Modeling Duration Patterns for Speaker Recognition

Luciana Ferrer *Harry Bratt* *Venkata R. R. Gadde* *Sachin Kajarekar*
Elizabeth Shriberg *Kemal Sönmez* *Andreas Stolcke* *Anand Venkataraman*

Speech Technology and Research Laboratory
SRI International, Menlo Park, CA, USA
http://www.speech.sri.com/

Abstract

We present a method for speaker recognition that uses the duration patterns of speech units to aid speaker classification. The approach represents each word and/or phone by a feature vector comprised of either the durations of the individual phones making up the word, or the HMM states making up the phone. We model the vectors using mixtures of Gaussians. The speaker specific models are obtained through adaptation of a "background" model that is trained on a large pool of speakers. Speaker models are then used to score the test data; they are normalized by subtracting the scores obtained with the background model. We find that this approach yields significant perfomance improvement when combined with a state-of-the-art speaker recognition system based on standard cepstral features. Furthermore, the improvement persists even after combination with lexical features. Finally, the improvement continues to increase with longer test sample durations, beyond the test duration at which standard system accuracy level off.

1. Introduction

Humans use several types of perceptual cues to identify a particular speaker from his or her voice. One set of cues reflect the spectral properties of speech. These cues are a direct function of the detailed anatomical structure of the speaker's vocal tract, and are modeled using very short time slices. A second set of cues, which we will refer to as higher-level cues, span longer temporal regions. These are behavioral in nature, and include idiosyncratic patterns in pronunciation, word usage, and prosody. Such behavioral cues are influenced by many factors, including dialect, cognitive style, pragmatics, and sociolinguistic context.

Current speaker recognition systems are based on cues of the first type, i.e., low-level acoustic information based on Mel-frequency cepstral coefficients (MFCCs) [1]. These systems are therefore quite sensitive to noise and channel mismatch. We expect that adding high level information could make such systems more robust. Furthermore, to the extent that high-level features encode independent information, speaker recognition accuracy under normal conditions should see considerable improvement by using both low- and high-level cues. In fact, higher-level lexical, prosodic, and pronunciation features extracted from conversation-length test samples have recently been shown to give dramatically improved speaker recognition accuracy when combined with traditional state-of-the-art systems [2].

In this work we exploit a novel high-level knowledge source for speaker recognition: the differences in detailed duration patterns exhibited by different speakers. Each word and each phone is modeled in terms of its durations and/or those of its constituent units. Speaker-specific models are used to compute scores for speaker recognition, which in turn can be combined with scores from other systems to give considerable accuracy improvements. Specifically, we investigate combinations of duration models with the MFCC-based scores used in a state-of-the-art speaker recognition system, as well as with models that capture speaker-specific word usage. We also investigate the effect of test set length on speaker recognition accuracy.

2. Duration Models

In this study we express duration patterns as vectors of durations of speech units. We looked at three types of feature vectors:

- **Word features:** The features for each word are the sequence of phone durations in the word. Thus word models have varying numbers of components, depending on the number of phones in their pronunciation. Each word pronunciation is treated separately. This approach is derived from an approach we have used for scoring speech recognition hypotheses [3].
- **Phone features (1-component):** The features are simply the durations of the phones and are one-dimensional vectors.
- **Phone features (3-component):** The features are the sequences of HMM state durations in the phone. Since our recognition system uses 3-state

phone HMMs throughout, all feature vectors are three-dimensional.

For the extraction of these features we use the state-level alignments from a large-vocabulary speech recognizer. To make the most effective use of each type of feature we use three different systems, one for each feature type. Each system generates a speaker recognition score for a given test sample, and the three systems are then combined at the score level. We find that this approach yields better results than an approach we used previously, in which word models would *back off* to phone-level models only when insufficient word-level training data was available [3]. We also suspect that modeling the phone durations in a separate system may yield better performance than adding phone durations as a fourth component to the three-component phone features.

3. Model training and adaptation

After feature extraction for the whole corpus, a subset of the data is used to create background models for each word and phone. This subset is composed of complete conversations from many speakers; hence the models obtained can be considered speaker-independent. Gaussian mixture models (GMMs) are trained for each word and each phone (both 1- and 3-component versions). The size of each of these GMMs is determined by the number of training samples available for that model in the background model data.

Once these models are trained, the training conversations for each target model are used to adapt the background model to the speaker. This way a speaker-specific model is robustly obtained for each word and each phone.

3.1. Pause context dependencies

Duration patterns often change when the speaker is about to pause. For this reason we train context-dependent models along with the context independent ones. The *pause context* model is trained for each word using the samples that are found before a pause longer than 200 msec. A *word context* model is trained for each word using the samples that are not followed by pause, or are followed by pauses shorter than 200 msec. For the phone models, the pause context models are trained using the phones that are found in the last syllable of a word followed by a pause longer than 200 msec. The rest of the phones are used to train the word context phone models.

4. Score Computation

Three different scores are obtained for each test, one for each set of features: word features, 1-component phone features and 3-component phone features. Each of these scores is computed as the sum of the likelihoods of the feature vectors in the test utterance given its models. This number is then divided by the number of components that were scored (this is better than dividing by the number of models because not all the models are formed by the same number of components in the case of the word features). The final score is obtained as the difference between the speaker-specific model score and that from the background model. This step makes the scores less dependent on what was spoken, and effectively cancels out score components that are not adapted to the speaker.

4.1. Scoring context-dependent models

As many of the context dependent models (especially the pause context models) are usually adapted to the speaker with very few samples, it is necessary to use a back-off strategy. When the context-dependent model corresponding to the current feature vector has not been adapted to the speaker with more than a certain number of samples (30 for the phone models, 15 for the word models), the context-independent model is used instead.

4.2. Avoiding poorly adapted models

For many of the word features, models may not be adapted to the speaker with enough samples for robust estimation. We find that if we only score those models that were adapted to the speaker with more than 5 samples, we obtain better results than when all test words are scored.

5. Lexical Features

We were also interested in whether duration features could offer any benefit to a system that uses not only standard features but also features from a speaker's word usage patterns. Following Doddington [4], we record speaker-specific word bigram frequencies, and score the test data with the likelihoods of observed bigram counts. The log probabilities thus obtained are normalized by subtracting scores from a background model trained on all the data, and then dividing by the number of observed words, similar to the way other scores are normalized.

6. Experiments

6.1. Methodology

The data used in these experiments is from the NIST 2001 speaker recognition evaluation extended-data task [5]. The dataset uses the entire Switchboard-I corpus [6], and is divided into six sets, or "splits", for jacknifing. We report performance of systems in terms of equal error rate (EER) for a training condition of 8 whole conversations and for different test lengths. EER is computed by moving a threshold over the receiver operating characteristic (ROC) curve. At every point, false-acceptance and false-rejection errors are calculated. EER is the point on the ROC where the two types of errors are equal.

System	True	Recog.
Baseline (MFCC features)	0.90%	
Lexical bigrams	8.65%	9.30%
3-comp. phone duration features.	3.71%	3.30%
1-comp. phone duration features	10.88%	8.82%
Word duration features	5.22%	6.22%

Table 1: Equal error rates for the baseline, the word bigram-based system and the three proposed duration-based systems. For each system, results using true words and automatically recognized words are reported.

Features used	EER
Baseline (MFCC features)	0.90%
Baseline + lexical features	0.57%
All duration features combined	2.59%
Baseline + duration features	0.40%
Baseline + lexical + duration features	0.29%

Table 2: Equal error rates for the baseline system and various combinations of lexical and duration-based models. All results are based on automatic word recognition.

To provide a baseline system for combination with our proposed new features, we extended SRI's conventional speaker recognition system, a system based on Gaussian mixture models [1] that 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 as is done for the duration models.

Lexical bigram features are obtained using the word hypotheses produced by a pared-down version of SRI's conversational speech recognizer [7]. The recognition word error rate on the data processed was 30%. Duration features were extracted using the phone- and state-level alignments of the recognizer output. To assess the effect of recognition errors we also tested systems that were based on the words and alignments of the reference transcripts.

The background model and the target models are always obtained on independent splits. For example: for target models in Splits 1, 2 and 3, the background model is computed on Splits 4, 5 and 6. Scores obtained with the baseline and lexical features are combined with duration scores, using a multilayer perceptron with 10 hidden layers. The classifiers are trained using 6-fold cross-validation. For example, for Split 1, the perceptron is trained using Splits 2, 3, 4, 5, and 6.

6.2. Results

Table 1 shows the EERs for the individual systems for two different test conditions:

- Features extracted using forced alignments for the true words
- Features extracted using alignments for automatically recognized words

In both cases full conversation sides are used for testing.

Note that models were trained and tested in matched conditions; in particular, the models used on recognition output were also trained on recognition output.

Clearly, none of the non-standard knowledge sources alone is competitive with the baseline system. Furthermore, systems based on word units (including lexical bigrams and word-level duration models) degrade somewhat as a result of recognition errors, as might be expected. However, the degradation is small (about 20% relative in the worst case) considering the relatively high word error rate. Somewhat surprisingly, the phone-level duration models actually improve when automatically recognized words are used. It is conceivable that speaker-specific misrecognitions (e.g., due to pronunciation variation) are captured and modeled by the phone duration models, and therefore benefit speaker recognition. Still, this result merits further investigation. For example, this may mean that phone models could benefit from the use of a simple phone recognizer instead of the automatic speech recognition system we currently use.

We now turn to systems that make use of combinations of knowledge sources. Table 2 shows results for several such systems, using automatic word recognition and full conversation sides as testing conditions. The three duration models combined give only a modest error reduction (20% relative) over the single best duration model, the three-component phone duration model, indicating that there is a large amount of redundancy among the three duration models. However, all combinations with the baseline system give significant error reductions relative to the baseline. Adding the duration features cuts EER by more than 50%. Adding both lexical and duration features reduces the EER to one third of the baseline. Relative to a baseline of combined cepstral and lexical features, the duration features again achieve an almost 50% error reduction.

6.3. Effect of test segment duration

Our duration models use much larger speech units than does the baseline system, and hence work with many fewer data points per unit time. (The same is true of the lexical model.) We were therefore interested in the relationship between test data duration (how much speech is available to identify the speaker by), and the effectiveness of our models. For this study we used a stripped-down baseline system that did not incorporate some advanced

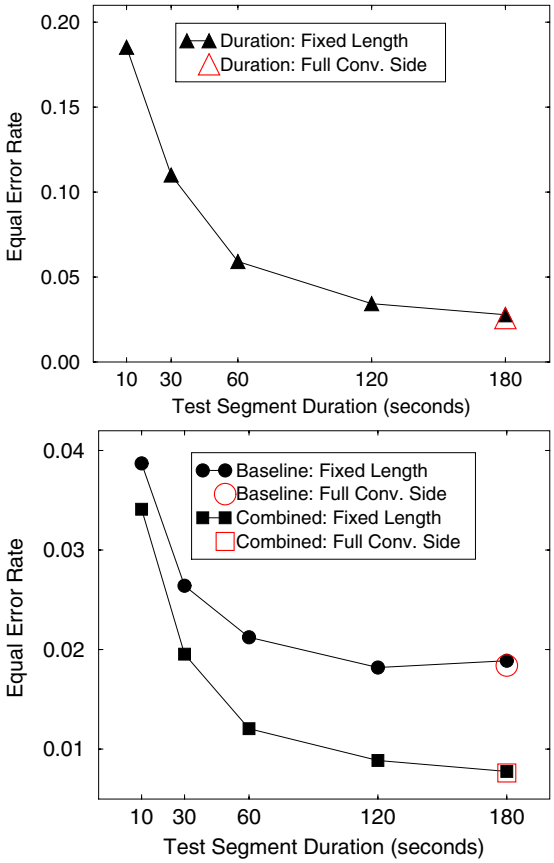

Figure 1: Performance for duration model, baseline model, and combined model, as a function of test sample length. Note the two graphs use different ordinate scales. "Full Conv. Side" = full conversation side; this result is plotted at 180 sec. because that is roughly the mean duration over full conv. sides.

functions, such as handset normalization. The resulting baseline EER was about 1.8%.

We tested the models on test segments of varying lengths, between 10 and 180 seconds. (The latter corresponds roughly to the mean test length if testing on full conversation sides. Although in principle one could examine lengths longer than 180 seconds, we run into a problem of data sparseness at high lengths because total conversation time was limited in the collection procedure for the present corpus.) To obtain these test sets we concatenated speech segments with only short embedded pauses, so that the nominal test set lengths corresponded mainly to speech, not silence. Similar to results in Table 2, we used automatic word recognition to obtain the duration features. Results are shown in Figure 1.

As expected, all models degrade on less test data, and this is (unsurprisingly) especially true for the duration system. Yet interestingly, even at the shortest test length there is still an improvement to be gained from duration modeling. Furthermore, while baseline system performance levels off at about 120s, the duration model continues to improve as more data is available. Thus, alternative models such as the duration model described here, may prove particularly useful for certain applications because they can take advantage of additional available test data.

7. Conclusions

We have shown that duration features based on automatic speech recognition are highly effective knowledge sources for speaker identification. Duration features are useful at the word, phone, and state level. When combined with information from a state-of-the-art conventional speaker recognition system, duration features reduce identification error by as much as 50%. This gain in performance does not disappear when other helpful features, such as word-usage based models, are also present. In addition, duration-based features have the ideal characteristic that their contribution to performance continues to increase as test sample length increases; in contrast, conventional systems saturate after test durations of a few minutes of speech and do not make use of additional available data.

8. Acknowledgments

We thank Doug Reynolds and Gary Kuhn for helpful discussions. This work was funded by a KDD supplement to NSF IRI-9619921 and by NASA Award NCC 2-1256. The views herein are those of the authors and do not reflect the policies of the funding agencies.

9. References

[1] D. Reynolds, "Speaker identification and verification using Gaussian mixture speaker models," *Speech Communication*, vol. 17, no. 1-2, pp. 91–108, Aug. 1995.

[2] D. Reynolds, W. Andrews, J. Campbell, J. Navratil, B. Peskin, A. Adami, Q. Jin, D. Klusacek, J. Abramson, R. Mihaescu, J. Godfrey, D. Jones, and B. Xiang, "The SuperSID project: Exploiting high-level information for high-accuracy speaker recognition," in *Proc. ICASSP*, Hong Kong, Apr. 2003, vol. 4, pp. 784–787.

[3] V. R. R. Gadde, "Modeling word durations," in *Proc. ICSLP*, B. Yuan, T. Huang, and X. Tang, Eds., Beijing, Oct. 2000, vol. 1, pp. 601–604, China Military Friendship Publish.

[4] G. Doddington, "Speaker recognition based on idiolectal differences between speakers," in *Proc. EUROSPEECH*, P. Dalsgaard, B. Lindberg, H. Benner, and Z. Tan, Eds., Aalborg, Denmark, Sept. 2001, pp. 2521–2524.

[5] National Institute for Standards and Technology, "The NIST year 2001 speaker recognition evaluation plan," http://www.nist.gov/speech/tests/spk/2001/doc/2001-spkrec-evalplan-v05.9.ps, 2001.

[6] J. J. Godfrey, E. C. Holliman, and J. McDaniel, "SWITCHBOARD: Telephone speech corpus for research and development," in *Proc. ICASSP*, San Francisco, Mar. 1992, vol. 1, pp. 517–520.

[7] A. Stolcke, H. Bratt, J. Butzberger, H. Franco, V. R. Rao Gadde, M. Plauché, C. Richey, E. Shriberg, K. Sönmez, F. Weng, and J. Zheng, "The SRI March 2000 Hub-5 conversational speech transcription system," in *Proceedings NIST Speech Transcription Workshop*, College Park, MD, May 2000.