The IESK-Magdeburg Speaker Detection system for the NIST 2006 Speaker Recognition Evaluation

Marcel Katz

IESK, Cognitive Systems
University of Magdeburg, Germany
marcel.katz@e-technik.uni-magdeburg.de

Abstract. In this paper we describe the speaker detection system developed by the *Institute of Electronics, Signal Processing and Communications* (IESK) of the University of Magdeburg for the NIST 2006 Speaker Recognition Evaluation. The system is based on a *Gaussian Mixture Model* (GMM) system which is used to model the low-level acoustic features and a *Support Vector Machine* (SVM) subsystem which is used to classify the high-level prosodic and durational features.

1 Datasets

The core task of the NIST 2006 SRE is a speaker detection task that contains a two-channel conversation of each speaker of approximately five minutes duration in the training and also two-channel conversations in the evaluation test. The evaluation set of the core test consists of 354 male and 462 female speaker sentences for the model training. There are more than 53000 trials in the evaluation test containing different languages as well as different transmission channels. We only submit results of this core test.

For the NIST 2006 evaluation we used four different datasets for the model training, the development test and the evaluation test of the IESK system. As

Table 1. Used Datasets in the NIST 2006 evaluation. The number of sentences is given for males and females

Dataset	UBM	T-Norm	Development	Evaluation
NIST SRE 2000	100/100	-	-	-
SWITCHBOARD Cell.	50/50	-	-	-
NIST SRE 2004	75/75	50/50	121/245	-
NIST SRE 2006	-	-	-	354/462

can be seen in table 1 most of the speakers are derived from the NIST 2000 evaluation which is composed of land-line telephone conversations only. Together

with a small fraction of the Switchboard Cellular dataset (gsm and cdma) and a part of the NIST 2004 dataset this forms a well balanced dataset between different transmission channels and handset types.

2 Front End Processing

The first step in the feature extraction is the detection of speech activity and we used two different approaches. The first one is the *Voice Activity Detection* (VAD) tool of the NIST SPQA package and the second one is a *Voiced Speech Detection* approach, which is based on a pitch frequency extraction of the speech. In the submitted system we only used voiced speech, so that the amount of speech data drastically decreased to nearly 25% of the original speech amount.

The speech data was bandlimited to the frequncy range 300Hz-3400Hz. Using a 20ms window and a window shift of 10ms 12 dimensional mel-cepstral (MFCC) feature vectors were extracted as well as the frame energy. Additionally the first and the second time differences where calculated and appended to the MFCCs resulting into a 39 dimensional feature vector. Also *cepstral mean substraction* (CMS) was applied to the feature vectors. The feature extraction was done by the HTK-Toolkit [1].

3 The GMM system

The GMM system is based on a *Universal Background Model* (UBM) approach. Two gender dependent UBMs were trained on background data taken from the NIST 2000, the Switchboard Cellular and NIST 2004 datasets. Each UBM consists of 1024 mixture components and is trained via the *Expectation Maximization* (EM) algorithm. The specific speaker models are derived from the UBMs using a one step *Maximum A-Posteriori* (MAP) adaptation [2]. Only the means of the mixtures were adapted with a relevance factor $\tau = 16$.

During the detection test only the top N-best mixture components with respect to the UBM model were used for scoring. In our experiments we set $N = 25$.

3.1 Feature Mapping

Due to the fact that different handset types (e.g. carbon, electret) and channels (e.g. land-line, cellular) are used in the evaluation a normalization of the feature vectors is performed by *feature mapping* [3]. First the *root*-UBM is trained on data from several different channels and handsets. We then adapted four channel-dependent GMMs (carbon, electret, gsm, cdma) from the root-UBM by MAP adaptation. Each utterance is then classified by the channel dependent GMMs and each feature vector is mapped into the channel independent space via the top 1 Gaussian mixture. These channel independent feature vectors are then used for the MAP adaptation of the speaker models.

3.2 Score Normalization

To compensate different shift and scales of the log-likelihood scores during testing, the so called *test normalization* (T-Norm) is applied to the ratio scores [4]. The T-Norm models were trained on impostor speakers of the NIST 2004 dataset. For each gender we used 50 speakers.

4 The SVM subsystem

Additionally to the GMM baseline system we investigated a SVM subsystem. This system models the characteristics of the pitch frequency, the energy and the duration of speech segments. For every voiced speech segment the mean and standard deviation of the log pitch frequency are calculated. Together with the logarithm of the segment energy and the segment duration they form a four dimensional feature vector. All the segments of a speaker are than used to train a SVM against a small fraction of the background data. The SVM training was performed using the well known Torch library [5]. The traditionally non-probabilistic output of the SVM is transformed to a class probability by the algorithm of Platt [6]. For a better fusion with the the GMM system this moderated SVM output is also normalized with the T-Norm.

5 Fusing and Performance Measure

The output scores of the two subsystems were combined by the weighted sum approach to form the final decision. For all trials a unique gender independent threshold is used. This decision threshold is set on the optimal DCF point estimated on the NIST 2004 development data.

The scores presented in our submission are log-likelihood ratio (LLR) scores. All trials that contain non-speech in a test- or a training segment are marked as *false* with a LLR of 0.0.

6 Hardware Environment and CPU timing

The experiments were performed on a Linux cluster consisting of 24 dual Xeon 2.0 GHz processors and 4 GB RAM. The CPU time for training and testing is given in table 2.

Table 2. CPU time of modelling and segment testing

SYSTEM	CPU-Time Train (min)	CPU-Time Test (min)
GMM	1508	742
SVM	1088	1766

References

1. S. Young, G. Evermann, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, *The HTK Book*, Cambridge University Engineering Department, 2002.
2. D. Reynolds, T. Quatieri, and R. Dunn, "Speaker verification using adapted gaussian mixture models," *Digital Signal Processing*, vol. 10, pp. 19–41, 2000.
3. D. Reynolds, "Channel robust speaker verification via feature mapping," in *International Conference on Acoustics, Speech, and Signal Processing*, vol. 2, 2003, pp. 53–56.
4. R. Auckenthaler, M. Carey, and H. Lloyd-Thomas, "Score normalization for text-independent speaker verification systems," *Digital Signal Processing*, vol. 10, pp. 42–54, 2000.
5. R. Collobert and S. Bengio, "SVMTorch: Support vector machines for large-scale regression problems," *The Journal of Machine Learning Research*, vol. 1, pp. 143–160, 2001.
6. J. Platt, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," in *Advances in Large-Margin Classifiers*, P. Bartlett, B. Schölkopf, D. Schuurmans, and A. Smola, Eds. Cambridge, MA, USA: MIT Press, oct 2000, pp. 61–74. [Online]. Available: http://research.microsoft.com/ jplatt/abstracts/SVprob.html