The 2004 MIT Lincoln Laboratory Speaker Recognition System•

D. A. Reynolds. W. Campbell, T. Gleason, C. Quillen, D. Sturim, P. Torres-Carrasquillo, A. Adami+
{dar,wcampbell,tpg,cbq,sturim,ptorres}@ll.mit.edu, adami@bme.ogi.edu

MIT Lincoln Laboratory, Lexington, MA USA, +OGI School of Science and Engineering Beaverton, OR USA

ABSTRACT

The MIT Lincoln Laboratory submission for the 2004 NIST Speaker Recognition Evaluation (SRE) was built upon seven core systems using speaker information from short-term acoustics, pitch and duration prosodic behavior, and phoneme and word usage. These different levels of information were modeled and classified using Gaussian Mixture Models, Support Vector Machines and N-gram language models and were combined using a single layer percepton fuser. The 2004 SRE used a new multi-lingual, multi-channel speech corpus that provided a challenging speaker detection task for the above systems. In this paper we describe the core systems used and provide an overview of their performance on the 2004 SRE detection tasks.

1. INTRODUCTION

Over the last several years, there has been continued interest in exploiting new levels of speaker information for improved speaker verification performance [1]. For the 2004 NIST speaker recognition evaluation (SRE), MIT Lincoln Laboratory continued efforts in this area with a submission built upon seven core systems using speaker information from short-term acoustics, pitch and duration prosodic behavior, and phoneme and word usage. These different levels of information were modeled and classified using Gaussian Mixture Models, Support Vector Machines and N-gram language models and were combined using a single layer percepton fuser. The 2004 SRE used a new multi-lingual, multi-channel speech corpus that provided a challenging speaker detection task for the above systems. In this paper we describe the core systems used and provide an overview of their performance on the 2004 SRE detection tasks.

2. 2004 NIST SPEAKER RECOGNITION EVALUATION

In an ongoing effort to support research and development in text-independent speaker recognition technologies, NIST has been conducting annual speaker recognition evaluations [2] The aim of these evaluations is to provide common framework (data, rules and scoring) to allow focused technology development and meaningful comparison of techniques and approaches. New for this year, a large suite of train/test conditions was provided and data from the bilingual cross-channel MIXER corpus was used.

As in past years, the core task for SRE04 was speaker detection: Given a model speaker, determine if that speaker is speaking in a given test segment. Performance is evaluated using Detection Error Trade-Off (DET) curves and the Decision Cost Function (DCF). There were 7 training conditions and 4 testing conditions for a total of 28 possible conditions. The train/test conditions covered varying amounts of data (10sec to 16 conversation sides) and contamination by other speakers (summed conversation sides). Additionally, there was an adaptation mode to the above tasks. A complete description of the SRE04 tasks and rules can be found at http://www.nist.gov/speech/tests/spk/2004/.

The data used in SRE04 was derived from the new MIXER corpus which was designed to support large multi-sessions training and to include cross-channel recordings and bi-lingual speakers [3]. A total of 3637 conversations involving 310 speakers were used, with bilingual speakers of Arabic, Mandarin, Russian, and Spanish with English. No development data from MIXER (or the related FISHER corpus) was provided. The MITLL system used development data from the aggregations of the Switchboard II phases 1-5 corpora.

3. CORE DETECTION SYSTEMS

3.1 Spectral Based

For the 2004 system we used two spectral based verification systems: a Gaussian Mixture Model-Universal Background Model (GMM-UBM) system and a Support Vector Machine (SVM) system.

3.1.1 GMM-UBM

The basic system used is a likelihood ratio detector with target and alternative probability distributions modeled by Gaussian mixture models (GMMs). A Universal background model GMM is used as the alternative hypothesis model and target models are derived using Bayesian adaptation [4]. The techniques of feature mapping [5], Tnorm [6] and a new form of Adapted Tnorm (ATnorm) were applied [7].

A 19-dimensional mel-cepstral vector is extracted from the speech signal every 10ms using a 20ms window. Bandlimitng is performed by only retaining the filterbank outputs from the frequency range 300Hz-3138Hz. Cepstral vectors are processed with RASTA filtering to mitigate linear channel bias effects. Delta cepstral are then computed over a +-2 frame span and appended to the cepstra vector producing a 38 dimensional feature vector. The feature vector stream is then processed through an adaptive, energy-based speech detector to discard low-energy vectors. The silence removed features are processed with feature mapping and, finally, normalized by removing the global mean and dividing by the standard deviation..

• This work is sponsored by the Department of Defense under Air Force Contract F19628-00-C-0002. Opinions, interpretations, conclusions and recommendations are those of the author and are not necessarily endorsed by the United States Government

The background model used for all targets is a gender-independent 2048 mixture trained using data from Switchboard II-phase1, Switchboard II-phase4 (cell), and the OGI National Cellular Database.

Target models are derived by Bayesian adaptation (a.k.a. MAP estimation) of the UBM parameters using the designated training data. Based on observed better performance, only the mean vectors are adapted. The amount of adaptation of each mixture mean is data dependent with a relevance factor of 16 used.

3.1.2 SVM

The Spectral SVM system uses a novel sequence kernel described in [8], that compares entire utterances using a generalized linear discriminant. For SRE04 the same front-end processing used for the GMM-UBM system was used for the SVM system.

The SVM used a Generalized Linear Discriminant Sequence kernel (GLDS) [8] with an expansion into feature space using a monomial basis. All monomials up to degree 3 were used, resulting in a feature space expansion of dimension 9139. We used a diagonal approximation to the kernel inner product matrix.

A "background" for the SVM consisted of a set of speakers taken from a database not used in the train/test set. Speakers from Switchboard II phases 1-5 were used as the background.

SVM training was performed as a two-class problem where all of the speakers in the background had SVM target -1 and the current speaker under training had SVM target +1. For each conversation in the background and for the current speaker under training, an average feature expansion was created. SVM training was then performed using the GLDS kernel implemented using SVMTorch[i].

For each test utterance the standard front end was used. An average feature expansion was then calculated. Scores for each speaker were an inner product between the speaker model and the average expansion.

3.2 Prosodic Based

A distribution based pitch/energy classifier and a pitch/energy sequence modeling system comprised the prosodic components used in the 2004 system.

3.2.1 Pitch and Energy GMM

The aim of this system is to capture the characteristics of the F0 and short-term energy features distribution. This system is based on a likelihood ratio detector that uses adapted GMMs for estimating the likelihoods.

The log F0 and log energy features are estimated every 10 ms from the speech signal using the RAPT - Robust Algorithm for Pitch Tracking proposed by Talkin [9]. Delta features, estimated over a 50ms window, are then appended. Only the feature vectors extracted for voiced speech regions are used in training and testing.. In addition, a speech activity detector is employed to discard feature vectors extracted from silence and noisy regions. Care is taken to handle feature extraction at discontinuities.

The UBM is a 512-component Gaussian Mixture Model trained with speech from Switchboard II. During recognition the target model scores are normalized using T-norm.

3.2.2 Slope and Duration n-gram

To capture prosodic differences in the realization of intonation, rhythm, and stress, we converted the F0 (acoustic correlate of pitch) and energy contours into a sequence of tokens reflecting the joint state of the contours (rising or falling) and then applied simple n-gram tools to model and classify distinctive token patterns from token sequences [10].

The joint-state classes estimation is divided into 5 steps: 1) compute the f0 and energy temporal trajectories, 2) compute the rate of change for each trajectory, 3) detect the inflection points (points at the zero-crossings of the rate of change) for each trajectory, 4) segment the speech signal at the detected inflection points and at the voicing starts or ends, and 5) convert the segments into a sequence of symbols by using the rate of change of both trajectory within each segment. We can also integrate the duration information in each segment class by adding an extra label with the duration information. All segments smaller than 30 ms are removed from the sequence of joint-state classes and we placed <bound> symbols around each utterance, defined as speech segments separated by silent gaps greater than 0.5 seconds.

The speaker detection score is computed using a conventional log-likelihood ratio test between the target-speaker model and the UBM averaged over all n-gram types [11,12]. Then, the target model scores are normalized using T-norm.

3.3 Phonetic Based

Two systems which operated on the phone stream were applied: One was based on standard N-gram modeling/scoring and one based on a new SVM classifier using N-gram counts.

Gender-independent phone recognition is performed using HTK 3.1.1[ii]. Six phone recognizers (English, German, Hindi, Japanese, Mandarin, and Spanish) were trained on phonetically marked speech from the OGI Multilanguage corpus. Output token streams from training and testing data were processed to produce a sequence of token symbols, removing silence phones and inserting <end> <start> tokens at utterance boundaries defined as silence gaps over a 0.5 sec duration.

3.3.1 Phone N-grams

The phone N-gram system operates similar to the one described in [12]. For each speaker, a bi-gram model is estimated for each of the 6 phone recognizers. In addition a UBM for each phone recognizer is trained using Switchboard II data. During testing likelihood ratio score between the target and UBM for each phone stream is computed and a final fused score is produced as a linear combination of the 6 likelihood ratio scores.

[i] http://www.idiap.ch

[ii] http://htk.eng.cam.ac.uk/

3.3.2 Phone SVM

The Phone SVM system uses a kernel for comparing conversation sides based upon methods from information retrieval. Sequences of phones are converted to a vector of probabilities of occurrences of terms and co-occurrences of terms (bag of unigram and bag of bi-grams). A weighting based upon a linearization of likelihoods is then used to compare vectors for SVM training [13].

Probabilities of each phone and its joint probability with other phones (unigram and bi-gram) in a given conversation were calculated with counts. These probabilities were then put in a large (sparse) vector for training. The SVM used a kernel derived from information retrieval methods and likelihood ratio scoring. This amounted to scaling individual entries in the vector of probabilities with a term weighting of $1/\mathrm{sqrt}(p(t_i))$, where $p(t_i)$ was the probability of the term over all conversations in the "background."

A "background" for the SVM consisted of a set of speakers taken from a database not used in the train/test set. Speakers from Switchboard II phases 1-5 were used as the background.

SVM training was performed as a two-class problem where all of the speakers in the background had SVM target -1 and the current speaker under training had SVM target +1. For each conversation in the background and for the current speaker under training, a term-weighted vector of probabilities was created. SVM training was then performed using a linear kernel in SVMTorch. Different models of the same speaker were constructed for each of the different languages.

For each utterance the standard front end was used in all 6 languages. The scores for each of the 6 target speaker language models were then found using a SVM. The scores were then fused with equal weighting.

3.4 Idiolectal Based

This system used the Idiolect word n-gram approach proposed in [11]. The idiolect system uses n-gram=2 (bigrams), discount=1 (full discounting), and minimum n-gram count (c_min) of 9. This setting performed best on the development data, as compared with discount=0, using trigrams, and using a higher minimum n-gram count (200). We did not vary the probability-smoothing factor of 0.001.

Word transcripts were derived from BBN Byblos 3.0 real-time recognition system[iii]. All of Switchboard II was processed and used to derive a UBM. For the evaluation data, all English and non-English data was processed by the English Byblos system.

During testing the likelihood ratio score between the target and UBM is computed and Tnorm is applied.

3.5 System Fusion

The scores from the systems were fused with a perceptron classifier using LNKnet[iv]. The perceptron architecture chosen has two input nodes, no hidden layers, and two output nodes. Input values to the perceptron were normalized to zero mean and unit standard deviation using parameters derived from the training data. The perceptron weights were trained using the entire development data. The classifier corresponding to the number of training conversations is then used to fuse scores from systems. The fusion classifier is trained using minimum DCF criterion. Prior probability for the target class in training and testing was set to 0.09 corresponding to the costs and priors (C_miss*P_tgt/(Cmiss*P_tgt + C_fa*(1-P_tgt))). The hard decision from the perceptron was used as the hard decision for the submission. The score for the test file was set to (s_tgt + (1-s_non))/2, where s_tgt and s_non are the perceptron scores for the target and nontarget classes, respectively.

4. RESULTS

Due to the limited space of this paper, we will limit our examination of results to the 1 side train / 1 side test and 8 side train / 1 side test conditions In Figure 1 we show the DET curves for the 7 core systems described above and the fusion system. As expected, the spectral based systems are providing the best performance, with phone-SVM system the best non-spectral based system. Since many of the non-spectral systems require more training data to learn speaker habits, it is not surprising that the fusion did not provide any improvement. We did, however, expect the fusion of the SVM and GMM systems to provide a boost in performance which we have seen on other data sets.

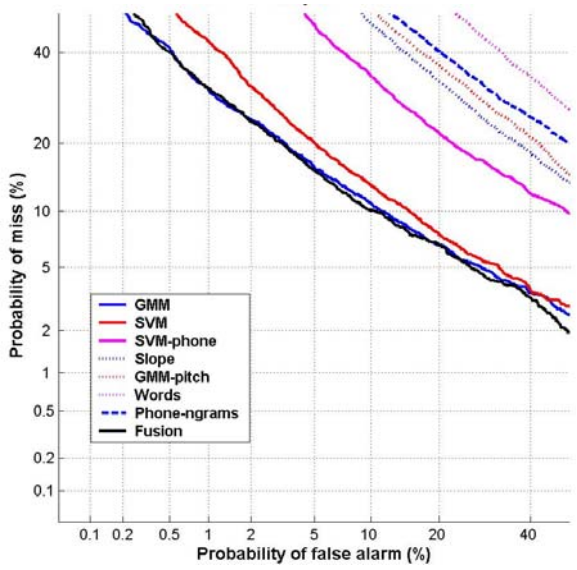

Figure 1 1side train / 1side test all score pooling DET curves for 7 core systems and fusion.

In Figure 2 we show the DET curves for the systems from the 8 side train / 1 side test condition. Again we observe that the spectral based systems are performing the best. Contrary to expectations, we found that fusion with higher-levels of information did not provide any gains. There are several possible reasons for this. First, all development data was derived from English data and so there may be a bias in the UBMs used in some systems, Second, the SRE04 data sets were designed to have more channel mismatch than in previous years making the

[iii] We would like to thank BBN for making this finely engineered recognition system available.
[iv] http://www.ll.mit.edu/IST/lknet

task more difficult and potentially masking gains from the high-level systems.

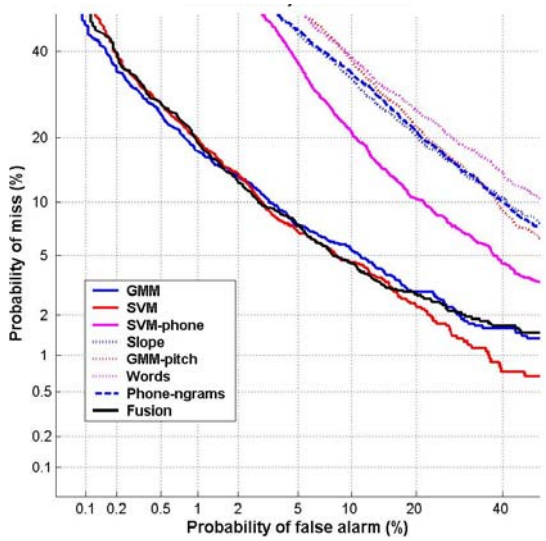

Figure 2 8 side train / 1side test all score pooling DET curves for 7 core systems and fusion.

To examine the contribution of the systems further, we searched for the best N-way system fusion for N=1-7 for two score pooling conditions: *All Pool*, using all scores, and *Common Pool*, where only scores from English train/test, handheld microphones and cell/landline are used.. The minimum DCF values for these cases are shown in Figure 3.

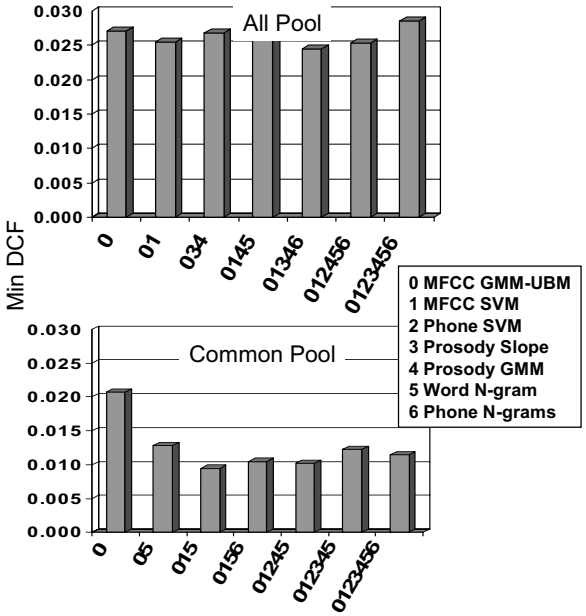

Figure 3 Minimum DCF values for best combinations of core systems from 8s/1s condition. Top plot is for All Pooling and bottom plot is for Common (English) Pooling.

We see for the Common Pool case, that the word n-gram system gives a decrease in error as has been seen in previous evaluations. Since language was matched for the train, test, UBM and STT system, this result appears to support the potential cross-lingual degradations effects to high-level systems. It is worth noting that, using other examinations, we did not see cross-lingual degradation effects with the spectral based systems.

5. CONCLUSIONS

We have presented a brief overview of the continuing efforts at MIT Lincoln Laboratory to further exploit new levels of information to better characterize and recognize a speaker. We described 7 core systems that used information from spectral, prosodic, phonetic and idiolectal sources, extracting different types of features for use in generative, discriminative and discrete classifiers. We presented results on the new and challenging MIXER corpus used in SRE04, showing that some previously successfully fused system may need to be better tailored to work in cross-lingual environments.

REFERENCES

[1] D.A. Reynolds, et. al, "The Supersid Project: Exploiting High-Level Information For High-Accuracy Speaker Recognition," ICASSP 2003.

[2] M. Przybocki and A. Martin, "NIST Speaker Recognition Evaluation Chronicles," Odyssey Workshop 2004.

[3] J.P. Campbell, et. al, "The MMSR Bilingual and Cross-channel Corpora for Speaker Recognition and Evaluation," Odyssey Workshop 2004.

[4] D.A. Reynolds, T.F. Quatieri and R.B, Dunn, "Speaker Verification using Adapted Gaussian Mixture Models," Digital Signal Processing, Vol 10(1-3), pp. 19-21, 2000.

[5] D.A. Reynolds, "Channel Robust Speaker Verification via Feature Mapping," ICASSP 2003.

[6] R. Auckenthaler, M. Carey and H. Lloyd-Thomas, *Score Normalization for Text-Independent Speaker Verification Systems,* Digital Signal Processing, 10 (2000), pp. 42-54.

[7] D.E. Sturim and D.A. Reynolds, "Speaker Adaptive Cohort Selection For Tnorm In Text-Independent Speaker Verification." ICASSP 2005.

[8] W. M. Campbell, "Generalized Linear Discriminat Sequence Kernels for Speaker Recognition," ICASSP 2002.

[9] D. Talkin, A Robust Algorithm for Pitch Tracking (RAPT), in Speech Coding and Synthesis, W. B. Kleijn and K. K. Paliwal, Eds. New York: Elsevier, 1995.

[10] A. Adami, R. Mihaescu, D. A. Reynolds, and J. Godfrey, "Modeling Prosodic Dynamics for Speaker Recognition," ICASSP 2003..

[11] G. Doddington, "Speaker Recognition based on Idiolectal Differences between Speakers," Eurospeech 2001.

[12] W. D. Andrews, M. A. Kohler, J. P. Campbell, J. J. Godfrey, and J. Hernandez-Cordero, "Gender-dependent Phonetic Refraction for Speaker Recognition," ICASSP '2002.

[13] W. Campbell, J, Campbell, D. Reynolds, D. Jones and T. Leek, "Phonetic Speaker Recognition with Support Vector Machines," NIPS 2003.