2006 NIST Speaker Recognition Evaluation:
Persay System Description

Ran Gazit, Nir Krause

1. Overview

Persay is submitting scores for four tasks, using three different speaker verification systems:

- An SVM-based classifier, working in the GMM model space
- A baseline GMM-UBM classifier
- A syllable–constrained version of the SVM-based system

Various combinations of these systems are used for different tasks, as described in the following sections.

2. SGM system (SVM classification in the GMM model space)

2.1 Features:

- 20 LPC cepstrum coefficients (LPCC) + 20 delta LPCC, with mean subtraction and variance normalization, computed over 250 msec frames with 125 msec overlap.
- 19 Mel frequency cepstrum (MFCC) + 19 delta MFCC including RASTA filtering, mean subtraction and variance normalization, computed over 250 msec frames with 125 msec overlap.
- Energy detection over the “other” side of each 4-wire conversation was used to discard silent segments. Additional silent frames were removed by an energy-based voice activity detector with adaptive threshold.

2.2 Modeling:

For each gender, five background models (BM) were trained:

- Cellular, GSM coding
- Cellular, CDMA coding
- Landline, carbon-button handset
- Landline, electret handset
- Multi-channel

Cellular calls come from NIST 2003 evaluation data. Landline calls come from NIST 1999 evaluation data. The multi-channel model is based on NIST 2005 data. Each BM is a 64-Gaussians GMM, based on at least 250 training and test calls from the relevant data set. The BMs were trained by means-only Bayesian adaptation of a root model, created from same-gender background data in all channels together.

Each train file was cut into 20sec long segments. A GMM was trained for each segment, using means-only Bayesian adaptation of the best-matching same-gender BM.

An SVM was trained for each training file, using as features the difference vector between the Gaussian means of each 20sec segment and the Gaussian means of the matching BM. Positive examples were taken from all 20sec segments of the training file. Negative examples were taken from the first 20sec segment of all same-gender files used for training the background models. Each example was rank normalized against the sorted list of negative examples scores. The whole process was done separately over the two feature sets (LPCC and MFCC), resulting in two separate classifiers per training file.

Additional details on SVM classification in the GMM model space can be found in [1]. The SVM classifier was implemented in SVMTorch [2], using a first-degree polynomial kernel.

2.3 Testing:

Each test file was cut into 20-sec long segments. A GMM was trained for each segment, using means-only Bayesian adaptation of the best-matching same-gender BM. The difference vector between the Gaussian means of each 10sec segment and the Gaussian means of the matching BM were used as features for the SVM classifier. The classifier computed the mean margin between all 20-sec segments and the separator obtained for the training file.

Each test file was also tested in a similar way against 100 models of same-gender speakers, trained from NIST 2004 evaluation data. Each test score was then T-normed by the statistics obtained over the 2004 data. For each test case two scores were derived using the two feature sets. These scores were linearly combined.

When the test segment was summed (2-wire), an external segmentation utility was used to divide the test segment to two sides. The external segmentation utility [3] is using two feature sets (FFT and LPC-cepstra) and two self-organizing–maps (SOM) classifiers in an iterative fashion to cut the summed file into two files which presumably hold the voice of only one speaker. Each such file was tested against a model created from the train segment. The highest score was selected as the score of this train-test pair.

2.4 Execution times

Preprocessing:	~ 17 sec for each 1conv4w (5 min) segment, including feature extraction (both MFCC and LPCC), BM selection and GMM adaptation for all 20sec segments
SVM training:	~ 18 sec for each 1conv4w (5 min) segment (both classifiers)
SVM testing:	~ 0.2 sec for each 1conv4w (5 min) segment (both classifiers)
Summed file separation:	~ 30 sec for each 1conv2w (5 min) segment

Processing was done on an Intel P4 with 1GB memory, running Linux.

3. GMM-UBM system

3.1 Features:

- 20 LPCC + 20 delta LPCC, with cepstral mean subtraction and variance normalization, computed over 250 msec frames with 125 msec overlap.
- Energy detection over the “other” side of each 4-wire conversation was used to discard silent segments. Additional silent frames were removed by an energy-based voice activity detector with adaptive threshold.

3.2 Modeling:

For each gender, four background models (BM) were trained:

- Cellular, GSM coding
- Cellular, CDMA coding
- Landline, carbon-button handset
- Landline, electret handset

Cellular calls come from NIST 2003 evaluation data. Landline calls come from NIST 1999 evaluation data. Each BM is a 256-Gaussians GMM, based on 300 training and test calls from the relevant data set. Models for each training or test segment are created through means-only Bayesian adaptation of the best-matching same-gender BM.

3.3 Testing

Two log-likelihood-ratio scores were computed for each training-test pair:

1. Scoring the test segment against a model created from the training segment
2. Scoring the training segment against a model created from the test segment

Each score was T-normed against 100 models of same-gender speakers from NIST 2005. The two normalized scores were linearly combined, and the result was scaled by the statistics of same-gender scores over the 2005 evaluation data.

3.4 Execution times

Training:	~1.5 sec for each 10sec4w segment, including BM selection
Testing:	~0.5 sec for each 10sec4w segment, including scoring against one target model and one BM

Processing was done on an Intel PIII 1.4GHz with 256MB memory, running Linux.

4. Syllable-constrained SVM system

This system is a based on segmentation of the audio into N-grams of broad phonetic classes, and classifying the audio corresponding to each N-gram using the SGM algorithm described in section 2. This concept is roughly based on the main idea in [4].

Phonetic transcription for all audio files is generated by the phoneme recognizer developed at Brno University of Technology[1]. This recognizer was successfully applied to tasks including language identification [5] and keyword spotting [6]. Phonemes were mapped into four broad classes: vowels, stops, fricatives and a fourth class containing nasals and semivowels.

The audio corresponding to each broad class was extracted. This created four audio segments from each original segment, where each new segment corresponds to a different broad phonetic class. Tri-grams of these broad phonetic classes were also generated, and the corresponding audio to each trigram was extracted, creating 64 additional audio segments from each original segment.

The classification of each new segment was done independently using the SGM algorithm described in section 2, including only the LPCC feature set, without cutting the audio into 20sec segments and without Tnorm scaling of the scores. From these 68 independent classifiers, only those who had valid scores for most of the tests were chosen.

5. Submission details

Scores derived by the SGM system are submitted for all four tasks. In the 10sec-10sec task, the SGM scores were linearly combined with the scores of the baseline GMM-UBM system. In the 1conv-1conv task, the SGM scores were fused with the scores of the syllable-constrained SVM system. Score fusion in this case was obtained by a first-degree polynomial kernel SVM.

Table 1 describes the various system combinations used for each task:

System #	PRS-1 (primary)	PRS-2	PRS-3
Task:			
1conv-1conv	SGM		Syllable + SGM
1conv-1conv2w	SGM		
1conv-10sec	SGM		
10sec-10sec	SGM + GMM	GMM	

Table 1: system combinations

[1] http://www.fit.vutbr.cz/research/groups/speech/index.php?id=phnrec

6. References

[1] N. Krause, R. Gazit , “SVM-based Speaker Classification in the GMM Models Space”, to appear in Proc. Odyssey 2006

[2] R. Collobert, S. Bengio, J. Mariéthoz, “Torch: a modular machine learning software library”, Technical Report IDIAP-RR 02-46, IDIAP, 2002.

[3] Y. Metzger, “Blind Segmentation of a Multi-Speaker Conversation Using Two Different Sets of Features”, in Proc. Odyssey 2001

[4] B. Baker, R. Vogt, S. Sridharan, “Gaussian Mixture Modelling of Broad Phonetic and Syllabic Events for Text-Independent Speaker Verification”, in Proc. Eurospeech2005, Sep, 2005

[5] P. Matejka et. al., "Phonotactic Language Identification using High Quality Phoneme Recognition", in Proc. Eurospeech2005, Sep, 2005

[6] I. Szoke et. al., "Comparison of Keyword Spotting Approaches for Informal Continuous Speech", in Proc. Eurospeech2005, Sep, 2005