IBM NIST 2010 SRE SYSTEM DESCRIPTION

Mohamed Kamal Omar and Jason Pelecanos

Natural Language Systems Group
IBM T.J. Watson Research Center
1101 Kitchawan Road, Yorktown Heights, NY 10598
{mkomar, jwpeleca}@us.ibm.com

1. SUBMISSION OVERVIEW

There were two systems submitted for NIST SRE 2010 core evaluation. The IBM primary submission is a score-level fusion of 6 core systems as follows:

- **1× Phonetically inspired NAP-GMM system**
- **3× Discriminatively trained NAP-GMM systems**, varied by choice of front-end features, data subsets and system configurations.
- **2× Factor analysis based systems** trained on different front-end features: LPCCs and MFCCs

The alternate submission consisted of just the "phonetically inspired NAP-GMM system".

This write-up discusses the data pre-processing and parameterisation components in Section 2. In Section 3 we discuss the core modelling components incorporated while Section 4 presents the system combination technique applied. Finally, Section 5 identifies the execution times of the major processing components.

2. DATA PRE-PROCESSING AND PARAMETERISATION

2.1. Data Pre-processing

For the NIST 2010 Core evaluation, we considered the data to be represented by three main components:

- Basic Telephony (cellular, landline)
- Telephony with Distant Microphone
- Interview data with Distant Microphone

Before any other processing, the conversational interview data was first pre-processed to remove the interviewer's speech from the audio. Here, NIST's ASR transcripts for the interviewer's lapel microphone were used to indicate when the interviewer was speaking. Subsequently these sections were removed from the audio. After this stage, all audio was treated in a similar manner.

For the systems using MFCCs or LPCCs, a fast dynamic energy noise floor tracking algorithm was incorporated (similar to [1]). If less than 30% of the audio was detected as non-silence the threshold was reduced to capture more non-silence audio. For the ASR inspired systems, information from an automatic speech recognition system provided the detected speech segments.

2.2. Feature Extraction

Three different feature sets were produced and are labelled as as MFCC36, LPCC36 and ASR40. These sets relate to Mel-Frequency Cepstral Coefficients (MFCCs), Linear Prediction Cepstral Coefficients (LPCCs) and automatic speech recognition inspired features. Basic details of each feature set follows below:

- **MFCC36** is based on 12 cepstral coefficients generated from 20 filterbanks. Delta and acceleration coefficients are also included. The filterbank features are calculated from 32ms frames with a 10ms frame advance. The filterbanks span a frequency range of 125-3800Hz.
- **LPCC36** has 12 LPCCs with delta and acceleration features added. The LPCCs were generated from 19 linear prediction coefficients. (32ms frames, 10ms frame shift)
- **ASR40** consists of 40 dimensional features generated from the IBM ASR system. These features are estimated from sequences of 13-dimensional perceptual linear prediction (PLP) features by using a linear discriminant analysis (LDA) projection, and then applying a maximum likelihood linear transformation (MLLT). The acoustic model consists of 250K diagonal-covariance Gaussian components. In the context of speaker-adaptive training, vocal tract length normalization (VTLN) and feature-space maximum likelihood linear regression (FMLLR) are used. An FMPE transform is applied on top of the utterance-specific FMLLR transforms.

The first two sets of features have feature warping (Gaussianization) [2] applied over all dimensions after silence removal was applied. Feature warping can mitigate linear channel and slowly varying additive noise effects. All LPCC features were generated with the HTK Toolkit [3]. An interesting aspect to note is that features generated with feature warping post-processing could be compressed to approximately one-quarter of the size (using bzip2) over our generic 4 byte float per parameter representation. This is very useful when working with large parameter sets.

3. CORE SYSTEMS

3.1. Phonetically Inspired UBM Modelling

This approach was first applied to nonnative speaker and accent detection in [4]. In this system, the UBM is estimated directly from the Gaussians of the acoustic model of the ASR system by using K-means clustering. A symmetric variant of the Kullback-Leibler

(KL) divergence between two Gaussian components is used as a distance measure in the K-means clustering algorithm to achieve the final clustering of the ASR acoustic model to a UBM of 1024 Gaussian components. Effectively, an ASR model of 250k Gaussians is transformed into a 1024 component GMM. This novel method for UBM construction is applied to ASR acoustic models trained in the feature-based minimum phone error (FMPE) feature space [5]. The front-end features also include VTLN, LDA, MLLT, and fMLLR. More information on the specifics of the approach is available in [6].

Using the transformed ASR-GMM statistics, speaker recognition is performed based on a NAP compensated GMM-UBM framework [7, 8]. As per our NIST 2008 SRE submission, the nuisance directions used for NAP are based on finding the direction of the largest within class covariance [9]. ZT-Norm [10] is also applied.

For this system, Switchboard II P3, NIST 2004, Dev 2008, Eval 2008[1] and Dev 2010 data were used for the NAP session compensation. For ZT-norm a combination of this data was applied.

3.2. Discriminatively Trained UBM-GMM

The UBM in speaker verification systems is typically a Gaussian mixture model (GMM) trained on a large amount of data using the EM algorithm. In this system we apply a discriminative method for training the UBM by adding a regularization term to the maximum likelihood objective function [6]. Here, the regularization term favors larger values for target trial scores and smaller values for imposter trial scores.

The objective function is,

$$O = L - \lambda_t \sum_{r=1}^{T} e^{a_t - b_t s_{tr}} - \lambda_i \sum_{j=1}^{J} e^{a_i + b_i s_{ij}}, \quad (1)$$

where λ_t is the target-trials regularization parameter, λ_i is the imposter-trials regularization parameter, s_{tr} is the rth target score, s_{ij} is the jth imposter score, a_t, b_t are the parameters of the target regularization function, a_i, b_i are the parameters of the imposter regularization function, T is the number of target scores, and J is the number of imposter scores. The parameters of the target and imposter regularization functions are estimated on a held-out set to provide proper conditioning of the target and imposter scores respectively. The target and imposter scores are the speaker recognition scores without NAP compensation and without ZT-normalization. The objective function specified in Equation 1 can be optimized using an E-M like algorithm. Further details are available in [6].

Using this approach, three discriminatively trained systems are constructed.

Interview only data using ASR based features:
This system uses the ASR based features with the UBM trained with 20 iterations on NIST Dev 2008, Eval 2008 and Dev 2010 data. The NAP approach utilizes the NIST Dev 2008, Eval 2008 and Dev 2010 data. ZT-Normalization is applied using all available data.

Expanded interview only data using ASR based features:
This system is the same as the system above except the NIST Eval 2006 data is also added as part of the UBM optimization. Additionally, only 10 iterations were performed.

[1]Note (here and throughout): The NIST 2008 SRE data was split into 2 parts. The first part consisted of two-thirds of the available speakers and the other part comprised the remainder. The larger portion was used for system development purposes while the remaining third was kept for fusion and calibration.

Interview data only using MFCC based features:
This system is based on the use of MFCC features with the UBM trained on NIST Dev 2008, Eval 2008 and Dev 2010 data. The NAP compensation is trained on Switchboard II Phase 3, NIST Dev 2008, Eval 2008 and Dev 2010. ZT-Norm is applied using all available data.

3.3. Factor Analysis Systems

There are two factor analysis based systems in this year's submission. The only difference between these two systems is type of features; the feature sets used are MFCCs and LPCCs as described in the feature extraction section. The factor analysis system is based on the factor analysis work of Kenny [11]. We use Gauss-Seidel iterative estimates (based on [12, 13]) to provide MAP estimates of the session and speaker contributions.

The statistics are based on the use of a 1024 component Gaussian mixture model. The factor analysis model is trained with relevance adaptation, 300 speaker factors and 100 session factors. The factor analysis system is trained first for relevance adaptation followed by interleaved iterations of speaker and session optimization. The speaker and session subspace optimization process is trained on the relevance adaptation residual. The FA system uses all available development data to train the factors.

Once trained on the development data, the factor analysis system is used to compensate for session variability in the utterance specific sufficient statistics from the NIST 2010 data. These statistics are then used to calculate an approximation of the log-likelihood ratio. That is, the log of the ratio of the likelihood of the test utterance given the compensated adapted target model to the likelihood of the test utterance given the UBM.

We also include ZT-Norm [10]. Here the ZT-Norm speaker set was chosen to dynamically match both the gender and channel type (interview microphone, "telephone microphone" and telephone) of the corresponding enrollment and test utterances for each trial. The role of the enrollment and test utterances is also switched and scored again to provide both the forward and reverse scores which is then used to provide a symmetric score.

4. SUBMITTED SYSTEMS

Two systems were submitted for this evaluation. The primary system submitted uses a combination of 6 systems based on the FoCal toolkit [14]. The optimization weights were determined from a subset of the NIST 2008 SRE. As mentioned earlier, two-thirds of the data was used for improving core systems while one-third was kept for optimizing the fusion weights. The bilinear version of the toolkit was employed to utilize the side labels relating to the broad audio recording types; interview microphone, telephone microphone and telephone. The operating threshold was selected based on plotting and finding the threshold of the minimum DCF over the range of thresholds on the held-out data. It aligned reasonably well with the theoretical log-likelihood ratio for the specified NIST cost function. An additional margin was added to the threshold to reduce the risk of high-cost false accepts.

The alternate system submitted is a single system designed for improved telephony speaker recognition performance with the new NIST minimum DCF criterion. Score calibration is performed in a similar manner to the primary system setup.

5. SYSTEM EXECUTION TIMES

System Component	Estimated Time on single CPU (hours)
Offline System Development	
Entire Parameterisation (ASR+non-ASR)	6000
Basic UBM training	400
Discriminatively trained UBM	1000
NAP Subspace Training	0.5
Factor Analysis Training	250
Online Evaluation System	
Entire Parameterisation (ASR+non-ASR)	1500
Sufficient Stats + NAP + Dot-Product Scoring	10
Factor Analysis Inspired Scoring	150

These statistics are estimates for a single processor on a 2.2GHz Pentium 4 machine.

6. REFERENCES

[1] D. Reynolds, "A Gaussian mixture modeling approach to text-independent speaker identification," Ph.D. dissertation, Georgia Institute of Technology, 1992.

[2] J. Pelecanos and S. Sridharan, "Feature warping for robust speaker verification," *A Speaker Odyssey, The Speaker Recognition Workshop*, pp. 213–218, 2001.

[3] S. Young et al, "HTK Toolkit," *HTK Version 3.3*, 2008.

[4] M. Omar and J. Pelecanos, "A novel approach to detecting non-native speakers and their native language," *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2010.

[5] D. Povey, B. Kingsbury, L. Mangu, G. Saon, H. Soltau, and G. Zweig, "fMPE: Discriminatively trained features for speech recognition," *IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 1, pp. 961–964, 2005.

[6] M. Omar and J. Pelecanos, "Training universal background models for speaker recognition," *Odyssey 2010 Speaker and Language Recognition Workshop*, 2010.

[7] A. Solomonoff, W. Campbell, and I. Boardman, "Advances in channel compensation for SVM speaker recognition," *IEEE ICASSP*, 2005.

[8] D. Reynolds, T. Quatieri, and R. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Processing*, vol. 10, no. 1/2/3, pp. 19–41, 2000.

[9] A. Hatch, S. Kajarekar, and A. Stolcke, "Within-class covariance normalization for SVM-based speaker recognition," *International Conference on Spoken Language Processing*, 2006.

[10] R. Auckenthaler, M. Carey, and H. Lloyd-Thomas, "Score normalization for text-independent speaker verification systems," *Digital Signal Processing*, vol. 10, no. 1/2/3, pp. 42–54, 2000.

[11] P. Kenny, "Joint factor analysis of speaker and session variability: Theory and algorithms (draft version)," *IEEE Speech, Acoustics and Language Processing*, 2006.

[12] R. Vogt, B. Baker, and S. Sridharan, "Modelling session variability in text-independent speaker verification," *Interspeech*, pp. 3117–3120, 2005.

[13] L. Burget, et al, "But system description: Nist sre 2008," *proc. of the 2008 NIST Speaker Recognition Evaluation Workshop*, 2008.

[14] N. Brummer, "Focal bilinear toolkit," *http://sites.google.com/site/nikobrummer/focalbilinear*, Accessed: May, 2010.