USING PROSODIC AND LEXICAL INFORMATION FOR SPEAKER IDENTIFICATION

Frederick Weber[1], *Linda Manganaro*[1], *Barbara Peskin*[1,now 2], *Elizabeth Shriberg*[2,3]

[1]Dragon Systems/Lernout and Hauspie, Newton, MA
[2]International Computer Science Institute, Berkeley, CA
[3]Speech Technology and Research Laboratory, SRI International, Menlo Park, CA

ABSTRACT

We investigate the incorporation of larger time-scale information, such as prosody, into standard speaker ID systems. Our study is based on the Extended Data Task of the NIST 2001 Speaker ID evaluation, which provides much more test and training data than has traditionally been available to similar speaker ID investigations. In addition, we have had access to a detailed prosodic feature database of Switchboard-I conversations, including data not previously applied to speaker ID. We describe two baseline acoustic systems, an approach using Gaussian Mixture Models, and an LVCSR-based speaker ID system. These results are compared to and combined with two larger time-scale systems: a system based on an "idiolect" language model, and a system making use of the contents of the prosody database. We find that, with sufficient test and training data, suprasegmental information can significantly enhance the performance of traditional speaker ID systems.

1. INTRODUCTION

The task of speaker identification in recorded speech is frequently pursued using purely acoustic techniques, commonly Gaussian Mixture Models (GMMs) [1]. The most obvious advantages of GMMs are their simplicity and robustness to short-length recordings. These characteristics reflect the model's assumption that every 10–20 msec frame of speech can be treated independently. This works well when the test recording is only a few seconds long. However, as the amount of test and training data increases it becomes attractive to make use of speaker-specific characteristics which involve larger time scales, such as prosodic patterns. In this paper we will describe two standard speaker ID systems: a Gaussian mixture model (GMM) approach and a system based on large vocabulary continuous speech recognition (LVCSR), which were developed as part of Dragon Systems' regular participation in the annual NIST Speaker ID evaluations [2]. We then examine how larger time-scale information can be added to enhance the performance of these baseline systems.

The data set used for this study comes from the Extended Data Task of the 2001 NIST Speaker ID evaluation, described in [3]. This task is based on the Switchboard-I (SWB-I) corpus of recorded telephone conversations [4], and consists of a series of trials, where the speaker ID system is presented with a test recording from an unknown speaker and training data from a hypothesized target speaker. The system is then required to provide a score reflecting its belief that the test came from the target. In this task, a "test" is an entire SWB-I conversation side, averaging roughly 3 minutes of speech, and between one and sixteen conversation sides (up to an hour of speech) of training data are provided from the hypothesized target speaker. This scale of test and training data has not been used in recent NIST evaluations and made this task a useful testbed for the study of the interaction between short- and long-time-scale features.

Another attraction of the Extended Data Task was the availability of SRI's prosody database of SWB-I conversations [5], making it possible to explore prosodic features for this study. The use of prosody for speaker ID is not a new idea. For example, SRI fielded a system incorporating prosodic features in the 1998 NIST speaker ID evaluation [6], and publications on prosody-based speaker ID go back at least 30 years [7]. The novelty of our system lies in our access to an unusually wide variety of prosodic indicators via SRI's database, coupled with the availability of test and training data based on entire conversation sides.

We contrast the results of our prosody-based speaker ID approach with a system which identifies speakers based on word usage, or "idiolect", alone. This language model (LM) approach was developed by G. Doddington [8], and he has kindly provided us his results for this study.

We begin with a brief description of our baseline GMM and LVCSR systems and their performance on the Extended Data Task. Doddington's "idiolect" approach is next described and interpolated with the GMM and LVCSR results. We then discuss our prosody-based system and its interaction with the other three systems.

2. BASELINE SYSTEMS: GMM AND LVCSR

Dragon Systems' GMM and LVCSR-based speaker ID systems are described in detail in [2]. We provide here an outline of their operation and performance on the Extended Data Task.

The GMM system consists of a single mixture of 4096 components representing a generic 10 msec frame of speech from a given target speaker. A universal background model (UBM) is first constructed from roughly 2 hours of gender-balanced speech taken from the Switchboard-II corpus of recorded telephone conversations. Speaker-specific target models are then generated using Baum-Welch adaptation of the UBM to the target's training data. For each hypothesized target-test combination specified by NIST's test protocol, the log likelihood score for the adapted target model and for the UBM are computed on the test data. An energy-based silence detector is applied to each test, and frames falling below an energy threshold are not included in the score computation. The raw score is computed as the difference between UBM and target model scores, normalized by the number of test frames. We further normalize the score to take into account several sources of variation, such as handset differences, using the HNORM technique described in [9].

The LVCSR-based system attempts to capture more about the structure of a target's speech characteristics, rather than modeling a "generic" speech frame. The system starts with a full recognition pass over the test and training data, using a slightly simplified version of Dragon's 1998 HUB5 evaluation recognizer [10]. From the recognition pass, we obtain errorful transcripts and associated forced time alignments, assigning each speech frame to a phoneme state. The speaker ID models are monophone acoustic models, and are used to rescore the test speech, given the frame labels from the recognition pass. As was done for the GMM, we construct a speaker-independent UBM from independent data, and train target speaker ID models using Baum-Welch adaptation. The speaker ID score for a target-test combination is then computed as the frame-averaged score difference between the UBM and target models, normalized using HNORM.

The results of the LVCSR-based system on the Extended Data Task are compared to the GMM results in Fig. 1, using the Detection Error Tradeoff (DET) curve format [11]. It is clear that the additional information available from the LVCSR model significantly improves speaker ID performance in the Extended Data Task environment. This is contrasted with earlier GMM and LVCSR results on 3 and 30 second tests from the NIST 1998 Speaker ID Evaluation [2], where the LVCSR system lags behind the GMM approach until at least 30 seconds of test data are available for analysis.

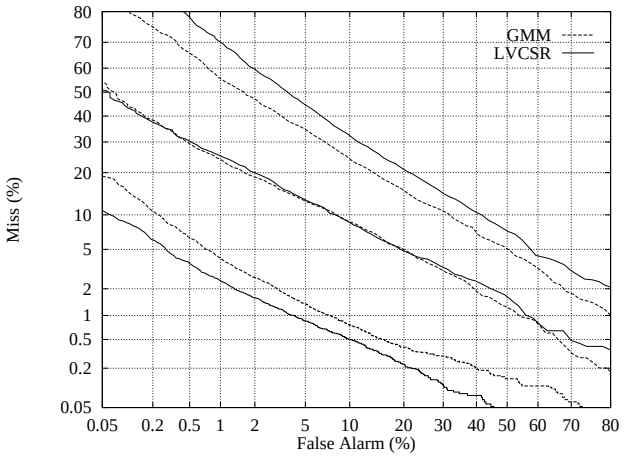

Fig. 1. GMM and LVCSR system results on Extended Data Task (lowest pair), contrasted with results on 30 sec tests (middle) and 3 sec tests (upper pair) from the 1998 NIST Evaluation.

3. SPEAKER ID USING A LANGUAGE MODEL

The length of the test segments and quantity of training data available in the Extended Data Task support the use of larger-scale structures for speaker ID. One possibility is to construct a speaker-specific, or "idiolect", language model.

G. Doddington has performed such an LM-based study, described in [8]. He constructed bigram language models from each target speaker's training data, and scored these models and the corresponding speaker-independent model on the target-test combinations in the Extended Data Task. The score he assigned was the log of the likelihood ratio of target and background values.

We have used logistic regression to determine the optimal linear interpolation weights of the GMM, LVCSR, and LM scores on the Extended Data Task. A summary of the results obtained from these approaches appears in Fig. 2. Not unexpectedly, the LM results alone, which use no acoustic information at all, significantly underperform both GMM and LVCSR results. More surprisingly, we see only a small gain in overall performance when the LM is interpolated with the GMM and LVCSR systems.

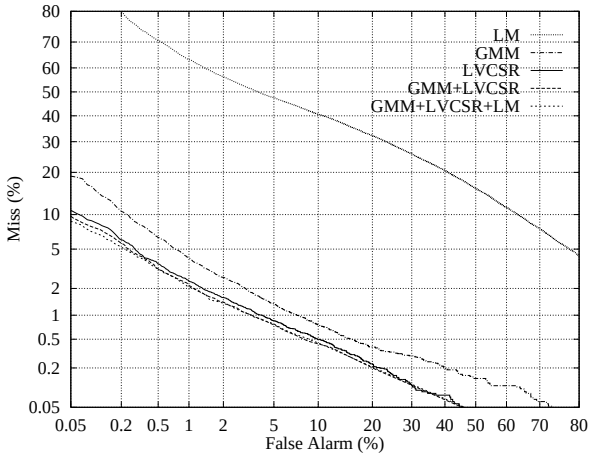

Fig. 2. Extended Data Task results for LM, GMM, and LVCSR separately, and successive interpolations of GMM+LVCSR, GMM+LVCSR+LM (Plot lines appear in order specified by key).

Our expectation was that the accuracy of the "idiolect" model would be a strong function of the amount of training data. We therefore considered two extremes of training data volume, namely for 1–2 training sides and for 8–16 training sides. The results of interpolating the GMM and LM systems for these two extremes are summarized in Fig. 3, and show clearly that the benefit of incorporating an LM increases substantially with larger training data volumes. A smaller gain is seen for the interpolation of the LM with the combined GMM+LVCSR system, but the trend with data volume is the same.

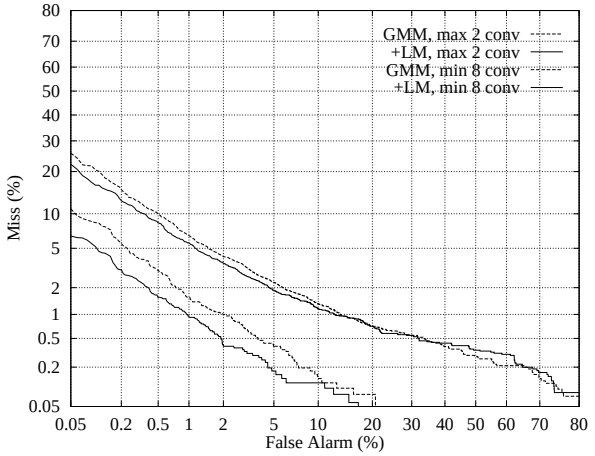

Fig. 3. Extended Data Task results from interpolation of LM scores with GMM system, broken out by amount of training data.

4. PROSODIC INDICATORS OF SPEAKER IDENTITY

To provide more speaker-specific characteristics, our next step was to incorporate prosodic information into our systems. We used a detailed prosodic feature database created by SRI, originally developed for an NSF project on utterance segmentation. Thus, the features were not optimized for Speaker ID. The database contains data drawn from two corpora, SWB-I and Broadcast News.

In SRI's database, the transcribed, time-aligned speech is associated with raw and derived prosodic features, including pause and segmental durations, voicing information and pitch-based features, with scales ranging from the subword level to the conversation level. Pitch information includes both actual F0 values and values based on a piecewise linear stylization. Besides prosodic variables, the database includes lexical annotations of phenomena such as disfluencies, yielding a total of approximately 120 variables per conversation side.

Our prosodic speaker ID system is based on a set of features, each of which is reduced to a single value per conversation side. Some quantities are taken as is from the SRI database; others are derived from its contents. The features that we use can be roughly divided into four types, some prosodic and others lexical (we will refer to them all as "prosodic"):

- Pitch-related features, such as the mean and standard deviation of raw and stylized versions of the pitch track;
- Duration-related features, such as the mean and standard deviation of the word and phone lengths;
- Indicator-word usage, as given by the relative frequencies of specific words like *I, okay, yeah, uhhuh, right*;
- Conversational-style features, such as pause and turn lengths, and the relative frequency of disfluency classes such as pause-fillers (e.g. *uh*, *um*), discourse markers (e.g. *you know*), backchannel expressions (e.g. *all right, sure*), editing markers (e.g. *I mean*), conjunctions, sentence fragments.

Altogether we use up to 48 predictors.

Our prosodic "model" for a target speaker simply uses the means of the predictor values over the assigned training conversations. For most predictors, the score for a given test conversation is defined as the normalized squared difference:

$$score = \sum_{predictors} w_i \frac{(x_i - \mu_i)^2}{\sigma_i^2} \qquad (1)$$

where x_i is the value of the ith predictor from the test conversation, μ_i is obtained from the target's training data, and weights w_i are obtained from logistic regression. Because of the paucity of training data, the standard deviations σ_i are computed from the deviations over the training data from *all* speakers in the corpus. (The σ_i could equally be absorbed into the weights w_i.)

The specific-word predictors are a special case. These can occur very rarely and are not well-modeled by equation (1), hence we use a Poisson distribution instead. The frequency of a word for a given speaker model is computed by aggregating all that speaker's training conversations before computing the mean predictor rate. If the word never occurs in the training data, we assign a count of 0.5. The score is then the log of the Poisson probability.

5. PERFORMANCE ON SWB-I DATA

SRI's database covers roughly half of the conversation sides in SWB-I. We built a subset of the Extended Data Task from the 2001 NIST evaluation, using only the SWB-I conversation sides present in the SRI database. We also removed any trial whose target model used one or more training conversation sides not present in the SRI database. We will subsequently refer to this reduced data set as the "prosody subset". Unfortunately, this process retained only 7679, or 13%, of the original 58642 trials. In particular, there was virtually no representation of trials involving target models built with 8 and 16 training conversation sides, where we expect to see the greatest benefit from suprasegmental information.

No independent development data was available in the Extended Data Task to train the weights w_i. We therefore derived them from the same data we used for our final tests, using a jackknifing approach to avoid "cheating". To train the weights, we removed any trials involving that target's data or data from any speaker used as an impostor for that target model. Unfortunately, this process has the highly undesirable effect of completely removing any speaker-characteristic information from the weights.

The results of the prosody system alone are compared to the LM results in Fig. 4, and can be seen to be comparable in discriminating power. The effect of linear interpolation of the LM and prosody scores is also displayed, and demonstrates that the two systems indeed capture some independent information.

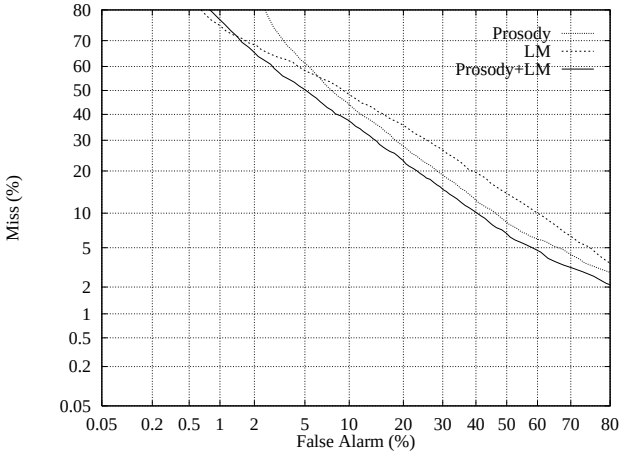

Fig. 4. Results of prosodic speaker ID system compared to and interpolated with the LM approach on the prosody subset.

We note that, as part of the logistic regression optimization, we allow the statistics package to select the "best" subset of predictors, both here and in the experiments below. Among the prosodic predictors routinely selected as valuable are

- the relative frequency of disfluencies of all types
- the average word duration
- the relative number of "long" pauses, defined as being over 150 msec in length
- the relative number of "sentence-like" boundaries [12]

The above features consistently appear as the top prosodic predictors, both in a prosody-only model and in combination with GMM, LVCSR, and LM scores. In addition, the predictors related to pitch and specific word identity are generally useful.

We repeated the interpolation via logistic regression, adding in the GMM and LVCSR model scores as well. A close-up of the results from sequentially adding GMM, LVCSR, LM, and prosody

together appears in Fig. 5. We can see evidence for a modest enhancement of performance with the interpolated GMM+LVCSR +LM+Prosody system over the GMM+LVCSR combination alone.

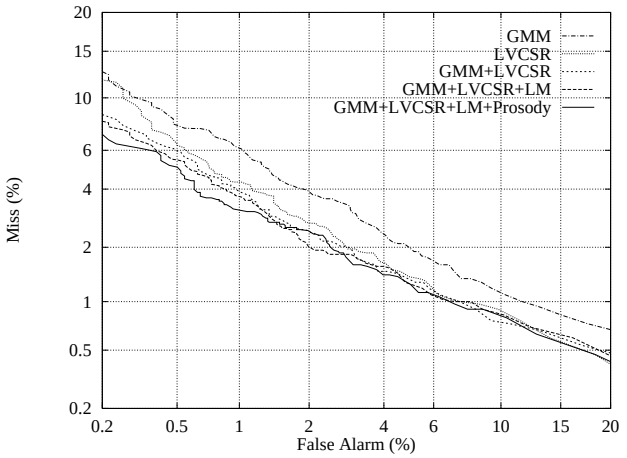

Fig. 5. Closeup of results of successive interpolation of GMM, LVCSR, LM and prosodic systems on the prosody subset of the Extended Data Task (Note: Plot lines appear in order specified by key).

Several factors encourage us to believe that the potential gains from prosody-based speaker ID are considerably larger than seen in this preliminary study. We believe the most significant performance limitations are the loss of the target models that are based on 8 and 16 training conversation sides and the lack of a development set with which to train speaker-specific feature weights w_i. Another constraint is the small size of the prosody subset, coupled with the small number of errors ($\approx 2\%$ equal error rate) made by the baseline GMM and LVCSR systems. This limits our statistical sensitivity to further performance improvements from the introduction of suprasegmental information.

Beyond overcoming the limitations imposed on the study by the data set, there are a number of ways the prosodic speaker ID system could be improved. The features of the SRI database were not selected with speaker ID in mind; we anticipate that further optimization is possible. We would also like to make better use of within-conversation distributions of prosodic predictors, rather than reducing them to single values for a conversation side. Improvements could be made in the robustness and accuracy of our pitch tracking system, particularly given the problems of background noise and narrow telephone bandwidth. Finally, more sophisticated approaches to combining information sources could be considered, rather than simple linear interpolation of the system scores.

6. CONCLUSIONS

We have found the Extended Data Task of the NIST 2001 Speaker ID evaluation to be an interesting testbed for comparing our GMM and LVCSR speaker ID systems with systems based on suprasegmental information. As we saw with the GMM and LM systems, the performance improvements from the introduction of large timescale information are greatest when large amounts of training data are available.

The constraints of the prosody subset limit the visibility of potential benefits of incorporating prosodic information. Despite these limitations, the incorporation of our preliminary prosodic system has already provided a modest enhancement of our baseline system performance. We believe that suprasegmental information will have a valuable role to play in future speaker ID efforts.

Acknowledgements: We would like to thank G. Doddington for generously providing his "idiolect" results, and K. Sönmez for his assistance in working with the SRI database.

7. REFERENCES

[1] D. Reynolds, "Robust text-independent speaker identification using gaussian mixture speaker models," *Speech Communication*, vol. 17, pp. 91–108, 1996.

[2] F. Weber et al., "Speaker recognition on single- and multi-speaker data," *Digital Signal Processing*, vol. 10, pp. 75–92, 2000.

[3] NIST 2001 Speaker ID Evaluation protocol, *http://www.nist.gov/speech/tests/spk/2001/index.htm*.

[4] J. Godfrey et al., "Switchboard: Telephone speech corpus for research and development," in *Proc. ICASSP*, San Francisco, CA, March 1992, pp. 517–520.

[5] E. Shriberg et al., "Prosody-based automatic segmentation of speech into sentences and topics," *Speech Communication*, vol. 32, pp. 127–154, 2000.

[6] K. Sönmez et al., "Modeling dynamic prosodic variation for speaker verification," in *Proc. ICSLP*, Sydney, Australia, December 1998, pp. 3189–3192.

[7] B. S. Atal, "Automatic speaker recognition based on pitch contours," *JASA*, vol. 52, pp. 1687–1697, 1972.

[8] G. Doddington, "Speaker recognition based on idiolectal differences between speakers," in *Proc. Eurospeech*, Aalborg, Denmark, September 2001, pp. 2521–2524.

[9] D. Reynolds, "Comparison of background normalization methods for text-independent speaker verification," in *Proc. Eurospeech*, Rhodes, Greece, September 1997, pp. 963–966.

[10] B. Peskin et al., "Improvements in recognition of conversational telephone speech," in *Proc. ICASSP*, Phoenix, AZ, March 1999, pp. 53–56.

[11] A. Martin et al., "The DET curve in assessment of detection task performance," in *Proc. Eurospeech*, Rhodes, Greece, September 1997, pp. 1895–1898.

[12] M. Meteer et al., "Disfluency annotation stylebook for the Switchboard corpus," Distributed by LDC, *ftp://ftp.cis.upenn.edu/pub/treebank/swbd/doc/DFL-book.ps*, February 1995.