"TalkPrinting": Improving Speaker Recognition by Modeling Stylistic Features

Sachin Kajarekar, Kemal Sönmez, Luciana Ferrer, Venkata Gadde, Anand Venkataraman, Elizabeth Shriberg, Andreas Stolcke, and Harry Bratt

Speech Technology and Research Laboratory
SRI International, Menlo Park, CA

Abstract. Automatic speaker recognition is an important technology for intelligence gathering, law enforcement, and audio mining. Conventional speaker recognition systems, which are based on independent short-term spectral samples, suffer from a lack of noise robustness and are unable to model a speaker's idiosyncratic stylistic features. This paper describes "TalkPrinting", a program of research aimed at adding such stylistic features to conventional systems. Results on three preliminary systems based on stylistic features demonstrate that (1) the new features alone carry significant speaker information; (2) they also carry significant complementary information compared to the conventional features; and (3) they provide increasing improvements in performance with increasing test durations.

1. Introduction

Automatic speaker recognition is the task of determining a speaker's identity from his or her speech. It is a crucial technology for finding and tracking conversations involving particular target speakers from the unwieldy amount of audio data captured each day by intelligence sources. It can be used to automatically prioritize conversations for further analysis by human listeners.

Conventional features for speaker recognition are based on the short-term spectrum[1] (e.g., cepstrum). However, they lack robustness to mismatched training and testing conditions, e.g., due to different telephone handsets [6]. To improve robustness, researchers are investigating into the features that reflect speaking *style*, such as idiosyncratic word usage, intonation, and timing patterns. Some of these features were recently investigated at a workshop at Johns Hopkins University (JHU) [7] using an acoustic-prosodic feature database developed at SRI for other work [8]. The workshop results showed significant improvements in the performance of a

[1] These features indirectly capture speaker's vocal tract shape and its movements.

conventional speaker recognition system by adding features that reflect higher-level and longer-term properties.

This paper describes further efforts to exploit speaking style features for speaker recognition, using an approach we call *TalkPrinting*[2]. We describe three types of features beyond those studied at the workshop: (1) higher-order cepstra modeling speaker fundamental frequency (F0), (2) language models capturing word usage, and (3) word duration models. Section 4 discusses results of individual systems and their combination. Section 5 describes the effect of different test sample durations and Section 6 describes conclusions from this work.

2. Overview of Task and Baseline System

We use data from the NIST 2001 speaker recognition evaluation extended-data speaker recognition task [5]. This is the entire Switchboard-I corpus [4] divided into six "splits", with 1 to 16 training conversations per split. Development on the whole task is computationally expensive, and we observed that meaningful results can be obtained by defining a "short" task, i.e., only two splits of 8-conversation training data. This report provides results on the *short* task[3]. We report performance of systems in terms of equal error rate (EER). EER is the point on the receiver operating characteristics curve at which false-acceptance and false-rejection errors are equal.

To provide a baseline system for combination with our proposed new features, we extended SRI's conventional speaker recognition system [9]. This Gaussian mixture model (GMM) [6] based system uses standard Mel-frequency cepstral coefficients (MFCCs) as features. The background GMM is trained with data from many speakers and the speaker GMM is adapted from the background GMM by using training data for the respective speaker.

3. High-Level Speaker Features

We investigated three types of TalkPrinting features: F0-based, language-based, and duration-based. For the latter two feature types, we report results on true words.

F0-based features. We used higher order linear frequency cepstra to represent F0, which allowed us to use the same modeling approach as for the baseline GMM system using MFCC features. This allowed us to model pitch information more explicitly than in the baseline system.

---

[2] This captures the notion that we are adding features related to voluntary stylistic patterns in how a person talks, rather than relying solely on features related to pre-determined vocal physiology

[3] Note that comparison experiments on the full task were similar or improved; results reported here thus *underestimate* our true performance.

Language-based features. We examined idiosyncratic word patterns, modeled using individual words and bigrams, following on previous work [1]. We characterized the stream of words by a single statistical language model (LM). This allowed us to optimize the vocabulary size of the LM, as well as the length of N-grams to be used.

Duration-based features. We adapted a duration model used for speech recognition [3] to capture individual differences in speaking rate, constrained by lexical information. We represented each word by a vector of the durations of the individual phones in the word. Using these vectors, background and speaker models (GMMs) were estimated using an approach similar to that used for the baseline system. During speaker adaptation, if a word had insufficient training data then we backed off to phone duration models.

4. Results

System scores from the high-level features and the baseline features were combined using linear discriminant analysis (LDA) [2]. Results on different system combinations on the short-task are shown in Table 1.

Table 1. Performance of different systems (and combinations) on the short task[4]

System	EER (%)
Baseline (S1)	2.6
F0-based (S2)	10.9
Language-model-based (S3)	12.6
Duration-based (S4)	10.6
S1+S2	2.5
S1+S3	2.1
S1+S4	1.5
S1+S2+S3	1.8
S1+S2+S3+S4	1.2

The EER of our baseline system (S1) is 2.6%. Using score normalization, performance improves to 1.3%. Due to practical limitations, we present system combinations *without* score normalization. Since our TalkPrinting features are not likely to be severely affected by channel variation, we believe that the gain in performance from score normalization will generalize to system combinations.

The F0-only system (S2) yields 10.9% EER and is the second-best performing system. However, it provides only a small improvement after combination with the baseline. This can be attributed to a correlation between the F0 features and baseline cepstrum features. For the LM (S3) system, best results were obtained with unigrams

[4] Note that they are from the suboptimal systems, please refer to end of this section for optimized results

over words occurring at least 300 times in the training data, for an EER of 12.6 on the full training set. The highest relative improvement is obtained using the duration-based system (S4), which provides the most complementary information.

As mentioned earlier, these results significantly underestimate the absolute performance level of our systems on the complete 8-conversation task. After various system enhancements, the baseline EER is 2.0% without score normalization and 0.9% with score normalization. Further, EER for the duration-based system is now 4.5% and EER for the LM-based system is now 8.5%. With a neural-network based combiner, overall EER reduces to 0.20%.

In summary, although individual systems based on the high-level features never outperform the baseline system, they provide a significant reduction in error when *combined* with that system. Duration modeling provides particularly useful complementary information, dramatically improving performance.

5. Effect of Test Segment Duration

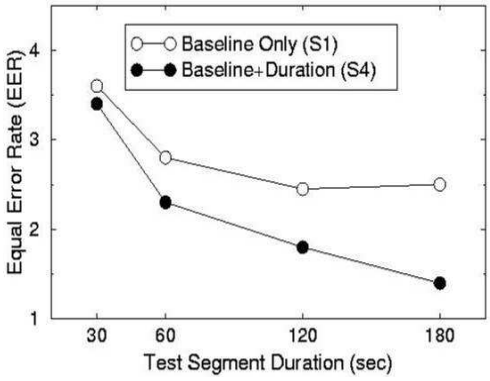

Fig. 1. % EER for different test durations using baseline system and its combination with duration-based system.

We also investigated the effect of test segment duration on performance, since for many applications more than a minute of data is likely to be available. We split our data into conditions using 30, 60, 120 and 180 second test segments. Figure 1 shows performance of the baseline system and various combined systems on these test sets. As can be seen, the baseline system does not take advantage of increasing test segment duration beyond approximately 120 seconds of speech[5], which the TalkPrint system continues to improve as test segment duration increases. These results suggest

[5] Note: the average duration of the original tests is about 180 seconds.

that TalkPrinting can be particularly beneficial to applications that have access to longer speech samples.

6. Conclusions

We investigated three high-level features for speaker recognition based on F0, word usage patterns, and timing patterns. Results show that the new features provide significant *complementary* information to standard speaker recognition systems, resulting in significant performance improvements after score combination. Furthermore, results show that unlike the baseline system, TalkPrint systems continue to improve as more data is available in testing, making their relative contribution even greater at longer test durations. TalkPrint features could therefore be of significant value to future intelligence applications.

Acknowledgments

This work was funded by a KDD supplement to NSF IRI-9619921. We thank Gary Kuhn for helpful discussion and technical suggestions.

References

1. Doddington, G.: "Some Experiments on Ideolectal Differences Among Speakers," http://www.nist.gov/speech/tests/spk/2001/doc/ (2001).
2. Fukunaga, K.: "Statistical Pattern Recognition," Academic Press, Indiana.
3. Gadde, V. R. R.: "Modeling Word Durations," *Proc. Intl. Conf. on Spoken Language Processing*, Beijing, (2000) 601-604.
4. Godfrey, J., Holliman, E., and McDaniel, J.: (1992) "SWITCHBOARD: Telephone speech corpus for research and development," *Proc. ICASSP*, (1992) 517-520.
5. NIST 2001, http://www.nist.gov/speech/tests/spk/2001/doc/2001-spkrec-evalplan-v05.9.ps
6. Reynolds, D.: "Speaker Identification and Verification Using Gaussian Mixture Speaker Models," *Speech Communication*, Vol. 17, No. 1-2, August (1995) 91-108.
7. Reynolds, D., et al.: "The SuperSID Project: Exploiting high-level Information for high-accuracy speaker recognition," To appear in *Proc. ICASSP*, Hong Kong (2003).
8. Shriberg, E., Stolcke, A., Hakkani-Tur, D., and Tur, G.: "Prosody-Based Automatic Segmentation of Speech into Sentences and Topics," Speech Communication, Vol. 32, No. 1-2, (2000) 127-154.

9. Sönmez, M. K., Heck, L., and Weintraub, M.: "Speaker Tracking and Detection with Multiple Speakers," *Proc. EUROSPEECH*, Vol. 5, Budapest, Hungary, (1999) 2219-2222.