

TEXT-CONSTRAINED SPEAKER RECOGNITION USING HIDDEN MARKOV MODELS

Kofi Boakye

EE225D Final Project Paper

ABSTRACT

This paper presents a possible application of a text-dependent speaker recognition system within the unconstrained domain of telephone conversation speech, as contained in the Switchboard I corpus. The system utilizes word HMMs to generate likelihood scores for key words among the backchannel, filled pause, and discourse marker categories. Results on tests using a variant of the NIST 2001 extended data task yield an EER of 2.87%

1. INTRODUCTION

Speaker recognition systems in which the utterances permissible for analysis are limited to a fixed number—termed “text-dependent”—can produce high performance, especially relative to text-independent systems. By fixing the utterance, more of the acoustic variation arises from speaker distinction rather than, say, the phones and from this the advantage arises. There are, however, a number of domains where constraining speakers is not permissible. Is it, perhaps, possible to capitalize on the advantages of text-dependent systems in these domains?

One idea is to limit the words of interest to those occurring with high frequency in the domain. For this project, the domain was selected to be conversational speech. Another useful attribute of the selected words would be high speaker-discriminative quality. For conversational speech, it has been suggested [1] that words which are highly spontaneous that represent habitual speaking style may possess this characteristic. The common discourse markers, filled pauses, and backchannels, then, are natural candidates for selection.

In current speaker recognition systems the standard practice is to generate speaker models using adapted Gaussian Mixture Models (GMMs) as described in [2]. These systems utilize a “bag of frames” approach in which input data frames are assumed to be independent. The systems, then, do not take advantage of sequential information which could be used to aid in recognition. A natural “sequence of frames” approach is the use of Hidden Markov Models (HMMs) for model generation. A text-independent version of such a system using a Large Vocabulary Continuous Speech Recognizer (LVCSR) has been examined [3], and the performance

of a text-dependent system would be of interest.

This paper describes such a text-dependent system. Here, the speaker models consist of word models represented by adapted word-level HMMs for a speech recognizer. In the sections to follow, the recognition task along with system design and implementation are detailed. Results are then described, and, finally, conclusions are drawn with possible future work being mentioned.

2. THE TASK

The recognition task for this project was based on the Extended Data Task of the 2001 NIST Speaker Recognition Evaluation [4]. In the NIST evaluation speaker models are trained using 1,2,4,8, and 16 complete conversation sides and are subsequently tested on a complete conversation side. The conversation sides are approximately 2.5 minutes in length. This marks a change from previous evaluations in which training occurred using only 2 minutes of speech and test segments averaged 30 seconds in length. The intention of the current task is to permit the inclusion of techniques, such as the use of higher-order prosodic features [5], which examine phenomena existing on longer timescales and as such rely on more training data. For the purposes of this project, the use of longer segments increases the probability that one of the words of interest will appear as well as the frequency of that appearance. This enables the speaker recognition system to utilize a constrained word set without constraining the speech.

The Extended Data Task uses the Switchboard I conversational telephone corpus in a cross-validation process whereby the data is partitioned into 6 sections of approximately equal size and each partition is tested independently; when one partition is being tested, data from the others can only be used for background models or normalization. For this project, only one partition (partition 1) was subject to testing and three others (4,5, and 6) were reserved for providing background model data. In addition, models were trained using 8 conversation sides, as 8 conversations would most likely provide the most amount of data for training before arriving at an undesirably small number of data points for the Detection Error Tradeoff (DET) curve and Equal Error Rate (EER) computation.

3. SYSTEM DESIGN AND IMPLEMENTATION

3.1. Theory: The likelihood ratio detector

The basic speaker recognition task is the determination of whether or not a putative target speaker is the speaker in a specified test segment of conversation data. It can thus be considered a binary detection problem, and the tool of interest is a likelihood ratio detector. The detector measures the likelihood that a test segment $X = \{x_1, x_2, \dots, x_T\}$ (where x_i is a feature vector with time index i) arose from speaker S rather than a general background speaker, often referred to as the Universal Background Model (UBM). In practice, $LLR(X)$, the log-likelihood ratio is used. The test is then:

$$LLR(X) = \log p(X|S) - \log p(X|UBM) \quad (1)$$

$$LLR(X) \begin{cases} \geq \theta & \text{choose speaker } S, \\ < \theta & \text{reject speaker } S. \end{cases} \quad (2)$$

where θ is the threshold parameter, which can be set to alter susceptibility to false alarms and missed detections. In general, these likelihoods are not known, but comparable quantities can be obtained. The accumulated log-probabilities from an HMM-based speech recognizer, for example, can be used. For the project, word-level HMMs of select words were trained using data from the three non-test partitions to serve as the UBM. To provide the speaker models, these HMMs were then adapted to each speaker in the test set using segments from the 8 conversations specified for training. In the event that the training data provided no instances of a given word for adaptation, the UBM was simply used for the speaker as well, causing the log-likelihood ratio to be 0 and effectively removing the influence of the word from the test score.

3.2. Word selection

The words used for recognition came from the set of common discourse markers, filled pauses, and backchannels and are as follows: **{actually, anyway, like, now, okay, right, see, uh, uhuh, um, well, yeah, yep}**. The 13 words listed account for approximately 6% of the total tokens. As previously mentioned, these words possess the qualities that they generally occur with great frequency in the conversation sides and that they may possess strong speaker-distinctive attributes given their spontaneous nature.

3.3. Word-level recognizer

Forced alignment was performed on the test segments using word-level recognizers for the selected words based on HMMs produced by the HMM Toolkit (HTK)[6]. The feature vectors for these HMMs consisted of 12 cepstra, their

first differences, and the zeroth order cepstrum as an energy parameter. The vectors represented 10ms frames of speech with a 25ms Hamming window. In order to isolate the sections of speech corresponding to words of interest, time alignments obtained from a forced alignment of the SRI recognizer with the true transcription were used. The words were then extracted either using built-in features of HTK (as in the case of the background models), or manually (as with adaptation and testing).

The prototype HMMs used for background model training were simple left-to-right HMMs with self-loops and no skips. Each state had an output distribution modeled as a mixture of four Gaussians with diagonal covariance matrices. The number of states for each model was the smaller of the number of phones in the word multiplied by 3 and the median duration, as expressed in frames, divided by 4.

The speaker models were obtained via adaptation of the background models using Maximum A Posteriori (MAP) adaptation of means as given by:

$$\hat{\mu}_{jm} = \frac{N_{jm}}{N_{jm} + \tau} \bar{\mu}_{jm} + \frac{\tau}{N_{jm} + \tau} \mu_{jm} \quad (3)$$

where τ is a weighting of the a priori knowledge to the adaptation speech, N is the occupation likelihood of the adaptation data, μ_{jm} is the speaker independent mean and $\bar{\mu}_{jm}$ is the mean of the observed adaptation data.

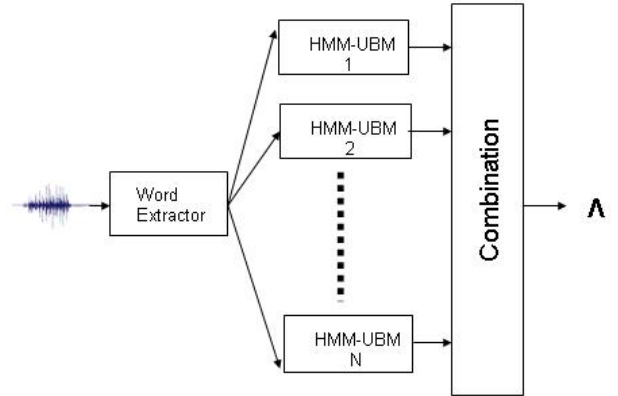


Fig. 1. Basic system layout

3.4. Detection procedure

The detection process involved a scoring for each test segment/target pair. For each of the selected words appearing in

the test segment, a target score was designated as the accumulated log-probability output from the appropriate HMM when forced alignment recognition was performed. The UBM score was similarly obtained by performing forced alignment with the non-adapted HMM. Scores were then combined to produce a final normalized score in one of three ways. For a frame-normalized score, the sum of all of the target scores was subtracted from the sum of all of the UBM scores and divided by the total number of frames. For a word-normalized score, frame normalization was performed at the word level and these scores were then averaged. Finally, for the n-best score, frame normalization was performed on the n best matching (i.e., highest log-probability) words. The basic system is indicated in figure 1.

4. RESULTS

The following table shows the equal error rate for scoring using various normalization methods.

Normalization method	EER
frame	2.87%
word	2.87%
2-best	17.23%
4-best	10.36%
8-best	6.96%

The first item of note is the equality of EER in the frame and word normalization methods, suggesting that either method is suitable. Also of interest is the trend of decreasing EER with increasing n in the case of n-best normalization. This suggests that the system's performance benefits from an increase in available data. The corresponding DET plots for these methods are shown in figure 2. In terms of relative performance, the most comparable system is a text-constrained GMM system described by Sturim et al. [7]. The system used a bank of GMM-UBM likelihood detectors rather than HMM detectors and yielded an EER of 1.3%. The word pool was significantly larger (50 words), however, and channel normalization was used. It is then possible that the HMM-UBM system could produce comparable performance with the appropriate modifications.

A rough analysis of the discriminative capability of the individual words in the set can be made by looking at the EER for each word. Figure 3 shows the sorted EERs for these words. For the majority of the words, the EERs produced lie within a small range of $\pm 3\%$ of 7%. This indicates that as a group these words share some qualities, which was indeed the initial supposition. It is only the last two words (**anyway**, and **yep**) which differ substantially, and this is likely partly due to the paucity of data observations for these words. It is of particular note that the word **yield-**

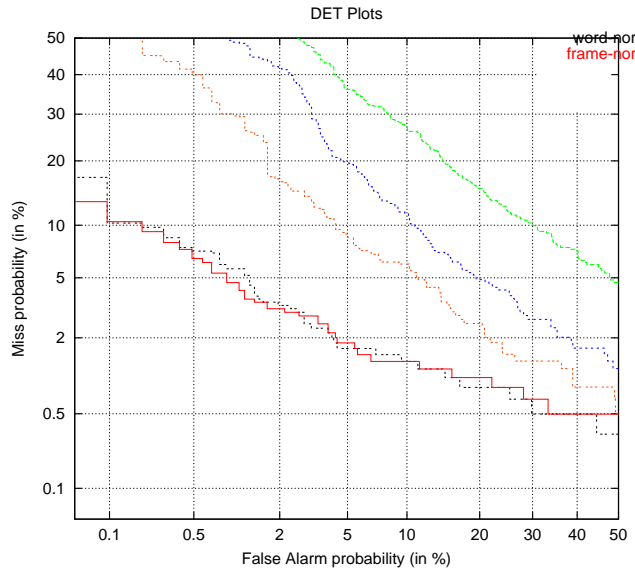


Fig. 2. DET plot for various normalization methods

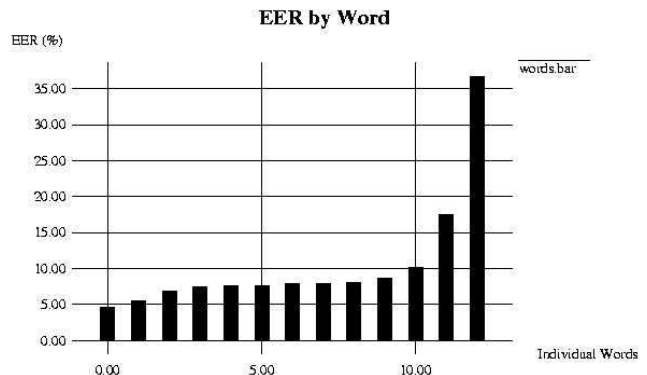


Fig. 3. EER for individual words

ing the best performance, **yeah**, produced an EER of 4.63% compared with 2.87% for the entire set.

5. CONCLUSIONS AND FUTURE WORK

In this paper, a text-dependent speaker recognition system using HMMs has been presented. Applying various normalization methods, best performance was shown to be obtained from the use of either frame or word normalization, both yielding an EER of 2.87%. The project demonstrates that it is indeed possible to apply a text-dependent system to an unconstrained corpus and obtain low EERs.

In the future, a number of modifications to the system could be made. For one, channel normalization via cepstral mean subtraction is likely to improve performance. In addition, the influence of word context could be examined. The word **well**, for example, occurs both as a discourse marker and within its standard adverbial context. It could be argued that these two realizations differ and that one may be more speaker-distinctive than the other. Also, the word list could be revised as the EERs for some words differed greatly from the majority, or perhaps expanded to increase data availability, from which the system appears to benefit.

6. ACKNOWLEDGEMENTS

I would sincerely like to thank Barbara Peskin, with whom I consulted extensively for this project. I would also like to thank Chuck Wooters and Yang Liu for their assistance with various technical aspects of the project.

7. REFERENCES

- [1] Larry Heck, "Integrating High-Level Information for Robust Speaker Recognition," presentation from WS2002 at JHU.
- [2] D.A. Reynolds, T.F. Quatieri, R.B. Dunn, "Speaker Verification Using Adapted Gaussian Mixture Models," Digital Signal Processing 10(1-3):19-41 (2000).
- [3] Frederick Weber, Barbara Peskin, Michael Newman, Andrés Corrada-Emmanuel, and Larry Gillick, "Speaker Recognition on Single- and Multispeaker Data," Digital Signal Processing 10(1-3):19-41 (2000).
- [4] NIST Speaker Recognition website
<http://www.nist.gov/speech/tests/spk/2001>
- [5] Doug Reynolds, Walter Andrews, Joseph Campbell, Jiri Navratil, Barbara Peskin, Andre Adami, Qin Jin, David Klusacek, Joy Abramson, Radu Mihaescu, Jack Godfrey, Doug Jones, Bing Xiang, "The SuperSID Project: Exploiting High-level Information for High-Accuracy Speaker Recognition," from WS2002 at JHU.
- [6] HTK Book is available at
<http://htk.eng.cam.ac.uk/docs/docs.html>
- [7] D.E. Sturim, D.A. Reynolds, R.B. Dunn, T.F. Quatieri, "Speaker Verification using Text-Constrained Gaussian Mixture Models," ICASSP, pp. 677-680, 2002.