GMM/SVM N-BEST SPEAKER IDENTIFICATION UNDER MISMATCH CHANNEL CONDITIONS

Ilija Zeljkovic, Patrick Haffner, Brian Amento and Jay Wilpon

AT&T Labs-Research
Florham Park, NJ 07932, USA

ABSTRACT

Under severe channel mismatch conditions, such as training with far-field speech and testing with telephone data, performance of speaker identification (SID) degrades significantly, often below practical use. But for many SID tasks, it is sufficient to recognize an N-best list of speakers for further human analysis.

We investigate N-best SID accuracy for matched (telephone/telephone) and mismatched (far-field/telephone) train/test channel conditions. Using an SVM-GMM supervector (GSV), pitch and formant frequency histograms (PFH) and cross-channel adaptation using cohorts, we reduced matched channel error rate by over 25% relative to the baseline (GMM-UBM), for top-1, and achieved mismatched N-best accuracy comparable to the baseline.

Index Terms— Speaker identification, GMM, SVM, formants, cohort speaker adaptation

1. INTRODUCTION

A typical Gaussian Mixture Model (GMM) based Speaker Recognition (SR) system performs well under matched channel and low-noise conditions. However, variability between training and testing speech drastically degrades performance of both SR techologies, speaker verification (SV) and speaker identification (SID). Dealing with speech variability is one of the most challenging problems in SR [1]. During the last decade, most of the research on robust SR was focused on speech variability caused by different telephone channels: ISDN, carbon-mic and cell. Augmenting SR with channel normalization and more robust features introduced through SVMs [2, 3, 4], enables adequate performance for telephone-based, speech applications.

Severe speech variations, due to different recording environments and microphones, are just starting to be addressed [5, 1]. Under this type of severe cross-channel mismatch condition, SR accuracy quickly deteriorates. In this paper, we are looking at one type of mismatch, low-quality recordings, such as far-field microphone speech recorded in noisy and reverberant rooms for training, with telephone speech for testing.

Our focus has been on improving SID for SCANMail, a system that provides a graphical user interface to voicemail messages, displaying a rich transcription that serves as an index into the original speech, along with additional information like the caller's name and extracted phone numbers [6]. With the prevalence of VOIP technologies, some of our customers have hands-free microphones, which produce low-quality, reverberant speech that is mismatched from traditional telephone speech. Using the techniques described in this paper we can now train SID models using any audio source to match against incoming telephone-based voicemail messages. This type of mismatch has also been investigated in the area of forensic science [1], where speech from a recorded telephone threat needs to be compared to the available microphone recordings of a set of potential suspects.

Fortunately, in these and many other situations, the SID system is one of a chain of analysis tools with human interfaces, and it is sufficient for the SID system to produce an N-best list of potential matches. When the results are presented in a rich user interface that enhances the N-best list with additional metadata from other analyses, a human can quickly assess the relevance of the candidate speakers, narrowing the choices to the best possible matches.

In this paper we present an analysis of N-best SID under matched and mismatched conditions. We demonstrate that a cross-channel cohort adaptation scheme can significantly improve SID accuracy and that, for an N-best list, it is comparable to matched channel accuracy. We also show that new features comprising pitch and formant histograms can improve matched channel accuracy and greatly contribute to an increase in performance under mismatched conditions.

Using a database of 770 speakers under matched conditions, the baseline accuracy (GMM-UBM system) is 87% and 98% for top-1 and top-50, respectively. The SVM GSV-PFH system increases the corresponding N-best accuracy to 92% and 99%, respectively. When far-field data is used for training and telephone data for testing, the cohort adaptation technique increases the GMM-UBM N-best accuracy, from practically zero, to 50% for top-50 candidates. Applying GSV-PFH further increased the N-best accuracy to about 90% for top-50.

2. TESTING AND TRAIN DATA DESCRIPTION

2.1. Telephone data

Existing datasets for SID proved to be inadequate to comprehensively evaluate our system. To properly validate the utility of our SID system, we collected a large dataset of telephone speech using an automated system that people accessed by calling in from their home, office or cell phones. Each of our 770 participants answered multiple questions across 10 different general interest topics until they exceeded five minutes of recorded speech. We processed almost 86 hours of collected speech using algorithms that automatically segmented the speech into utterance-like segments of approximately 15 seconds. We combined these utterances into one- and two-minute training sets that were used to adapt a Universal Background Model (UBM). For testing of our baseline system, we reserved an average of 25 utterances per speaker.

2.2. Far-field microphone data

To analyze mismatched train/test conditions, we collected speech samples recorded on far-field microphones. Because this collection was very time consuming, we selected a random subset of 5% of the 770 speakers, and asked them to answer questions across the same general interest topics from the telephone collection. We recorded these sessions with a distant microphone in rooms with differing acoustic properties and microphone placement. We then trained speaker models with this microphone data and compared the top-N identification accuracy against the corresponding telephone-based models using the same telephone data for testing in both conditions.

3. GMM-UBM SID SYSTEM

Our baseline system is a text-independent SID system using a 1024-mixture UBM built from the speech of several hundred speakers, unrelated to this task. We used AT&T's Watson speaker-independent ASR engine to recognize each utterance and provide speech/non-speech segmentation. We then computed spectral features every 10ms over a window of 50ms. The basic acoustic features we used consisted of twelve cepstral coefficients derived from 12th order LPC and log-energy. Together with their first and second derivatives, we formed a 39-component feature vector for each time frame. For the cepstral mean subtraction and energy normalization, we applied a three-second look-ahead window. Finally, we generated the speaker-dependent GMMs using MAP adaptation of the UBM means on the speech portion of each training utterance. Additional details of model generation and likelihood calculation for speaker identification matching scores are described in [7, 6].

4. SUPPORT VECTOR MACHINES (SVM) SYSTEMS

SVMs are becoming a standard technique in speaker recognition because of their inherently discriminative training property, and the ability to easily combine different types of features. In our system, we combine a continuous-value GSV supervector with discrete-value pitch and formant histograms. These distinct feature streams are separately preprocessed, as described below, then combined with equal weights.

4.1. SVM-GMM supervector (GSV)

Our initial experiments compared the baseline GMM-UBM system with an SVM system that incorporates a GMM-mean supervector. For each speaker utterance, we adapted the means of a 1024 mixture UBM using MAP adaptation, concatenating them into a 39*1024=39936 supervector (GSV). We standardized all features to a unit variance on the training set, a widely used technique in SVMs that gives equal importance to each feature. Optimization of a soft-margin SVM includes the choice of the metaparameter C, which weights the relative importance of misclassification on the training set compared to the regularization (margin) term. In the case of linear SVMs, considerable experimental evidence suggests that, provided the input samples are normalized to 1, $C = 1$ is a good value to start with. In practice, we found that normalizing to 1 saved us considerable time, as our first choice $C = 1$ yielded the best performance on a validation dataset.

4.2. GSV-PFH: GSV and pitch/formant histograms

Pitch and formants are basic speech and speaker properties and they have been successfully employed in SID from the early days of the field [8] to most recently in [9]. Formants are well defined only in vowel and vowel-like speech regions. Even in those regions, their accurate and continuous estimation is difficult, thus stand-alone formant-based SID systems are inferior to well-established cepstrum-based systems. Fundamental frequency (F_0), which is a channel-independent feature, is not preserved in cepstral coefficients. Though formant frequencies and bandwidths are modeled by cepstrum, only formant frequencies are robust to channel characteristics and noise, especially in the telephone 300-3500Hz region that is common to all speech channels.

Formant frequencies for each phone carry speaker characteristics. In a particular speech sample, formant frequency distribution histograms also characterize the phonetic speech content, representing a crude speaker-dependent language model. Dynamic pitch and formant properties are modeled by histograms of their first derivatives.

In our current system, we compute pitch and formant frequency supervectors (F_{0123}) every 10 ms (using Xwaves) for voiced speech. The total size of F_{0123} supervector is 2866 and its components are F_0, estimated in 50-450Hz range and quantized with 2Hz resolution (201 values), and dF_0, quantized with 4Hz resolution (101 values kept). Respective values for F_1 are: 100-1250Hz, 2Hz (576), 5Hz (231). For F_2: 350-3250Hz, 3Hz (968), 5Hz (387). For F_3: 1350-3350Hz, 10Hz (201), 2Hz (201). Adding PFH supervector increases the size of the original GSV by 7% only.

Using histograms as input to SVMs was first studied for image histogram classification problems [10]. For distribution-

type supervectors, the Hellinger distance approximates the KL divergence between two probability distributions better than the Euclidean distance. The Hellinger distance simply amounts to the Euclidean distance where feature values are replaced with their square roots. As with image histogram classification problems [10], this simple transformation significantly improved our results.

In our current system, we first concatenated all pitch and formant histograms into a large histogram (PFH) supervector. Then, we combined the GSV and PFH supervectors into a single supervector. By tuning the weights for GSV and PFH, we were able to generate a minor improvement in SID accuracy. But, for generality, we continued to use equal weights.

5. CROSS-CHANNEL COHORTS FOR TARGET SPEAKER ADAPTATION

5.1. Filtering and noise equalization

To mitigate severe channel mismatch conditions, we investigated the utility of adapting our microphone-based speaker models to the telephone test data using speaker cohorts from the telephone-based speaker models. Ideally, we would like to have a significant number of speakers recorded under identical far-field microphone conditions (same room, microphone and microphone distance) to determine cohorts within each sub-group of speakers. With such data, we can assume that cohorts in the matched far-field condition will also be cohorts in the telephone condition.

Since we have not yet collected ample microphone data, we need to find cohorts within the telephone data for each microphone speaker. To reduce the microphone and telephone data mismatch, we had to perform several pre-processing steps. We filtered microphone speech using a *microphone-to-telephone* filter that approximated the difference between the microphone and telephone transfer functions. Then, we further processed each utterance by applying a Wiener filter estimated from the utterance-level background noise and adapted every 10ms.

The average signal-to-noise ratio (SNR) of the telephone speech was 26dB, which is much higher than the SNR of microphone speech even after the Wiener filter was applied. In order to reduce the SNR difference, we added fixed room noise computed from average microphone recordings across room types, to bring the utterance SNR up to about 20dB. Since some of the microphone recordings contained a noticeable echo, we experimented with room reverberation simulation on the telephone speech to enhance cohort selection, but found no additional gains.

5.2. Cohort selection procedure

To select cohorts using our hybrid datasets, we first trained GMM-UBM models using pre-processed telephone speech for each of the 770 speakers, excluding the 36 speakers that we also had microphone data for. We filtered the training utterances for each of the 36 microphone-trained speakers as described above, then scored the utterances against all telephone models, using Maximum Likelihood (ML). For each speaker, we sorted the models based on the number of closest matched utterances and selected the first model that had at least 80% top-1 accuracy to be the best cohort. Selecting a model with high accuracy is essential since the speakers with low-quality telephone data were often good matches for the microphone speech model but were not desirable for adaptation. This cohort estimation has already proved valuable in increasing the accuracy of our SID system, but we will continue to focus on improved algorithms for finding better adaptation-cohorts, under matched and mismatched channel conditions.

6. EXPERIMENTAL RESULTS

6.1. SID results on matched, telephone/telephone data

Our initial results in the matched condition showed predictably high accuracy. Figure 1 contains the text-independent N-best SID results on our set of 770 speakers for matched (telephone/telephone) channels for the GMM-UBM and the SVM system. The baseline accuracies of the GMM-UBM system are for one and two minutes of training. When increasing training data from one minute to two minutes, the error rate reduction is 50% for the first choice, and 70% for the top-10 and larger lists. Training with larger amounts of speech data results in minimal additional improvement.

Using an SVM classifier, trained on GMM means only, reduces the error rate for the first candidate by 25% and over 50% for top-5 and larger lists. The error reduction rate for one-minute and two-minute trained systems is essentially identical. When the two-minute trained SVM classifier is augmented with pitch and formant frequencies, the error rate is further reduced by about 14%.

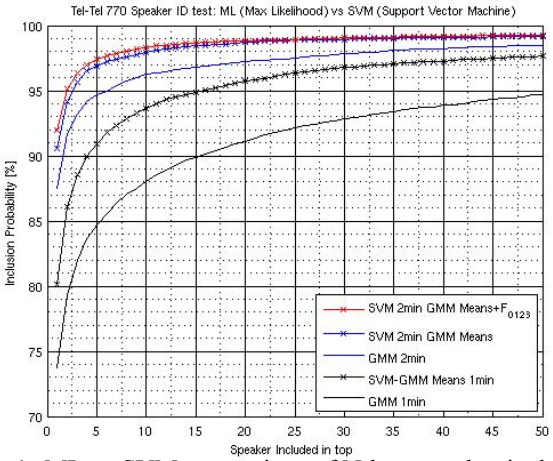

Fig. 1. ML vs SVM comparison of N-best speaker inclusion rate for one and two minutes of training data

6.2. SID results on mismatched channels

For the mismatched condition, we compared telephone test data to a composite dataset consisting of the 36 far-field microphone speaker models, adapted with two minutes of speech from the corresponding telephone cohort and the remaining 734 telephone speaker models. Figure 2 contains a graph of these results. The top curve shows the ultimate matched channel (telephone/telephone) conditions for these 36 speakers - in the set of 770 speakers. The lowest curve (*GMM-Cohort ML-test*) shows the ML-derived results using cohort-adapted models, instead of telephone models, for the set of 36 speakers. The improvement in N-best accuracy is impressive considering that the N-best accuracy for un-adapted models (with all signal processing above) only reaches 10% (not shown in the graph).

Supervectors for SVM classifiers were calculated for each target speaker microphone utterance and each of its cohort telephone training utterances. When the SVM classifier was used with either adapted GMM means or F_{0123} features alone, the recognition accuracy for each feature stream was below the accuracy of the GMM-ML accuracy. When both GMM means and F_{0123} were combined with equal weights, the SVM classifier significantly outperformed the GMM-ML system as shown in Figure 2 with error reduction going from 14% for top-1 to over 60% for top-15. We found these results very promising and believe that a better cohort selection algorithm will yield further improvements.

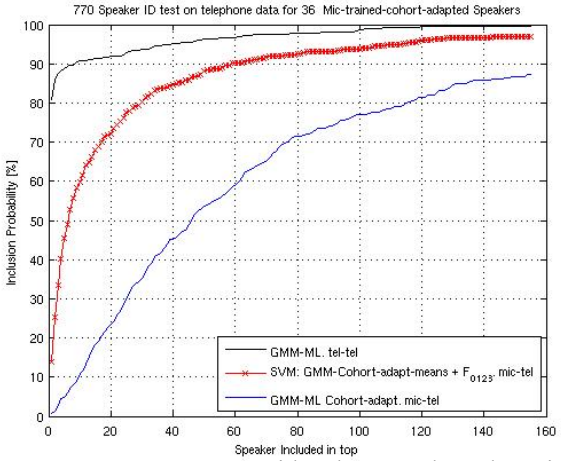

Fig. 2. ML vs SVM error with cohort speaker adaptation

7. CONCLUSIONS

In this paper we have shown N-best results for text-independent SID under matched train/test (telephone/telephone) and severely mismatched (far-field/telephone) channels on a standard GMM-UBM system and SVM system that combines the GMM-UBM supervector with pitch and formant histogram supervectors. We proposed and tested cohort's matched channel speech to adapt target speaker models that are trained on the mismatched speech only. For the scenario we are focusing on, an N-best list presented in an interface that integrates external metadata allows users to quickly locate the best speaker match. Future work will focus on better cohort selection algorithms and a robust feature space to obtain more robust cohort selection, to ensure that cohorts selected under one channel condition are valid cohorts under test/target channel conditions.

8. REFERENCES

[1] Sturim, D.E. Campbell, W.M Reynolds, D.A. Dunn, R.B. Quatieri, T.F., "Robust Speaker Recognition with Cross-Channel Data: MIT-LL Results on the 2006 NIST SRE Auxiliary Microphone Task," in *ICASSP 2007*.

[2] W.M. Campbell, D.E. Sturim, W. Shen, D.A. Reynolds, and J. Navratil, "The MIT-LL/IBM 2006 Speaker Recognition System: High-Performance Reduced-Complexity Recognition," in *Processing ICASSP 2007.*, vol. 4.

[3] L. Ferrer, E. Shriberg, S.S. Kajarekar, A. Stolcke, K. Sonmez, A. Venkataraman, and H. Bratt, "The Contribution of Cepstral and Stylistic Features to SRI's 2005 NIST Speaker Recognition Evaluation System," in *ICASSP 2006 Proceedings*, vol. 1.

[4] S. Fine, J. Navratil, and R.A. Gopinath, "A hybrid gmm/svm approach to speaker identification," in *Proceedings of ICASSP 2001*.

[5] Jin, Q.; Schultz, T.; Waibel, A., "Far-Field Speaker Recognition," in *Audio, Speech and Language Processing, IEEE Transactions on , Volume 15, Issue 7, Pages: 2023 - 2032* , Sept. 2007.

[6] Aaron Rosenberg, Julia Hirschberg, Michiel Bacchiani, S. Parthasarathy, Philip Isenhour, and Larry Stead, "Caller identification for the SCANMail voicemail browser," in *EUROSPEECH-2001*, 2001.

[7] Aaron E. Rosenberg, S. Parthasarathy, Julia Hirschberg, and Stephen Whittaker, "Foldering voicemail messages by caller using text independent speaker recognition," in *EUROSPEECH-2000*, 2000.

[8] M. Sambur, "Selection of acoustic features for speaker identification," in *Acoustics, Speech, and Signal Processing [see also IEEE Transactions on Signal Processing], IEEE Transactions on*, Apr 1975.

[9] de Wet, F. Cranen, B. de Veth, J. Boves, L., " Comparing acoustic features for robust ASR in fixed and cellularnetwork applications," in *ICASSP 2000*.

[10] O. Chapelle, P. Haffner, and V. Vapnik, "SVMs for Histogram-Based Image Classification," *IEEE Transaction Neural Networks*, 1999.