USING PROSODIC AND CONVERSATIONAL FEATURES FOR HIGH-PERFORMANCE SPEAKER RECOGNITION: REPORT FROM JHU WS'02

Barbara Peskin[1], Jiri Navratil[2], Joy Abramson[3], Douglas Jones[4], David Klusacek[5], Douglas A. Reynolds[4], Bing Xiang[6]

[1]International Computer Science Institute, [2]IBM T.J. Watson Research Center, [3]York University, [4]MIT Lincoln Laboratory, [5]Charles University, [6]Cornell University

ABSTRACT

While there has been a long tradition of research seeking to use prosodic features, especially pitch, in speaker recognition systems, results have generally been disappointing when such features are used in isolation and only modest improvements have been seen when used in conjunction with traditional cepstral GMM systems. In contrast, we report here on work from the JHU 2002 Summer Workshop exploring a range of prosodic features, using as testbed NIST's 2001 Extended Data task. We examined a variety of modeling techniques, such as *n*-gram models of turn-level prosodic features and simple vectors of summary statistics per conversation side scored by k^{th} nearest-neighbor classifiers. We found that purely prosodic models were able to achieve equal error rates of under 10%, and yielded significant gains when combined with more traditional systems. We also report on exploratory work on "conversational" features, capturing properties of the interaction across conversation sides, such as turn-taking patterns.

1. INTRODUCTION

State-of-the-art text-independent speaker recognition systems have traditionally used short-term low-level acoustic features, such as cepstra, with modeling via gaussian mixture models (GMMs) for the speakers [1]. Such systems perform very well, especially given limited training and test data. However, they fail to model information about the speaker at many levels that might contribute to speaker recognition, such as word usage, prosodic characteristics, etc. While there is a long tradition of exploring higher-level features for speaker recognition – especially the use of pitch and other prosodic markers (see, e.g., [2,3,4,5]) – systems incorporating them generally require significantly more data for adequate training (or impose other constraints, such as text-dependency). Consequently, their effectiveness has been limited in evaluations such as NIST's annual series of speaker recognition tests, where systems have only 2 minutes of training data and 3-30 seconds of test.

In 2001, in response to growing interest in the use of higher-level features, NIST introduced the Extended Data task [6] based on the Switchboard-I corpus of conversational telephone speech. Unlike the traditional speaker recognition tasks, the Extended Data task provided multiple whole conversation sides for speaker training (for up to about 45 minutes of speech) and tested on whole conversation sides, thus enabling research on larger-scale features. As described in the overview paper [7], a working group at the Johns Hopkins 2002 Summer Workshops – the SuperSID team – assembled to systematically explore a wide range of features for Speaker ID using the Extended Data task as its testbed.

This paper and the companion paper [8] describe the group's explorations of prosodic features. Here we focus on investigations of a diverse collection of prosodic features spanning many types of pitch, energy, and duration predictors. We also include a look at modeling conversational patterns. The companion paper [8] describes experiments specifically exploring the modeling of pitch dynamics.

Our work on prosodic features was made possible in large part through the availability of SRI's Prosodic Feature Database. This rich resource, originally designed for work on prosodic predictors for topic and sentence segmentation [9], includes frame-level pitch information for the full Switchboard-I corpus (raw pitch values, as well as SRI's piecewise linear stylization of pitch contours and their lognormal tied-mixture (LTM) models for pitch – see [10]). In addition, it provides parallel arrays giving per-word values for a wide range of prosodic features, based on "truth" transcriptions force-aligned to the speech stream and on automatic speech recognition (ASR) output.

2. PROSODIC SUMMARY STATISTICS

To explore the value of various prosodic indicators to the speaker recognition task, we conducted a simple experiment using statistics collected on a per-conversation-side basis.

The main idea was, for each conversation side in the Switchboard-I corpus, to obtain a vector of N features capturing various prosodic characteristics. Each such vector – i.e. each conversation side's statistics – can then be viewed as a point in an N-dimensional feature space, and the T training conversations per target speaker form a "cloud" of T such points. Given a test conversation, we then compute the distance of its corresponding point to this target cloud and employ a k^{th} nearest-neighbor classifier, comparing the distance to the target speaker cloud vs. the average distance to the data clouds for a collection of cohort (impostor) speakers. The test/trial score is then the log likelihood ratio of target and average cohort distances.

2.1. Prosodic Features

We examined a total of 19 prosodic features, some directly provided in the SRI database (designed for speaker

normalization, but consequently providing good summary statistics for speakers' prosodic baselines) and others derived from the word-level data provided. The features fell into three main groups:

6 related to word, phone, and segmental durations:
- log(#frames/word) averaged over all words
- log(#phones/word) averaged over all words
- log(#frames/phone) averaged over non-silence phones
- (#frames/phone)/(corpus average for that phone), averaged over non-silence phones
- average length of word-internal voiced segments
- average length of word-internal unvoiced segments

5 related to pause durations and frequency:
- relative frequency of "medium" (7-15 frame) pauses
- relative frequency of "long" (16-99 frame) pauses
- log(pause duration) averaged over "long" pauses
- relative frequency of "turn-size" (>= 100 frame) pauses
- log(pause duration) averaged over "turn-size" pauses

8 related to pitch:
- log(mean_F_0) averaged over all words with (good, i.e. not halved or doubled) voiced frames
- log(max_F_0) averaged as for log(mean_F_0)
- log(min_F_0) averaged as above
- log(range_F_0) averaged as above
- pitch "pseudo-slope": (last_F_0 - first_F_0) / (#frames in word), averaged as above
- average (per word) slope over all segments of piecewise linear stylization of F_0
- model mean for log(F_0) in the LTM model for pitch
- the triple of weights (prob halving, prob whole, prob doubling) in the LTM pitch model.

For those features described as "averages", we computed both mean and standard deviation. The distance metric used in the nearest-neighbor calculation was then a symmetrized Kullback-Leibler distance computed from these statistics.

2.2. Experimental Results

We first explored the value of the various features in isolation, using only splits 1-3 of the Extended Data task (with splits 4-6 used to provide the background/cohort speakers) and using only k=1 in the nearest-neighbor calculation. The results are provided in Table 1, giving equal error rates (EER) for each of the features for the 8-training-side condition.

We then examined various combinations of features. Table 2 gives EER's for various groupings of features using all 6 splits combined. For the results in the table, fusion of individual features was performed at the score level for each split, using a single-layer perceptron with weights trained from the 5 held-out splits. The table shows the merged result for the 6 splits for the 8-training-side condition, using k=3 in the nearest-neighbor computation, which improved slightly over the k=1 result. (Splits 1-3 used speakers in splits 4-6 for their cohort set and vice versa.)

feature	*EER (%)*
log(#frames/word)	30.7
log(#phones/word)	40.4
log(#frames/phone)	24.1
normalized #frames/phone	29.7
voiced segment length	30.7
unvoiced segment length	33.9
medium pause rate	43.3
long pause rate	36.1
log(long pause duration)	38.9
turn-size pause rate	39.7
log(turn pause duration)	40.5
log(mean_F_0)	19.4
log(max_F_0)	20.2
log(min_F_0)	19.6
log(range_F_0)	31.5
pitch "pseudo-slope"	31.8
PWL slope average	32.0
mean of LTM pitch model	21.7
half/whole/double weights	30.8

Table 1. EER's for individual prosodic features, 8-conv training (splits 1-3, 1st nearest-neighbor)

feature set	*EER (%)*
6 word, phone, segment durations	18.9
5 pause durations and rates	25.2
8 pitch features	14.8
11 duration features (6 speech + 5 silence)	15.2
all 19 features	8.1

Table 2. EER's for prosodic feature combinations, 8-conv training (splits 1-6, 3 nearest-neighbors)

We examined several different forms of fusion (the single-layer perceptron used above, simple linear average of scores with various weightings, feature vector concatenation, etc.) and different values of *k* in the nearest-neighbor computation. While the actual EER values varied somewhat with the configuration, the general patterns held firm.

The results provided above used features drawn from the version of the SRI database based on truth transcripts and forced alignments. For completeness, we also ran the analogous experiments for the ASR version of the database. Not surprisingly, there was very little change in most results, since these features are largely independent of the actual word identities, using words primarily to chop up the data into units over which pitch and duration features were averaged. Most EER's based on ASR output fell within a point of those from truth transcripts. Overall the EER from the entire 19-feature ensemble went from 8.1% for the truth transcripts to 8.9% for the ASR-based features.

2.3. Discussion

Although the value of individual features, as presented in Table 1, varies a great deal (with pitch features generally doing

best and pause features worst), taken together the set of pitch features and the set of duration features (including word, phone, segment, and pause) performed about the same: roughly 15% EER for each class. Further, the information provided by the pitch and duration ensembles was highly complementary; when the two sets were combined the error was nearly halved, achieving about 8% EER altogether. Clearly, even features that on their own are not strong predictors may contribute to what is in aggregate a powerful speaker discrimination system.

While these results are not as strong as those of more traditional systems using cepstral features or phone decodings, the use of prosodic features clearly adds new and useful information for the speaker recognition task. Indeed, as described in the overview paper [7], this system was often a contributor to the most profitable fusion systems, adding novel information not captured in more traditional acoustic approaches.

3. CONVERSATIONAL PATTERNS

We were also interested in the question of whether the system could learn a speaker's "conversational style" – how he interacts with his conversational partners, through turn-taking patterns, prosodic features within a turn, etc. – rather than using a simple flat vector of summary statistics accumulated over the full conversation side.

3.1. Motivation: A Turn-Taking Model

Our explorations in this area were motivated by work of Douglas Jones of MIT/LincolnLab, reported at NIST's 2002 Speaker Recognition Workshop. Jones modeled duration of speaker turns, alternating target talker with nontarget speaking partner (inferring turn-length for nontarget by the length of silence between target turns).

These turn durations were quantized using a log scale and then the conversation as a whole was encoded as a sequence of tokens of the form $T\text{-}D_1$ $N\text{-}D_2$ $T\text{-}D_3$ $N\text{-}D_4$ ..., alternating target and nontarget durations, labeled by side. The target model was trained as a simple bigram model on such tokens, collected over the target's training conversations, and the speaker recognition system employed a log likelihood ratio of the target bigram score vs. the score for a background model trained from the merged data from a collection of "background" speakers.

This basic model could be enhanced by expanding the target tokens to include a tag for the (log-scaled, quantized) number of words (W) and/or number of characters (C) in the target turn transcripts. (Such information is unavailable for the nontarget partner under the rules of the evaluation.) This yields an enriched sequence of the form: $T\text{-}D_1W_1C_1$, $N\text{-}D_2$, $T\text{-}D_3W_3C_3$, $N\text{-}D_4$, $T\text{-}D_5W_5C_5$, ...

This simple turn-taking model was able to achieve an EER of 26% on the Extended Data task, using a bigram token model with 8-conversation training. (The EER was 3-4% higher if using only the duration without the text content expansion.)

3.2. Technical Approach

For the Summer Workshop, we examined a much larger range of features under a somewhat expanded technical plan. As in the turn-taking model above, we continued to focus on turn-level values and on *n*-gram modeling.

We examined a set of 33 features extracted on a per-turn basis, including features characterizing

- *turn-length*: number of words, phones, frames, non-silence frames, and voiced frames
- *speaking rate*: phones per second, both for the whole turn and excluding silences
- *durations*: average and standard deviation of word and silence lengths in frames, number of phones per word
- *pitch and energy values*: average and standard deviation for all frames, and for rising or for falling frames
- *pitch and energy dynamics*: number of frames where pitch (or energy) is rising, number of segments in the piecewise linear stylization, etc.

Feature values were smoothed using linear interpolation between turn value and conversation-side average.

This 33-dimensional feature vector was then reduced to a lower-dimensional one. We first examined an approach using linear discriminant analysis, but found that we were better served by a naïve search of various feature combinations.

Next, the resulting vectors were quantized. Again, we examined different approaches: using deciles in the feature distributions for each component to quantize the data, or clustering feature vectors using a GMM trained from held-out data and then assigning each vector the index of its maximum likelihood component. We explored GMMs ranging from 4 to 128 components in order to study the trade-off between feature-space resolution and *n*-gram model size.

The resulting quantized token sequences were then used to create *n*-gram models. We examined unigrams, bigrams, and trigrams, and looked at sequences only involving target turns and at target-nontarget alternations. (For the latter – unlike in formal evaluations – we allowed ourselves to look at the wavefile for the nontarget side of the conversation.)

3.3. Experimental Results

Table 3 shows the features we found (through heuristic search) to perform best, alone and in various combinations. The results are provided for the decile-quantized data, and unigram target-only models.

feature	*EER (%)*
phones/sec (excluding silence)	26.9
average log energy	28.6
average log pitch	30.2
average word length	34.2
average number of words	35.6
log pitch + log energy	21.0
log pitch + log energy + phone rate	16.7
log pitch + log energy + word length	18.0
all 5 features above	18.3

Table 3. EER's for best-performing turn-level features, 8-conv training (test on split 1, with 4-6 for background)

Table 4 compares unigram, bigram, and trigram results, as well as the cost of decreasing the amount of training data, for the best

feature combination seen above: the pitch + energy + phone rate combination. The results are again for target-side *n*-grams only and here use a 128-component GMM for vector quantization.

n-gram	*2-conv*	*4-conv*	*8-conv*
unigram	20.9	17.3	15.2
bigram	20.7	16.2	15.7
trigram	---	18.0	16.3

Table 4. EER's for turn-level *n*-grams, as *n* and amount of training vary (test on split 1, with 4-6 as background)

We also explored various ways of using information from the nontarget side, both through alternating *n*-grams as in Jones' work and by building models for target features conditioned on features from the nontarget side. However, in the experiments during the Workshop we were unable to improve on the pure-target results: models incorporating nontarget features generally performed several percentage points worse than pure-target models.

3.4. Discussion

The above provides only a sampling of the many experiments performed over the course of the JHU Workshop. Other experiments included an investigation of word usage models conditioned on the conversational partner's usage and experiments with higher-order *n*-grams and with nontarget conditioning using binary-tree models, but none thus far have improved on these results.

This work provides a better turn-level feature baseline, but not yet a benefit from incorporating turn-taking patterns relative to that baseline. This may be due in part to the insufficient sampling of the "conversational space" available in Switchboard – too few speaking partners, or conversations too short or artificial to elicit natural speaking partner influences. It may also suffer from the automatic turn labeling provided in the SRI database, which did not conform to the turn units as labeled by human transcribers. In this respect, Jones' use of silences may be a more promising mechanism for automatic turn delimiting. Nonetheless, we find the area of style-related features an intriguing area for research and believe that it offers significant promise both for the speaker recognition task and for the finer-grained task of identifying particular conversations or conversational segments of interest.

4. CONCLUSIONS

The work of the SuperSID team on prosodic and conversational features surveyed a variety of prosodic predictors and of speaker modeling techniques. We were able to achieve equal error rates in the single digits on NIST's Extended Data task using purely prosodic models involving simple vectors of conversation-level summary statistics.

The experiments reported here represent a preliminary look at prosodic modeling using easily extracted (or already surfaced) features from the SRI database. However, that database was designed for applications to sentence and topic segmentation and so is primarily focused on word-level (actually, "between word") phenomena rather than on characterizing speakers' overall patterns. Despite this limitation, we found the database a wealth of useful information and were able to demonstrate significant progress by incorporating prosodic features into the speaker recognition task. But we believe there are many more features that may be of interest and that should hold greater value for speaker recognition.

Our exploration of turn-taking patterns and conversational "style" is still very much a work in progress. While we were able to markedly reduce the error rate from Jones' initial study by enriching the set of turn-level prosodic features, we have not yet demonstrated any improvement from incorporating knowledge based on interaction with the conversational partner. However, we believe that this is fertile ground for future work. We look forward to continued work in this area and to the collection of new speaker recognition corpora geared toward speaking style that might better support this type of research.

5. REFERENCES

[1] 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.

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

[3] M.J. Carey, E.S. Parris, H. Lloyd-Thomas, and S. Bennett, "Robust prosodic features for speaker identification," *Proc. ICSLP-96*, Philadelphia, Nov 1996.

[4] K. Sonmez, E. Shriberg, L. Heck, and M. Weintraub, "Modeling dynamic prosodic variation for speaker verification," *Proc. ICSLP-98*, Sydney, Dec 1998.

[5] F. Weber, L. Manganaro, B. Peskin, and E. Shriberg, "Using prosodic and lexical information for speaker identification," *Proc. ICASSP-02*, Orlando, May 2002.

[6] NIST 2001 Speaker Recognition Evaluation website: *http://www.nist.gov/speech/tests/spk/2001/index.htm*.

[7] D.A. Reynolds, et al. "The SuperSID Project: Exploiting high-level information for high-accuracy speaker recognition," *Proc. ICASSP-03*, Hong Kong, Apr 2003.

[8] A. Adami, R. Mihaescu, D. Reynolds, and J. Godfrey, "Modeling prosodic dynamics for speaker recognition," *Proc. ICASSP-03*, Hong Kong, Apr 2003.

[9] E. Shriberg, A. Stolcke, D. Hakkani-Tur, and G. Tur, "Prosody-based automatic segmentation of speech into sentences and topics," *Speech Communication*, vol. 32, no. 1-2, pp. 127-154, 2000.

[10] M.K. Sonmez, L. Heck, M. Weintraub, and E. Shriberg, "A lognormal tied mixture model of pitch for prosody-based speaker recognition," *Proc. Eurospeech'97*, Rhodes, Sept 1997.