Fusing High- and Low-Level Features for Speaker Recognition•+

Joseph P. Campbell, Douglas A. Reynolds, Robert B. Dunn

MIT Lincoln Laboratory
Lexington, Massachusetts 02420 USA
{jpc, dar, rbd}@ll.mit.edu

ABSTRACT

The area of automatic speaker recognition has been dominated by systems using only short-term, low-level acoustic information, such as cepstral features. While these systems have produced low error rates, they ignore higher levels of information beyond low-level acoustics that convey speaker information. Recently published works have demonstrated that such high-level information can be used successfully in automatic speaker recognition systems by improving accuracy and potentially increasing robustness. Wide ranging high-level-feature-based approaches using pronunciation models, prosodic dynamics, pitch gestures, phone streams, and conversational interactions were explored and developed under the SuperSID project at the 2002 JHU CLSP Summer Workshop (WS2002): http://www.clsp.jhu.edu/ws2002/groups/supersid/. In this paper, we show how these novel features and classifiers provide complementary information and can be fused together to drive down the equal error rate on the 2001 NIST Extended Data Task to 0.2%—a 71% relative reduction in error over the previous state of the art.

1. INTRODUCTION

Humans rely on several different types or levels of information in the speech signal to recognize others from voice, alone. We can roughly categorize these features into a hierarchy running from low-level information, such as the sound of a person's voice (related to physical traits of the vocal apparatus), to high-level information, such as particular word usage or idiolect (related to learned habits and style). While all of these levels appear to convey useful speaker information, automatic speaker recognition systems have relied almost exclusively on low-level information via short-term features related to the speech spectrum.

A concerted research effort to advance these underutilized high-level information sources was undertaken at the 2002 JHU Summer Workshop on Human Language Technology [1]. The time was ripe for this undertaking based on early successful evaluations [2, 3, 4, 5, 6], recent advances in tools to reliably extract features for high-level characterization (e.g., phone and speech recognizers), grand-scale applications providing vast amounts of speech from a speaker to learn speaking habits (e.g., audio mining), large development corpora, and plentiful computational resources. Details of the various approaches undertaken in the project can be found in the companion papers related to the SuperSID project [7, 8, 9, 10, 11] and on the SuperSID website [12].

These new features hold the promise of improving basic recognition accuracy by adding complementary information and, possibly, robustness to acoustic degradations from channel and noise effects, to which low-level features are highly susceptible. After conceiving of and extracting various high-level features, they need to be combined to reinforce each other. This paper describes the fusion techniques developed at MIT Lincoln Laboratory and the use of these techniques to fuse the scores of the component systems developed in the SuperSID project. The resulting fusion system dramatically reduces the error rate.

2. EXTENDED DATA TASK

The focus of the SuperSID project was on text-independent speaker detection using the extended data task from the 2001 NIST Speaker Recognition Evaluation [6]. This task was introduced to allow exploration and development of techniques that can exploit significantly more training data than is traditionally used in NIST evaluations. Speaker models are trained using 1, 2, 4, 8, and 16 complete conversation sides (where a conversation side is nominally 2.5 minutes long) as opposed to the normal 2 minutes of training speech used in other NIST evaluations. A complete conversation side was used for testing. The 2001 Extended Data Task used the entire Switchboard-I conversational telephone speech corpus and was selected for the project because of the availability of several Switchboard-I annotated resources providing features and measures related to high-level speaker information, which are outlined in [1].

To supply a large number of target and nontarget trials and speaker models trained with up to 16 conversations of training speech (~40 minutes), the evaluation used a crossvalidation processing of the entire corpus. The corpus was divided into 6 partitions of ~80 speakers each. All trials within a partition involved models and test segments from within that partition, only; data from the other 5 partitions were available for background model building, normalization, etc. The task consists of ~500 speakers with ~4,100 target models (a speaker had multiple models for different amounts of training data) and

• This work is sponsored by the Department of Defense under Air Force Contract F19628-00-C-0002 and the National Science Foundation under Grant No. 0121285. Opinions, interpretations, conclusions, and recommendations are those of the authors and are not necessarily endorsed by the United States Government.
+ The authors gratefully acknowledge the CLSP group at JHU for organizing and hosting WS2002.

~57,000 trials for the testing phase, containing matched and mismatched handset trials and some cross-sex trials. The crossvalidation experiments were driven by NIST's speaker model training lists and index files indicating which models were to be scored against which conversation sides for each partition.

Scores from each partition are pooled and a detection error tradeoff (DET) curve is plotted to show system results at all operating points. The equal error rate (EER), where the false acceptance rate equals the missed detection rate, is used as a summary performance measure for comparing systems. Each approach formed a likelihood ratio detector by creating a speaker model using training data and a single speaker-independent background model using data from the held-out splits. For some systems, a set of individual background speaker models from the held-out set was used as cohort models. During recognition, a test utterance is scored against the speaker and background model(s) and the ratio (or difference in the log domain) is reported as the detection score for DET plotting and for fusing.

3. HIGH-LEVEL APPROACHES

In this section, we survey some approaches to exploit high-level speaker information. The nine numbered systems were selected for fusion. The reader should consult the referenced papers for more details on the systems developed at the Workshop.

3.1 Acoustic Features

1. Acoustic Baseline (GMM-UBM cepstral features) [13]: Although this project purposely avoided using standard acoustic frame-level signal processing features such as cepstra, we wanted to establish a baseline of standard approaches on the extended data set. The acoustic system was a standard GMM-UBM system using short-term cepstral-based features with a 2048 mixture UBM built using data from the Switchboard-II corpus. This system produces an EER ranging from 3.3% for 1-conversation training to 0.7% for 8-conversation training.

3.2 Prosodic Features

2. Pitch and Energy Distributions [11]: As a baseline, a simple GMM classifier using a feature vector consisting of per-frame log pitch, log energy, and their first derivatives was developed which produced an EER of 16.3% for 8-conversation training.

3. Pitch and Energy Track Dynamics [11]: The aim was to learn pitch and energy *gestures* by modeling the joint slope dynamics of pitch and energy contours. A sequence of symbols describing the pitch and energy slope states (rising, falling), segment duration, and phoneme or word context is used to train an n-gram classifier. Using only slope and duration produced an EER of 14.1% for 8-conversation training, which dropped to 9.2% when fused with the absolute pitch and energy distributions, indicating it is capturing new information about the pitch and energy features. Although not purely a prosodic system, adding phoneme context to duration and contour dynamics produces an EER of 5.2%. (Also, examining pitch dynamics by dynamic time warp matching of word-dependent pitch tracks using 15 words or short phrases produced an EER of 13.3%.)

4. Prosodic Statistics [10]: Using the various measurements from the SRI prosody database, 19 statistics from duration and pitch related features, such as mean and variance of pause durations and F0 values per word, were extracted from each conversation side. Using these feature vectors in a k-nearest neighbor classifier on 8-conversation training produced an EER of 15.2% for the 11 duration-related statistics, 14.8% for the 8 pitch-related statistics and 8.1% for all 19 features combined.

3.3 Phone Features

5. Phone N-grams [14]: In this approach, the time sequence of phones coming from a bank of open-loop phone recognizers is used to capture some information about speaker-dependent pronunciations. Multiple phone streams are scored independently and fused at the score level. Using the 5 PPRLM phone streams and the *bag-of-n-grams* classifier, an EER of 4.8% was obtained for 8-conversation training.

6. Phone Binary Trees [8]: This approach also aims to model the time sequence of phone tokens, but a binary tree model is used instead of an n-gram model. With a binary tree, it is possible to use large context without exponential memory expansion and the structure lends itself to some adaptation and recursive smoothing techniques important for sparse data sets. Using a 3-token history (equivalent to 4-grams) and adaptation from a speaker-independent tree, an EER of 3.3% is obtained for 8-conversation training. The main improvement with this approach is robustness for limited training conditions. For example, it obtains an EER of 11% for 1-conversation training compared to 33% for the n-gram classifier.

7. Cross-stream Phone Modeling [7]: While the above phone approaches attempt to model phone sequences in the temporal dimension, this approach examines capturing cross-stream information from the multiple phone streams. The phone streams are first aligned and then co-occurrence of the different language phones are modeled via n-grams. This produces an EER of 4.0% for 8-conversation training. Cross-stream and temporal systems can be fused together to produce an EER of 3.6%. In general, this technique can be expanded using graphical models to simultaneously capture both cross-stream and temporal sequence information.

8. Pronunciation Modeling [9]: The aim here is to learn speaker-dependent pronunciations by comparing constrained word-level automatic speech recognition (ASR) phoneme streams with open-loop phone streams. The phonemes from the SRI ASR word transcripts are aligned on a per-frame level with the PPRLM open-loop phones. Conditional probabilities for each open-loop phone, given an ASR phoneme, are computed per speaker and for a background model. This technique produces an amazing 2.3% EER for 8-conversation training.

3.4 Lexical Features

9. Word N-grams [15]: Although not an active focus in the project, an n-gram idiolect system was implemented and used to examine the effects of using errorful word transcripts at various word error rates (WER). The automatic transcripts were selected to provide a range of WERs and do not reflect fundamental differences in the suppliers' technologies. The 8-conversation training EERs using the different transcripts are as follows: Manual 9%, Dragon 11% (20% WER), SRI 12% (30% WER), and BBN 16% (50% WER). The approach is relatively robust, even at 50% WER using BBN's real-time system.

4. FUSION

Given the pallet of new features and approaches outlined above, we next set out to examine fusion of the different levels of information to see if they are indeed providing complementary information to improve speaker recognition accuracy.

4.1 Fusion System Design

Many classifiers for fusing speaker recognition systems were explored at Lincoln Laboratory prior to the JHU Workshop. These systems included GMM-UBM, text-constrained GMM-UBM, word n-gram, phone n-gram, part-of-speech n-gram, and conversational-pattern n-gram methods spanning EERs from 1% to 30% [16]. The LNKnet pattern classification software [17] was used to design and compare multilayer perceptron, radial basis function, Gaussian, Gaussian mixture, k-nearest neighbor, binary tree, and support vector machine classifiers. The Lincoln fusion approach was adopted and adapted for use at the Workshop because of the related and similarly accurate Lincoln and Workshop systems to be fused.

Various fusion systems were developed in two stages using the NIST SRE'01 Switchboard-I corpus. NIST partitioned the data into 6 splits so that no speaker occurs in more than one split to allow crossvalidation jackknifing. In stage 1, splits 1, 2, and 3 are combined for training and splits 4, 5, and 6 are combined for testing. This resulted in low statistical significance because of too few errors; additional testing was needed. In stage 2, 6-fold crossvalidation using the SRE'01 splits with a leave-one-out technique was used. The perceptron and the GMM were the two best classifiers found with respect to minimizing the total number of errors over the training set. The perceptron has no hidden layer and is similar to a linear discriminant with a sigmoidal output squashing function.

Finally, the fuser can be fine tuned by adjusting its configuration, normalization, and prior probabilities settings. For example, separate classifiers can be created and used for each training condition and split. Also, LNKnet can minimize the errors for a given operating point, e.g., EER, by adjusting the priors [16].

4.2 Decision Region Comparison

Fusions using a perceptron or a GMM were found to have lower error rates than the component systems and other fusions (multilayer perceptron, radial basis function, Gaussian, k-nearest neighbor, binary tree, and support vector machine). Figure 1 shows a comparison of the decision regions created by the GMM and perceptron fusion classifiers. Four components were fused in this comparison: baseline GMM-UBM, text-constrained GMM-UBM, word n-gram, and the phone n-gram systems. Two-dimensional slices are shown for 2604 test patterns using 8 training conversations per speaker model in each of the systems being fused. The error rates are extremely low (~0.6%), with the perceptron committing 15 errors and the GMM 18 errors. This insignificant difference did not cause us to favor one classifier over the other. We chose the perceptron fuser because its simpler decision regions in sparse data areas, as shown in Figure 1, appear to be more reasonable and, thus, it is likely to be more robust.

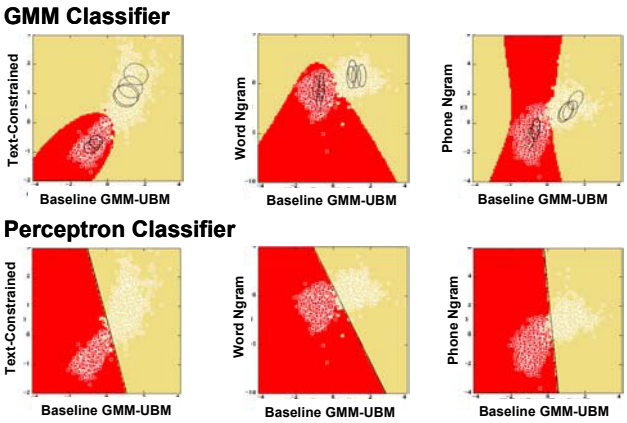

Figure 1. Decision region comparison of GMM vs. Perceptron.

4.3 Perceptron Fusions

For the Workshop, we chose a single-layer perceptron to fuse the scores from the various systems. This perceptron has a layer consisting of inputs for each system (and a bias term), no hidden layer, and an output layer using sigmoids on the target and nontarget nodes. A separate perceptron was trained for each of the six splits using the five held-out splits. The priors were adjusted in LNKnet to minimize detection cost [6, 16].

We fused the nine best performing individual systems covering acoustic, prosodic, phone, and lexical approaches. The systems and their EERs are given in Table 1. Each system used 8 conversations to train each speaker's model. These models were scored against messages according to the NIST control file to determine EERs and provide inputs to the fusion system. The GMM cepstra (1) and Pronunciation modeling (8) systems gave the best and next best individual EERs, respectively.

Table 1. The nine component systems to be fused. EERs are from the 8-conversation training condition.

System	EER
1. Acoustic baseline (GMM-UBM cepstral features)	0.7
2. Pitch and energy distributions	16.3
3. Pitch and energy slopes + durations + phoneme context	5.2
4. Prosodic statistics	8.1
5. Phone n-grams (5 PPRLM phone sets)	4.8
6. Phone binary trees (5 PPRLM phone sets)	3.3
7. Phone cross-stream + temporal (5 PPRLM phone sets)	3.6
8. Pronunciation modeling (SRI prons + 5 PPRLM phone sets)	2.3
9. Word n-grams/idiolect (Dragon transcripts)	11.0

In Figure 2, we show a DET plot with three curves from the fusion experiment. The top two, with EER=0.7%, are for the GMM cepstra system, alone, and from fusing all but the GMM cepstra system (Fuse 8). The fusion of all 9 systems produces the bottom curve with EER=0.2%—a 71% relative reduction. In this 8-conversation training condition, 272 speaker models were trained and scored against 3,813 target and 6,564 impostor messages. Binomial and bootstrap tests show at least 95% confidence that the fusion system exceeds the baseline, clearly showing that the new features and classifiers are supplying complementary information to the baseline acoustic system.

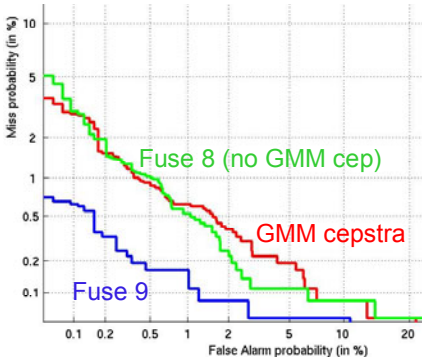

Figure 2. DET plot showing three curves. Using only GMM-cepstra (EER=0.7%), fusing 8 systems without GMM-cepstra (EER=0.7%), and fusing all 9 systems (EER=0.2%).

We also conducted experiments examining fusing subsets of the systems. Table 2 shows the best subset of *m* systems to fuse from among all 9 systems. Fusions of *m*=1 to 5 systems are shown, beyond which accuracy plateaus at 0.22% EER. For all *m,* the GMM-cepstra acoustic baseline system (1) is selected and, thus, is most powerful with respect to this fusion. The next most powerful feature is seen to be pitch gesture with phoneme context (3)—when fused with system 1, it yields the lowest 2-way fusion EER of 0.3%. It is intuitively appealing to see that a system covering both prosodic and phonemic information is the best one to fuse with conventional acoustics. Selecting system 3 tends to knock out phone-based systems that are individually more accurate, but lacking prosodic information. One of the two best 3-way fusions shown in the table adds idiolect to the best 2-way fusion system—another intuitively appealing result.

Table 2. Best *m*-way system perceptron fusions for 8 training conversations (☑ denotes a tie – choose either system).

Systems	*m*=1	*m*=2	*m*=3	*m*=4	*m*=5
1. Acoustic baseline	✓	✓	✓	✓	✓
2. Pitch and energy distributions			☑		
3. Pitch and energy slopes + durations + phoneme context		✓	✓	✓	✓
4. Prosodic statistics				✓	✓
5. Phone n-grams					
6. Phone binary trees					✓
7. Phone x-stream + temporal					
8. Pronunciation modeling				☑	
9. Word n-grams/idiolect			☑	☑	✓
EER (%)	0.7	0.3	0.27	0.24	0.22

We investigated fusion subsets excluding the baseline acoustic system to focus on high-level features. The best two non-GMM-cepstra systems to fuse, with an EER of 1.2%, are the pronunciation (8) and pitch-energy slopes (3) [12]. The best three non-GMM-cepstra system combinations gave an EER of 0.9%. There were three combinations that produced this EER: Systems (8, 4, 3), (8, 4, 9) and (8, 3, 9). In each case, the pronunciation system (8) is chosen with addition of two out of these three systems: pitch-energy slope (3), prosodic statistics (4), and word n-gram (9) [12]. The sampling of different levels of information in these combinations is again intuitively appealing.

5. CONCLUSIONS, FUTURE, AND ACKNOWLEDGEMENTS

We have shown here and in companion papers that the SuperSID project succeeded in exploiting high-level information to improve speaker recognition accuracy. Accuracy is improved by fusing complementary features and the information they cover is intuitively appealing. Even at extremely low error rates, there is still significant benefit in combining complementary types of information. The accuracy of the baseline acoustic GMM-cepstra system was matched by fusing 8 high-level SuperSID systems together. Fusing all 9 systems together yielded a new record for accuracy on this task with astonishingly low error rates.

Our error analysis is expanding to understand which errors remain and what types of features, models, and fusers can correct them. We are investigating advanced methods of feature selection and combination, including incorporating confidence measures to know when different types of features and systems are reliable. The generality of these techniques is being examined on other corpora to determine the relative robustness of the new features, models, and fusers to factors such as noise, channel variability, speaking partners, topics, language, and reduced training.

The authors greatly appreciate Richard Lippmann's guidance on classifier design and Pedro Torres-Carrasquillo's LNKnet runs. The SuperSID team created the systems to be fused and we are grateful to: Walter Andrews, Jiří Navrátil, Barbara Peskin, Andre Adami, Qin Jin, David Klusáček, Joy Abramson, Radu Mihaescu, John Godfrey, Douglas Jones, and Bing Xiang.

6. REFERENCES

[1] D. Reynolds, W. Andrews, J. Campbell, J. Navrátil, B. Peskin, A. Adami, Q. Jin, D. Klusáček, J. Abramson, R. Mihaescu, J. Godfrey, D. Jones, B. Xiang, "The SuperSID Project: Exploiting High-level Information for High-accuracy Speaker Recognition," *ICASSP* 2003.
[2] G. Doddington, "Some Experiments on Idiolectal Differences among Speakers," Available: http://www.nist.gov/speech/tests/spk/2001/doc/, 14 November 2000.
[3] W. Andrews, T. Kohler, J. Campbell, "R523: Extended Data Task," NIST Speaker Recognition Evaluation Workshop, Linthicum, Maryland, 15 May 2001.
[4] D. Sturim, B. Dunn, D. Reynolds, "MIT Lincoln Laboratory Site Presentation: Extended Data Task," NIST Speaker Recognition Evaluation Workshop, Linthicum, Maryland, 15 May 2001.
[5] F. Weber, L. Manganaro, B. Peskin, "Speaker ID on the Extended Data Set," NIST Speaker Recognition Evaluation Workshop, Linthicum, MD, 15 May 2001.
[6] "NIST Speaker Recognition – Eval 2001," Available: http://www.nist.gov/speech/tests/spk/2001/.
[7] Q. Jin, J. Navrátil, D. Reynolds, J. Campbell, W. Andrews, J. Abramson, "Combining Cross-Stream and Time Dimensions in Phonetic Speaker Recognition," *ICASSP* 2003.
[8] J. Navrátil, Q. Jin, W. Andrews, J. Campbell, "Phonetic Speaker Recognition Using Maximum Likelihood Binary Decision Tree Models," *ICASSP* 2003.
[9] D. Klusáček, J. Navrátil, D. Reynolds, J. Campbell, "Conditional Pronunciation Modeling in Speaker Detection," *ICASSP* 2003.
[10] B. Peskin, J. Navrátil, J. Abramson, D. Jones, D. Klusáček, D. Reynolds, B. Xiang, "Using Prosodic and Conversational Features for High-Performance Speaker Recognition: Report from JHU WS'02," *ICASSP* 2003.
[11] A. Adami, R. Mihaescu, D. Reynolds, J. Godfrey, "Modeling Prosodic Dynamics for Speaker Recognition," *ICASSP* 2003.
[12] "Workshop 2002 – SuperSID: Exploiting High-level Information for High-performance Speaker Recognition," Available: http://www.clsp.jhu.edu/ws2002/groups/supersid/.
[13] D. Reynolds, T. Quatieri, R. Dunn, "Speaker Verification Using Adapted Gaussian Mixture Models," *Digital Signal Processing,* 10 (1-3), p. 19-41, 2000.
[14] W. Andrews, M. Kohler, J. Campbell, J. Godfrey, J. Hernández-Cordero, "Gender-Dependent Phonetic Refraction for Speaker Recognition," *ICASSP,* vol. I, p. 149-152, 2002.
[15] G. Doddington, "Speaker Recognition based on Idiolectal Differences between Speakers," *Eurospeech,* vol. 4, p. 2521-2524, 2001.
[16] J. Campbell, D. Reynolds, D. Sturim, D. Jones, B. Dunn, "MIT Lincoln Laboratory Site Presentation: Extended Data," NIST Speaker Recognition Evaluation Workshop, Vienna, Virginia, 22 May 2002.
[17] R. Lippmann, et al., LNKnet. Available: http://www.ll.mit.edu/IST/lnknet/.