SRI'S 2004 NIST SPEAKER RECOGNITION EVALUATION SYSTEM

Sachin S. Kajarekar[1], Luciana Ferrer[1], Elizabeth Shriberg[1,2], Kemal Sonmez[1], Andreas Stolcke[1,2], Anand Venkataraman[1], Jing Zheng[1]

[1]SRI International, Menlo Park, CA, USA
[2]International Computer Science Institute, Berkeley, CA, USA

ABSTRACT

This paper describes our recent efforts in exploring longer-range features and their statistical modeling techniques for speaker recognition. In particular, we describe a system that uses discriminant features from cepstral coefficients, and systems that use discriminant models from word n-grams and syllable-based NERF n-grams. These systems together with a cepstral baseline system are evaluated on the 2004 NIST speaker recognition evaluation dataset. The effect of the development set is measured using two different datasets, one from Switchboard databases and another from the FISHER database. Results show that the difference between the development and evaluation sets affects the performance of the systems only when more training data is available. Results also show that systems using longer-range features combined with the baseline result in about a 31% improvement with 1-side training over the baseline system and about a 61% improvement with 8-side training over the baseline system.

1. INTRODUCTION

Automatic speaker recognition is the task of identiying 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 (GMM) [1]. This approach, while successful in matched acoustic conditions, suffers significant performance degradation in the presence of handset variability. Furthermore, since spectral information is not modeled as a sequence, the approach fails to capture longer-range stylistic features 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 speaker information to the conventional system [2, 3].

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 2004 SRE. This system uses a number of long-range features and has one of the lowest error rates of all official submissions on the NIST scoring metric for the core condition (1-conversation training, 1-conversation test). We also describe the effects of using different types of development sets and significant post-submission improvements. Performance is measured using equal error rate (EER) and decision cost function (DCF). EER assumes that false acceptance is as important as false rejection, and target and impostor priors are equal. DCF is a NIST-specific measure described in [4].

2. EVALUATION SETUP

The 2004 NIST SRE dataset (referred to as EVAL2004) is part of the conversation speech data recorded in the Mixer Project. The speech was mostly in English and was recorded over telephone (landline and cellular) channel. The evaluation consists of twenty-eight conditions differing in the amount of training and test data [4]. The core condition is defined as the one where one conversation side (about 5 min including silence) was used as training and test data. The primary evaluation subset is defined as that data recorded in English and over telephone channel with a handheld instrument (excluding cordless handsets).

We submitted results for two conditions – 1-side training, 1-side testing, and 8-side training, 1-side testing. The stylistic features are estimated using an automatic speech recognition (ASR) system, which is trained only on English language data. We submitted results for the primary evaluation subset. Table 1 shows the number of trials for each condition. In this paper, we report results for trials satisfying the primary evaluation criterion.

Table 1 Trials for different conditions in EVAL2004

Data length		English-English trials		Other trials	Total
Training	Test	**Primary**	Other		
1-side	1-side	**5202**	10596	10423	26224
8-side	1-side	**2433**	8851	5696	16980

3. DEVELOPMENT SETS

The systems described in this paper were developed using two different devsets: Switchboard (SWB) and FISHER.

3.1. SWB devset

The NIST 2003 evaluation dataset (EVAL2003) was divided into two halves – one set using splits 1-5 and a second set using splits 6-10. For the baseline cepstral system, the background model is trained using equal amount of landline (NIST 2001 extended evaluation data) and cell data (NIST 2001 cellular development data). The same background model is used for both

halves. For the other systems, the background model for one set is trained using the speakers from the other set. In each set, scores for each system are normalized using TNORM, and TNORM speakers are also obtained from the other set. From this data, we trained systems with 1 conversation side and 8 conversation sides.

The same background model from the baseline development system is used for the EVAL2004 set. For the other systems, the background model trained on split 6-10 is used. A combiner [5] trained on splits 1-5 is used to combine the scores from different evaluation systems. Different combiners are trained for 1 and 8 conversation side training conditions.

3.2. FISHER devset

The FISHER devset is created from the FISHER database, which is collected and distributed by the LDC for the DARPA EARS program. We selected two nonoverlapping sets of speakers from this data. In one set, speakers participated in only one recording; in the other set, speakers participated in multiple recordings. Each set is balanced with respect to different genders and handsets. The first set was used to create the background models. The second set was divided into two equal splits. For 1-side training, we created development sets that were used as devsets for the evaluation. Table 2 shows the statistics for target models and trials for the two splits.

As with SWB devset, two sets of development systems are trained on these splits. For each split, TNORM speakers from the other split are used. The system combination strategy is similar to that described in Section 3.1. The difference is that for the FISHER devset we have used a single combiner, trained with 1-conversation side, for both evaluation conditions.

Table 2 Model and trial statistics for FISHER devset

Split 1		Split 2	
Models	Trials	Models	Trials
734	14488	617	10665

For the EVAL2004 set, the background model and TNORM speakers are obtained from split 2, and the combiner is trained on split 1.

4. ACOUSTIC-FEATURE-BASED SYSTEMS

We used two acoustic features sets derived from 13 Mel frequency cepstral coefficients (MFCCs) with delta and double-delta coefficients. These features were preprocessed via mean normalization and feature transformation [6].

4.1 Cepstral-GMM System (Baseline)

This system uses the 2048-component GMM and is described in detail in [7]. Table 3 shows baseline results for different devsets and the EVAL2004 set. It shows that the FISHER devset is more difficult than the SWBD devset, and the EVAL2004 set is the most difficult for the baseline system. It also shows that the improvement using more training conversation sides falls from 67% on the SWBD set to 41% on the EVAL2004 set. We observe that EVAL2004 baseline performance is similar across two devsets for 1-side training. For 8-side training, however, systems trained on the FISHER devset give significantly better performance than those trained on the SWB devset.

4.2 HLDA Cepstral-GMM System

The HLDA cepstral-GMM system applies an additional heteroscedastic linear discriminant analysis (HLDA) transform to the channel-normalized features, projecting the original 39-dimension feature onto 25 dimensions. The HLDA transform is estimated using the baseline background GMM as a reference model, with individual Gaussians as classes and modeling the 14 nuisance dimensions with a global Gaussian distribution [8]. The transformed features are used to train a new background GMM. The target model estimation and verification is performed as in the baseline system. On the devsets, this system performs slightly worse than the baseline system and shows improvement in combination with it. However, the improvement does not hold for the EVAL2004 set.

Table 3 Performance of Cepstral GMM system

Dataset		**%EER**/DCF	
		1-side	8-side
SWBD		**6.87** / 0.247	**2.25** / 0.092
FISHER		**8.06** / 0.321	-
EVAL2004	SWBD	**11.61** / 0.445	**9.09** / 0.337
	FISHER	**11.27** / 0.449	**6.54** / 0.234

5. STYLISTIC-FEATURE-BASED SYSTEMS

Stylistic features used in this work are based on pitch, pause, word usage, and other statistics estimated from the data. We describe first the ASR system and follow with a brief description of the individual systems.

5.1. ASR System for Decoding SWB Devset

The long-term, higher-level features used in the SWB devset are generated using the first pass of SRI's conversational telephone speech recognition (CTS) system [9]. The acoustic models are those developed for the NIST RT-03 CTS evaluation, trained on SWB1 and CallHome data. The language model was a bigram trained on those sources, as well as broadcast and web data. The word-level 1-best recognition output, as well as word-, phone-, and state-level time alignments, was then used in the speaker ID systems. For transcribed portions of SWB2 phase 2 made available for RT-03 (but not used in our speaker ID system), the word error rate (WER) is about 38%.

5.2. ASR System for FISHER and EVAL 2004 Data

For the FISHER and EVAL2004 data, transcriptions were generated with SRI's 5xRT CTS recognition system, using improved models developed for the NIST RT-03F evaluation. Additional training material was drawn from transcribed SWB2 phase 2 and 3 data; no Fisher data was used in training the ASR system. Two different versions of ASR hypotheses and alignments were produced and used for speaker modeling. The first one corresponds to the output of the first-pass bigram decoding, similar to the decoder used on the SWB devset. The second recognition pass incorporated more global constraints, such as a 4-gram language model, word-level duration models,

and self-supervised speaker-level acoustic adaptation The WER on RT-03 evaluation data was 29% and 21%, respectively, for the two passes.

5.3. Word N-gram Language Modeling System

Our N-gram-based language model systems are based on the technique first used by Doddington [3]. The vocabulary (set of bigrams) is drawn from the same set of conversation sides used to estimate the background model. The model is a simple bag of N-grams. The score is calculated as the difference between the log-likelihood of the trial with respect to the target and the background models.

Table 4 shows the performance of the word N-gram LM system (Row 1) and its combination with the baseline (Row 2). The system (Row 1) gives about a 12% improvement over the baseline on the FISHER devset. However, it does not improve over the baseline on 1-side training from the EVAL2004 set. Our preliminary investigation shows that this disparity is due to a mismatch in the vocabulary. However, on 8-side training from the EVAL2004 set, the word N-gram system gives about a 20% improvement over the baseline.

5.4 Word N-gram Support Vector Machine system

This system uses a support vector machine (SVM) with a linear kernel [10] to separate true and imposter speakers. A training or test conversation side provides a single feature vector of the raw relative frequencies of word N-grams. In contrast to the language model system, the relative frequencies are unsmoothed and unboosted. We used first-, second-, and third-order N-gram counts as features with a minimum cutoff of 2. The bias against false positives was set to 500.

Table 4 shows the performance of the N-gram SVM system (Row 2). Results show that this system gives significantly better performance than the N-gram LM system (Section 5.3). This improvement is also observed when the SVM system is combined with the baseline (Row 4). Thus, the SVM-based approach is more efficient than the LM approach in modeling speaker information from N-grams.

Table 4 %EER of LM and SVM N-gram systems

System	FISHER	EVAL 1-side	EVAL 8-side
1) LM	18.07	27.81	16.36
2) SVM	14.68	23.06	12.36
3) Baseline + 1	7.08	11.44	5.09
4) Baseline + 2	6.62	10.03	3.27

5.5 Duration Feature System

Three duration features – state, phone, and word level – are used in this system [7]. Phone and state features are obtained using the noncrossword alignments from the recognizer, while the word features are obtained using both the crossword and the noncrossword versions. (Note: the crossword version is available only for the FISHER-dev systems.) This choice is based on previous results that showed better performance for both phone and state models using noncrossword alignments, and better performance for word models using the crossword alignments.

Table 5 shows the performance of different duration systems and of their combination with the baseline system. Of the three duration systems, the system using durations of the three states within a phone gives the best performance. The combination of duration systems with the baseline system gives about a 27% improvement with 1-side training and about a 44% improvement with 8-side training from the EVAL2004 set.

Table 5 %EER of different duration systems

System	FISHER	EVAL 1-side	EVAL 8-side
1) State (S)	12.65	15.14	9.09
2) Phone (P)	18.07	19.01	13.45
3) Word (W)	20.48	21.30	10.18
4) Baseline + 1 + 2 + 3	6.47	8.27	3.63

5.6 Pause-to-pause Feature System

This system includes a sample subtype of a large set of nonuniform extraction region features (NERFs) [11]. This particular system uses regions between pauses of 500 ms or more. A feature vector comprising various F0, energy, and duration features is extracted for each region. Features are modeled using GMMs. Due to the undefined features present, statistical modeling is modified as explained in [11].

Table 6 shows the performance of the pause-to-pause NERF system alone (Row 1) and in combination with the baseline system (Row 3). The performance of this system degrades significantly from the devset to the EVAL2004 set and also gives a smaller improvement when combined with the baseline system. Our primary investigation shows that the degradation in performance is related to the pitch features, and we are investigating this further.

5.7 Syllable NERFs

This system uses a type of NERFs where the extraction region is defined by automatically estimated syllable boundaries. These features are described in detail in [12]. For each syllable in the utterance, several duration-based, F0, and energy features are extracted. These features are then quantized to create sequences of different lengths (unigrams, bigrams and trigrams). Features can also include quantized pauses. The final features are then computed as the N-gram counts for a particular sequence normalized by the total number of syllables in that conversation side. These features are modeled using SVMs as for the word N-gram SVM system.

Table 6 %EER of pause-to-pause and syllable NERFs

System	FISHER	EVAL 1-side	EVAL 8-side
1) PAU	16.04	27.46	22.56
2) SYL	14.60	20.10	12.00
3) Baseline + 1	7.00	10.74	6.54
4) Baseline + 2	7.30	8.98	4.00

Table 6 shows the performance of the SNERF-based system (Row 2) and its combination with the baseline system (Row 4). This system gives a significant improvement – around 20% for 1-side training and around 40% for 8-side training. These features are similar to duration features and might have similar speaker information. However, our experiments show that a system using these features provides significant improvements when combined with both the duration and baseline systems.

6. SYSTEM COMBINATION RESULTS

Twelve individual systems are used for score-level combination. These include two acoustic-feature-based systems and seven stylistic-feature-based systems using either noncrossword (NC) or crossword (C) alignments. As mentioned earlier, these systems are combined at the score level, using the Neural Network classifier [5]. Table 7 shows results for different system combinations. Combinations 1 and 2 compare the effect of devsets on the combination of acoustic-feature-based systems (baseline and HLDA) with duration- and N-gram-based systems. Results show that the choice of devsets does not have a significant effect with 1-side training. With 8-side training, however, systems trained on the FISHER devset perform significantly better than those trained on the SWB devset. This result is consistent with the baseline result (Table 3). This combined performance is further improved by adding NERF systems. Combination 3 shows about a 31% improvement with 1-side training and about a 61% improvement with 8-side training over the baseline.

7. SUMMARY AND CONCLUSIONS

We have described twelve different speaker recognition systems. Two of them are based on commonly used cepstral features, and others are based on stylistic features. Apart from baseline and duration systems, we used three new systems: HLDA cepstrum, SVM N-grams, and SNERFs. These systems were developed on two development sets from the Switchboard and FISHER databases. The HLDA system gave significant improvements in combination with the baseline on the devset but did not improve performance on the EVAL2004 set. We described an SVM framework for modeling N-grams, which performed better than the LM framework. Finally the SNERF system also gave a significant improvement in combination with the baseline.

We extended the NERF system by modeling the undefined NERFs efficiently in a single model. This system did not give significant improvement in combination with the baseline.

Our experiments with different alignments (crossword and noncrossword) for stylistic features showed that these systems give better performance with less accurate transcription. This shows that speaker-specific errors in the transcription made by the ASR system are useful for stylistic features.

Some open issues remain regarding exact differences between the devset and EVAL 2004 set, and the generalizability of a system combination. However, the significant improvements from stylistic features – 31% with 1-side training and 61% with 8-side training – show the importance of stylistic features for robust speaker recognition performance.

8. ACKNOWLEDGMENT

This work was funded by a DoD KDD award via NSF IRI-9619921. The views herein are those of the authors and do not reflect the views of the funding agencies.

9. 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] 2001 JHU summer workshop report, "SuperSID: Exploiting High-level Information for High-performance Speaker Recognition", http://www.clsp.jhu.edu/ws2002/groups/supersid/supersid-final.pdf

[3] G. Doddington, "Some Experiments on Ideolectal Differences Among Speakers", http://www.nist.gov/speech/tests/spk/2001/doc/ , 2001.

[4] NIST 2004 Speaker Recognition Evaluation plan, http://www.nist.gov/speech/tests/spk/2004/SRE-04_evalplan-v1a.pdf.

[5] MIT Lincoln Laboratory, LNKNet, http://www.ll.mit.edu/IST/lnknet/

[6] D. A. Reynolds, "Channel Robust Speaker Verification via Channel Mapping", *Proc. IEEE ICASSP*, vol. 2, pp. 53-56, Hong Kong, 2003.

[7] 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.

[8] M.J.F. Gales, "Semi-Tied Covariance Matrices for Hidden Markov Models", *IEEE Transactions on Speech and Audio Processing*, May 1999.

[9] A. Stolcke, H. Bratt, J. Butzberger, H. Franco, V. R. Rao Gadde, M. Plauche, C. Richey, E. Shriberg, K. Sonmez, F. Weng, and J. Zheng, "The SRI March 2000 Hub-5 Conversational Speech Transcription System", *Proc. NIST Speech Transcription Workshop*, College Park, MD, 2000.

[10] V. Vapnik, "The Nature of Statistical Learning Theory", Springer, 1995.

[11] S. Kajarekar, L. Ferrer, K. Sonmez, J. Zheng, E. Shriberg, and A. Stolcke, "Modeling NERFs for Speaker Recognition", *Proc. Odyssey 2004: The Speaker and Language Recognition Workshop*, Toledo, Spain, 2004.

[12] E. Shriberg, L. Ferrer, A. Venkataraman, and S. Kajarekar, "SVM Modeling of SNERF-Grams for Speaker Recognition", *Proc. ICSLP*, South Korea, September, 2004.

Table 7 Combination of stylistic-feature-based systems with two acoustic-feature-based (baseline and HLDA) systems. NC=noncrossword alignments, C=crossword alignments and * = systems without TNORM.

Combination (devset)	(Acoustic +) Stylistic							Results (%EER/DCF)	
	Duration			Word-based		NERFs		EVAL 1-side	EVAL 8-sides
	S	P	W	LM	SVM	PAU	SYL		
1 (SWB)	NC	NC	NC	NC	NC*			8.2 / 0.358	5.4 / 0.226
2 (FISHER)	NC	NC	NC	NC	NC*			8.2 / 0.346	2.9 / 0.119
3 (FISHER)	NC	NC	C	NC & C	NC & C	C	C	**7.7 / 0.325**	**2.5 / 0.100**