Speaker Detection Using Phoneme Specific Hidden Markov Models

Edvin Pakoci, Nikša Jakovljević, Branislav Popović, Dragiša Mišković, and Darko Pekar

Faculty of Technical Sciences, University of Novi Sad, Serbia
{edvin.pakoci,darko.pekar}@alfanum.co.rs,
{jakovnik,bpopovic,dragisa}@uns.ac.rs

Abstract. The paper presents a speaker detection system based on phoneme specific hidden Markov model in combination with Gaussian mixture model. Our motivation stems from the fact that the phoneme specific HMM system can model temporal variations and provides possibility to ponder the scores of specific phonemes as well as efficient pruning. The performance of the system has been evaluated on speech database which contains utterances in Serbian from 250 speakers (10 of them being the target speakers). The proposed model is compared to a system based on Gaussian mixture model - universal background model, and showed a significant improvement in detection performance.

Keywords: Speaker detection, Hidden Markov models, Gaussian mixture models.

1 Introduction

Nowadays, when the amount of available multimedia data is huge and rapidly growing, it is becoming vital to classify such data automatically, in order to speed up data search. One of the interesting areas in this process is automatic speaker detection and indexing in audio stream, which corresponds to the task of speaker recognition. Traditionally, there are two basic approaches in speaker recognition: verification and identification. Speaker verification (or speaker authentication) is a binary decision problem, determining whether or not an unknown voice belongs to the particular (claimed) speaker. Speaker identification is the task of choosing an unknown speaker from the predefined set of speakers, who has the voice closest to the unknown voice. Speaker detection is the task of detecting one or more specific speakers (target speakers) in an audio stream [1]. If the number of target speakers is negligible compared to the total number of speakers presented in the data, the detection task is similar to the verification task, i.e. the task is to decide whether the speech segment belongs to the one of target speakers. On the other hand, if only target speakers are included in the data to be searched, or if they represent the majority, the detection task can be treated as an identification task [1].

The paper investigates the use of phoneme models in a speaker detection task. Previous experiments showed that the best system based on Gaussian mixture models (GMMs) outperforms the systems based on Hidden Markov models

A. Ronzhin et al. (Eds.): SPECOM 2014, LNAI 8773, pp. 410–417, 2014.

(HMMs), due to the ability of the GMM to better exploit the training data and a poor HMM parameters estimation [2]. While identical front ends were used, no gain in performance was achieved by the use of temporal information captured in the HMMs. Nonetheless, the HMM based systems that incorporate the knowledge about phonemes or phoneme classes outperform the conventional GMM based systems [3], especially in case where there is channel mismatch between training and testing data. Individual phonemes carry various amounts of speaker discriminating capability, and they each show different levels of performance, depending on the level of adaptation used to build the phoneme-specific target models [4]. Unlike the phoneme-based systems, the conventional speaker detection as well as speaker recognition systems rely on short-term spectral and prosodic features extracted from speech, e.g. [5–7]. These features are often unable to capture long-term speaker characteristic, even when they are expanded with their first and second order derivatives [8].

GMM in combination with universal background model (GMM-UBM) is a referent speaker identification system used in this paper. It has become a dominant modeling approach in speaker recognition applications [9]. The proposed HMM-GMM system uses phonetic information for creating the model of a speaker. Our motivation stems from the fact that the HMM system allows standard pruning procedures (pruning is more efficient in case of phonetic models) and the possibility to ponder the specific phonemes. This is an important advantage, having in mind the differences in reliability of log-likelihood rates for different phonemes (e.g. the log likelihood rates for vowels are much more reliable than the log-likelihood rates for plosives). We combine short-term spectral features with phoneme information, using the continuous speech recognition system described in [10], in order to improve the accuracy of our speaker identification system. Hidden Markov model toolkit (HTK) is used for speaker adaptation [11], while the rest of the system was built from scratch. The results are provided for both context-dependent and context independent phonetic models.

The paper is organized as follows. Section 2 provides the detailed description of the referent and proposed HMM-GMM system, as well as database used for performance evaluation. Section 3 gives the results of our experiments. Section 4 contains conclusions based on considerations from the previous sections.

2 Experimental Setup

All of the systems analyzed in this paper have the same front-end. Our feature vectors include 12 Mel-frequency cepstral coefficients, normalized energy and their first and second order derivatives, which are extracted on 30 ms frames with a 10 ms shift.

The referent system is GMM-UBM model where both the target speaker model and the universal background model (UBM) are based on GMMs with diagonal covariance matrices. The UBM is trained using the EM algorithm and the speech data from a large number of speakers (in this paper more than 400 speakers and 12 hours of speech). Maximum likelihood linear regression (MLLR)

[11] of the mean vectors is used to adapt the UBM for a target speaker. In the recognition phase, models for silence and noise are used to remove non-speech segments from the test sequence. The remaining speech frames are used in scoring procedures, where the logarithm of likelihood ratio of the target model and the UBM is computed for each frame. The final score for a test utterance is calculated by averaging the log-likelihood ratios on some portion of the best scored frames. There are 2 variants of a system based on the GMM-UBM used in this paper, which differ only in the number of Gaussians. The one containing 400 Gaussians will be referred as "GMM 400" and the other containing 800 Gaussians will be referred as "GMM 800".

The main difference between the proposed HMM-GMM system and the baseline system is in the way of modeling target speaker, i.e. in the proposed model each phoneme is represented with one or more HMMs. The number of HMM states is proportional to the average duration of the phoneme instances in the training set, and the number of Gaussians per state is estimated using the cross-validation procedure described in [13]. Since each phoneme corresponds to many phones, three different phoneme modeling units have been examined: monophones, biphones and triphones. The monophone model is the most general one, since it includes many phonemes which can be acoustically significantly different. In order to reduce these within-class variations, context dependent models (biphones and triphones) are introduced. Specifically, triphone model takes into consideration both (left and right) contexts, and biphone model takes only closer one. In case of triphone models, tree based clustering procedure was used to tie similar states, such that the total number of states is 1000. Low decoding computational complexity is the main reason for this small number of different HMM states in a model. It is important to note that a set of phoneme specific HMMs was built for each gender separately, in order to improve phoneme recognition accuracy.

To enroll target speaker adaptation based on MLLR is applied. Only mean vectors are adapted, leaving variances unchanged. Gender dependent model with higher likelihood on training data for a given speaker was used as an initial model in the adaptation procedure. Both supervised and unsupervised adaptation procedures were performed, since the assumption was that besides audio data, their transcriptions could also be known. The enrolment procedure consists of 3 adaptations, one (the first one) with global transformation and two with regression tree based transformations [11]. The regression tree consists of 128 leaves and about 1000 observations were used per each transformation matrix. In the case of unsupervised adaptation, the procedure is almost the same as the previous one, but it has an additional step in which phonetic (unfortunately erroneous) transcriptions of training data are automatically generated using speech recognition module. It is important to note that the speaker enrollment procedures are the same for all HMM variations which differ in phoneme modeling units.

The decoding procedure is the same for all HMM-GMM variations. It starts with Viterbi decoding with unconstrained language model (it allows transitions from any phoneme, silence and noise model, into any phoneme, silence and noise model) followed by scoring with UBM system (in this case “GMM 400”). After that, for each frame, the log-likelihood ratio of the target model and the UBM is computed, and all non-speech segments, as well as segments corresponding to the sequence of phonemes without any vowel are discarded from further scoring.

Besides capturing temporal information, as opposed to the baseline GMM-UBM approach, additional motive for phonetic approach was the ability to assign different weights to different groups of phonemes. Although we could not completely rely on the phonetic recognition results, the system at least recognized the correct group of phonemes (vowels, nasals, fricatives, stops...) which turned out to be very useful. Assigning different weights to different groups of phonemes improved the results significantly, compared to the average metric computed over the whole utterance. Postprocessing performs one additional task. It discards desired percentage of the worst scores from the results. Since our tests were performed on recordings which included both target speaker and imposter (or two imposters), it was perfectly reasonable to try this approach. All the mentioned coefficients (phoneme weights and percentage of utterance to be discarded), were carefully trained and optimized in order to maximize accuracy.

The speech corpus used in this paper is based on the speech corpora described in [14], where the utterances in original corpora are combined in the way that each audio file contains utterances of two different speakers with minimum duration of 30 s per speaker. On the other hand, the data intended for training speaker specific model contains only 30 s of speech of a single speaker. The database encompasses voices of 250 different speakers, where only 10 of them are target speakers (7 male and 3 female). The test data contains about 15 hours of audio recordings split into about 160 files, but only 50 of them contains voices of target speakers. The full spectrum speech data from “S70W100s120”, “AN_Books” and “AN_Broadcast Speech” [14] is filtered and re-sampled to 8 kHz to reproduce telephone channel data.

3 Experimental Results

Figure 1 shows detection error trade-off (DET) curves of the baseline systems. Using 50 % of the best scoring frames resulted in performance gain for both baseline systems. Since the difference in the performance between “GMM 800 top 50” and “GMM 400 top 50” is negligible, the GMM-UBM models with more than 800 Gaussians were out of consideration. The “GMM 800 top 50” system was used as the referent one for the all HMM-GMM-UBM systems. The speaker scoring for both systems was performed only on the detected speech frames by averaging log likelihood ratios in 2 variants. In the first variant all speech frames were used (“all”) and in the second variant only 50 % of the best scoring frames were used (“top 50”).

Figure 2 shows DET curves for HMM-GMM systems with different level of phoneme model generalizations when both audio data and their transcriptions are used in speaker adaptation . Using scores obtained on all valid speech frames significantly deteriorates system performance, which was expected, since the test

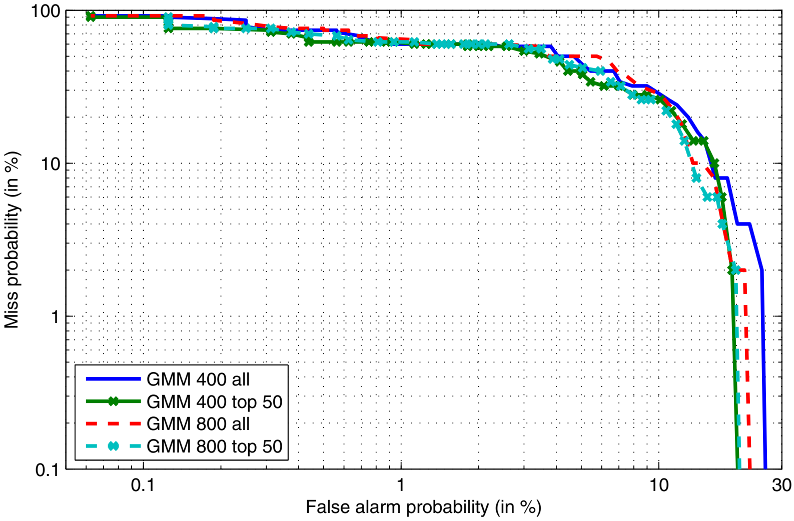

Fig. 1. DET curve for GMM-UBM systems

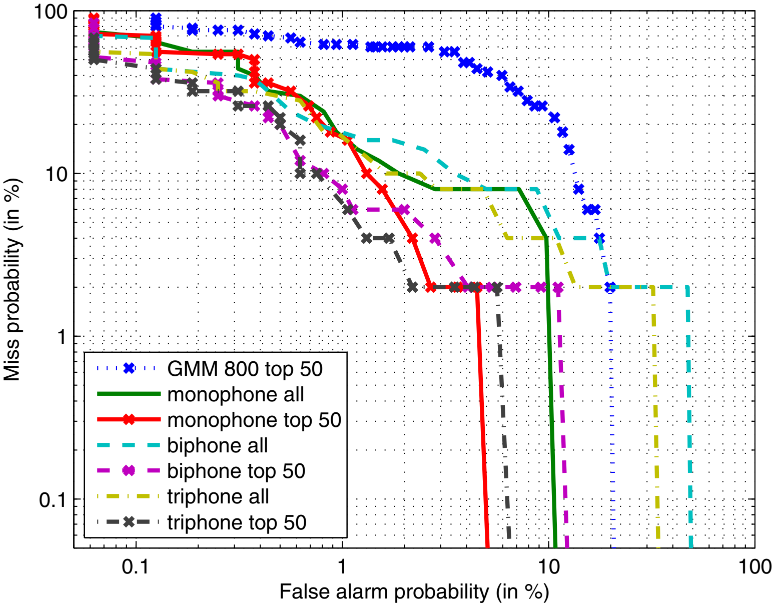

Fig. 2. DET curve for HMM-GMM systems if both audio data and their transcriptions are used in speaker adaptation

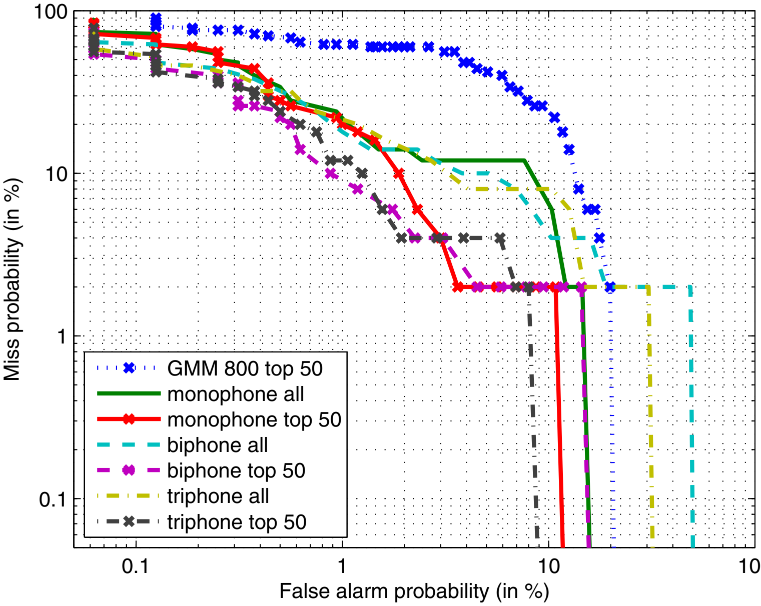

Fig. 3. DET curve for HMM-GMM-UBM systems if only audio data are used in speaker adaptation

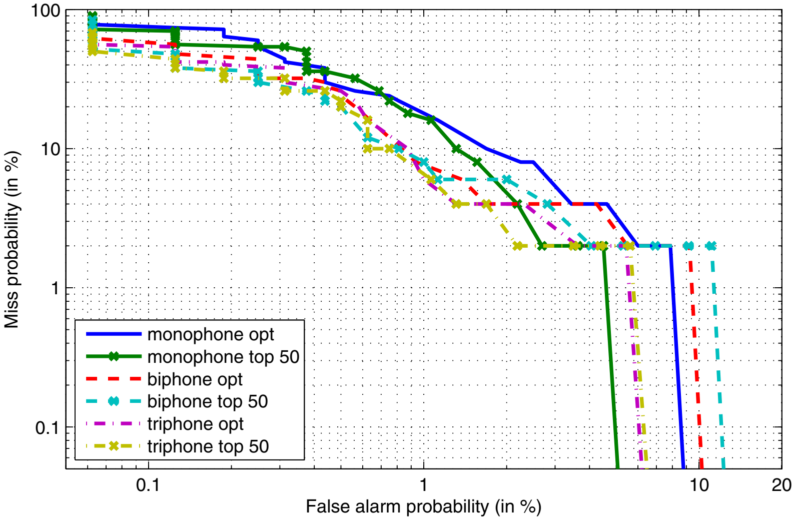

Fig. 4. DET curve for HMM-GMM-UBM systems if both audio data and their transcriptions are used in speaker adaptation, and the phoneme weights and percentage of utterance to be discarded are optimized

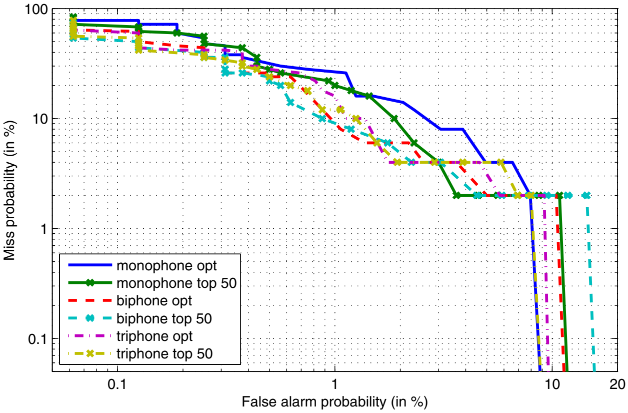

Fig. 5. DET curve for HMM-GMM-UBM systems if only audio data are used in speaker adaptation, and the phoneme weights and percentage of utterance to be discarded are optimized

utterances with target speaker voice include the voice of another speaker. Scores on the frames belonging to imposter decrease the total score and in that way mask the target speaker. One can see that the all HMM-GMM systems outperforms the baseline system ("GMM 800 top 50"). Since the test set has only 50 audio files containing target speaker, resolution of miss probability is 2 %, thus DET curves for small values of miss probability are unreliable and should be ignored in further analysis. Both HMM systems with context dependent models outperform the system with monophones, but the number of Gaussians in monophone model is significantly smaller (1716 compared to 14404 for biphones and 12732 for triphones), therefore there is no reliable explanation. Similar results are obtained for the HMM-GMM systems if adaptation is based only on audio data (see Fig. 3).

The performance of these systems in case of additional phoneme weighting and automatic discarding of potential imposter frames are shown in Fig. 4 and Fig. 5. One can see that the optimization procedure gives results that are comparable with those obtained when 50 % of frames are discarded. We expected better results, but small validation set used for the parameter (phoneme weights and discard percent) estimation as well as evolutionary optimization procedure can lead to these results.

4 Conclusions and Future Work

This paper presents a comparison between GMM-UBM and phoneme specific HMM based speaker detection systems. The proposed HMM based systems outperform GMM-UBM systems, since they use phonetic informations to filter frames log-likelihood ratios (exclude unreliable frames or segments from scoring).

Context dependent HMMs show a slightly better performances than the context independent ones, which means that the longer matching between unknown sequence and mode reduce classification errors. The proposed model should be tested on the larger database containing utterances in different languages.

Acknowledgements. This research work has been supported by the Ministry of Education, Science and Technological Development of the Republic of Serbia, and it has been realized as a part of the research project TR 32035.

References

1. Beigi, H.: Fundamentals of Speaker Recognition. Springer (2011)
2. Auckenthaler, R., Parris, E., Carey, M.: Improving a GMM speaker verification system by phonetic weighting. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 1999), vol. 1, pp. 313–316. Phoenix, Arizona (1999)
3. Kajarekar, S., Hermansky, H.: Speaker verification based on broad phonetic categories. In: A Speaker Odyssey - The Speaker Recognition Workshop (2001)
4. Hansen, E., Slyh, R., Anderson, T.: Speaker recognition using phoneme-specific GMMs. In: ODYSSEY 2004-The Speaker and Language Recognition Workshop, pp. 179–184 (2004)
5. Dunn, R., Reynolds, D., Quatieri, T.: Approaches to speaker detection and tracking in conversational speech. Digit. Signal Process. 10, 93–112 (2000)
6. Kinnunen, T., Li, H.: An Overview of Text-Independent Speaker Recognition: From Features to Supervectors. Speech Commun 52, 12–40 (2010)
7. Scheffer, N., Ferrer, L., Graciarena, M., Kajarekar, S., Shriberg, E., Stolcke, A.: The SRI NIST 2010 Speaker Recognition Evaluation System. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2011), pp. 5292–5295. Prague, Czech Republic (2011)
8. Antal, M.: Phonetic Speaker Recognition. In: 7th International Conference COMMUNICATIONS, pp. 67–72 (2008)
9. Reynolds, D., Quatieri, T., Dunn, R.: Speaker Verification Using Adapted Gaussian Mixture Models. Digit. Signal Process. 10, 19–41 (2000)
10. Delić, V., Sečujski, M., Jakovljević, N., Janev, M., Obradović, R., Pekar, D.: Speech Technologies for Serbian and Kindred South Slavic Languages. In: Advances in Speech Recognition, pp. 141–165 (2010)
11. Young, S.J., Evermann, G., Gales, M.J.F., Hain, T., Kershaw, D., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., Woodland, P.C.: The HTK Book, version 3.4 (2006)
12. Gales, M., Young, S.: The Application of Hidden Markov Models in Speech Recognition. Foundations and Trends in Signal Processing 1(3), 195–304 (2007)
13. Jakovljević, N., Miškovic, D., Janev, M., Sečujski, M., Delić, V.: Comparison of Linear Discriminant Analysis Approaches in Automatic Speech Recognition. Elektronika Ir Elektrotechnika 19(7), 76–79 (2013)
14. Delić, V., Sečujski, M., Jakovljević, N., Pekar, D., Mišković, D., Popović, B., Ostrogonac, S., Bojanić, M., Knežević, D.: Speech and language resources within speech recognition and synthesis systems for serbian and kindred south slavic languages. In: Železný, M., Habernal, I., Ronzhin, A. (eds.) SPECOM 2013. LNCS, vol. 8113, pp. 319–326. Springer, Heidelberg (2013)