AFRL-HE-WP-TP-2006-0027

AIR FORCE RESEARCH LABORATORY

The 2005 AFRL/HEC One-Speaker Detection Systems

Raymond E. Slyh
Eric G. Hansen

Human Effectiveness Directorate
Warfighter Interface Division
Wright-Patterson AFB OH 45433-7022

Brian M. Ore

General Dynamics
5200 Springfield Pike, STE 200
Dayton OH 45431

February 2006

Human Effectiveness Directorate
Warfighter Interface Division
Wright-Patterson AFB OH 45433

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*
February 2006

2. REPORT TYPE
Proceedings

3. DATES COVERED *(From - To)*

4. TITLE AND SUBTITLE
The 2005 AFRL/HEC One-Speaker Detection Systems

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
*Raymond E. Slyh, *Eric G. Hansen, **Brian M. Ore

5d. PROJECT NUMBER
7184

5e. TASK NUMBER
08

5f. WORK UNIT NUMBER
71

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) AND ADDRESS(ES)
**General Dynamics
5200 Springfield Street, STE 200
Dayton OH 45431

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
**Air Force Materiel Command
Air Force Research Laboratory
Human Effectiveness Directorate
Warfighter Interface Division
Wright-Patterson AFB OH 45433-7022

10. SPONSOR/MONITOR'S ACRONYM(S)
AFRL/HECP

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
AFRL-HE-WP-TP-2006-0027

12. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES
This will be published in the Proceedings of Odyssey 2006: The Speaker and Language Recognition Workshop. The clearance numbers are AFRL/WS-06-0430 and AFMC/PAX-06-067, cleared 17 February 2006.

14. ABSTRACT

This paper describes the one-speaker detection systems submitted by AFRL/HEC for several of the training and testing conditions in the 2005 NIST Speaker Recognition Evaluation. For each condition, the overall system score was the weighted combination of scores from several component systems. The component systems were based on (1) mel-frequency cepstral coefficients (MFCCs) and Gaussian mixture models (GMMs); (2) MFCCs and phoneme-specific GMMs (PS-GMMs); (3) linear-prediction-based cepstral coefficients (LPCCs) from closed-phase analysis; (4) formant center frequencies, formant bandwidths, and fundamental frequency)FMBWFO); and (5) word language modeling (WLM). The score combination was done using single-layer perceptrons, with the grouping of the component systems depending on the lengths of the training and testing files. For some of the testing and/or training conditions involving ten-second speech files, the system performance improved from the inclusion of the FMBWFO and LPCC systems, while the MFCC/PS-GMM system provided additional benefits in the one-conversation testing conditions involving larger amounts of training data.

15. SUBJECT TERMS
Speaker Detection Systems

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON** Raymond E. Slyh
a. REPORT UNC	**b. ABSTRACT** UNC	**c. THIS PAGE** UNC	SAR	10	**19b. TELEPHONE NUMBER** *(include area code)* (937) 255-9248

The 2005 AFRL/HEC One-Speaker Detection Systems

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

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

Abstract

This paper describes the one-speaker detection systems submitted by AFRL/HEC for several of the training and testing conditions in the 2005 NIST Speaker Recognition Evaluation. For each condition, the overall system score was the weighted combination of scores from several component systems. The component systems were based on (1) mel-frequency cepstral coefficients (MFCCs) and Gaussian mixture models (GMMs); (2) MFCCs and phoneme-specific GMMs (PS-GMMs); (3) linear-prediction-based cepstral coefficients (LPCCs) from closed-phase analysis; (4) formant center frequencies, formant bandwidths, and fundamental frequency (FMBWF0); and (5) word language modeling (WLM). The score combination was done using single-layer perceptrons, with the grouping of the component systems depending on the lengths of the training and testing files. For some of the testing and/or training conditions involving ten-second speech files, the system performance improved from the inclusion of the FMBWF0 and LPCC systems, while the MFCC/PS-GMM system provided additional benefits in the one-conversation testing conditions involving larger amounts of training data.

1. Introduction

This paper describes the speaker recognition systems submitted by AFRL/HEC for the (four-wire) one-speaker detection conditions in the 2005 Speaker Recognition Evaluation (SRE) sponsored by the National Institute of Standards and Technology (NIST) [1].[1] One of the recent trends in speaker recognition is the fusion or combination of the output scores from several systems such as in [3] to provide an overall score, and our system was similar in this regard. For each condition, the overall system score was the weighted combination of scores from several component systems. The component systems were based on (1) mel-frequency cepstral coefficients (MFCCs) and Gaussian mixture models (GMMs); (2) MFCCs and phoneme-specific GMMs (PS-GMMs); (3) linear-prediction-based cepstral coefficients (LPCCs) from closed-phase analysis and GMMs; (4) formant center frequencies, formant bandwidths, and fundamental frequency with GMMs (denoted here by FMBWF0); and (5) language modeling on the words from speech recognition transcripts (denoted here by WLM). For testing or training conditions involving short speech files, the scores from the MFCC, FMBWF0, and LPCC systems were combined using a single-layer perceptron (SLP). For testing and training conditions involving larger amounts of speech data, the score combination was done in two stages. First, the scores from fifteen PS-GMM systems were combined using an SLP. Then, the output score from the SLP was combined with the scores from the MFCC, FMBWF0, LPCC, and WLM systems to yield the final score.

We show that, compared to the baseline MFCC/GMM system, the inclusion of the FMBWF0 and LPCC systems improved the performance for some of the training and/or testing conditions involving ten-second speech files, while the inclusion of the MFCC/PS-GMM system improved the performance for the training and testing conditions involving larger amounts of data.

An outline of the paper is as follows. The next section briefly describes the 2005 evaluation conditions considered in this paper. Section 3 describes the component systems as well as the speech activity detector (SAD) used with some of the GMM-based systems, while Section 4 describes the development of the score combination system. Section 5 presents the evaluation performance results, and Section 6 presents the results of some post-evaluation experiments aimed at improving the use of the PS-GMM system. Finally, Section 7 presents the conclusions.

2. The NIST 2005 Speaker Recognition Evaluation

The NIST 2005 SRE consisted of 20 distinct tasks [1]. Here, we consider the eight standard one-speaker detection tasks, consisting of four training conditions by two testing conditions. The training conditions all involved four-wire (two-channel) conversations and were defined by the following amounts of data: (1) an excerpt estimated to contain approximately 10 seconds of speech of the target on its designated side (designated as 10sec4w),

Opinions, interpretations, and conclusions are those of the authors and are not necessarily endorsed by the United States Air Force.

[1] The AFRL/HEC system submitted for the conditions requiring speaker segmentation and clustering is described in [2].

(2) one conversation of approximately five minutes total duration with the target speaker (designated as 1conv4w), (3) three conversations involving the target speaker (designated as 3conv4w), and (4) eight conversations involving the target speaker (designated as 8conv4w). The testing conditions involved either (1) 10 seconds of speech (designated as 10sec4w) or (2) one five-minute conversation (designated as 1conv4w) as in the 10sec4w and 1conv4w training conditions, respectively.

In addition to the speech files, NIST provided transcripts produced by an English-language speech recognition system from BBN with word error rates typically in the range of 15–30% for English conversational telephone speech. English language transcripts were provided for all files, despite the fact that some of the files contained speech in other languages—namely, Arabic, Mandarin, Russian, and Spanish.

NIST compares system performance in two major ways. First, NIST uses a detection cost function, C_D, defined as a weighted sum of miss and false alarm probabilities:

$$C_D = C_M P_{M|T} P_T + C_{FA} P_{FA|NT}(1 - P_T),$$

where C_M is the cost of a miss (chosen by NIST as 10), C_{FA} is the cost of a false alarm (chosen by NIST as 1), P_T is the *a priori* probability of a target (chosen by NIST as 0.01), $P_{M|T}$ is the probability of a miss given a target trial, and $P_{FA|NT}$ is the probability of a false alarm given a non-target trial. $P_{M|T}$ and $P_{FA|NT}$ are a function of system performance and the chosen detection threshold. For a given system, chosen costs, and *a priori* target probability, there is a threshold that yields a minimum value of C_D; we refer to this minimum value of C_D as the minDCF value. Second, NIST uses plots of $P_{M|T}$ versus $P_{FA|NT}$, called Detection Error Trade-off (DET) plots [4], to show how system performance varies for a wide range of operating points. In addition to these two presentations of performance, we will also use the equal error rate (EER), the value of $P_{M|T}$ (or $P_{FA|NT}$) when $P_{M|T} = P_{FA|NT}$.

3. Component Systems

The overall system consisted of various combinations of the scores from five component systems, depending on the length of the training and testing files. Four of the component systems were based on GMMs, while the WLM system involved language modeling. The next subsection discusses the GMM-based systems, while Subsection 3.2 discusses the WLM system.

3.1. GMM-Based Systems

This section discusses the various GMM-based systems. Common aspects of the systems are presented in the next subsection, while the unique aspects of each feature set are discussed in their respective subsections.

3.1.1. Overview of GMM-Based Systems

The GMM-based systems, regardless of feature set, all used Version 2.1 of the MIT Lincoln Laboratory (MIT-LL) MFCC/GMM system [5] with 2048 mixtures per model and diagonal covariance matrices for each mixture.

All of the GMM-based systems used a common speech activity detector (SAD),[2] 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[3] using 64 mixtures per state. The energy-based detection was performed using the MIT-LL *xtalk* program from their MFCC/GMM speaker recognition system.

The cepstral-coefficient-based systems (*i.e.*, all of the GMM-based systems except the FMBWF0 system) shared a number of additional similarities. Each set of cepstral coefficients had RASTA filtering [6] applied and included the deltas of the features. After the RASTA filtering of the cepstral features and the deltas were added, feature mapping [7] was also used; however, the channel was always chosen using the channel determined by the MFCCs. Finally, the mapped features were normalized to have zero mean and unit variance.

Gender-dependent T-norm [8] 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 conditions. For the 10sec4w-10sec4w training/testing 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 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[4] (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. The T-norm model data came from NIST 2001–2003 evaluation data.

[2] The MFCC/PS-GMM system only used this SAD if the SAD from the SONIC speech recognizer failed to find any speech frames.

[3] Available at: http://htk.eng.cam.ac.uk/

[4] See: http://cslu.cse.ogi.edu/corpora/corpCurrent.html

3.1.2. The MFCC/GMM System

Nineteen MFCCs were computed using the MIT-LL GMM system [5] in the bandwidth of 300–3138 Hz every 10 msec. RASTA filtering was applied to the MFCCs and deltas were then calculated. Only frames labeled as speech by the SAD (discussed in Section 3.1.1) were used. The remaining processing was performed as discussed in Section 3.1.1. In building target and T-norm models, only the mixture means were adapted from those of the background model.

3.1.3. The LPCC System

The LPCC system calculated 16 cepstral coefficients (excluding the 0^{th} cepstral coefficient) from the linear prediction (LP) parameters derived from smoothed closed-phase analysis as described in [9]. The cepstral coefficients were computed from the LP parameters using the recursion outlined in [10]. The features were only calculated for voiced speech frames, where the voicing was determined using the *get_f0* program from the Entropic Signal Processing System (ESPS). RASTA filtering was applied, and the feature set included the deltas of the features. The remaining processing was performed as discussed in Section 3.1.1. In building target and T-norm models, only the mixture means were adapted from those of the background model.

3.1.4. The FMBWF0 System

The FMBWF0 system was similar to that of [11]. First, F0 and the probability of voicing were determined every 10 msec using the ESPS *get_f0* command, which implements the pitch tracking algorithm described in [12]. Next, the first three formant center frequencies (F1–F3) and the first three formant bandwidths (B1–B3) were determined from Snack Version 2.2.2 from KTH.[5] Each F0 value was converted to log scale. Each formant center frequency and bandwidth value was converted to radians.

Extracted frames had (1) to be declared to be speech by the SAD, (2) to be voiced; (3) to have F0 < 250 Hz; and (4) to have (F1, F2, F3) $\neq$ (500 Hz, 1500 Hz, 2500 Hz). Condition (3) was imposed because the pitch extractor was found to output pitch-doubled frames at times, while condition (4) was imposed to eliminate frames where the formant tracker failed (at which point it would output the default values of 500, 1500, and 2500).

These features were used in the GMM system, and T-norm was applied as discussed in Section 3.1.1. Target and T-norm models were adapted from the background model by updating the weights, means, and variances.

3.1.5. The MFCC/PS-GMM System

The basic idea of the MFCC/PS-GMM system is to assign each feature vector a phoneme label, build a GMM for each phoneme for each speaker, score each labeled feature vector against the proper phoneme-specific model, and combine the phoneme-specific scores to form a single output score. The MFCC/PS-GMM system was similar to the system described in [13] that used phoneme-only adaptation, but with some notable changes. First, this year's system used MFCCs that were computed as in the MFCC/GMM system described in Sections 3.1.1 and 3.1.2, including the use of feature mapping, which was not used in the system of [13]. Second, each feature vector was associated with a phoneme label as output by the SONIC speech recognizer (Version 2.0-beta2) from the University of Colorado at Boulder [14, 15], whereas the system of [13] used phoneme labels from speech recognition transcripts provided by Stanford Research Institute for the NIST 2003 Extended Data Task. Thus, the phoneme alignments were constructed from the state file output by SONIC, and the feature vectors for a given phoneme were then scored with a GMM built for that phoneme. Third, in contrast to the system of [13], the GMM for each phoneme used 2048 mixtures. Finally, only phonemes from the following set were used: {AE, AH, AX, AY, DH, EH, EY, IH, IY, L, M, N, OW, S, Y}, in contrast to the larger set used in [13].

There are some additional points worth noting. First, SONIC has its own SAD, so the SAD described in Section 3.1.1 was only used if the SONIC SAD returned no speech frames. Second, the acoustic and trigram language models used with SONIC were trained using land line data from the Switchboard database.[6] Third, target and T-norm phoneme-specific models were adapted from the background phoneme-specific models by updating only the means. Finally, the scores for each phoneme (after the phoneme-dependent T-norm was applied) were combined with a perceptron neural net that was trained using the MIT-LL LNKnet package.[7] The neural net used no hidden layers, and the output nonlinearity was a standard sigmoid. The neural net was trained using data from the NIST 2004 Evaluation.

3.2. The WLM System

The WLM system is motivated by the original work done by Doddington on idiolectal differences between speakers [16]. The CMU-Cambridge Language Modeling Toolkit[8] (Version 2.05) formed the basis of this system. The words from the NIST-supplied transcripts were assembled into pseudo sentences, where a pause greater than one second between words defined a sentence break.

[5] Available at: http://www.speech.kth.se/snack

[6] See http://www.ldc.upenn.edu

[7] Available at: http://www.ll.mit.edu/IST/lnknet

[8] Available at: http://svr-www.eng.cam.ac.uk/ prc14/toolkit.html

Using no sentence breaks, where each conversation side became one sentence, yielded worse performance than using pseudo sentence breaks when tested on previous NIST evaluations.

Bigram language models with back-off 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 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 Switchboard II.

To compute a score using the WLM system, the sentences from a test file were tested against a claimant model and the background model. The score for a given test file and claimant model pair was computed as follows. Let B_C be the set of bigrams in the claimant model, C; B_B be the set of bigrams in the background model; and B_T be the set of bigrams in a test file, T. Let $B_{TCB} = B_T \cap B_C \cap B_B$, and let N_{TCB} be the number of bigrams in B_{TCB}. Let $P_{b,C}$ be the probability of bigram b in model C and $P_{b,B}$ be the probability of bigram b in the background model. The score for T against the claimant model C was computed as:

$$s(T, C) = \frac{1}{N_{TCB}} \sum_{b \in B_{TCB}} \log(P_{b,C}) - \log(P_{b,B}).$$

Thus, unknown or non-matching bigrams were ignored.

One final step was taken with the inclusion of a gender-independent T-norm. Fifty male and fifty female models were built using two conversation sides of data from Switchboard II with transcripts supplied by NIST that were generated by a BBN speech recognizer.[9]

4. System Combination and Thresholds

For all of the one-speaker detection training and testing conditions, the component system scores were combined using SLPs 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/target model pairs) from the 2004 evaluation was split into ten disjoint parts. In other words, there were no test file/target pairs common to two or more parts (thus, one could concatenate the parts to recover all of the test file/target model pairs from 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 target models. Using the ten split training files, ten SLPs were built and applied to the system scores for their respective split control files. The score combination results for the splits were concatenated, and the thresholds to be applied for the 2005 evaluation were determined. Then, new SLPs 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.

5. Evaluation Results

This section presents the performance results of the individual component systems as well as that of the overall submitted system for each of the one-speaker detection conditions. The performance is shown using DET plots, and in some cases, the corresponding system minDCF and EERs are given.

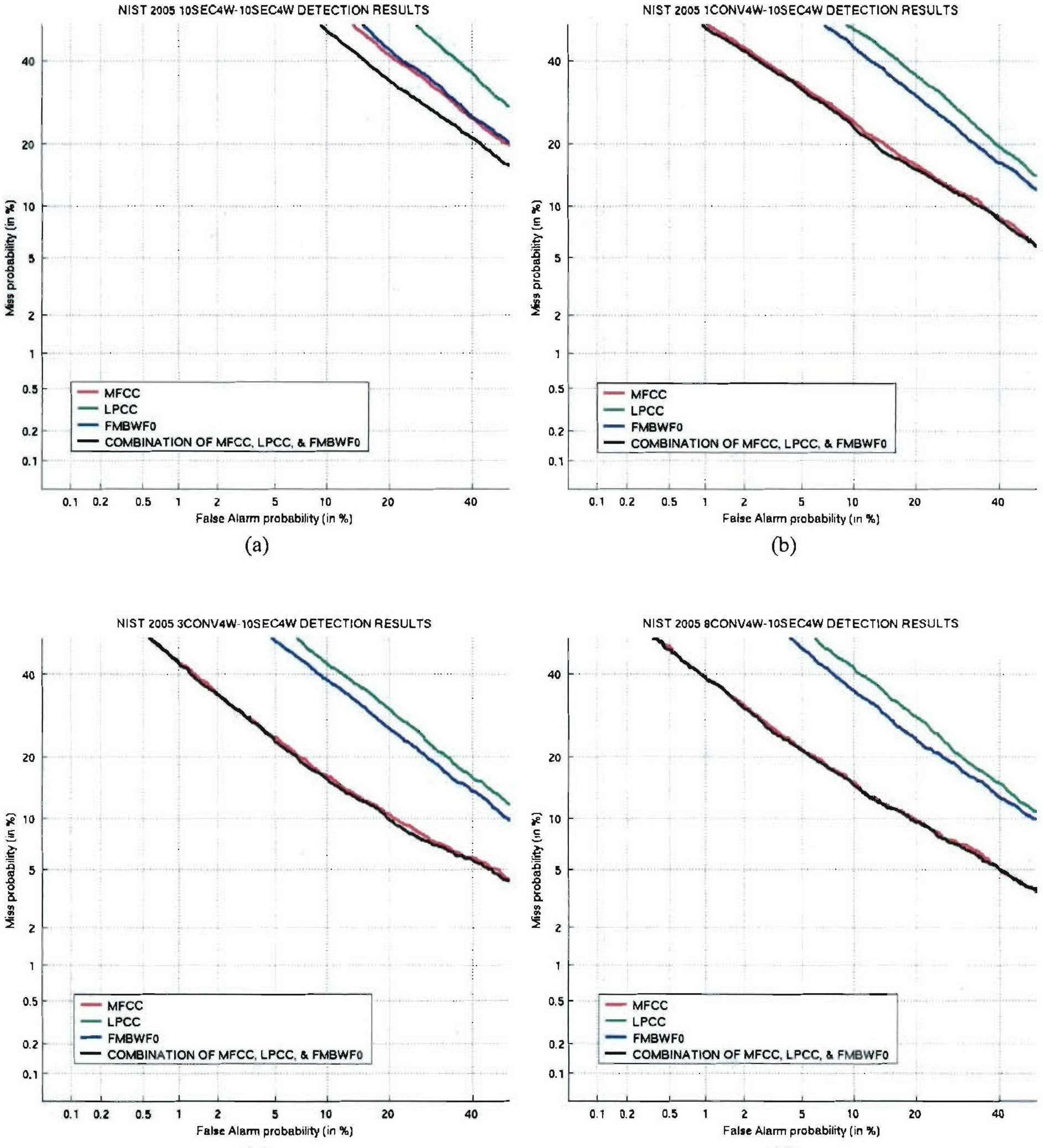
5.1. 10sec4w Testing Conditions

Figures 1(a)–(d) show the performance of the component and combined systems for 10sec4w testing with the training conditions of 10sec4w, 1conv4w, 3conv4w, and 8conv4w, respectively. The minDCF values for the combined systems for the 10sec4w, 1conv4w, 3conv4w, and 8conv4w training conditions were 0.0860, 0.0590, 0.0522, and 0.0485, respectively, while the EERs were 28.20%, 17.02%, 13.30%, and 12.65%, respectively. From the plots, one can see that the combination of the LPCC, FMBWF0, and MFCC/GMM systems leads to substantial performance improvement relative to that of the standard MFCC/GMM system for the 10sec4w training condition; however, the combination systems do not yield any substantial performance improvement for the 1conv4w, 3conv4w, and 8conv4w training conditions. The FMBWF0 and LPCC systems combine with the MFCC/GMM system to yield a 6.0% relative improvement in minDCF and a 10.9% relative improvement in EER over those obtained solely with the MFCC/GMM system (minDCF = 0.0915, EER = 31.65%). Also, in the 10sec4w training condition, the FMBWF0 system performs almost as well as the MFCC/GMM system.

5.2. 1conv4w Testing Conditions

Figure 2(a) shows the performance for the 10sec4w-1conv4w training/testing condition. Note that for the 10sec4w-1conv4w condition, the roles of the training and testing files were reversed. Thus, models were built using the 1conv4w files from the test list, and the frames of the 10sec4w training files were scored against these models. With this role reversal, this condition was similar to the 1conv4w-10sec4w condition shown in Figure 1(b). Figure 2(a) shows that the FMBWF0 and LPCC systems provide a benefit over the MFCC/GMM system alone, improving the minDCF from 0.0708 to 0.0675 and improving the EER slightly from 20.75 to 19.93%.

[9] Note that the recognizer used to generate transcripts for Switchboard II does not appear to be of the same vintage as that used to generate transcripts for the 2005 Evaluation data.



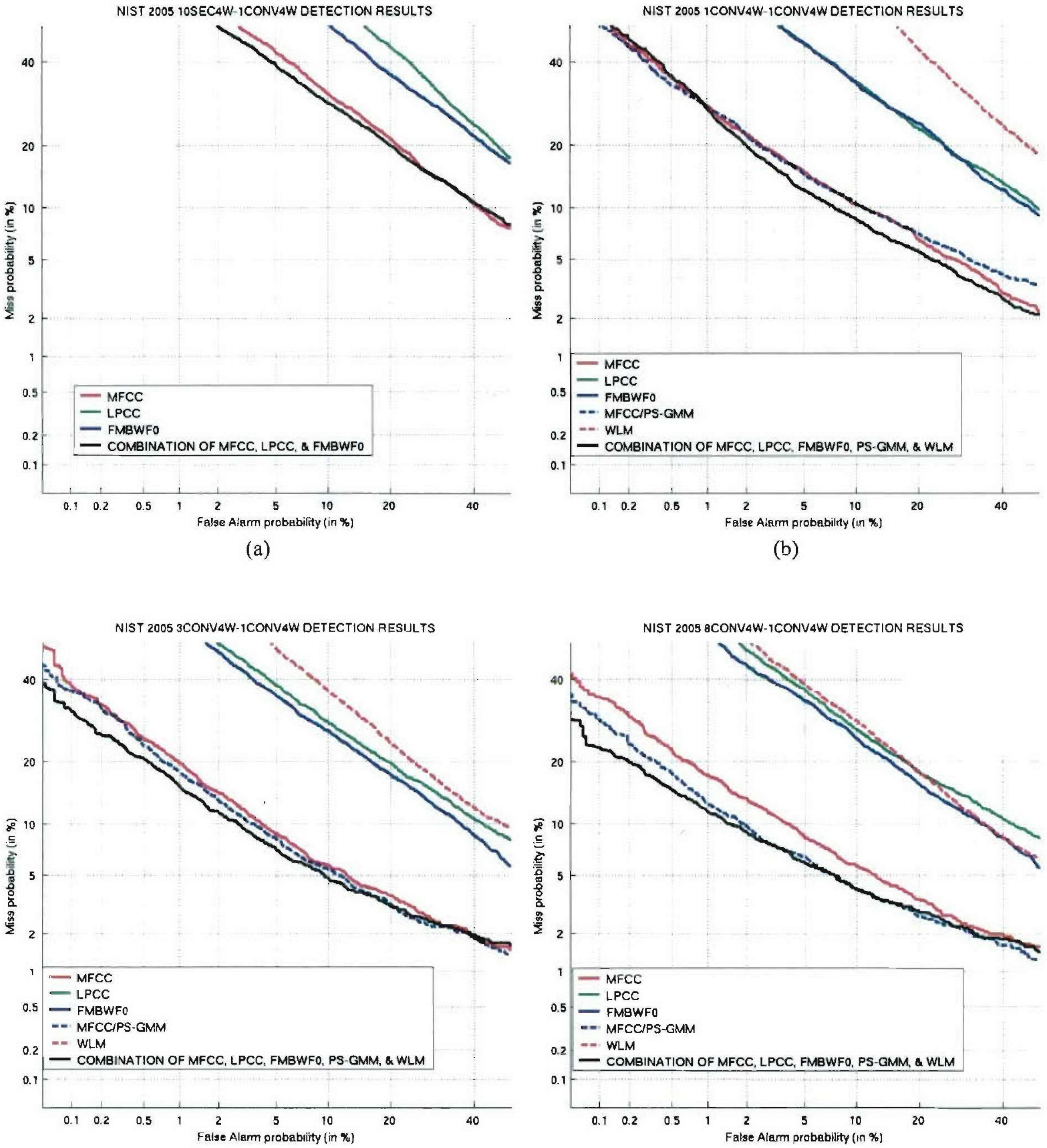
(a)

(b)

(c)

(d)

Figure 1: DET plots for the 10sec4w testing conditions showing the performance of the LPCC, FMBWF0, and MFCC/GMM systems with that of the combined systems for the (a) 10sec4w, (b) 1conv4w, (c) 3conv4w, and (d) 8conv4w training conditions.



(a)

(b)

(c)

(d)

Figure 2: DET plots for the 1conv4w testing conditions showing the performance of the LPCC, FMBWF0, MFCC/GMM, MFCC/PS-GMM, and WLM systems with that of the combined systems for the (a) 10sec4w, (b) 1conv4w, (c) 3conv4w, and (d) 8conv4w training conditions.

	1conv4w		3conv4w		8conv4w	
System	**minDCF**	**EER**	**minDCF**	**EER**	**minDCF**	**EER**
MFCC/GMM	0.0383	10.33%	0.0293	7.11%	0.0265	7.09%
MFCC/PS-GMM	0.0381	10.33%	0.0271	6.70%	0.0220	5.75%
C1: Original Combination	0.0371	9.15%	0.0249	6.22%	0.0197	5.66%
C2: Combination Without Sigmoid	0.0335	8.88%	0.0239	6.02%	0.0187	5.30%
Relative Improvement from C1	3.2%	11.4%	15.0%	12.5%	25.7%	20.2%
Relative Improvement from C2	12.5%	14.0%	18.4%	15.3%	29.4%	25.2%

Table 1: EER and minDCF for the MFCC/GMM, the MFCC/PS-GMM, and two combination systems for 1conv4w testing with 1conv4w, 3conv4w, and 8conv4w training. The first combination system, C1, is the original combination system discussed in Section 4; the second combination system, C2, is the combination system with the sigmoid removal discussed in Section 6. Also shown are the improvements in the performance of the combination systems, C1 and C2, relative to that of the MFCC/GMM systems for each condition.

Figures 2(c)–(d) show the performance of the systems for 1conv4w testing over the training conditions of 1conv4w, 3conv4w, and 8conv4w, respectively. The MFCC/PS-GMM system outperforms the MFCC/GMM system for the 8conv4w training condition, but doesn't significantly outperform it for the 1conv4w and 3conv4w training conditions. The combination system outperforms the MFCC/GMM system alone for the 3conv4w and 8conv4w training conditions and for the 1conv4w training condition with $1\% < P_{FA|NT}$. Table 1 shows the EERs and minDCF values for the MFCC/GMM, MFCC/PS-GMM, and the original combination system, designated C1, for the 1conv4w, 3conv4w, and 8conv4w training conditions (along with the performance of a second combination system, designated C2, that uses a sigmoid removal procedure to be discussed in Section 6). Also included are the relative improvements in the performance of the combination systems over that of the MFCC/GMM systems. One can see that the C1 combination system outperforms the MFCC/GMM system by 11.4–25.7% in minDCF and EER for these conditions, except for the minDCF for 1conv4w training, which is only improved by 3.2%.

6. Post-Evaluation Experiments

After the evaluation, additional experimentation was conducted in an effort to improve the utilization of the MFCC/PS-GMM system scores in the overall combined system. After training the first-stage SLP applied to the MFCC/PS-GMM scores, the output sigmoid of the SLP was removed. The combined MFCC/PS-GMM score without the sigmoid was then used along with the four other system scores to train the second-stage SLP.

Figure 3 shows the result of removing the sigmoid for 1conv4w and 8conv4w training with 1conv4w testing, while Table 1 shows the corresponding minDCF and EERs (as well as the values for the 3conv4w training case) in the row designated as "C2: Combination Without Sigmoid." Relative to the original C1 score combination method, the sigmoid removal yields substantial performance improvement in the 1conv4w training condition in the low false alarm region, resulting in a 9.7% relative improvement in minDCF. In contrast, the sigmoid removal yields only modest additional improvement over that provided by the C1 combination system for the 3conv4w and 8conv4w training conditions.

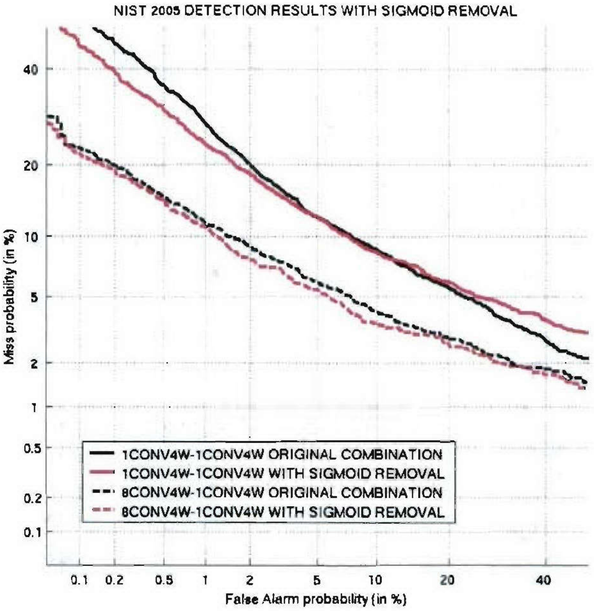


Figure 3: DET plot comparing the MFCC/GMM and combined system performance with and without the sigmoid for 1conv4w testing under the 1conv4w and 8conv4w training conditions.

7. Conclusions

We have discussed the details and presented the performance of the AFRL/HEC one-speaker detection systems submitted for the 2005 NIST Speaker Recognition Evaluation. It was shown that the FMBWF0 and LPCC systems combined with the MFCC/GMM system to improve the performance relative to that of the MFCC/GMM system in some of the conditions involving speech files on the order of 10 sec. The MFCC/PS-GMM provided additional performance benefit over that provided solely by the MFCC/GMM system for conditions involving longer training and testing files, especially when combined using the two SLPs with the sigmoid removed from the first SLP after training.

8. Acknowledgements

The authors would like to acknowledge the considerable support from MIT Lincoln Laboratory by providing their MFCC/GMM system and the LNKnet neural network package. Thanks go to Prof. Bryan Pellom of the University of Colorado at Boulder for the SONIC speech recognizer and for help with its use. Thanks go to Barbara Peskin of ICSI for suggesting the removal of the output sigmoid after training of the SLP used to combine the scores of the PS-GMM systems. Finally, Mr. Anthony Halley, formerly of General Dynamics Advanced Information Systems, ran the SONIC speech recognizer to generate the transcripts used in the PS-GMM system.

9. References

[1] NIST, *The NIST Year 2005 Speaker Recognition Evaluation Plan,* Version 6, 29 March 2005. (Available at: http://www.nist.gov/speech/tests/spk/2005/sre-05_evalplan-v6.pdf)

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

[3] D. Reynolds, *et al.*, "The SuperSID project: Exploiting high-level information for high-accuracy speaker recognition," in *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP),* (Hong Kong), April 2003.

[4] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki, "The DET curve in assessment of detection task performance", in *Proc. of EuroSpeech '97,* (Rhodes, Greece), September 1997.

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

[6] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE Transactions on Speech and Audio Processing*, vol. 2, pp. 578–589, October 1994.

[7] D. Reynolds, "Channel robust speaker verification via feature mapping," in *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP),* (Hong Kong), April 2003.

[8] 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.

[9] R. Slyh, E. Hansen, and T. Anderson, "Glottal modeling and closed-phase analysis for speaker recognition," in *Proceedings of Odyssey 2004: The Speaker and Language Recognition Workshop,* (Toledo, Spain), May–June 2004.

[10] J. Deller, J. Proakis, and J. Hansen, *Discrete-Time Processing of Speech Signals,* MacMillan: New York, 1993.

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

[12] 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.

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

[14] B. Pellom and K. Hacioglu, "Recent improvements in the CU SONIC ASR system for noisy speech: The SPINE task", in *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP),* (Hong Kong), April 2003.

[15] B. Pellom, *SONIC: The University of Colorado Continuous Speech Recognizer,* University of Colorado, Technical Report TR-CSLR-2001-01, (Boulder, Colorado), March 2001.

[16] G. Doddington, "Speaker recognition based on idiolectal differences between speakers," in *Proceedings of EUROSPEECH,* (Aalborg, Denmark), September 2001.