Text-Constrained Speaker Recognition on a Text-Independent Task

Kofi Boakye and Barbara Peskin

International Computer Science Institute
Berkeley, CA 94704 – USA
{kaboakye,barbara}@icsi.berkeley.edu

Abstract

We present an approach to speaker recognition in the text-independent domain of conversational telephone speech using a text-constrained system designed to employ select high-frequency keywords in the speech stream. The system uses speaker word models generated via Hidden Markov Models (HMMs) — a departure from the traditional Gaussian Mixture Model (GMM) approach dominant in text-independent work, but commonly employed in text-dependent systems — with the expectation that HMMs take greater advantage of sequential information and support more detailed modeling which could be used to aid recognition. Even with a keyword inventory that covers a mere 10% of the word tokens and a system that does not yet incorporate many standard speaker recognition normalization schemes, this approach is already achieving equal error rates of 1% on NIST's 2001 Extended Data task.

1. Introduction

Speaker recognition systems in which the permissible utterances are limited to a fixed inventory — so-called "text-dependent" systems — generally achieve much higher accuracy than text-independent systems. By fixing the utterance, much more detailed modeling is possible, with acoustic variation much more likely to arise from speaker differences than from the variability due to the phonetic content of the utterance. Despite this well-known performance advantage, there are numerous applications where constraining the speakers' utterances is neither feasible nor desirable. Examples include speaker indexing of audio archives, background verification during commercial interactions, and forensic and security applications involving found speech. This paper addresses the question: Is it possible to capitalize on the advantages of text-dependent systems in such domains?

One possibility is to limit the words of interest to those occurring with high frequency in the domain, so that they are likely to be encountered in unconstrained speech. For this paper, in which we consider NIST's 2001 Extended Data task as a testbed, the domain of interest is conversational speech. Another desirable characteristic of the selected words is that they have high speaker-discriminative power. For conversational speech, it has been suggested [1] that words which are highly spontaneous, representing habitual speaking style, may possess this characteristic. We therefore chose to focus attention on common discourse markers (*you know*, *like*), filled pauses (*uh*, *um*), and backchannels (*uh-huh*, *right*), as natural candidates.

In current text-independent speaker recognition systems the standard practice is to generate speaker models using Gaussian Mixture Models (GMMs), with target models adapted from a Universal Background Model, as described in [2]. These systems utilize a "bag of frames" approach in which input speech frames are assumed to be essentially independent. Such systems simply model a "generic" speech frame, and so fail to take advantage of sequential information and of more focused modeling, which might be used to aid in recognition. A natural choice to capture greater sequential information and more tightly focused speech states is to use Hidden Markov Models (HMMs) for model generation. HMMs have been employed in the context of text-independent speaker verification systems before (e.g. [3] [4] [5] [6]), but these systems are generally based on simple monophone models or on broad phonetic classes in order to ensure coverage of the full large-vocabulary, unconstrained speech domain as well as adequate training of the speaker models. In contrast, in the work reported here we build word-specific models for a highly limited subset of the words. This allows us to create much more focused models — in the style of text-dependent, password-based systems — to assess the benefits of more traditional "text-constrained" approaches.

This paper describes a first attempt to build such a "text-constrained" system, where the speaker models consist of whole-word models represented by adapted word-level HMMs for a speech recognizer. We have attempted to keep as much as possible the same general framework as in the standard text-independent GMM systems, but now employing HMM-based whole-word models. In particular, we use the same front-end processing, train speaker-independent background models and derive the target models via adaptation, etc., in order to provide the cleanest comparison.

In the sections that follow, we first review the Extended Data speaker recognition task and then describe the design and implementation of our text-constrained system. We then present a series of experiments detailing the development stages of the current system, and indicating the value of the various enhancements. Fusion of scores with those of other systems is examined next. Comparisons of this system to other closely related approaches, such as a GMM-based "text-constrained" system [7] for this domain, are then discussed, and next steps are outlined.

2. The NIST Extended Data task

Experiments for the proposed system were based on the Extended Data task of the 2001 NIST Speaker Recognition Evaluation [8], a text-independent single-speaker detection task using data obtained from the Switchboard-I corpus. The corpus consists of recordings of approximately 2400 telephone conversations among 543 speakers (302 male, 241 female).

In the evaluation, speaker models are trained using 1, 2, 4, 8, and 16 complete conversation sides and are subsequently tested on a complete conversation side, where each conversation side is 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 Extended Data task is to permit (indeed, to encourage) the investigation of techniques which examine phenomena existing on longer timescales (e.g., prosodic features, word usage, etc. [9]) and those involving longer-term statistics, and as such rely on more training data. In our case, the use of extended data 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.

For training and testing, a jack-knifing process is used whereby the data is partitioned into 6 sections (or "splits") of approximately equal size and each partition is tested independently; when one partition is being tested, data from the others can be used for background modeling and for normalization. In this paper, we present results for development stages of the system on split 1, using splits 4, 5, and 6 for background modeling. Results over all six splits using the above-mentioned jacknifing are reported for the final system.

In addition, although the evaluation allows for a varying number of training conversations, we focus here on models trained using 8 conversation sides. This has become the standard testing ground for data-hungry techniques, as it provides the most training data while still involving a reasonably-sized speaker population. Performance is reported both in terms of the usual Detection Error Tradeoff (DET) curve [10] as well as the simple summary statistic, Equal Error Rate (EER).

3. System design and implementation

The general procedure for the speaker recognition task is to compute a score for each test/target trial that represents a likelihood of the test segment having been uttered by the putative target talker. This typically is captured through a log-likelihood ratio for the test segment X, $LLR(X)$, given by

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

where S represents a speaker-specific target model and UBM is a Universal Background Model. The proposed system generates scores based on the accumulated log-probabilities from a collection of HMM-based word models. Whole-word models are trained for select keywords using data from the held-out "splits" to serve as a universal background model. Target versions of these models are then created by adaptation to each speaker using word instances extracted from the target talker's 8-conversation-side training. The use of the log-likelihood ratio for scoring and the UBM and target-adapted speaker models is designed to parallel as much as possible the standard GMM framework.

3.1. Feature extraction

The feature vectors for the HMMs in the baseline system consisted of the first 12 mel-cepstral coefficients plus the zeroth-order cepstrum, and their first differences. In order to isolate the sections of speech corresponding to words of interest, time alignments were obtained from a forced alignment of the truth transcripts to the speech stream, as generously provided by SRI's Decipher speech recognizer [11] through the prosodic feature database supplied by SRI to the JHU Summer Workshop WS'02 [9]. (note: We also report on the contrast using ASR output rather than truth transcripts, below.) The frame sequences corresponding to the words of interest, as identified through the forced alignment times, were then extracted to be used for the remaining processing.

3.2. Word selection

The words used for recognition in the baseline system came from the set of common discourse markers, filled pauses, and backchannels, and are as follows: {*actually, anyway, like, now, okay, right, see, uh, uhhuh, um, well, yeah, yep*}. These 13 words account for only about 6% of the total tokens in the corpus. As previously mentioned, these words possess the qualities that they are likely to occur with great frequency in the conversation sides and that they may possess strong speaker-distinctive attributes given their habitual, spontaneous nature. It should be noted, however, that many of these words — especially discourse markers such as *well*, *like* and *see* — also commonly occur in other roles (e.g., *He did well*, *I'd like to see that*) and no attempt was made to select only those occurrences functioning as discourse markers. The hope was that the modeling and scoring would be sufficiently robust to this mixed population, as further discussed below.

3.3. Training and adaptation

Once frame sequences for the keywords have been extracted, a speaker-independent UBM is trained for each of these words. The prototype whole-word HMMs were simple left-to-right state sequences with self-loops and no skips. Each state has an output distribution modeled as a mixture of four Gaussians with diagonal covariance. The choice of four Gaussians was made with the objective of keeping the number small enough to give good speaker focus, but large enough to account for the natural acoustic variation of the word, including variations due to potentially different word usage, as noted above. The number of states for each model was heuristically defined to be the smaller of the number of phones in the standard pronunciation of the word, multiplied by 3, and the median duration, as expressed in frames, divided by 4. The somewhat ad-hoc nature of these model design parameters certainly merits further analysis.

The speaker models were obtained via adaptation of the background models using Maximum A Posteriori (MAP) adaptation of the model means. In the event that the training data for a speaker provided no instances of a given word for adaptation, the unadapted UBM was simply used for the speaker model as well, causing the two scores (speaker and UBM) to cancel and effectively removing the influence of the word from the overall test score. All keyword modeling and subsequent scoring was performed using the HMM Toolkit, HTK [12].

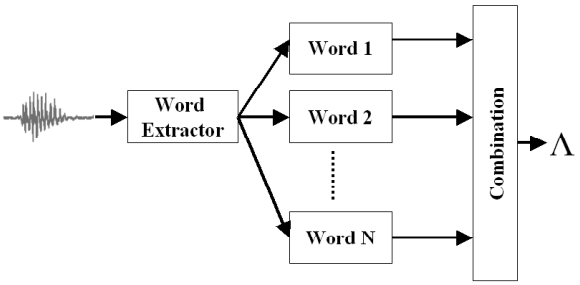

Figure 1: *System architecture.*

3.4. Scoring

The detection process must generate a score for each test/target trial. For each of the selected words appearing in the test segment, a target score was computed as the accumulated frame scores over all test instances when the appropriate speaker-adapted HMM was force-aligned to the extracted frame sequence. The UBM score was similarly obtained by performing forced alignment against the unadapted HMM. Scores were then combined over all keywords to produce a composite score, the difference between the composite target and composite background score was computed, and the resulting value was frame-normalized by dividing by the total number of frames in the word instances. The basic system is indicated in figure 1.

4. Experiments

In this section, we describe a sequence of experiments performed on the 2001 Extended Data testset. The experiments proceed as a series of modifications to the baseline system with a view to analyzing the influence of different components on performance. Results are summarized in table 1, giving equal error rates. As mentioned previously, development results are reported over split 1 only, with the final system results being given over all six splits.

System	EER (%)
baseline	2.87
base + additional words	2.53
base + higher cepstra	1.88
base + CMS	1.35
combined (true trans)	1.01
combined (ASR output)	1.01
final (true trans)	1.06
final (ASR output)	1.25

Table 1: *System performance. The first six entries give results for split 1 alone and the last two entries are for all splits.*

4.1. Experiment 1: baseline system

This initial experiment evaluates the baseline system as described in section 3. The EER achieved by the system is 2.87%, surprisingly good performance given the relative simplicity of the system and the small percentage of data utilized.

4.2. Experiment 2: additional words

In this experiment we expand the keyword list to include a number of common backchannel and discourse marker word bigrams. Added to the list are: {*you_know, you_see, i_think, i_mean, i_see, i_know*}. With these 6 additional bigrams the total coverage of tokens increases to 10% from 6%. Each of these word pairs was treated as a single entity with a single HMM "word" model for each. As a consequence of these additions, the EER is reduced from 2.87% to 2.53%. The DET curves are compared in figure 2.

A rough analysis of the discriminative capability of the individual words (and phrases) in the set can be made by looking at the EER for each when tested in isolation, in conjunction with its frequency of occurrence, as shown in figure 3. The two statistics should be viewed jointly, as the individual EERs alone convolve speaker-characterizing power with word frequency. For the majority of the words, the EERs produced lie

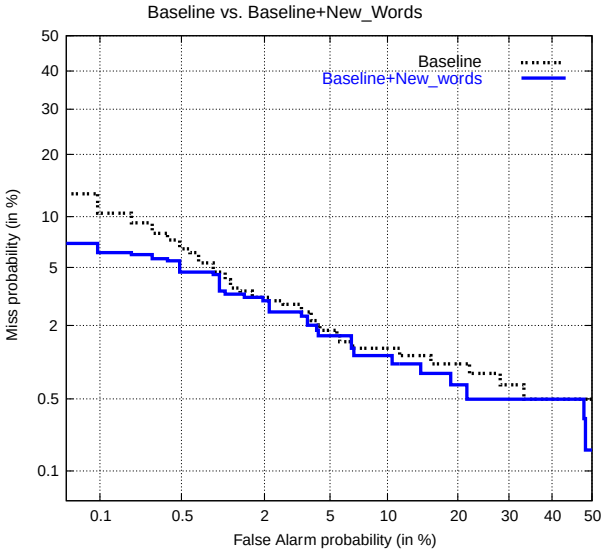

Figure 2: *Baseline versus additional words.*

within a small performance range at around 7% even though the word frequencies vary significantly. Only the last two entries in both the single-word and word-pair groupings have performance differing markedly, and this is likely due to the paucity of data observations for these words, as indicated. It is of particular note that the word yielding the best performance, *yeah*, alone produced an EER or 4.63% compared with 2.53% for the entire set.

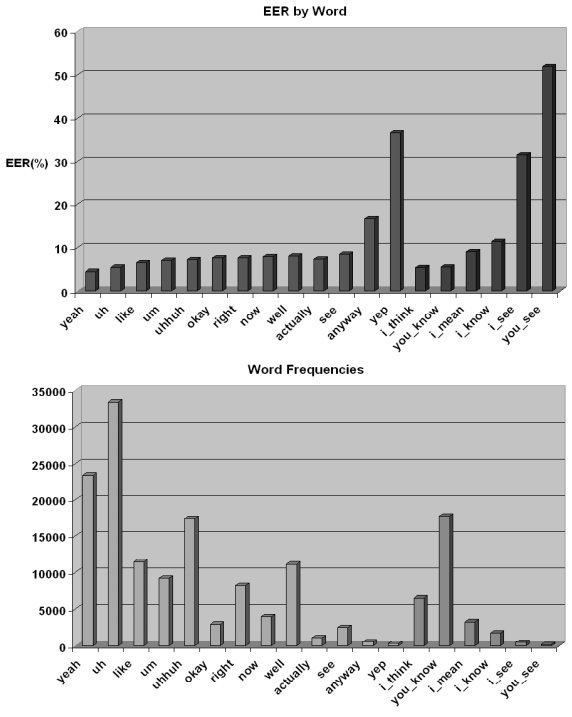

Figure 3: *Individual word/phrase EERs and frequencies.*

4.3. Experiment 3: higher-order cepstra

The inclusion of higher-order cepstral coefficients in the acoustic feature vector has been shown to improve performance in numerous speaker recognition systems. These coefficients may possess more speaker-sensitive information (e.g. regarding pitch) so we next assessed their impact on the baseline system. The input features were extended to include 19, rather than 12, mel-cepstra, and their first differences. The result is a reduction in EER of about 1% from the baseline to 1.88%, as shown in figure 4.

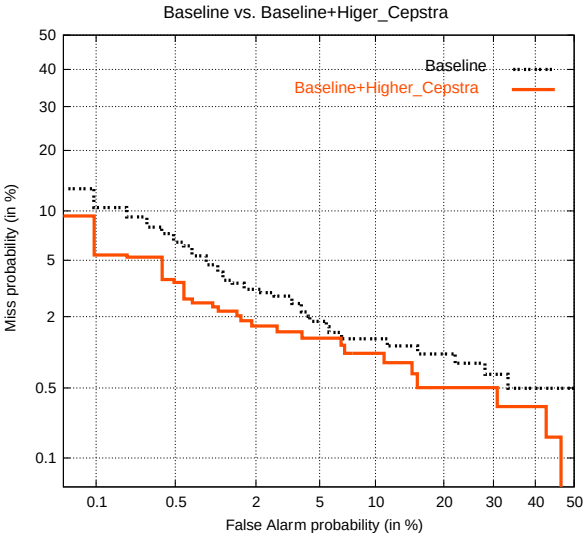

Figure 4: *Baseline versus higher order cepstra.*

4.4. Experiment 4: cepstral mean subtraction

The baseline system did not include any form of channel normalization. In this experiment we process the features by performing cepstral mean subtraction (CMS) on a per conversation-side basis. CMS seeks to reduce undesirable variability introduced by the channel (e.g., the same speaker on different handsets) and as a result we see a large reduction in EER: from 2.87% to 1.35%. However, examination of the DET curves displayed in figure 5 shows that the performance does degrade in the very low false alarm region. This is likely due to the very small number of test trials represented in this region of the curve, as corroborated by later experiments involving all splits.

4.5. Experiment 5: combined system

This experiment involves the combination of the previous modifications into a single system, incorporating the expanded keyword list, higher-order cepstra, and cepstral mean subtraction. The resulting EER is 1.01%, indicating that the information obtained through the various modifications is, to some extent, complementary. The composite DET curve, along with the contributing stages, is provided in figure 6. This overall performance is especially impressive in light of the standard system features that have not yet been employed (e.g. various normalization schemes such as Z-norm, H-norm, or T-norm), the design choices that have not yet been optimized, and the small percentage of each conversation (only about 10% of the word tokens) that contributed to the system scoring.

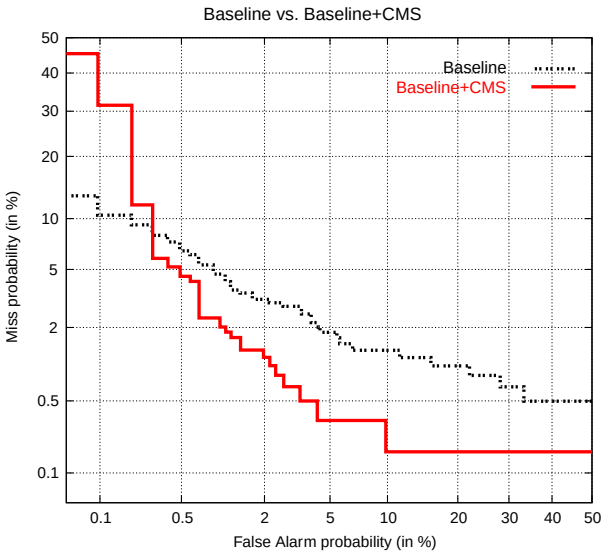

Figure 5: *Baseline versus CMS.*

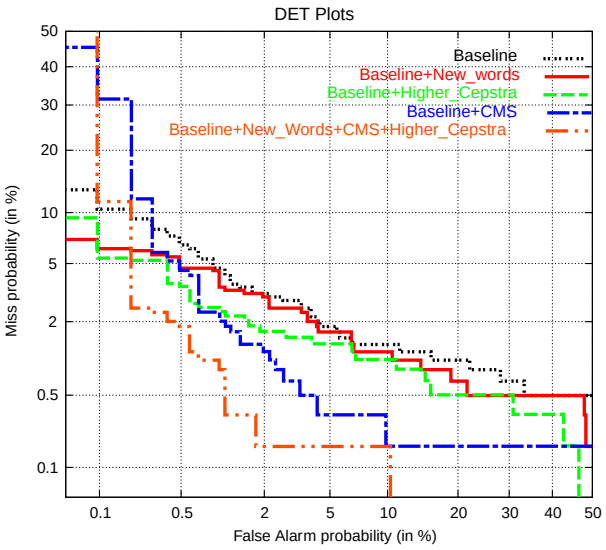

Figure 6: *DET curves for all experiments.*

4.6. Experiment 6: ASR transcription

In all of the experiments reported above, the word extraction was based on a forced alignment of the speech stream using the true (i.e., human-generated) transcription. While such an analysis is useful for validating the technical approach, any real-world implementation would necessarily rely on automatic speech recognition (ASR) output rather than expert human transcription. In this experiment we repeat the combined system experiment above, with the modification that ASR output is used to identify the keyword intervals both in training and in test. This system employs recognition output made available by SRI for the JHU 2002 Summer Workshop, using a somewhat simplified version of their Switchboard-trained recognizer [11], which achieved a word error rate of approximately 30% on the Extended Data conversations. The speaker recognition EER that results is equivalent to that obtained previously, although the

DET curves do differ somewhat (see figure 7), suggesting good performance of such a system even in the case of fully automatic transcription.

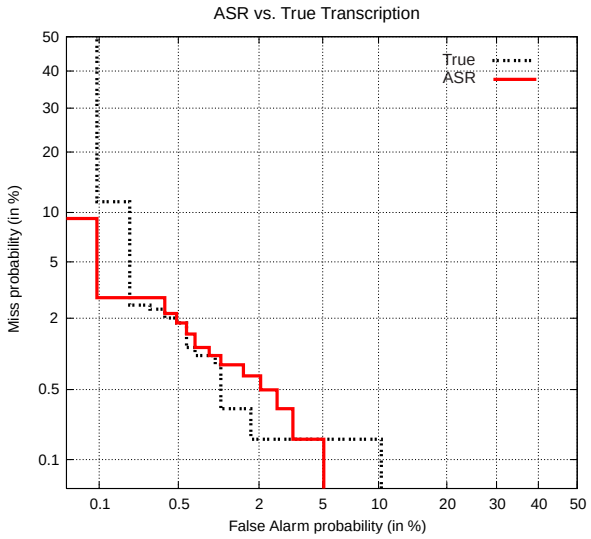

Figure 7: *ASR versus true transcription.*

4.7. Experiment 7: Final system performance

For this experiment, evaluation of the final combined system was expanded to cover all six splits of the corpus. This is contrasted with analysis of only split 1 for the previous experiments, which was considered sufficient for development. Figure 8 shows the DET curves for the system using both the ASR and true transcriptions. The EER for ASR output is 1.25% and that for the true transcriptions is 1.06% (each compared with 1.01% for split 1 alone). Again we see little difference between ASR and true transcription results. Also note the poor performance in the low false alarm region is no longer evident, suggesting the phenomenon was related to the smaller number of trials for a single split.

5. Score Fusion

We next examine how our keyword system would combing with markedly different knowledge sources. Table 2 shows the fusion results for three other systems with our own. Here the fusion consisted of a simple linear combination of scores. The

System	EER alone	EER w/ fusion
GMM	0.90	0.52
Full LM	9.71	0.81
Keyword LM	18.43	0.98

Table 2: *Fusion performance. The second column gives system EER percentage in isolation and the third gives percentages when fused with the keyword system. Results are reported over all splits.*

GMM system is a version provided by SRI and appears to benefit considerably from the fusion. The Full Language Model (LM) system listed is the bigram modeling developed by Doddington in his idiolect work [13]. What is referred to as the

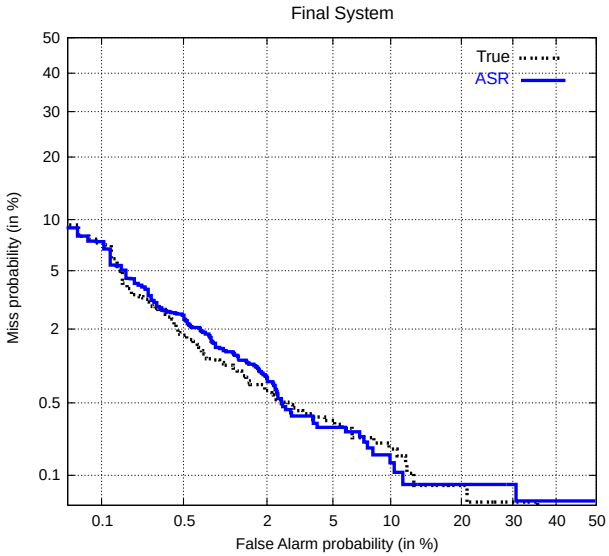

Figure 8: *Final system (ASR and true transcription).*

"Keyword LM" in the table is a simple model involving only relative frequencies among the keywords. This seeks to capture speaker preference among these keywords. Both of these systems fuse effectively. Indeed, the language information they contain is considerably different from the acoustic information of either the GMM or our proposed system; the LM systems model word preferences, the keyword system the acoustics of what a speaker sounds like when he says them. Such results provide further motivation for the exploration of extended data techniques.

6. Discussion

Our keyword system is still preliminary, with many system features not yet incorporated and many design choices not yet optimized. But despite its provisional nature, the system already gives evidence of the speaker-characterizing power of certain common, highly habitual words and phrases.

A few variations were explored to address some of the perceived shortcomings of the design. For example, we tried weighting the contribution of individual word scores to the composite score by word rather than by frame, with essentially no change in performance. We also experimented with using only the top N best-matching word tokens (for various values of N) to try to address, among other issues, the problem of mismatch due to different word contexts and uses (such as the use of *like* as a discourse marker vs. as a verb), but this generally degraded performance, presumably because the system is already fairly starved for data and the modeling supported enough freedom to handle variations.

It is interesting to compare this system to the text-constrained GMM system introduced by Sturim et al. [7]. The latter system also uses a shortlist of keywords from which it extracts acoustic frames and uses only those frames in building and scoring more conventional GMM models. It is difficult to compare directly the performance of the two systems, since they employed largely different word sets with different frequencies of occurrence, so results are somewhat conflated with coverage statistics. However, the performance seems generally comparable: in the 1% EER range for this testbed. More careful com-

parison, using the same wordlists, signal processing, and normalizations, as well as an exploration of which types of words are most valuable to each system would be illuminating.

Our approach also bears comparison to other HMM-based systems such as [5] [6]. These systems use HMMs for speaker scoring, but rather than whole-word models, the systems employ monophone models in order to cover all the words in the large-vocabulary task and to allow adequate adaptation of the speaker models. Such systems have the advantage of using all the words in the test data, whereas whole-word models are only feasible for a subset of sufficiently high-frequency words, though they do support more finely focused modeling for this subset.

Finally, it should be noted that this work was inspired in part by work at the JHU 2002 Summer Workshop [14] which also investigated the acoustic match of high-frequency discourse markers, filled pauses, and backchannels. That work used pitch trajectory information, via a smoothed version of F0, as the sole feature and used dynamic time warping (DTW) for the scoring.

7. Future Work

There are a number of respects in which the current system should be improved. These include:

- augmenting the keyword list in a number of different ways, such as using more words from these classes, highest-frequency words in the domain regardless of role, and/or words and phrases that are particularly characteristic for each individual target speaker;
- filtering the keyword occurrences to use only the intended functions (discourse marker, filled pause, backchannel) or building separate word models for the separate functions (as in the *like* example above);
- exploring different choices for HMM topology and for mixture model make-up.

These variations may potentially interact. For example, it may be that if we filter the words by usage, building separate models for each, we can profitably build word models which are even more tightly focused, e.g. employing fewer Gaussians in the speaker's mixture models.

Other system fusions are also planned. For example, we would like to combine this keyword-limited system with a text-independent HMM-based system such as described in [5] [6]. The keyword system allows for much more focused modeling of a handful of words but the latter would augment that by contributing at least some information from all the words.

8. Conclusions

With the availability of the Extended Data task, it has become possible to explore the potential for exploiting the acoustics of common habitual words in speaker recognition systems. In the work reported here, we bring the power of text-dependent modeling to work in a text-independent context, by focusing on frequent, reflexive words and word pairs and modeling them via whole-word HMM models. While this is a first, and still highly preliminary, exploration of these techniques, the very low error rates obtained indicate the promise of such an approach. We look forward to enhancing the current system and to combining it with systems capturing other sources of speaker-characteristic knowledge.

9. Acknowledgments

Special thanks to SRI for the use of their prosodic feature database and GMM system scores. Additional thanks to Dan Gillick of ICSI for assistance with score fusion. This research was supported in part by NSF award IIS-0329258.

10. References

[1] L.P. Heck, "Integrating High-Level Information for Robust Speaker Recognition," presentation at WS'02, Johns Hopkins University, July 2002.

[2] D.A. Reynolds, T.F. Quatieri, and R.B. Dunn, "Speaker Verification using Adapted Gaussian Mixture Models," *Digital Signal Processing*, vol. 10, no. 1-3, pp. 19-41, 2000.

[3] M.J. Carey, E.S. Parris, H. Lloyd-Thomas, and S. Bennett, "Robust Prosodic Features for Speaker Identification," *Proc. ICSLP-96*, vol. 3, pp. 1800-1803, 1996.

[4] J.L. Gauvain, L.F. Lamel, and B. Prouts, "Experiments with Speaker Verification over the Telephone," *Proc. Eurospeech'95*, vol. 1, pp. 651-654, 1995.

[5] M. Newman, L. Gillick, Y. Ito, D. McAllaster, and B. Peskin, "Speaker Verification through Large Vocabulary Continuous Speech Recognition," *Proc. ICSLP-96*, vol. 4, pp. 2419-2422, 1996.

[6] F. Weber, B. Peskin, M. Newman, A. Corrada-Emmanuel, and L. Gillick, "Speaker Recognition on Single- and Multispeaker Data," *Digital Signal Processing*, vol. 10, no. 1-3, pp. 75-92, 2000.

[7] D.E. Sturim, D.A. Reynolds, R.B. Dunn, and T.F. Quatieri, "Speaker Verification using Text-Constrained Gaussian Mixture Models," *Proc. ICASSP-02*, vol. 1, pp. 677-680, 2002.

[8] NIST 2001 Speaker Recognition website: http://www.nist.gov/speech/tests/spk/2001.

[9] D.A. Reynolds, *et al.*, "The SuperSID Project: Exploiting High-level Information for High-Accuracy Speaker Recognition," *Proc. ICASSP-03*, vol. IV, pp. 784-787, 2003.

[10] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki, "The DET Curve in Assessment of Detection Task Performance," *Proc. Eurospeech'97*, vol. 4, pp. 1895-1898, 1997.

[11] A. Stolcke, *et al.*, "The SRI March 2000 Hub-5 Conversational Speech Transcription System," *Proc. NIST Speech Transcription Workshop*, College Park MD, 2000.

[12] HMM Toolkit (HTK): http://htk.eng.cam.ac.uk/

[13] G. Doddington, "Speaker Recognition based on Idiolectal Differences between Speakers," *Proc. Eurospeech'01*, vol. 4, pp. 2521-2524, 2001.

[14] A. Adami, R. Mihaescu, D.A. Reynolds, and J.J. Godfrey, "Modeling Prosodic Dynamics for Speaker Recognition," *Proc. ICASSP-03*, vol. IV, pp. 788-791, 2003.