SPEAKER RECOGNITION USING PROSODIC AND LEXICAL FEATURES

Sachin Kajarekar[1], *Luciana Ferrer*[1], *Anand Venkataraman*[1], *Kemal Sonmez*[1], *Elizabeth Shriberg*[1,2], *Andreas Stolcke*[1,2], *Harry Bratt*[1], *Ramana Rao Gadde*[1]

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

ABSTRACT

Conventional speaker recognition systems identify speakers by using spectral information from very short slices of speech. Such systems perform well (especially in quiet conditions), but fail to capture idiosyncratic longer-term patterns in a speaker's habitual speaking *style*, including duration and pausing patterns, intonation contours, and the use of particular phrases. We investigate the contribution of modeling such prosodic and lexical patterns, on performance in the NIST 2003 Speaker Recognition Evaluation extended data task. We report results for (1) systems based on individual feature types alone, (2) systems in combination with a state-of-the-art frame-based baseline system, and (3) an all-system combination. Our results show that certain longer-term stylistic features provide powerful complementary information to both frame-level cepstral features and to each other. Stylistic features thus significantly improve speaker recognition performance over conventional systems, and offer promise for a variety of intelligence and security applications.

1. INTRODUCTION

Speaker recognition systems based on short-term spectral features model a speaker's use of the resonances of his or her vocal tract as determined by its physical dimensions. In a conventional speaker recognition system, these features are modeled by a Gaussian Mixture Model (GMM) and scored with respect to a universal background model trained from many speakers [e.g., 1]. Such conventional systems result in good speaker detection performance under favorable acoustic conditions. Yet because these systems depend heavily on features affected by spectral variation, they degrade in noise or unmatched acoustic conditions. More generally, and perhaps more importantly, conventional approaches fail to capture habitual *stylistic* patterns in the way a person talks. For example, they do not directly capture prosodic variations in speaking rate, pausing, or intonation, because these patterns occur on a scale larger than the frame. Humans use such patterns to identify speakers, for example, when listening to a conversation through a wall (or, as a more entertaining example, to discern the targets of comedians' impersonations). Because such habits are behavioral, rather than physiologically based, they can provide complementary information to standard GMM cepstral features. Furthermore, because certain longer-range prosodic patterns (such as duration and pausing) are invariant to channel variation, they may offer robustness in the face of difficult or unmatched acoustic conditions.

In recent years, a number of approaches for using stylistic features for speaker modeling have been investigated. Early work explored the modeling of prosodic variation based on a parameterization of stylized pitch contours and pause and voiced segment durations [2], and modeling of idiosyncratic lexical features [3]. More recently, a group at the JHU Summer Workshop experimented with prosodic and lexical features, and found that the performance of conventional systems can be greatly improved by adding longer-term features [4].

In this work we describe research on the modeling of prosodic and lexical features for the NIST Speaker Recognition Evaluation (SRE) 2003 extended data task. The extended data task provides considerably more training data per speaker (as many as 16 conversation sides) than the standard one-speaker task, thus facilitating the modeling of long-range features. We report results for (1) systems based on individual feature types alone, (2) systems in combination with a state-of-the-art frame-based baseline system, and (3) an all-system combination.

1-7803-7980-2/03/$20.00 © 2003 IEEE

Authorized licensed use limited to: International Computer Science Inst (ICSI). Downloaded on July 15,2010 at 22:07:28 UTC from IEEE Xplore. Restrictions apply.

The paper is organized as follows: Section 2 introduces the task; Section 3 details the segmentation, transcription, and system combination; Section 4 describes our baseline system. Sections 5, 6, 7, and 8 describe individual systems. The results for the individual systems and the combination of all systems are presented in Section 9. Finally, Section 10 presents our conclusions.

2. 2003 NIST EXTENDED-DATA ONE-SPEAKER RECOGNITION EVALUATION (SRE)

The 2003 NIST Extended-data SRE is a text-independent one-speaker detection task based on data drawn from Switchboard 2 (SWB2) phase 2 and 3 databases. These databases have two-speaker conversations recorded over telephone channels. The main task is divided into three parts based on the amount of data used for training speaker models. In each part, 4, 8, or 16 conversation sides are used for training speaker models and one conversation side is used for testing. In general, speaker recognition performance improves by increasing the training data. However, the 16-conversation (training) condition has fewer speaker models and a smaller test set compared to the 8-conversation condition. Therefore, in this paper we focus on the 8-conversation condition.[1]

For each training condition, the evaluation data are divided into 10 non-overlapping splits. The evaluation paradigm uses the splits for N-fold cross validation. N-1 split(s) are used for development and the results are applied on the remaining one split.

Two metrics are used to evaluate a system: (1) equal-error rate (EER), and (2) detection cost function (DCF). These measures are computed using false-acceptance (FA) rate, false-rejection (FR) rate, the cost for both types of errors, and the target and impostor priors. EER assumes that the costs and priors are equal. DCF assumes that FA is 10 times more costly than FR, and impostors are 10 times more likely than target speakers.

3. SEGMENTATION, TRANSCRIPTION AND SYSTEM COMBINATION

The evaluation data are segmented into speech and non-speech segments using a two-state hidden Markov model (HMM). The resulting speech segments are processed by DECIPHER, SRI's conversational speech recognition system [5], to generate word-, phone-, and state-level transcriptions. The recognition system is trained using the SWB1 database. The system uses Mel-frequency cepstral coefficient (MFCC) features, which are normalized using vocal-tract length normalization, and then transformed using heteroscedastic linear discriminant analysis. The acoustic models are gender dependent within-word triphone models, trained using a maximum mutual information estimation criterion. The models are adapted to each speaker using open-loop maximum likelihood linear regression. The language model uses word bigrams, and is trained on a 37K-word plus 3K-multiword vocabulary. The recognition system runs at 3xRT on a Pentium Xeon processor and yields a word error rate of approximately 38% on the transcribed portion of the SWB2 database.

Information from different systems is combined at the score level using LNKnet software from MIT Lincoln Laboratory [6]. The combiner is a single-layer perceptron. It has two complementary output classes: true speaker and impostor. Class priors are adjusted to minimize the DCF metric. Combiners are trained using N-fold cross-validation for each training condition.

4. MFCC-GMM BASELINE SYSTEM

Our baseline system uses a 39-dimensional feature vector: 13 MFCCs (C1-C13) after mean subtraction, 13 delta and 13 double delta coefficients. This vector is normalized using feature normalization as described in [7]. For normalization, gender and handset models are trained using the NIST 1997 SRE data.

Distribution of the normalized features is modeled using a GMM with 2048 Gaussian components. The background model is trained using the same data used for feature normalization. Speaker models are adapted from the background model using MAP adaptation, and the score is computed as the log-likelihood ratio of test data with respect to the speaker and the background model. The score is normalized to compensate for variation due to different test durations (TNORM, [8]). Normalization data for splits 1-5 is obtained from splits 6-10, and vice versa. This system results in an EER of 1.9% and a DCF of 0.91×10^{-03}. The performance of our baseline system was thus superior to the baseline scores made available by NIST for the evaluation.

[1] The 8-conversation training condition uses roughly 1200 speaker models and 23,000 test trials.

Authorized licensed use limited to: International Computer Science Inst (ICSI). Downloaded on July 15,2010 at 22:07:28 UTC from IEEE Xplore. Restrictions apply.

5. CONDITIONAL PHONE PRONUNCIATION SYSTEM

We developed a conditional phone pronunciation system by adopting the approach used in the system that was first developed in the JHU summer workshop [4,9]. The main idea is to model phonemic variation across different speakers. The phonemic variation is captured using conditional probability of an open-loop phone given the phone identity obtained by an automatic speech recognition (ASR) system. These probabilities are used to train background and speaker models. The score is computed as the likelihood ratio of the observed pairs of {open-loop phone, ASR phone} with respect to the speaker and background models.

The open-loop phone transcriptions are obtained from five recognition systems for a set of languages – English, German, Spanish, Japanese, and Mandarin. For each language, conditional phone pronunciation systems are developed separately and they are combined at the score level using the neural network combiner (see Section 4). Results show that this system gives the best performance among non-baseline systems with EER=2.25% and DCF=8.870×10^{-3}. However, it does not provide any improvement when combined with the baseline system, suggesting that it does not capture complementary information.

6. DURATION-BASED SYSTEM

This system aims to capture speaker-specific duration patterns. To this end, each word and each phone are represented by feature vectors comprised by the durations of their component phones or ASR phone model states. These vectors are then modeled by GMMs. For a detailed description of this system see [10].

6.1 Features

Three different duration features are used and for each of them a different system is created:

- **Word features:** Sequence of phone durations in the words in the utterance. The number of components depends on the word pronunciation.
- **Phone features (1comp):** Duration of the phones in the utterance. These are scalar values.
- **Phone features (3comp):** Sequence of state durations in the phones in the utterance. These are three-component vectors.

In all cases the features are obtained from alignments of recognized words.

6.2 Training and adaptation

Once the features are computed, speaker-independent background GMMs for each word and each phone (one-component and three-component models) are estimated. The speaker models are later obtained through MAP adaptation of means and weights of the background model. Because speakers typically increase duration before pausing ("prepausal lengthening") for both grammatical and hesitation pauses, we train context-dependent models along with the context-independent ones. ***Pause context*** models are trained using samples that occur before a pause longer than 200 ms; ***word context*** models are trained using the remaining samples.

6.3 Scoring procedure and results

Three separate scores are obtained, one for each set of models: words, phones-1comp, phones-3comp. Each score is computed as the sum of the log-likelihoods of the corresponding feature vectors in the test utterance according to target speaker models. This sum is then divided by the total number of scored components and normalized by the background model score.

Table 1: Performance of different duration systems

System	Without TNORM		With TNORM	
	EER (%)	DCF ($x10^{-3}$)	EER (%)	DCF ($x10^{-3}$)
Phone model (3-comp)	8.01	53.4	5.70	34.6
Phone model (1-comp)	12.47	87.3	8.64	49.8
Word model	10.93	60.8	9.12	43.4
Combination	7.07	42.0	4.62	25.5

During scoring, if the context-dependent model corresponding to a feature vector was not adapted to the speaker with more than a certain number of samples, the context-independent model is used instead. In addition, only models adapted to the speaker with at least five samples are used for scoring.

Authorized licensed use limited to: International Computer Science Inst (ICSI). Downloaded on July 15,2010 at 22:07:28 UTC from IEEE Xplore. Restrictions apply.

Results for the three sets of models are presented in Table 1. We observed that scores from duration models are sensitive to the test segment length. Therefore, scores were normalized using TNORM where TNORM statistics were estimated in the same way as they are for the baseline system scores. As shown in Table 1, normalizing the score using TNORM yields considerable performance improvements. Also as shown, performance improves when all duration systems are combined.

7. NEW EXTRACTION REGION FEATURES

We also defined a large set of features based on various new regions of interest, which we will refer to as "New Extraction Region Features (NERFs)." Work on these features is preliminary, thus we report results for only two sample subtypes of NERFs. NERFs define a sliding temporal region based on either a fixed window length or boundaries defined by the presence or values of other features (see Figure 1).

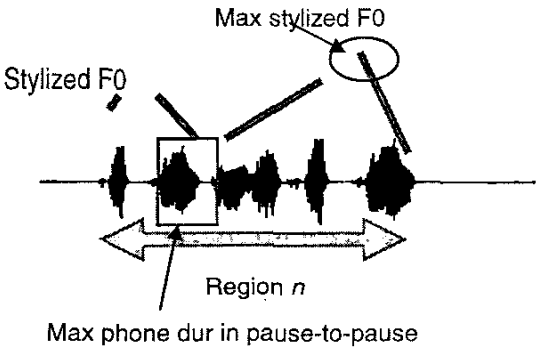

Figure 1: Sample NERFs

The regions are defined to be larger than a frame, to capture longer-term behaviors, and to be smaller than a whole conversation side, to yield more samples. Both the region definitions and the features defined inside regions are motivated (albeit loosely) by psycholinguistics. Note that the duration features described earlier can also be thought of as NERFs, where regions are defined as words.

7.1 Pause-to-pause region

This particular system's region stretches from a pause to the next pause, using a certain minimum pause duration threshold (500 ms for the system used in these experiments). Four types of prosodic features are extracted for each region: pitch, duration, pause, and energy.

Duration features are extracted from the time alignments of the phones inside the region. An example of these features is the duration of the longest phone or vowel in the region, normalized by the average duration for that vowel overall.

Pitch features are obtained from pitch tracks extracted from the signal and then post-processed using an improved version of the approach in [2], where pitch contours are "stylized." Features such as the maximum stylized pitch, the last slope, and the mean pitch in the region are then computed.

Energy features are obtained using a stylized version of the raw energy. A linear approximation of the energy contour is obtained over each segment of the stylized pitch, and features like the energy range and the last or first slope in the region are computed.

Pause features, such as the average pause in the region, the pause before and after the region, or the maximum pause inside the region, are also computed.

After extraction, features are whitened and then modeled using GMMs. The rest of the recognition algorithm is similar to that used for the baseline system.

7.2 Stylizer segment region

In this system, the extraction regions are the segments of the stylized pitch. These types of units were successfully used in the JHU workshop [4]. Several feature vectors are defined using start, end, and mean values of pitch and energy in the segment, their slopes, and the duration of the segment. These features are then modeled using GMMs. The individual EER for these systems is between 22% and 32%.

In contrast with the findings of the JHU workshop, we found that none of these systems gave substantial improvements when combined individually with the baseline. This may be due to two main differences in the systems. First, JHU workshop approaches included phone identity along with duration, and slopes of pitch and energy. We excluded this information since it is already modeled in our duration system. Second, unlike our approach, the JHU workshop approach used discretized versions of these features and modeled them using language models.

Authorized licensed use limited to: International Computer Science Inst (ICSI). Downloaded on July 15,2010 at 22:07:28 UTC from IEEE Xplore. Restrictions apply.

8. LANGUAGE MODEL BASED SYSTEM

This system uses word sequences from a spoken utterance as features and is an extension of [3]. In this work, the word sequences are obtained from the recognizer described in Section 3. The sequence of words is modeled as a bag of bigrams. Background and target speaker models are the estimated probabilities of these bigrams. The score is computed as the log-likelihood ratio of a set of bigrams extracted from test utterance with respect to background and speaker models.

We observed that the distribution of scores varies with the test duration length. This variation was compensated for by normalizing the scores using TNORM [8]. TNORM statistics were estimated in the same way as they are for baseline. This system gives an EER of 9.2% and a DCF of 41.31×10^{-3}.

9. RESULTS

Table 2 shows the results of two-way combinations of the baseline system with each of the non-spectral systems. Results show that combination with the duration-based system yielded the greatest improvement in performance (about 40%, for both EER and DCF). Combination with the LM-based system also gives large improvements over the baseline system. Among NERFs, pause-to-pause features give about 15% improvement, and stylizer segments give about 5% improvement in both DCF and EER. This is an encouraging first result, especially given the low density of the NERFs.

Table 2: Results of two-way combination with baseline

Systems	%EER	DCF ($\times 10^{-3}$)
Baseline (1)	2.35	9.144
(1) + Duration system	1.40	5.319
(1) + LM system	1.61	5.817
(1) + Pause to pause NERFs	2.03	7.643
(1) + Stylizer segment NERFs	2.22	8.655

Table 3 and Figure 2 present combination performance results for five sets of systems: (1) MFCC-GMM baseline (BASELINE), (2) NON-SPECTRAL systems without TNORM (NON-SPEC), (3) NON-SPECTRAL systems with TNORM (NON-SPEC_TNORM), (4) BASELINE and NON-SPEC (ALL), and (5) BASELINE and NON-SPEC_TNORM (ALL_TNORM). All the systems in each set are combined using the single-layer perceptron described in Section 4.

Table 3: EERs and DCF values for different systems

Systems	EER(%)	DCF ($\times 10^{-3}$)
BASELINE	2.35	9.144
NON-SPEC	2.50	13.670
NON-SPEC_TNORM	1.76	8.544
ALL	1.33	5.286
ALL_TNORM	1.06	3.820

These results show that performance of the non-spectral system combination (NON-SPEC) improves by using TNORM on duration and LM system scores. This improvement is also reflected in the combination of scores from all systems (ALL versus ALL_TNORM). Note that the NON-SPEC_TNORM result is better than the BASELINE for both EER and DCF values and it gives significant improvement when combined with the BASELINE.

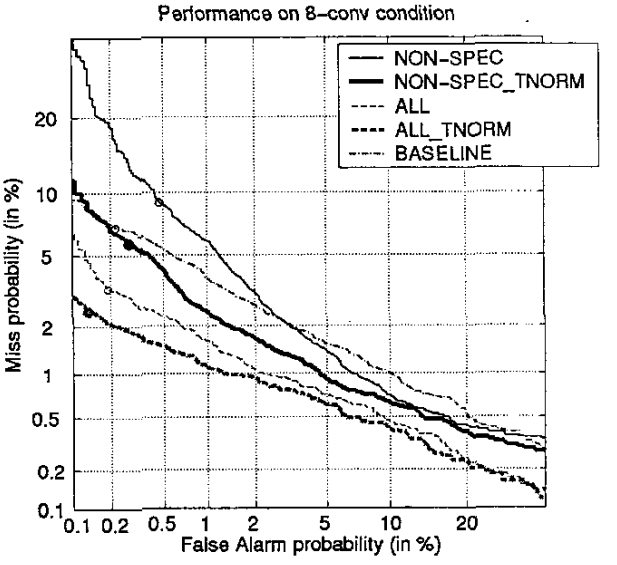

Figure 2: DET curves for different sets of systems

10. SUMMARY AND CONCLUSIONS

We explored the contribution of stylistic features based on prosodic and lexical patterns to speaker recognition in the NIST 2003 extended data task. The systems using these

Authorized licensed use limited to: International Computer Science Inst (ICSI). Downloaded on July 15,2010 at 22:07:28 UTC from IEEE Xplore. Restrictions apply.

features were evaluated independently and in combination with each other in the context of the NIST 2003 extended data task evaluation. The baseline system was a state-of-the-art MFCC-GMM system, which was superior in performance to the baseline scores distributed by NIST.

Combination results showed that the duration-based system provided the most complementary information to the baseline system. This was followed by the LM-based system and the prosodic NERF system. In addition to providing information complementary to the baseline, each of these new stylistic features provided information that was complementary to the other new features. Additional gains were achieved by applying TNORM to various systems. Overall, these results suggest that modeling of longer-range features such as prosodic and lexical patterns can improve speaker recognition performance in contexts and applications in which extended training and test data are available.

11. ACKNOWLEDGMENTS

We thank Doug Reynolds, Gary Kuhn, and Barbara Peskin for useful discussions. We also thank Doug Reynolds and Walter Andrews for assistance with the JHU workshop conditional phone system data. This work was funded by a DoD KDD award via NSF IRI-9619921. Additional support came from NASA award NCC2-1256. The views herein are those of the authors and do not reflect the views of the funding agencies.

12. REFERENCES

[1] D. Reynolds, T. Quatieri, R. Dunn, "Speaker Verification Using Adapted Mixture Models", *Digital Signal Processing*, vol. 10, pp.181-202 (2000).

[2] K. Sonmez, E. Shriberg, L. Heck, M. Weintraub, "Modeling Dynamic Prosodic Variation for Speaker Verification", *Proc. Intl. Conf. on Spoken Language Processing*, vol. 7, pp. 3189-3192, Sydney, Australia (1998).

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

[4] 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

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

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

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

[8] R. Auckenthaler, M. Carey, H. Lloyd-Thomas, "Score Normalization for Text-independent Speaker Verification Systems", *Digital Signal Processing*, vol. 10, no. 1-3, January 2000.

[9] D. Klusacek, J. Navratil, D. A. Reynolds, J. P. Campbell, "Conditional Pronunciation Modeling in Speaker Detection", *Proc. IEEE ICASSP*, vol. 4, pp. 804-807, Hong Kong (2003).

[10] L. Ferrer, H. Bratt, V. R. R. Gadde, S. Kajarekar, E. Shriberg, K. Sonmez, A. Stolcke, A. Venkataraman, "Modeling Duration Patterns for Speaker Recognition", in Eurospeech (Geneva), September 2003.

Authorized licensed use limited to: International Computer Science Inst (ICSI). Downloaded on July 15,2010 at 22:07:28 UTC from IEEE Xplore. Restrictions apply.