COMBINING CROSS-STREAM AND TIME DIMENSIONS IN PHONETIC SPEAKER RECOGNITION

Qin Jin[1], Jiri Navratil[2], Douglas A. Reynolds[3], Joseph P. Campbell[3], Walter D. Andrews[4], Joy S. Abramson[5]

[1]Interactive Systems Lab., CMU, [2]IBM T.J. Watson Research Center, [3]MIT Lincoln Lab., [4]Department of Defense, [5]York University
qjin@cs.cmu.edu, jiri@us.ibm.com, dar@ll.mit.edu, jpc@ll.mit.edu, waltandrews@ieee.org, joy@yorku.ca

ABSTRACT

Recent studies show that phonetic sequences from multiple languages can provide effective features for speaker recognition. So far, only pronunciation dynamics in the time dimension, i.e., n-gram modeling on each of the phone sequences, have been examined. In the JHU 2002 Summer Workshop, we explored modeling the statistical pronunciation dynamics across streams in multiple languages (cross-stream dimension) as an additional component to the time dimension. We found that bigram modeling in the cross-stream dimension achieves improved performance over that in the time dimension on the NIST 2001 Speaker Recognition Evaluation Extended Data Task. Moreover, a linear combination of information from both dimensions at the score level further improves the performance, showing that the two dimensions contain complementary information.

1. INTRODUCTION

Previous speaker recognition techniques rely almost exclusively on features representing short-term acoustic information extracted from speech signals, such as spectral energy-based features [1]. These features convey information about the physical traits of the speaker's vocal apparatus, but are known to have a relatively high sensitivity to noise and channel mismatch. In contrast, humans recognize speakers using not only low-level features, but also high-level features, such as word usage (idiolect), pronunciation, prosody, laughter, and other idiosyncratic supra-segmental information. These high-level features often represent learned traits of a speaker related to the speaker's socio-economic status, personality type, education, etc. Higher-level features are expected to be less affected by noise or channel mismatch. For example, people are unlikely to change their idiolectal word usage, their pronunciation idiosyncrasies and their accent when the channel or background noise changes. Furthermore, high-level features should supply complementary information to the low-level features and potentially improve overall recognition accuracy. However, these high-level features are not effectively being exploited in current automatic speaker recognition systems.

Recently published research in [2] described using speaker's idiolectal word usage for speaker recognition. Phonetic speaker recognition approaches explored how to model a speaker's pronunciation patterns using n-grams on phone sequences [3, 4]. In these approaches, pronunciation dynamics were modeled using n-grams on each of the phone streams emitted by each language-specific phone recognizer, which means that these approaches work principally in the time dimension. However, there may be speaker information in the patterns of phone co-occurrences across the multiple phone streams from various languages. In this paper we present an approach and results aimed at modeling the statistical pronunciation patterns across multiple phone streams, which we refer to as phonetic information in the cross-stream (cross-language) dimension.

2. PHONETIC SPEAKER RECOGNITION IN THE TIME DIMENSION

Generally phonetic speaker detection in the time dimension using a single-language phone recognizer is performed in three steps: Firstly, the phone recognizer processes the test speech utterance to produce a phone sequence. Secondly, the test phone sequence is compared to a previously trained hypothesized Speaker Phonetic Model (SPM) and a Universal Background Phonetic Model (UBPM) to compute likelihood scores. Finally, the log of the ratio of the two likelihood scores is computed as the detection score. This process can be expanded to use multiple phone sequences from a parallel bank of phone recognizers trained on different languages. In this case, each phone stream is independently scored and the scores are combined together forming a single weighted detection score.

In all experiments shown in this paper, the phone sequences are produced by open-loop phone recognizers created by Zissman for language identification via Parallel Phone Recognition with Language Modeling (PPRLM) [5]. We used gender-dependent phone streams processed by Kohler in five languages: English (EG), German (GE), Japanese (JA), Mandarin (MA), and Spanish (SP) [3]. All experiments in this paper are conducted on the corpus used for the NIST 2001 Speaker Recognition Evaluation Extended Data Task. Leave-one-out cross-validation is used on a pool of six splits, which are different partitions of the entire Switchboard-I corpus to ensure an adequate number of tests. The detailed description of the experimental setup can be found in [6].

2.1 Speaker Phonetic Model (SPM)

A speaker's language-dependent phonetic model (SPM) is generated using a n-gram language modeling technique. The SPMs used here are bi-gram models created using the CMU-Cambridge Statistical Language Modeling Toolkit (CMU-SLM) [7]. Unlike typical Gaussian Mixture Model-Universal Background Model (GMM-UBM) systems, the n-gram speaker phonetic models are not adapted from the universal background phonetic model, but instead are estimated directly from the speaker's available training data.

2.2 Universal-Background Phonetic Model (UBPM)

The Universal Background Phonetic Model (UBPM) is generated using the NIST control file, which provides a list of hypothesized and test speakers for exclusion from the UBPM training [6]. All

0-7803-7663-3/03/$17.00 ©2003 IEEE

the phone sequences decoded from all of the conversations for the non-excluded speakers were used to build the UBPM using n-gram modeling. Five language-dependent UBPMs trained on the corresponding phone stream are used. The SPMs and UBPMs used in all experiments in this paper, if not mentioned, are bigram models.

2.3 Speaker Detection

Detection is performed using log-likelihood ratios (LLR). Formula (1) defines the LLR detector for a single-language phonetic speaker recognition system, where L_{S_i} is the likelihood score of test sequence X for speaker i's phonetic language model SPM_i and L_U is the likelihood score of test sequence X for the universal background model $UBPM$. The recognition/detection score is the log of the ratio of these two likelihood scores.

$$L_{S_i} = P(X|SPM_i) \qquad L_U = P(X|UBPM) \qquad (1)$$

$$Score_i = \log\left(\frac{P(X|SPM_i)}{P(X|UBPM)}\right) = \log L_{S_i} - \log L_U$$

For a multilingual phonetic speaker recognition system, the scores from each of the languages are fused using a linear combination, such as in formula (2), where k is used to index multiple languages.

$$Score_i^k = \log L_{S_i}^k - \log L_U^k \qquad Score_i = \frac{1}{5}\sum_{k=1}^{5} Score_i^k \qquad (2)$$

2.4 Experimental Results in the Time Dimension

Figure 1 shows the phonetic speaker recognition performance in the time dimension for different training conditions (1, 2, 4, 8, or 16 training conversations [6]). The Equal Error Rate (EER) is 8.4% for the 8-conversation training condition. The 8-conversation training condition is the most representative and statically significant condition in the extended task [6]. The comparison of performance from different approaches will mainly focus on this training condition in following sections

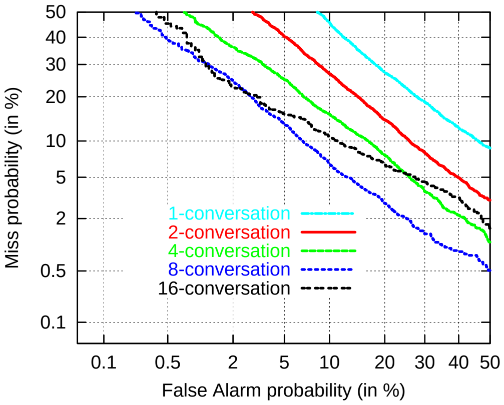

Figure 1: Experimental Results of Phonetic Speaker Recognition in the Time Dimension

3. PHONETIC SPEAKER RECOGNITION IN THE CROSS-STREAM DIMENSION

Since we use multiple tokenizers (phone recognizers) to tokenize the same speech utterance, for the same speaker-specific sound, ideally there should be some fixed set of tokens from each of the multiple languages to represent it, which means that there should exist token dependencies across multiple languages at a given time instance. If we assume that these token dependencies are related to how different speakers realize phonemes, then we can code speaker-dependent pronunciation dynamics across multiple-language phone sequences. Similar to the time dimension, the SPM and the UBPM are created in the cross-stream dimension and detection is done based on the log likelihood ratio. The detailed procedure of phonetic speaker recognition in the cross-stream dimension is described in the following sections.

3.1 Cross-stream Alignment

To discover the underlying dependencies of phones from multiple languages, we need first to align the multiple phone sequences. This alignment is done simply by aggregating all time boundaries from all phone sequences. As illustrated in Figure 2, the phones are duplicated to the smallest time slot in each language in order to unify the boundaries across languages. According to the smallest time overlap across the three languages, the EG phone S originally in [*t1, t3*] is duplicated two times into time slots [*t1, t2*] and [*t2, t3*] and the JA phone B originally in time slot [*t1, t4*] is duplicated three times into time slots [*t1, t2*], [*t2, t3*] and [*t3, t4*]. Similarly, other phones are duplicated into their smallest time slots across the three languages.

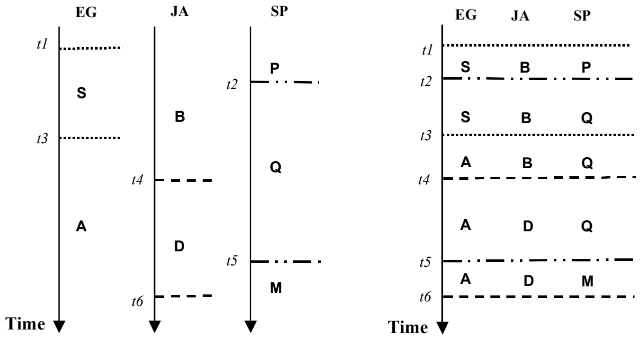

Figure 2: Temporal Alignment of Multiple Phone Sequences

3.2 N-gram Modeling in the Cross-stream Dimension

A straightforward way to model the pronunciation dynamics in the cross-stream dimension is to model the statistical dependencies across streams. For this, as in the time dimension, we use n-grams by treating the aligned phones at each time slot as one input "sentence" for the n-gram modeling. In the above example, we will have five "sentences" to train the n-gram model: "S B P," "S B Q," "A B Q," "A D Q" and "A D M." Since we want to model the bigram dependencies across all streams, it would be better to model all possible pair dependencies. From the above alignment, however, bigrams can only model the dependencies of EG-JA pairs and of JA-SP pairs, but not of EG-SP pairs. Therefore, we simply permute the aligned phones at each time slot, thus modeling all possible pairs from all languages at a given time.

Figure 3 shows the phonetic speaker recognition performance in the cross-stream dimension with different training conditions. Figure 4 compares the performance in the cross-stream vs. time dimension with the 8-conversation training condition. In the cross-stream dimension, experimental results with and without permutation after alignment are shown. For the 8-conversation training condition, the cross-stream system achieves 4.0% EER with permutation and 5.1% EER without permutation; both significantly outperformed the time dimension system, where the EER was 8.4%.

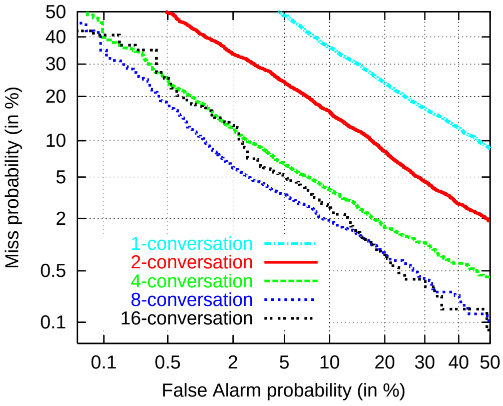

Figure 3: Experimental Results of Phonetic Speaker Recognition in the Cross-stream Dimension

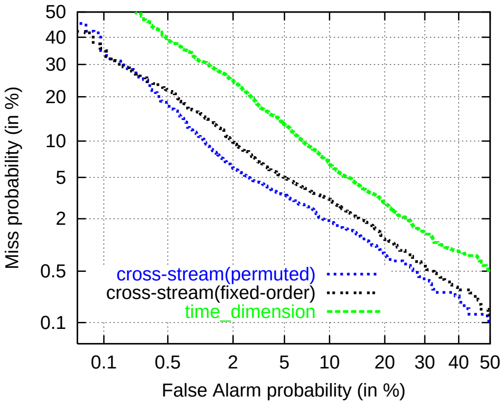

Figure 4: Comparison of Cross-stream vs. Time dimension

3.3 Binary-Tree Modeling in the Cross-stream Dimension

To further explore the cross-stream dependencies, a comparative experiment was carried out using an alternative to the bigram, namely statistical binary-decision tree (BT) models. The BT models were successfully applied to other tasks such as speech recognition [8], language identification [9], and speaker recognition in the time dimension [10]. The training objective is to find a tree structure and a set of binary questions that maximize the training likelihood (or equivalently, minimize the average model prediction entropy). Each binary question involves a predictor; i.e., a context variable from a selected set of variables pre-determined by the user. In time-dimension modeling, the predictors are typically chosen to be the phones preceding the modeled token. When modeling cross-stream dependencies of a particular phone, however, we choose the predictor set to cover the aligned phone tokens from other languages. At each tree node, the BT training algorithm selects from this set the predictor that minimizes the prediction entropy [10]. Preliminary experiments were conducted on the same task and data as with the bigram system to investigate possible future avenues for a more flexible cross-stream information extraction. The number of predictors was varied from two to six and the predictor set was varied to include predictors from all other streams, and alternatively the time slots of the same, one preceding, and one subsequent. Preliminary experimental observations indicate more or less strong degradation in performance of these configurations compared to the bigrams. The degradation seems stronger with purely cross-stream predictor sets and with their maximum number 4.

This finding is in sharp contrast with results obtained with BT models applied in the time dimension where they seem to consistently outperform the bigram performance due to a wider context modeled and the advantage of structural flexibility [10]. To gain insight, we compared average entropy values of token distributions in the permuted bigrams and the BT models on the same training data. The comparison reveals a stark difference: While the standard bigrams provide token distributions with entropy averaging between 2 bits to 4 bits, the purely-cross-stream BT models tend to be significantly “sharper” in predictions with entropy values between 0.5 bits to 1 bit. Obviously, the BT structures adapt to dependencies inherent across the stream in such a way that observing certain contexts causes very accurate prediction in the modeled stream, which indeed is the objective of these models. General dependencies across the phonetic streams, however, may not necessarily imply strong speaker characteristics. In fact, if we could assume that the five tokenizers’ outputs are completely correlated, then the BT will describe these correlations for all speakers identically, i.e., the models will lose speaker discriminancy despite strong dependencies and thus high data likelihood. This leads to an interesting conclusion regarding the cross-stream modeling: The tokenizer-specific phone errors (i.e., substitutions, insertions, and deletions), reflecting each speaker in a different and partially de-correlated way, seem to be the contributing factor in the system. Due to the fixed structure of n-grams and the fact that the bigrams model only pairs of streams, these models do not adapt to the actual (and in this case noisy) data dependencies and hence are not as sensitive to the problem mentioned. This observation further suggests that the original maximum likelihood objective of the BT models, or any adaptive-structure model for that matter, needs to be customized appropriately to reflect the fact that speaker-specific phone dependencies may be contaminated by dominant components carrying irrelevant tokenizer correlations.

4. COMBINATION OF CROSS-STREAM AND TIME DIMENSIONS

Modeling pronunciation dynamics in the cross-stream dimension is expected to carry complementary information to that in the time dimension and, hence, potentially can improve performance when combined. A simple linear combination was used to fuse the recognition scores from both systems. Figure 5 compares the performance of the time dimension system, cross-stream dimension system, and combination systems for the 8-conversation training condition. Bigrams were used in both systems. The EER of the combination is reduced to 3.6%, compared to 8.4% in the time dimension alone and 4.0% in the cross-stream dimension alone. Figure 6 shows the performance of the system using the BT models in the time dimension alone [10], the performance of the system using bigrams in cross-stream dimension alone, and the performance of combining both systems for the 8-conversation training condition. The EER of the combination is further reduced to 3.0%, compared to 3.4% in the time dimension and 4.0% in the cross-stream dimension. Both experimental results indicate that the two dimensions do contain complementary information.

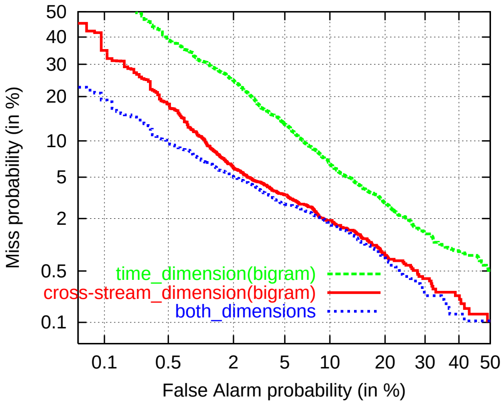

Figure 5: Combination of the Cross-stream Dimension (bigram) with the Time Dimension (bigram)

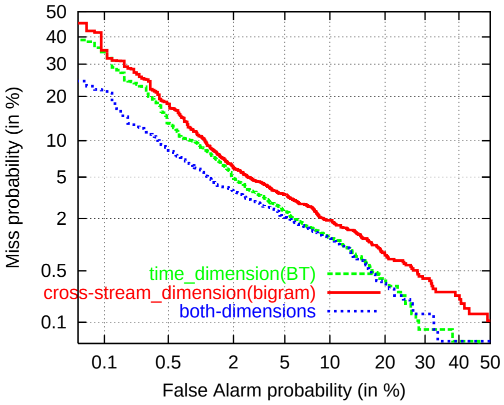

Figure 6: Combination of the Cross-stream Dimension (bigram) with Time Dimension (BT)

As described in previous sections, we modeled the speakers' pronunciation patterns in the time and cross-stream dimensions independently under the assumption that token dependencies are related to how a speaker pronounces specific sounds. However, there may exist token or phone dependencies across both the time and the cross-stream dimensions, such as the EG phone at time t+1 strongly dependent on the JA phone at time t-1. Moreover, our linear combination experiments showed that both dimensions contain complementary knowledge. Therefore, an appropriate modeling approach is desired to more efficiently capture a speaker's distinguishing pronunciation dynamics in both dimensions simultaneously. Graphical models provide a general framework for capturing information from both dimensions simultaneously. Future investigations will apply graphical models for learning the pronunciation dynamics in both dimensions and discovering what underlying dependencies graphical models capture.

5. CONCLUSIONS

In this paper we introduced the concept of phonetic speaker recognition in the cross-stream dimension. Bigram modeling of the phone dependencies across tokenizers in multiple languages achieves 4% EER, a significant improvement over 8.4% EER in the time dimension on the NIST 2001 Speaker Recognition Evaluation Extended Data Task. A linear combination of systems in both dimensions at the score level reduces the EER to 3%, which indicates that the information captured in the cross-stream dimension is complementary to that in the time dimension. Therefore, using graphical models, among others, there is potential for improving the performance further by modeling the information in both dimensions simultaneously. Our preliminary experiments involving binary tree models with adaptive structures applied to the cross-stream dimension indicate the need for a modified training objective as an alternative to the standard maximum likelihood criterion to focus on the speaker discriminative information in this dimension.

6. REFERENCES

[1] D. Reynolds and R. Rose, "Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models," IEEE Trans. on Speech and Audio Processing, Vol. 3, No.1, Jan. 1995.

[2] G. Doddington, "Speaker Recognition based on Idiolectal Differences between Speakers," In Proc. Eurospeech, Aalborg, Denmark, September 2001, Vol. 4, p. 2521-2524.

[3] W. Andrews, M. Kohler, J. Campbell, J. Godfrey, and J. Hernandez-Cordero, "Gender-Dependent Phonetic Refraction for Speaker Recognition," In Proc. ICASSP, Orlando, May 2002, Vol. 1, p. 149-152.

[4] Q. Jin, T. Schultz and A. Waibel, "Speaker Identification Using Multilingual Phone Strings," In Proc. ICASSP, Orlando, May 2002, Vol. 1, p. 145-148.

[5] M. Zissman, "Language Identification Using Phone Recognition and Phonotactic Language Modeling," In Proc. ICASSP, Detroit, MI, May 1995, Vol. 5, p. 3503-3506.

[6] D.A. Reynolds, et al., "The SuperSID Project: Exploiting High-level Information for High-accuracy Speaker Recognition," To appear in Proc. ICASSP, Hong Kong, April 2003.

[7] P. Clarkson and R. Rosenfeld, "Statistical Language Modeling Using the CMU-Cambridge Toolkit," In Proc. Eurospeech 1997.

[8] L. Bahl, P. Brown, P.V. DeSouza, and R.L. Mercer, "A Tree-based Statistical Language Model for Natural Language Speech Recognition," IEEE Trans. on Acoustics, Speech and Signal Processing, July 1989, Vol. 37, p. 1001-8.

[9] J. Navratil, "Spoken Language Recognition – A Step toward Multilinguality in Speech Processing," IEEE Trans. on Speech and Audio Processing, September 2001, Vol. 9, No. 6.

[10] J. Navratil, Q. Jin, W. Andrews, and J. Campbell, "Phonetic Speaker Recognition Using Maximum-Likelihood Binary-Decision Tree Models," To appear in Proc. ICASSP, Hong Kong, April 2003.