Improved Novelty Detection for Online GMM based Speaker Diarization

Konstantin Markov[1,2], *Satoshi Nakamura*[1,2]

[1]ATR Spoken Language Communication Research Labs, Japan
[2]Spoken Language Communication Group, NICT, Japan
konstantin.markov@atr.jp, satoshi.nakamura@atr.jp

Abstract

Detection of speakers which have not been seen before is an essential part of every online speaker diarization system. New speaker detection accuracy has direct impact on the overall diarization performance. In our previous system, for novelty detection we used global GMM likelihood ratio (LR) threshold. However, as the system analysis showed, the optimal threshold depends on the speaker gender as well as on the number of registered speakers. In this paper, we present the results of this analysis and the approach we have taken to solve this problem. First, we use different thresholds for male and female speakers, and second, for each gender before the thresholding we apply likelihood ratio mean and variance normalization. This greatly reduced the threshold dependency on the number of speakers and allowed to use fixed threshold for each gender. The LR distribution statistics are collected online and updated every time new likelihood ratio is calculated. Experiments on the TC-STAR database showed that compared with the previous global threshold method, the new novelty detection approach reduces the speaker diarization error rate up to 35%.

Index Terms: speaker diarization, novelty detection, never-ending learning, likelihood ratio normalization.

1. Introduction

Identifying and labeling the sound sources within a spoken document is the task of audio diarization. A main part of this process is the speaker diarization where the goal is to find out “who spoke when”.

In the speaker diarization task, the number of speakers, i.e. classes, is not known in advance, and this requires automatic systems to be capable of some form of unsupervised learning. Agglomerative clustering is the method used by the wast majority of the current systems[1]. Initially, every speech segment is assigned to a different cluster and then, at each iteration the two closest clusters are merged into one. It is assumed that speech segments belonging to the same speaker are closer to each other than segments belonging to different speakers. After some stopping criterion is met, the remaining clusters are supposed to represent individual speakers. Widely used distance measure and stopping criterion are the generalized likelihood ratio (GLR) and the Bayesian information criterion (BIC) [2, 3]. Several variations of this method have also been proposed [4, 5], but they are still based on the same bottom-up clustering technique. Although, quite successful, agglomerative clustering approach has several drawbacks that limit the potential use of the speaker diarization systems in the real-world, real-time applications. First, it requires all the speech segments to be available before the clustering starts and, therefore, makes on-line processing impossible. Second, the computational load increases almost exponentially with the number of segments[6]. Finally, the performance is greatly affected by the stopping criterion which is considered as a critical part of the algorithm [1].

A sequential algorithm based on the leader-follower clustering [7] and suitable for on-line operation has been proposed several years ago [6]. However, as in the agglomerative clustering method, the speech segments are modeled by a single Gaussian distribution and the GLR is used as a distance metric. This reduces the clustering accuracy for short segments and delays the decision until the whole segment is received. In consequence, the system latency becomes dependent on the segment’s length which can be up to 30 sec. or even longer.

Recently, we proposed a new speaker diarization system which operates on-line, in less than real time, and has low latency of up to few seconds [8]. What makes it significantly different from the other systems is the way the segment clustering is performed as well as the overall operating algorithm, which is based on the Never-Ending Learning (NEL) principle [9]. In our system, when assigning speaker label to a given segment, first, it is decided whether it belongs to one of the known speakers or to a new speaker. Then, in the former case, speaker identification is performed and the winning speaker label is assigned to the segment. In the latter case, new speaker is registered to the system and his/her model is created. Each speaker is represented by a GMM which is learned on-line every time it has been a winner. New speaker’s GMM is created by spawning speaker independent GMM trained in advance. In addition, each speaker GMM has a time counter which is set to zero whenever it wins the identification. Otherwise, the counter is incremented by the current segment length. Models whose counter reaches some threshold T, are deleted from the system. This way, the system can operate indefinitely, adapting itself to the environment changes, i.e. changes in the number of speakers and their characteristics in an unsupervised manner, and this makes it a NEL system. Essential part of the system is the novelty detection which is based on hypothesis testing and is implemented as likelihood ratio thresholding. In our previous experiments, we used fixed global threshold tuned on held-out data set. However, detailed performance analysis showed that the optimal threshold depends on the number of registered speakers as well as on the speaker gender.

In this paper we present the results of the performance analysis and the approach we have taken to reduce the threshold dependency on the current speaker gender and the number of known speakers.

The next section gives a formal definition of the speaker diarization task as statistical pattern recognition problem where we assume that the underlying probability density function is in fact time-varying. In section 3 we provide brief description of the system, followed by details about the novelty detection algorithm. Section 5 summarizes the experiments and obtained results. The conclusions are presented in the last section.

2. Statistical task definition

Traditionally, most pattern recognition tasks are defined as a maximization problem of the following form:

$$\hat{S} = \max_i P(S_i|X) = \max_i P(X|S_i)P(S_i) \tag{1}$$

where X is the input pattern, $S = \{S_i\}$ represents the classes of interest, $P(X|S_i)$ is an unknown pdf which we approximate with some set of models (one for each S_i) trained with data samples coming from this pdf, and $P(S_i)$ is the prior probability of observing S_i. In the speaker identification task, for example, S_i corresponds to a speaker ID and the $P(S_i)$ is usually ignored. In some more complicated tasks like speech recognition, S_i would be a sequence of classes, i.e. words, X - the sequence of their realizations, and $P(S_i)$ - the language model. The above definition, however, is based on several assumptions: 1) the pdf is constant; 2) the number of classes is fixed and known; and 3) the test patterns X come from the same pdf as the training data.

Given the specifics of the speaker diarization task, it is clear that we cannot apply Eq.(1) because the number of speakers, i.e. classes, is unknown, and furthermore, it changes from one audio document to another. The way we solve this problem is to assume that the pdf is time varying and to define the task as[1]:

$$\hat{S} = \max_i P_t(S_i^t|X) = \max_i P_t(X|S_i^t)P_t(S_i^t) \tag{2}$$

where S_i^t is the i^{th} sequence of speakers from the set of speakers S^t at time t. Without any additional knowledge about speakers it is reasonable to assume that all possible speaker sequences are equally probable and therefore we can drop the term $P_t(S_i^t)$ from the equation above. Obviously, to approximate time varying pdf we need time varying models. But such models cannot be trained in advance as we usually do. The alternative is to learn the models online and keep learning all the time tracking the pdf changes. To do this, the speaker diarization system should be capable of:

- **unsupervised adaptive learning** - to automatically learn variations in speakers voice characteristics.
- **novelty detection** - to detect previously unseen speakers.
- **knowledge preservation** - to preserve the knowledge about the old speakers while learning a new speaker.
- **gradual forgetting** - to delete the knowledge about irrelevant speakers.

A system satisfying these four requirements is called *Never-Ending Learning* system and is suitable not only for the speaker diarization task, but for any other task involving unknown time varying pdf. And since most of the natural processes evolve and change in time, modeling them with varying pdfs might be a key to many challenging real world problems.

3. System operation

In this section, we briefly describe the our system and how the above four requirements are satisfied. More details are given in [8]. The system works on-line and and its operation is schematically shown in Fig.1. The speech segments and their reference speaker labels are at the top of the figure. The bottom part shows the speaker models and how they change in time. For each speech segment, there is a winning model indicated by a thick border line. At the beginning, there are only three GMMs: one for pause (not shown for clarity) and two for each speaker gender. They are trained in advance from some labeled data. For the first segment, the speaker gender is identified (male in the figure) and a new GMM is created from the male GMM. It is learned on-line with the segment's data, and from this point it becomes the GMM for Speaker 1 (SP1 in the figure). The next segment is from the same speaker, so the SP1 GMM will be the winner. It is again learned on-line with the second segment's data. The third segment comes from a female speaker and the same procedure is repeated resulting in a set of two speaker GMMs. This way, the system generates a set of speaker models on the fly. If some GMM (SP1 in the figure) has not been a winner for a long time, it is deleted from the system (indicated by an "X" on the figure). Such operating mode allows the system to work indefinitely.

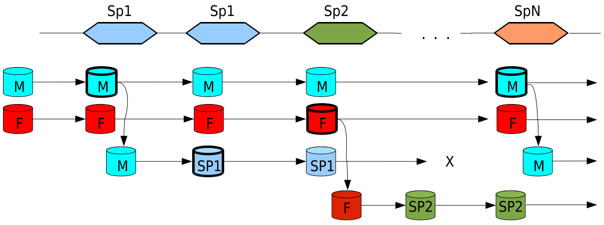

Figure 1: System operation. For each speech segment, the winning GMM is denoted by bold border lines.

4. Novelty detection

The ability to detect new speakers is essential for the online speaker diarization system. Ideally, the speaker independent gender GMM would be a better match for every new speaker (from the same gender) than any of the old speaker models and when it gives the highest likelihood it can be considered that current speech segment comes from an unseen speaker. In practice, however, likelihood ratio (LR) based hypothesis testing is more robust and intelligent approach. It is formulated as follows:

$$X \in \begin{cases} \omega_0, & \text{if } \mathrm{L(X)} > \theta \\ \omega_1, & \text{if } \mathrm{L(X)} < \theta \end{cases} \tag{3}$$

where X is the speech segment and ω_0 is the class corresponding to the hypothesis H_0, i.e. old speaker. Respectively, ω_1 corresponds to H_1, i.e. new speaker. The likelihood ratio is:

$$L(X) = \frac{p(X|\omega_0)}{p(X|\omega_1)} \tag{4}$$

There are various ways to define $p(X|\omega_i)$. Considering the available set of GMMs, a straightforward approach is to define them as:

$$p(X|\omega_0) = P_{sp} \;= \max_{\lambda_j \in \Lambda} p(X|\lambda_j) \tag{5}$$

$$p(X|\omega_1) = P_{gen} = \begin{cases} P(X|\lambda_M), & \text{if } gen(sp) = M \\ P(X|\lambda_F), & \text{if } gen(sp) = F \end{cases}$$

where $\Lambda = \{\lambda_j\}$ is the current set of speaker GMMs and $gen(sp)$ is the gender of the winning speaker. Another approach, often used in speaker verification is to define $p(X|\omega_1)$ as:

$$p(X|\omega_1) = P_{ave} = \frac{1}{n-1}(\sum_j p(X|\lambda_j) - P_{sp}) \tag{6}$$

[1] Actually, X is also time dependent, but since its time index may not be related to that of the pdf, we have dropped it for clarity.

i.e. the average of all model likelihoods except for the winning model. Here $n = |\Lambda|$ is the size of the speaker set. We can also combine the two approaches and in this case the likelihood ratio is:

$$L(X) = \frac{P_{sp}^2}{P_{gen}P_{ave}} \tag{7}$$

In practice, the optimal threshold θ is determined on a held-out data set and this was the approach we used in our previous system. However, the analysis of the system performance presented in Section 5 showed that the optimal threshold is different for male and female speakers. Furthermore, it changes with the number of registered speakers. In order to reduce the threshold variability, this time, we applied mean and variance normalization of the likelihood ratio values [10]. The normalized LR value is calculated as:

$$L^{norm}(X) = \frac{L(X) - \mu_L}{\sigma_L} \tag{8}$$

where μ_L and σ_L are the mean and standart deviation of the LR values. There are two ways of obtaining the LR statistics. One is to estimate them on some development data and the other is to do this online during the system operation. The first approach gives reliable but fixed estimates which may not match the actual LR distribution of the evaluation data. The later is better, but at the beginning, when there are only few LRs available, the estimation error may be too big. As it usually turns out, the combination of the two approaches is the best. Thus, we use the mean and standart deviation obtained from the development set as initial values for the incremental online statistics estimation.

5. Experiments

5.1. Database and pre-processing

For the system evaluation, we used the data released for the TC-STAR 2007 evaluation campaign [11]. The data consists of recordings of the European Parliament plenary speeches. From the training part of the database, for the gender dependent GMMs we used about 2 min. of speech from each of 20 male and 15 female speakers. The official development set was used as development data ("dev"), and the evaluation set ("eval") from the TC-STAR 2006 campaign was used for the final system evaluation.

All audio data were transformed into 26 dimensional feature vectors consisting of 12 MFCC coefficients, power and their first derivatives. The frame length and rate were 20 and 10 ms. respectively.

5.2. Baseline agglomerative clustering system

A system based on the standard agglomerative clustering approach was built to provide a baseline performance for comparison. In this system, each cluster is represented by a single Gaussian function with full covariance matrix and generalized likelihood ratio (GLR) was used as intercluster distance measure. The clustering procedure is stopped when the change in the Bayesian information criterion statistic (ΔBIC) turns positive. The GRL and ΔBIC are defined as follows:

$$GLR_{x,y} = \frac{|\Sigma_{x\cup y}|^{N_{x\cup y}/2}}{|\Sigma_x|^{N_x/2}|\Sigma_y|^{N_y/2}} \tag{9}$$

$$\Delta BIC = \log GLR_{x,y} - \alpha\left(\frac{d(d+3)}{4}\right)\log N_{x\cup y}$$

where x and y are the two clusters to be merged, N is the number of frames in the cluster, d is the feature vectors dimension, and α is a free parameter tuned on the development data.

The baseline system uses the same voice activity detector (VAD) and gender identification module as our online system and therefore the difference in performance comes only from the different speaker segmentation algorithms. For each segment detected by the VAD the speaker gender is identified and agglomerative clustering is performed on the pool of male and female speaker segments separately. Table 1 shows the speaker diarization error rate (DER) for the baseline system when the forgiveness collar around the reference segments boundaries is set to 0.0 or 0.25 sec.

Table 1: DER (%) for the baseline system with α tuned on the "dev" data set.

Collar = 0.0		Collar = 0.25	
"dev"	"eval"	"dev"	"eval"
10.9	9.5	8.4	7.6

5.3. Online system with global threshold

Essential parameter of any online system is the system latency, i.e. the time delay needed for making decision which in our case is labeling the input segment with a speaker ID. In both the "dev" and "eval" data sets the segments length varies from less than a second to more than 10 sec. To achieve consistent latency we force the system to make decision using only the initial part of each segment long as much as the desired delay. If the segment's length is less, then the whole segment is used. In our previous system [8], the optimal global threshold was obtained on a separate data set. Now, we used the "dev" data to tune it which resulted in small improvement for the "eval" data. Table 2 shows the DER for both test sets when the system latency is set in the range from 1 to 5 sec.

Table 2: DER (%) for the online system with fixed global threshold tuned on the "dev" data set.

System	Collar = 0.0		Collar = 0.25	
latency	"dev"	"eval"	"dev"	"eval"
1 sec.	14.9	20.3	12.4	18.5
2 sec.	9.6	14.8	6.9	12.9
3 sec.	6.8	13.4	4.2	11.5
4 sec.	6.3	11.9	3.6	9.9
5 sec.	6.2	9.8	3.6	7.9

5.4. Optimal threshold variability analysis

In order to find out how the optimal threshold changes depending on the speaker gender and the number of registered speakers, we did the following experiment. We assigned the correct labels to each segment of the "dev" data and by random shuffling the segments order 50 times, we obtained 50 sets of data. Then the system was run on each of these sets and likelihood ratio values were collected and split into different clusters depending on the number of registered speakers. Then from the data in each cluster two gaussian distributions were estimated: one using LR values when the input speaker was old and the other when the input speaker was new. The optimal threshold value was then calculated at the intersection point of the two gaussians. These values for each number of registered speakers are plotted in the upper part (the positive values) of Fig.2 when the system latency was set to 3 seconds. For all other latency values we obtained very similar results. The square points show the thresholds for male speakers and circle points correspond to female speakers. To show the general change tendency, a second order polynomial fit lines are also plotted in the figure. It is clear that the threshold varies depending on both the speaker gender and the speaker number. The lower part (the negative values) of Fig.2 shows the same thresholds obtained from normalized likelihood ratio values as describes in Section 4. The mean and standard deviation statistics were calculated from LR values in each cluster separately. As can be seen from the plot, likelihood ratio normalization greatly reduces the threshold variability with respect to both speaker number and speaker gender.

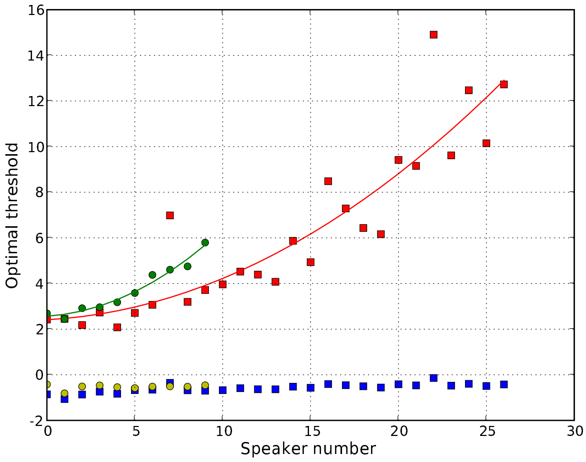

Figure 2: Optimal threshold values (upper part) and normalized optimal threshold values (lower part) given the number of speakers for the "dev" data set. Square point correspond to male speakers and circle points - to female speakers. System latency is 3 seconds.

5.5. Online system using normalized likelihood ratio

Based on the findings above and practical considerations presented in Section 4, in our final system we applied the following likelihood normalization scheme. The LR statistics for the case of one speaker estimated in the previous experiment were used as initial statistics values at the beginning of the system operation. Then, for each input segment the LR value was used to update the statistics and then normalization was applied. Normalized LR is compared with a fixed threshold and novelty decision is made. We used separate thresholds for male and female speakers and fine tuned them on the "dev" set. However, there was no big difference between them. Final DER results are summarized in Table 3. Comparing the results with those from Table 2 we can see that LR normalization is really effective and that the performance gain for the "eval" set is much bigger than for the "dev". This shows that the system performance dependency on the test data is reduced by the new novelty detection approach.

Table 3: DER (%) for the online system with gender dependent threshold and normalized likelihood ratio.

System	Collar = 0.0		Collar = 0.25	
latency	"dev"	"eval"	"dev"	"eval"
1 sec.	12.5	14.3	10.1	11.9
2 sec.	8.3	10.7	5.9	8.3
3 sec.	5.8	8.5	3.3	6.1
4 sec.	5.3	7.9	3.0	5.4
5 sec.	5.1	7.6	2.9	5.3

6. Conclusions

In this paper we presented our investigations on how to improve the novelty detection performance of our online speaker diarization system. Since the decision whether the input segment comes from a new speaker or not is based on hypothesys testing implemented as likelihood ratio thresholding, it is important to ensure that the distributions of the LR values for each hypothesis do not depend too much on the environment and input data. This was achieved by using LR normalization and the experimental results showed not only better performance but also higher system stability with respect to different test sets.

7. References

[1] S. Tranter and D. Reynolds, "An Overview of Automatic Speaker Diarization Systems," *IEEE Trans. ASLP*, vol. 14, no. 5, pp. 1557–1565, Sept. 2006.

[2] L. Lamel, J.-L. Gauvain, G. Adda, C. Barras, E. Bilinski, O. Galibert, A. Pujol, H. Schwenk, and X. Zhu, "The LIMSI 2006 TC-STAR transcription system," in *Proc. TC-STAR Workshop on Speech-to-Speech Translation*, Barcelona, June 2006, pp. 123–128.

[3] F. Brugnara, D. Falavigna, D. Giuliani, R. Gretter, D. Pineda, D. Seppi, and G. Stemmer, "The ITC-irst transcription systems for the TC-STAR-06 evaluation campaign," in *Proc. TC-STAR Workshop on Speech-to-Speech Translation*, Barcelona, June 2006, pp. 117–122.

[4] D. Liu and F. Kubala, "Fast speaker change detection for broadcast news transcription and indexing," in *Proc. Eurospeech*, Sept. 1999, pp. 1031–1034.

[5] M. Ben, M. Betser, F. Bimbot, and G. Gravier, "Speaker diarization using bottom-up clustering based on a parameter-derived distance between adapted GMMs," in *Proc. ICSLP*, Oct. 2004, pp. 2329–2332.

[6] D. Liu and F. Kubala, "Online Speaker Clustering," in *Proc. ICASSP*, May 2004, pp. 333–336.

[7] R. Duda, P. Hart, and D. Stork, *Pattern Classification*, John Wiley & Sons, Inc., Second edition, 2001.

[8] K. Markov and S. Nakamura, "Never-Ending Learning System for Online Speaker Diarization," in *Proc. ASRU*, Dec. 2007, pp. 699–704.

[9] K. Markov and S. Nakamura, "Never-Ending Learning with Dynamic Hidden Markov Network," in *Proc. INTERSPEECH*, Aug. 2007, pp. 1437–1440.

[10] D. Reynolds, T. Quatieri, and R. Dunn, "Speaker Verification using Adapted Gaussian Mixture Models," *Digital Signal Processing*, vol. 10, no. 1-3, pp. 19–41, Jan. 2000.

[11] TC-STAR, "Technology and Corpora for Speech to Speech Translation," Online: http://www.tc-star.org/.