The AFRL/HEC 2006 Speaker Recognition Systems

Eric G. Hansen[1], Brian M. Ore[2], and Raymond E. Slyh[1]

[1]Air Force Research Laboratory, Human Effectiveness Directorate, Wright-Patterson AFB OH
[2]General Dynamics Advanced Information Systems, Dayton OH

1. Overview

The table below shows the component systems that went into making the overall primary system submitted for each training/testing condition. The various systems were combinations of the component systems fused with a single-layer perceptron trained using NIST 2004 data. The colors in the table indicate systems that have the same underlying components and processing strategy.

		TESTING		
		10sec4w	**1conv4w**	**1conv2w**
TRAINING	**10sec4w**	FMBWF0/GMM MFCC/GMM MFCC/SVM		
	1conv4w	FMBWF0/GMM MFCC/GMM MFCC/SVM	FMBWF0/GMM MFCC/GMM MFCC/SVM MFCC/PS-GMM WLM	FMBWF0/GMM MFCC/GMM MFCC/SVM
	3conv4w	FMBWF0/GMM MFCC/GMM MFCC/SVM	FMBWF0/GMM MFCC/GMM MFCC/SVM MFCC/PS-GMM WLM	FMBWF0/GMM MFCC/GMM MFCC/SVM
	8conv4w	FMBWF0/GMM MFCC/GMM MFCC/SVM	FMBWF0/GMM MFCC/GMM MFCC/SVM MFCC/PS-GMM WLM	FMBWF0/GMM MFCC/GMM MFCC/SVM
	3conv2w		FMBWF0/GMM MFCC/GMM MFCC/SVM	FMBWF0/GMM MFCC/GMM MFCC/SVM

The various feature sets and classifiers are as follows:

- FMBWF0/GMM: Formant center frequencies, formant bandwidths, and F0 with Gaussian Mixture Models (GMMs)
- MFCC/GMM: Mel-frequency cepstral coefficients (MFCCs) with GMMs
- MFCC/SVM: MFCCs with Support Vector Machines (SVMs)
- MFCC/PS-GMM: MFCCs with phoneme-specific GMMs
- WLM: Word language modeling

The systems submitted this year for conditions involving only four-wire training and testing were similar in many respects to the systems submitted last year (and discussed in [1]) with the following differences: (1) the LPCC system used last year was not used this year, (2) the FMBWF0/GMM system added feature mapping this year, (3) the MFCC/SVM system was added for this year, (4) word-level transcripts (used in the WLM system) and phoneme alignments (used in the MFCC/PS-GMM system) were generated this year using Version 2.0-beta5 of the SONIC

speech recognizer from the University of Colorado at Boulder, and (5) the fusion used in the MFCC/PS-GMM system this year used the sigmoid removal discussed in Section 6 of [1].

The systems submitted this year for conditions involving two-wire (3conv2w) training and/or two-wire (1conv2w) testing were similar to those submitted last year (and discussed in [2]), but had two major differences: (1) this year's systems used the likelihood ratio clustering discussed in Section 5.2 of [2] and (2) the FMBWF0/GMM and MFCC/SVM systems were added this year and fused with the MFCC/GMM system.

In addition to the combined systems submitted as primary systems, the MFCC/GMM system was submitted by itself, both with and without unsupervised adaptation, for 1conv4w testing with 1conv4w, 3conv4w, and 8conv4w training conditions and was denoted as system HEC2 for these conditions. The unsupervised adaptation system was very similar to that discussed in [3] with Adaptation Factor 4.

2. Speech Activity Detection

A number of the systems used a common speech activity detector (SAD), which worked in three stages. The first stage utilized a two-state speech/non-speech Hidden Markov Model (HMM) with MFCCs as the features. The second stage refined the HMM output by applying an energy-based detector. The final stage post-processed the output by reclassifying as non-speech any segments labeled as speech that were less than 20 msec in duration. The MFCC/HMM portion of the SAD was built using HTK from Cambridge University using 64 mixtures per state. The energy based detection was performed using the MIT Lincoln Laboratory *xtalk* program from Version 2.1 of their MFCC/GMM speaker recognition system [4]. In the remainder of this report, we refer to this entire three-stage process as the MFCC/HMM SAD.

3. Gaussian Mixture Model Systems

All of the component systems except the MFCC/SVM and WLM systems used GMMs built with Version 2.1 of the MIT Lincoln Laboratory GMM system [4] with a mixture size of 2048 and diagonal covariance matrices. Only means were adapted from the background model using MAP adaptation. For the MFCC-based systems, RASTA filtering [5] was applied to the coefficients, and the feature set included the deltas of the features. For both the MFCC- and FMBWF0-based systems, feature mapping [6] was used; however, the channel was always chosen using the channel determined by the MFCCs. Finally, the mapped MFCC features were normalized to have zero mean and unit variance.

Gender-dependent T-norm [7] was applied (using 120 models for each gender), with the exception that gender-independent T-norm (with 240 models) was used in the 10sec4w training condition. For the 10sec4w-10sec4w condition, T-norm models were built from 30 seconds of data. For the other training conditions, T-norm models were built using approximately two minutes of data. The T-norm model data came from the NIST 2001–2003 evaluation data sets.

The background model data consisted of approximately 16 hours of speech from a variety of sources, including the NIST 2001–2003 evaluations (for carbon button land line data, electret microphone land line data, and digital cellular data) and the OGI National Cellular Database (for analog cellular data). The background model data were balanced for gender and the four previously mentioned channel types, and these channels were the ones used in the feature mapping.

3.1. FMBWF0 Features

Seven features were used composed of formant center frequencies, formant bandwidths, and F0 calculated every 10ms. The formant values were computed using Version 2.2.2 of the Snack toolkit from KTH. The F0 values were computed using the *get_f0* command from the Entropic Signal Processing System (ESPS); the *get_f0* command is an implementation of the pitch tracking algorithm of [8]. The frequency components of each file were calculated with a moving 10 msec window. The following information was generated:

- Formant Center Frequencies 1–3 (F1–F3)
- Formant Bandwidths (B1–B3)

- Fundamental Frequency (F0)
- Voicing status

Each F0 value was converted to log scale. Each formant center frequency and bandwidth value was converted to radians. Only extracted frames satisfying the following constraints were used:

- Must be labeled as speech by the MFCC/HMM SAD
- Must be labeled as voiced by *get_f0*
- F0 < 250 Hz (See footnote [1] below)
- (F1 != 500 Hz) && (F2 != 1500 Hz) && (F3 != 2500 Hz) (See footnote [2] below)

See [1, 9] for more details on this feature set. As previously noted, for this year, the FMBWF0 features had feature mapping applied, but with the channel chosen by the MFCC feature set.

3.2. MFCC Features

The MFCCs were computed using Version 2.1 of the MIT Lincoln Laboratory MFCC/GMM speaker recognition system [4]. Nineteen MFCCs (with RASTA filtering) and deltas were calculated in the bandwidth of 300–3138 Hz from the speech waveform and output every 10 msec. Only frames labeled as speech by the MFCC/HMM SAD were used. Feature mapping and mean and variance normalization were applied to the features.

3.4. MFCC Features in the PS-GMM System

This system used MFCCs that were computed exactly as in the standard MFCC system except that each feature vector was then associated with a phoneme label as output by Version 2.0-beta5 of the SONIC speech recognizer from the University of Colorado at Boulder [10, 11]. SONIC was run as an English-language speech recognizer using acoustic models (provided by CU Boulder) that were trained on Fisher data and a trigram language model trained on hand-generated transcripts of Switchboard data. The phoneme alignments were constructed from the state file output by SONIC. The feature vectors for a given phoneme were then scored with a GMM model built for that phoneme. We used the phone-only adaptation method detailed in [12] to build the phoneme-specific models for each claimant model. The phonemes used were from the following set:

{AE, AH, AX, AY, DH, EH, EY, IH, IY, L, M, N, OW, S, Y}

The scores from the individual phoneme systems were combined using a single layer perceptron trained on NIST 2004 evaluation data using LNKnet from MIT Lincoln Laboratory. Once the weights of the network were determined in training, the output sigmoid was removed from the perceptron, so the final combined score for this system was not nonlinearly compressed by a sigmoid. See [1] for further discussion of this sigmoid removal.

4. MFCC/SVM System

This system used the same feature-mapped MFCCs as used in the MFCC/GMM system, but the classifier was an SVM with a generalized linear discriminant sequence kernel [13]. The classifier used the SVM portion of v2_0.060217 of the MIT Lincoln Laboratory speech tools, which in turn used SVMTorch.

5. Word Language Modeling (WLM)

The CMU-Cambridge Language Modeling Toolkit formed the basis of this system. Version 2.0-beta5 of SONIC was used to generate the transcripts. The words from the transcripts were assembled into pseudo sentences, where a pause greater than one second between words defined a sentence break. Using no sentence breaks, where each conversation side became one sentence, yielded worse performance than using pseudo sentence breaks.

Bigram language models were trained with the following parameters set in the toolkit: top 20,000 words, Witten-Bell discounting, and zero cut-offs. Target models were trained by concatenating all the sentences for each of the

[1] This constraint was due to the fact that the pitch extractor was found to output pitch-doubled frames at times.
[2] These were the defaults output by the formant tracker when it failed to find suitable formant candidates.

conversations allowed for each model, while the background model was built in a similar way, but with all the sentences from all the files that made up the background model. The background model data came from transcripts of the data used in the background model for the GMM-based systems, except that transcripts of the OGI data were not used.

To evaluate the language modeling system, each group of test file sentences was tested against its list of claimant models, and the background model. The final score for a given test file and claimant model pair was:

$$\frac{1}{K}\sum_{k=1}^{K} \log(\text{Prob}_{\text{Claimant}}(k)) - \log(\text{Prob}_{\text{Background}}(k)) ,$$

where K is the number of matching bigrams. Unknown or non-matching bigrams were ignored. One final step was taken with the inclusion of gender-independent T-norm. Fifty male and fifty female models were built using two conversation sides of data from Switchboard II with transcripts generated by SONIC.

6. System Fusion and Thresholds

For all of the four-wire training and testing conditions, the component systems scores were combined using a single-layer perceptron built from the 2004 evaluation data using LNKnet. For each training and testing condition, the test control file (*i.e.,* the list of test file/claimant model pairs) from the 2004 evaluation was split into ten "disjoint" parts. By disjoint, we mean that there were no test file/claimant pairs common to two or more parts (thus one could concatenate the parts to recover the original control file). Further, all of the test files from a given speaker were contained in a single split control file. For each split control file, a training control file was constructed from the original control file such that it had no speakers in common with the split control file either in terms of test files or in terms of claimant models. Using the ten split training files, ten perceptrons were built and applied to the system scores for their respective split control files. The fused results for the splits were concatenated, and the best perceptron parameters were determined as well as the thresholds to be applied for the 2005 evaluation. With the best set of perceptron parameters and the thresholds, new perceptrons were built from the entire 2004 control file for each condition to be applied to the 2005 evaluation, but using the thresholds determined from the combination of the splits. Note that the output scores from the fusion system are *not* meant to be posterior probabilities.

7. Two-wire Training and Testing Conditions

7.1. 3conv2w Training

The procedure used for 3conv2w training conditions is presented in detail in [2] and briefly discussed here. In order to complete the 3conv2w training task, each speech file was first classified as having either same-gender (male-male or female-female) or opposite-gender (male-female) speakers. This was accomplished as follows. First, each speech file was scored against male and female GMMs (with 2048 mixtures and diagonal covariance matrices) built using the background model data and with MFCCs band limited to 300–3138 Hz (with RASTA filtering and deltas, but without feature mapping or mean and variance normalization). If the target speaker was male, then for the conversation to be classified as opposite-gender, the background score minus the male score had to be greater than a threshold; otherwise the speech file was classified as same-gender. The threshold was determined from the NIST 2004 evaluation data such that the probability of misclassifying a conversation as opposite-gender was low (approximately 3%). The procedure worked in a likewise fashion if the target speaker was female.

If a speech file was classified as opposite-gender, then the target speaker was extracted as follows. First, the MFCC/HMM SAD was used to determine the speech and non-speech segments. Next, each speech segment was scored against the male and female GMMs. The top 90% of the speech segments in which the desired gender scored highest were labeled as being produced by the target speaker.

If a speech file was classified as same-gender, an agglomerative clustering method was used to cluster similar speech segments. In order to determine which segments should be clustered together, a GMM system was used. A 64-mixture GMM was trained using all of the vectors classified as speech, and then the weights of this model were adapted to fit the characteristics of each speech segment (thus creating a separate model for each segment). In each stage of clustering, a likelihood ratio metric was calculated and the two closest segments were merged [14]. This

process was repeated until three sets of segments were left (presumably, a set of segments for each of the two speakers and a set of "garbage" segments). The feature set used for the clustering consisted of MFCCs band limited to 200–2860 Hz with deltas, but without RASTA filtering, feature mapping, or mean and variance normalization.

The procedure for determining the common speaker from all three files was varied based on how many of the files were classified as opposite-gender. If none of the files for a target speaker were classified as opposite-gender, the following clustering procedure was applied. First, the means of the background model (with 2048 mixtures and diagonal covariance matrices) were adapted to fit the characteristics of each of the nine speech clusters (thus creating a separate model for each of the three clusters from each of the three files). In each stage of clustering, the feature vectors for each speech segment were scored against all of the models and the highest scoring feature vector/model pair was merged. This process was repeated to select a single speech cluster from each of the speech files. The feature set used in clustering consisted of MFCCs band limited to 300–3138 Hz with RASTA filtering, deltas, feature-mapping, and mean and variance normalization.

If all three of the speech files were classified as opposite-gender, the target speaker clusters from each file were pooled to train the speaker model. If one or two of the speech files were classified as opposite-gender, the target speaker clusters from those files were used to train an initial seed model. The clustering procedure discussed in the previous paragraph was then repeated across the same-gender files, with the restriction that the seed model was always one of the merged models.

7.2. 1conv2w Testing

The 1conv2w testing condition used the same procedure as applied to each of the speech files from the 3conv2w training condition. If the speech file was classified as opposite-gender, only the speech segments of the desired gender were scored; otherwise, each of the three speech clusters were tested against the target model, and the highest score was taken as the overall score.

7.3. Score Combination for Two-Wire Conditions

For the two-wire conditions, the component system scores were combined with a linear weighting that was determined experimentally using the NIST 2004 evaluation data. For all cases involving two-wire training or testing with the exception of the 8conv4w-1conv2w condition, the combination was:

$$score_{Combined} = 0.5*score_{MFCC/GMM} + 0.4*score_{MFCC/SVM} + 0.1*score_{FMBWF0/GMM}$$

For the 8conv4w-1conv2w condition, the combination was:

$$score_{Combined} = 0.4*score_{MFCC/GMM} + 0.5*score_{MFCC/SVM} + 0.1*score_{FMBWF0/GMM}$$

8. References

[1] R. Slyh, E. Hansen, and B. Ore, "The 2005 AFRL/HEC one-speaker detection systems," *Proceedings of Odyssey 2006: The Speaker and Language Recognition Workshop,* (San Juan, Puerto Rico), June 2006.

[2] B. Ore, R. Slyh, and E. Hansen, "Speaker segmentation and clustering using gender information," *Proceedings of Odyssey 2006: The Speaker and Language Recognition Workshop,* (San Juan, Puerto Rico), June 2006.

[3] E. Hansen, R. Slyh, and T. Anderson, "Supervised and unsupervised speaker adaptation in the NIST 2005 speaker recognition evaluation," *Proceedings of Odyssey 2006: The Speaker and Language Recognition Workshop,* (San Juan, Puerto Rico), June 2006.

[4] D. Reynolds, T. Quatieri, and R. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Processing*, vol. 10, nos. 1–3, pp. 19–41, 2000.

[5] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 4, pp. 578–589, October 1994.

[6] D. Reynolds, "Channel robust speaker verification via feature mapping," *Proceedings of ICASSP 2003*, (Hong Kong), April 2003.

[7] R. Auckenthaler, M. Carey, and H. Lloyd-Thomas, "Score normalization for text-independent speaker verification systems," *Digital Signal Processing*, vol. 10, nos. 1–3, pp. 42–54, 2000.

[8] D. Talkin, "A robust algorithm for pitch tracking (RAPT)," in *Speech Coding and Synthesis*, W. B. Kleijn and K. K. Paliwal, eds., Elsevier: New York, 1995.

[9] E. Hansen, R. Slyh, and T. Anderson, "Formant and F0 features for speaker recognition," *Proceedings of A Speaker Odyssey: The Speaker Recognition Workshop*, (Chania, Crete, Greece), June 2001.

[10] B. Pellom and K. Hacioglu, "Recent improvements in the CU SONIC ASR system for noisy speech: The SPINE task," *Proceedings of ICASSP 2003*, (Hong Kong), April 2003.

[11] B. Pellom, "SONIC: The University of Colorado Continuous Speech Recognizer," University of Colorado, Technical Report #TR-CSLR-2001-01, Boulder, Colorado, March 2001.

[12] E. Hansen, R. Slyh, and T. Anderson, "Speaker recognition using phoneme-specific GMMs," *Proceedings of Odyssey 2004: The Speaker and Language Recognition Workshop*, (Toledo, Spain), May–June 2004.

[13] W. Campbell, "Generalized linear discriminant sequence kernels for speaker recognition," *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, (Orlando FL), May 2002.

[14] L. Wilcox, F. Chen, D. Kimber, and V. Balasubramanian, "Segmentation of speech using speaker identification," *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, (Adelaide, South Australia), April 1994.