The SuperSID Project: Exploiting High-level Information for High-accuracy Speaker Recognition[•+]

Douglas Reynolds[1], Walter Andrews[2], Joseph Campbell[1], Jiri Navratil[3], Barbara Peskin[4], Andre Adami[5], Qin Jin[6], David Klusacek[7], Joy Abramson[8], Radu Mihaescu[9], Jack Godfrey[2], Doug Jones[1], Bing Xiang[10]

(1) MIT LL (2) DoD (3) IBM (4) ICSI (5) OGI (6) CMU (7) Charles Univ. (8) York Univ. (9) Princeton Univ. (10) Cornell Univ.

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 indeed produced very low error rates, they ignore other levels of information beyond low-level acoustics that convey speaker information. Recently published work has shown examples that such high-level information can be used successfully in automatic speaker recognition systems and has the potential to improve accuracy and add robustness. For the 2002 JHU CLSP summer workshop, the SuperSID project (http://www.clsp.jhu.edu/ws2002/groups/supersid/) was undertaken to exploit these high-level information sources and dramatically increase speaker recognition accuracy on a defined NIST evaluation corpus and task. This paper provides an overview of the structure, data, task, tools, and accomplishments of this project. Wide ranging approaches using pronunciation models, prosodic dynamics, pitch and duration features, phone streams, and conversational interactions were explored and developed. In this paper we show how these novel features and classifiers indeed 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

What is it in the speech signal that conveys speaker identity? This is one of the central questions addressed by automatic speaker recognition research. From self-observation and experience, it is pretty clear that we (humans) rely on several different types or levels of information in the speech signal to recognize others from voice alone. These can be the deep bass and timber of a voice, a friend's unique laugh, or the particular repeated word usage of a colleague. Roughly we can categorize these 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 (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. With the continual advancement of tools, such as phone and speech recognition systems, to reliably extract features for high-level characterization, the increase in applications (like audio mining) allowing for relatively large amounts of speech from a speaker to learn speaking habits, the availability of large development corpora and plentiful computational resources, the time is right for a deeper exploration into using these underutilized high-level information sources. These new sources of information hold the promise not only for improvement in basic recognition accuracy by adding complementary knowledge, but also the possibility for robustness to acoustic degradations from channel and noise effects, to which low-level features are highly susceptible. Furthermore, previous work examining certain high-level information sources has provided strong indications that potential gains are possible (for example see recent papers [1,2,3,4]).

Inspired by these factors, the **SuperSID** project for the exploitation of high-level information for high-performance speaker recognition was undertaken as part of the 2002 JHU Summer Workshop on Human Language Technology [5]. The JHU WS2002 is one in a series of 6-week workshops hosted by the CLSP group at JHU with the aim of bringing together researchers to focus on challenging projects in the areas of speech and language engineering. The authors of this paper constituted the team members for the SuperSID project representing a diverse group of senior researchers from academia, commercial, independent and Government research centers, as well as graduate and undergraduate students. The aim of the SuperSID project was to analyze, characterize, extract, and apply high-level information to the speaker recognition task. The goals were to develop new features and classifiers exploiting high-level information, show performance improvements relative to baselines on an established evaluation data and task, and demonstrate that new features and classifiers provide complementary information.

This paper provides an overview of the framework and overall accomplishments of the SuperSID project. Details of the various approaches undertaken in the project can be found in the companion papers related to the SuperSID project [6,7,8,9,10] as well as on the SuperSID website [11].

2. TASK, DATA AND TOOLS

The focus for the SuperSID project was on text-independent speaker detection using the extended data task from the 2001 NIST Speaker Recognition Evaluation [12]. This task was introduced to allow exploration and development of techniques that can exploit significantly more training data than is

[•] This work is sponsored by the Department of Defense under Air Force Contract F19628-00-C-0002.and the CLSP/JHU workshop was supported by NSF and DoD fudning. 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.

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. To supply a large number of target and non-target trials and speaker models trained with up to 16 conversations of training speech (~40 minutes), the evaluation used a cross-validation 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 ~4100 target models (a speaker had multiple models for different amounts of training data) and ~57,000 trials for the testing phase, containing matched and mismatched handset trials and some cross-sex trials. The cross-validation 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[i].

The 2001 extended data task 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.

- *SRI prosody database [13]*: The SRI database provides frame-level pitch and energy tracks (in raw and stylized forms) as well as a wealth of word-level prosodic features derived both for "truth" transcripts and for speech recognizer output, time-aligned to the speech stream at the phone level. Features include pause and segmental durations, voicing and stress information, pitch statistics, and much more.

- *Four word transcriptions of varying word error rates (WER)*: Manual transcripts from ISIP, automatic transcripts from Dragon Systems (~20% WER), automatic transcripts from SRI's Decipher (~30% WER), and automatic transcripts from BBN's real-time Byblos (~50% WER)[ii].

- *Two sets of open-loop (i.e., no language models in decoder) phone transcripts in various languages:* From MIT's PPRLM system, we had phone transcripts in English, German, Japanese, Mandarin, and Spanish. From CMU's GlobalPhone system, we had phone transcripts in Chinese, Arabic, French, Japanese, Korean, Russian, German, Croatian, Portuguese, Spanish, Swedish, and Turkish.

- *Articulatory feature transcripts [14]*: (pseudo-)articulatory classes automatically extracted from the speech signal and designed to capture characteristics of speech production such as consonantal place of articulation, manner of articulation, voicing, etc.

We also assembled a suite of models to apply to features we extracted from the above data sets. These included standard n-gram tools found in the CMU-CU language modeling toolkit[iii] as well as a "bag-of-n-grams" classifier as described in [2], a discrete token binary tree classifier [7], a discrete HMM classifier[iv], a continuous GMM classifier[v], and a MLP fusion tool[vi].

These models were used to form likelihood ratio detectors 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 were used as cohort models. During recognition, a test utterance is scored against the speaker and background model(s) and the ratio (or in the log domain, difference) is reported as the detection score for sorting.

3. APPROACHES

In this section we survey some of the highlights of approaches developed to exploit high-level speaker information. The reader should consult the referenced papers for more details.

3.1 Acoustic Features

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 [15] 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

- Pitch and Energy Distributions [10]: 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.

- Pitch and Energy Track Dynamics [10]: 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 phone 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 phone context to duration and contour dynamics produces an EER of 5.2%. Examining pitch dynamics by

[i] Due to the limited number of speakers/models, the results for the 16-conversation training condition were found to have high statistical variation so we will generally cite results only up to the 8-conversation training condition.

[ii] These automatic transcripts were selected to provide a range of WERs and do not reflect fundamental differences in the supplier's technology.

[iii] http://svr-www.eng.cam.ac.uk/~prc14/toolkit.html

[iv] http://www.cfar.umd.edu/~kanungo/software/software.html

[v] From MITLL's GMM-UBM speaker recognition system

[vi] http://www.ll.mit.edu/IST/lnknet/

dynamic time warp matching of word-dependent pitch tracks using 15 words or short phrases produced an EER of 13.3%.

- Prosodic Statistics [9]: 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

- Phone N-grams [4]: 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.

- Phone Binary Trees [7]: This approach also aims to model the time sequence of phone tokens, but instead of an n-gram model a binary tree model is used. 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.

- Cross-stream Phone Modeling [6]: 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.

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

3.4 Lexical Features

Although not an active focus in the project, an n-gram idiolect system like that described in [2] was implemented and used to examine the effects of using errorful word transcripts. The 8-conversation training EERs for the different transcripts are as follows: Manual 9%, Dragon 11%, SRI 12%, BBN 16%. So the approach appears to be relatively robust even as WER increases to 50%.

3.5 Conversational Features

In this approach, we examined whether there was speaker information in turn-taking patterns and conversational style. The motivation of this work is from results in the 2002 NIST evaluation where n-grams of speaker turn durations and word density were able to produce an EER of 26% for 8-conversation training. A system was developed using feature vectors containing turn-based information about pitch, duration and rates derived from the SRI prosody database. These feature vectors were converted into a sequence of turn-based tokens from which n-gram models were created to capture turn characteristics [9]. On split 1 for 8-conversation training the best system EER was 15.2%. We also examined conditional word usage in speaker turns with the idea that a speaker may adapt his/her word usages based on his/her conversational partner, but found this produced > 26% EER.

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 performance. For the workshop we used a simple single layer perceptron with sigmoid outputs for fusing system scores. A fuser was trained for each split using the five held out splits. There are no doubt better fusion approaches for combining information sources, but the aim here was merely a proof of concept. For the fusion experiment we selected the 9 best performing individual systems covering acoustic, prosodic, phonetic and lexical approaches. The EERs for the individual systems are shown in Table 1. After the GMM cepstra system the best performing system is the one based on pronunciation modeling.

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 + phone context	5.2
4. Prosodic statistics	8.1
5. Phones 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 (Dragon transcripts)	11.0

In Figure 1 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. Based on the number of trials, this is a statistically significant improvement. These results clearly show that the new features and classifiers are supplying complementary information to the baseline acoustic system.

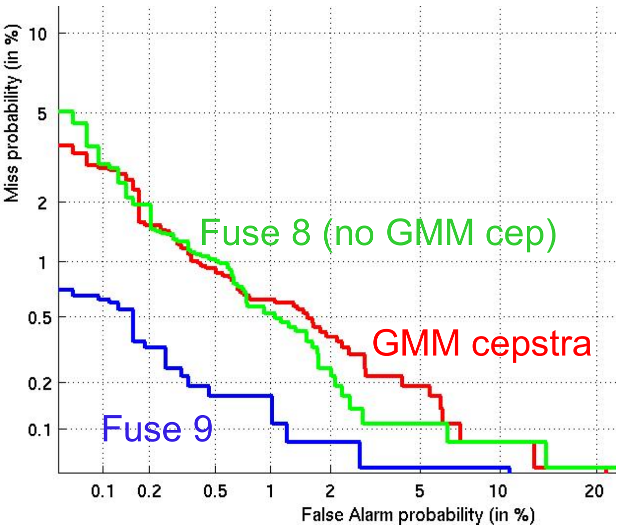

Figure 1 DET plot showing three curves. Using only GMM-cepstra (EER=0.7%), fusing 8 systems without GMM-cepstra (EER=0.7%), and fusing all 9 systems (EER=0.2%).

We also conducted experiments examining fusing subsets of the systems. The single best system to fuse with the GMM cepstral system (system 1 in table) is the pitch/energy slope system (system 3), yielding an EER of 0.3%. It is intuitively appealing to see that a system that covers both prosodic and phone information was the best one to fuse with the standard acoustics. The best two non-GMM-cepstral systems to fuse, with an EER of 1.2%, were the pronunciation (system 8) and pitch/energy slopes (system 3). The best three non-GMM-cepstral 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 included with addition of the pitch/energy slope (3), the prosodic statistics (4), and/or the word n-gram (9) systems. The sampling of different levels of information in these combinations is also intuitively appealing and again confirms that the systems are indeed providing complementary information.

5. CONCLUSIONS AND FUTURE DIRECTIONS

From the results presented in this paper and in the companion papers, it is clear that the SuperSID project achieved the aim of exploiting high-level information to improve speaker recognition performance. Even at extremely low error rates, it was shown that there is still significant benefit in combining complementary types of information. However, this is just the beginning of truly exploiting these sources of speaker information, with many open avenues to explore. First, the results need to be validated on a different corpus to show they indeed generalize. Current work is underway to implement these approaches on the Switchboard-II corpus, which has a higher acoustic error rate. Second, we need to expand our error analysis to understand which errors are left and what features can address them. Third, we need to examine better ways of feature selection and combinations perhaps incorporating confidence measures to know when different types of features/systems are reliable. Finally, we need to examine the relative robustness of the knowledge sources to factors like noise, channel variability, speaking partners, topics and language.

6. REFERENCES

[1] K. Sonmez, E. Shriberg, L. Heck, and M. Weintraub, “Modeling Dynamic Prosodic Variation for Speaker Verification”, ICSLP, Vol. 7, pp. 3189-3192, 1998

[2] G. Doddington, “Speaker Recognition based on Idiolectal Differences between Speakers,” Eurospeech, Vol. 4, pp. 2517-2520, 2001

[3] F. Weber, L. Manganaro, B. Peskin, and E. Shriberg, “Using Prosodic and Lexical Information for Speaker Identification,” ICASSP 2002

[4] W. Andrews, M. Kohler, J. Campbell, J. Godfrey, and J. Hernandez-Cordero, “Gender-dependent Phonetic Refraction For Speaker Recognition,” ICASSP 2002

[5] JHU WS2002 website http://www.clsp.jhu.edu/ws2002/

[6] Q. Jin, J. Navratil, D. Reynolds, J. Campbell, W. Andrews, J. Abramson, “Combining Cross-Stream And Time Dimensions In Phonetic Speaker Recognition,” ICASSP 2003

[7] J. Navratil, Q. Jin, W. Andrews, and J. Campbell, “Phonetic Speaker Recognition Using Maximum Likelihood Binary Decision Tree Models,” ICASSP 2003

[8] D. Klusacek, J. Navratil, D. Reynolds, and J. Campbell “Conditional Pronunciation Modeling In Speaker Detection,” ICASSP 2003

[9] B. Peskin, J. Navratil, J. Abramson, D. Jones, D. Klusacek, D. Reynolds, B. Xiang, “Using Prosodic and Conversational Features for High-performance Speaker Recognition: Report from JHU WS'02,” ICASSP 2003

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

[11] SuperSID Project website http://www.clsp.jhu.edu/ws2002/groups/supersid/

[12] NIST Speaker Recognition website http://www.nist.gov/speech/tests/spk/2001/

[13] E. Shriberg, A. Stolcke, D. Hakkani-Tur, G. Tur, "Prosody-Based Automatic Segmentation of Speech into Sentences and Topics", Speech Communication, Vol. 32, No. 1-2, pp. 127-154, 2000

[14] K. Kirchhoff, “Robust Speech Recognition Using Articulatory Information,” PhD thesis, University of Bielefeld, Germany, July 1999

[15] D. Reynolds, T. Quatieri and R. Dunn, “Speaker Verification Using Adapted Mixture Models”, Digital Signal Processing, Vol. 10, pp. 181-202, 2000