN								1.47
	NSC								1.31

Table 4: Same as Table 3 but for 8-side training condition.

N	Combiner	Cep Sv	Sty Sn	Cep MI	Sty Wn	Sty Wd	Cep Gm	Sty Sd	DCF x100
1	-								1.03
2	NN								0.75
	NSC								0.74
3	NN								0.66
	NSC								0.64
4	NN								0.61
	NSC								0.58
5	NN								0.59
	NSC								0.55
6	NN								0.59
	NSC								0.54
7	NN								0.60
	NSC								0.56

Table 5: Performance on SRE04 data using 2004 and 2005 systems with NN and NSC combiners trained on Fisher data.

System/Combiner used	1-side training		8-side training	
	DCFx100	%EER	DCFx100	%EER
SRE04/NN	3.15	7.73	1.60	3.50
SRE05/NN	2.20	5.27	0.91	2.91
SRE05/NSC	2.18	4.85	0.91	2.62
Rel. improvement	31%	37%	43%	25%

6. SUMMARY AND CONCLUSIONS

We have described our submission to the 2005 NIST Speaker Recognition Evaluation. Results show a relative improvement over last year's performance (on last year's data) of more than 25% relative. This improvement was achieved mainly by the introduction of two new cepstral systems, the improvement of the SNERF system and the use of a new class-dependent combination method. We have focused our analysis of results on the relative importance of the cepstral and stylistic systems being combined. It was found that improvements over the baseline cepstral system when combining all subsystems range from 47% to 67%, with larger improvements for the 8-side condition. This justifies and encourages the development of those nonstandard systems that utilize prosodic or lexical features, or which model the spectral features in a manner different from GMMs. Analysis of the order in which systems are chosen for the best combination of increasing number of systems shows an alternate pattern of stylistic and cepstral features, with higher priority for the stylistic features when more training data is available.

7. ACKNOWLEDGMENTS

This work was funded by NSF IIS-9619921 and IIS-0544682. The views herein are those of the authors and do not reflect the views of the funding agencies. We thank our colleagues at ICSI for fruitful discussions.

8. REFERENCES

- [1] D. Reynolds, T. Quatieri, and R. Dunn, "Speaker Verification Using Adapted Mixture Models," *Digital Signal Processing*, vol. 10, pp.181-202 (2000).
- [2] D. Reynolds et al., "The SuperSID Project: Exploiting High-level Information for High-accuracy Speaker Recognition," *Proc. IEEE ICASSP*, Hong Kong, pp. 784-787, 2003.
- [3] G. Doddington, "Speaker Recognition Based on Idiolectal Differences Between Speakers," *Proc. Eurospeech*, Aalborg, pp. 2521-2524, 2001.
- [4] W. Campbell, J. Campbell, D. Reynolds, D. Jones, and T. Leek, "High-Level Speaker Verification with Support Vector Machines," *Proc. IEEE ICASSP*, Montreal, 2004.
- [5] S. Kajarekar, "Four Weightings and a Fusion: A Cepstral SVM System for Speaker Recognition," *Proc. IEEE ASRU Workshop*, Cancun, 2005.
- [6] A. Stolcke, L. Ferrer, S. Kajarekar, E. Shriberg, and A. Venkataraman, "MLLR Transforms as Features in Speaker Recognition," *Proc. Eurospeech*, Lisbon, pp. 2425-2428, 2005.
- [7] NIST 2005 Speaker Recognition Evaluation plan, http://www.nist.gov/speech/tests/spk/2005/sre-05_evalplan-v6.pdf.
- [8] S. Kajarekar, L. Ferrer, E. Shriberg, K. Sonmez, A. Stolcke, A. Venkataraman, and J. Zheng, "SRI's 2004 NIST Speaker Recognition Evaluation System," *Proc. IEEE ICASSP*, Philadelphia, pp. 173-176, 2005.
- [9] R. Auckenthaler, E. Parris, and M. Carey, "Improving a GMM Speaker Verification System by Phonetic Weighting," *Proc. IEEE ICASSP*, Phoenix, Arizona, 1999.
- [10] E. Shriberg, L. Ferrer, S. Kajarekar, A. Venkataraman, and A. Stolcke, "Modeling Prosodic Feature Sequences for Speaker Recognition," *Speech Communication* 46(3-4), 455-472, 2005.
- [11] L. Ferrer, H. Bratt, V. R. Gadde, S. Kajarekar, E. Shriberg, K. Sonmez, A. Stolcke, and A. Venkataraman, "Modeling Duration Patterns for Speaker Recognition," *Proc. Eurospeech*, Geneva, pp.2017-2020, September, 2003.
- [12] L. Ferrer, K. Sönmez, and S. Kajarekar, "Class-based Score Combination for Speaker Recognition," *Proc. Eurospeech*, Lisbon, 2005.