ICSI System Description for SRE2008 Submission

Howard Lei, David Van Leeuwen

1. Introduction

The three ICSI systems involved in the evaluations are the keyword HMM supervector system [1], the GMM supervector system, and the keyword phone lattice N-grams system [2], which we enhanced by including prosodic N-grams. Descriptions of the keyword HMM supervector and keyword phone lattice N-grams + prosodic N-grams systems will be discussed in sections 3 and 4. A description of the GMM supervector system can be found in TNO's system descriptions.

2. Data and preprocessing

We applied the ICSI wiener filter to all speech data prior to any processing. We used a total of 4,378 Fisher, Switchboard II, and SRE04 conversation sides for background model training, 2,726 SRE05 microphone conversation sides for Nuisance Attribute Projection (NAP) [3] training based on microphone data, 2,742 Switchboard II conversation sides for NAP training based on telephone data, and 11,045 SRE06 conversation sides for development. There are roughly 210,000 trials used for SRE06, of which roughly 11,000 are true speaker trials.

In order to implement our keyword-based systems, we are provided with ASR decodings for all conversation sides by SRI, obtained via the DECIPHER recognizer [4]. We used MFCC features (C0-C19 plus deltas) with cepstral mean subtraction, obtained via HTK [5]. Note that we've attempted feature warping, but discovered no significant improvements after the application of NAP. Hence, feature warping is omitted for the SRE08 results.

3. System 1: HMM Supervector System

For each conversation side, this system trains left-to-right keyword HMM models with 8 gaussian mixtures per state [6] on the MFCC features for 16 different word unigrams and bigram – *but, have, just, like, not, really, right, so, that, think, uh, uhhuh, um, was, yeah, you_know* – using ASR from the SRI's DECIPHER recognizer. A background keyword HMM is first trained for each keyword using 1,553 conversation sides from Fisher and Switchboard II. Then, for each conversation side, keyword HMM models are trained via MAP adaptations from the set of corresponding background keyword HMMs. Note that only the Gaussian mixture means are altered. If a keyword does not exist in a conversation side, its keyword HMM is replaced by the corresponding background keyword HMM.

The Gaussian mixture means for each state for all keyword HMM models of a conversation side are concatenated into an HMM supervector for that conversation side. Rank-normalization is performed on these supervectors, followed by NAP. For the short2-short3.ndx and long-long.ndx conditions, NAP is trained using the 2,742 Switchboard II telephone conversation sides for telephone-telephone trials; NAP is trained using the 2,726 SRE05 microphone conversation sides if the trials are not telephone-telephone trials. For the 8conv-short3.ndx condition, NAP is trained using only the 2,742 telephone conversation sides. The supervectors are classified via the linear kernel SVM (implemented using SVMlight [7]) to obtain scores.

The CPU execution times are approximately 41 hours for creating short2+long+8conv models, 25 hours for creating long models, 35 hours for creating 8conv models, 25 hours for processing segments in short2-short3.ndx + 8conv-short3.ndx, and 20 hours for processing segments in long-long.ndx.

4. System 2: Keyword Phone N-grams system with Prosodic N-grams features

This system extracts keyword-constrained phone N-gram counts from phone lattice decodings for each conversation side, which are obtained using SRI's DECIPHER recognizer. The phone N-gram counts are concatenated into feature vectors for each conversation side, which are classified via the SVM (implemented using SVMlight [7]) to obtain scores. Here, a set of 52 keywords are used – *a, about, all, and, are, be, because, but, do, for, get, have, i, if, in, is, it, just, know, like, mean, my, no, not, of, oh, okay, on, one, or, people, really, right, so, that, the, there, they, think, this, to, uh, uhhuh, um, was, we, well, what, with, would, yeah, you.* If a keyword does not exist in a conversation side, its phone N-gram counts for that particular keyword will be assigned to 0.

In addition, pitch prosodic feature sequences (f_0 mean), where the feature frames are 40 ms in length and non-overlapping, are extracted for each conversation side. Each prosodic feature is classified into one of 8 bins, and uni-,bi-,and tri-grams are formed from prosodic feature sequences with respect to their bin labels for each keyword. Note that the boundaries for the 8 bins are trained using the prosodic feature distribution from the 1,553 Fisher and Switchboard II conversation sides, and the same set of 52 keywords is used. Hence, for each conversation side, a set of prosodic N-grams are obtained for each keyword, and these N-grams are concatenated with the phone N-gram counts for that conversation side to form the final feature vectors. The SVM classifier with a linear kernel is used to classify the feature vectors.

The CPU execution times are approximately 38 hours for creating short2+long+8conv models, 80 hours for processing segments in short2-short3.ndx + 8conv-short3.ndx, and 60 hours for processing segments in long-long.ndx.

5. Systems 3: GMM Supervector system

A description of this system can be found in TNO's system descriptions.

6. System 4: SRI GMM-UBM system

We additionly fused our systems with a GMM-UBM system from SRI. A description of this system can be found in SRI's system descriptions.

7. System combination

We used Niko Brummer's Focal combiner [8] to fuse the various systems. The bilinear fusion technique is applied, where the side-information consists of whether a trial is English or non-English, male or female, telephone-telephone, telephone-microphone, or microphone-microphone. Overall, there are 12 potential classes of side information (2 for English versus non-English, 2 for gender, and 3 for channel type). For some submissions, only 4 classes are used, where all non-English trials are grouped together into one class, and male and female trials are also grouped together. Subsets of the SRE06 trials are created for training and testing the combinations. Note that the keyword HMM supervector and keyword phone N-grams + prosodic N-grams systems are only run for the English trials, while the remaining systems are run for all trials.

8. SRE08 submissions

We submitted results for the short2-short3.ndx, 8conv-short3.ndx, and long-long.ndx conditions. Because the long-long.ndx condition involves longer training and testing conversation sides for which more keyword instances would appear, we believe that the long-long.ndx condition would benefit our systems. The advantage of the long-long.ndx condition over the 8conv-short3.ndx condition is that the test conversation sides of the long-long.ndx condition are also extended to potentially give more keyword instances. Note that the 8conv-short3.ndx condition uses un-wiener filtered data and NAP trained using the telephone conversation sides only.

Denote the keyword HMM supervector system as S1, the keyword phone N-grams + prosodic N-grams system as S2, the GMM supervector system as S3, and SRI's GMM-UBM system as S4. Table 1 lists the systems used for each submission (3 per condition), along with the number of side-info classes. Note that the system S1u refers to keyword HMM supervector system using un-wiener filtered data and telephone NAP training, and one or more side-info classes may contain no trials with respect to the conditions. Scores for all conditions except for the long-long.ndx condition may be interpreted as log-likelihood ratios.

Condition	Submitted systems	Side-info classes
short2-short3.ndx	S1+S3	12
short2-short3.ndx	S1+S3	4
short2-short3.ndx	S1+S3+S4	12
8conv-short3.ndx	S1+S2	12
8conv-short3.ndx	S1+S2+S4	12
8conv-short3.ndx	S1+S4	4
long-long.ndx	S1	12
long-long.ndx	S1u	12
long-long.ndx	S2	12

Table 1: NIST SRE08 submission conditions

9. References

[1] Lei, H., Mirghafori, N., "Word-Conditioned HMM Supervectors for Speaker Recognition", in Proc. of Interspeech, 2007.

[2] Lei, H., Mirghafori, N., "Word-Conditioned Phone N-grams for Speaker Recognition," in Proc. of ICASSP, 2007.

[3] Solomonoff, A., Campbell, W. M., Boardman, I., "Advances in Channel Compensation for SVM Speaker Recognition," in Proc. of ICASSP, 2005.

[4] Kajarekar, S., Ferrer, L., Venkataraman, A., Sonmez, K., Shriberg, E., Stolcke, A., Gadde, R.R., "Speaker Recognition using Prosodic and Lexical features", in Proc. of the IEEE Automatic Speech Recognition and Understanding Workshop, pp. 19-24, 2003.

[5] HMM Toolkit (HTK): http://htk.eng.cam.ac.uk

[6] Boakye, K., "Speaker Recognition in the Text-Independent Domain Using Keyword Hidden Markov Models", Masters Report, University of California at Berkeley, 2005.

[7] Joachims, T., "Making Large Scale SVM Learning Practical", in Advances in kernel methods - support vector learning, B. Schoelkopf, C. Burges, and A. Smola, Eds. MIT-press, 1999.

[8] Brummer, N., Focal Bilinear Toolkit, http://niko.brummer.googlepages.com/focalbilinear.