Improvement Speaker Clustering Using Global Similarity Features

Konstantin Biatov, Joachim Köhler

NetMedia Center
Fraunhofer Institute for Media Communication, Sankt Augustin, Germany
konstantin.biatov@imk.fraunhofer.de, joachim.koehler@imk.fraunhofer.de

Abstract

In this paper global similarity features that improve speaker clustering based on standard bottom-up clustering are proposed. The novelty of this approach lies in the fact that it exploits the hypothesis that audio segments belonging to the same speaker cluster should demonstrate similar global behavior, exhibit the same similarity and dissimilarity with all the other segments. Every segment is represented by a global similarity vector whose components are encoded by the distance between that segment and each of the other segments to be clustered. The distance between global similarity vectors is used for pre-selection of segment pairs having high global similarity for further merging. In this paper inter-segment distance for global similarity vectors based on Bayesian Information Criterion (BIC) and based on adapted cross likelihood ratio (CLR) are investigated. The evaluation, performed on radio programs, shows that the proposed approach represents an improvement in comparison with the baseline clustering.

Index Terms: speaker clustering, global similarity, adapted cross likelihood ratio

1. Introduction

Speaker clustering is a process of partition of audio data into speaker homogeneous segments. The segments produced by the same speaker are labeled as one cluster. The task of speaker clustering is useful for speaker adaptation and for speaker identification. Currently speaker clustering is used in spoken documents indexing applications. In most situations in the speaker clustering task the number of speakers, the channel and speaker characteristics are unknown and must be estimated from the data. In such cases unsupervised speaker clustering is required. The main problem in this task is that sometimes very few data per speaker is available and robust speaker model from this data can not be obtained. In order to avoid this problem speaker model is mostly obtained by adapting a prior speaker model. The adaptation is usually achieved by using maximum a posteriori (MAP) adaptation of universal background model (UBM) based on a large collection of prior available audio data used in the training phase. MAP adaptation of Gaussian Mixture Models (GMM) for cluster models was described in some recent publications [1], [2].

In [3] our approach based on global similarity constraint for speaker clustering was described. In standard agglomerative clustering of audio segments in every step each pair of segment models are compared and then the two most similar segments are merged. After updating the model of merged segments the process is repeated and then is stopped according to the stop criterion. In [3] we have suggested global similarity metrics for compared segments. The global similarity depends on how a segment is similar to all other segments taking part in the clustering. The global similarity features of each segment are presented as a vector. Each component of this vector is the local distance between this segment and each other segment. For local inter-segment similarity ΔBIC was used. In the process of the clustering we are looking for a pair of segments having distance between global similarity vectors less than the predefined threshold. Then the best pair with the minimal distance is selected to be merged. The process of the clustering stops when no more pairs satisfying these global similarity constraints exist.

In this paper we improve performance described in [3] algorithm using adapted CLR as local inter-segment distance for global similarity vectors generation.

The presented paper has the following structure. Section 2 describes the baseline speaker clustering algorithm, section 3 describes suggested algorithm based on adapted CLR, evaluation criterion and experimental setup are presented in section 4, followed by some conclusions.

2. Baseline speaker clustering system

The baseline clustering algorithm is carried out in a series of steps: feature selection, silence frame filtering, speech/non-speech classification, gender recognition, speaker segmentation and speaker clustering. This section describes each of these steps.

2.1. Basic features

- For speech/non-speech classification, the feature vectors consist of 12 mel-cepstral coefficients plus energy extended with delta mel-cepstral coefficients and delta energy, delta-delta mel-cepstral coefficients and delta-delta energy.
- For gender recognition, only the 12 mel-cepstral coefficients are used.
- For speaker segmentation and clustering, 12 mel-cepstral coefficients plus energy are used extended with delta mel-cepstral coefficients plus delta energy.
- For all cases, a 30 ms analysis window and 10 ms step size are used.

2.2. Silence frame filtering

Before clustering the noise floor is calculated by finding minimal energy in the signal. Then the frames with energy above the noise floor plus a small threshold are discarded. The threshold was selected experimentally using development data.

2.3. Segmentation with BIC

Segmentation via BIC was initially proposed in [4]. In general way BIC is defined as

$$BIC(M) = \log L(X,M) - \lambda \frac{\#(M)}{2} \log(N) \quad (1)$$

where *log L(X,M)* denotes segment *X* likelihood given by the model *M, N* is the number of data points, *#(M)* is the number of free parameters in the model and λ is a tuning parameter.
Only λ=1 corresponds to the strict definition of BIC. In practice, the value of λ giving the best segmentation performance is different than 1 and depends on the features used. In order to estimate speaker's turn point between two segments c_i and c_j that have n_i and n_j frames respectively, the ΔBIC value is computed as:

$$\Delta BIC = \frac{1}{2} n_i \log|\Sigma_i| + \frac{1}{2} n_j \log|\Sigma_j| - (n_i + n_j)\log|\Sigma_{ij}| + \lambda P \quad (2)$$

where d is the dimension of the feature vector, Σ_{ij} is the covariance matrix of the data points from two segments c_i and c_j, Σ_i is the covariance matrix of the data points from the segment c_i, Σ_j is the covariance matrix of the data points from the segment c_j, and P is:

$$P = \frac{1}{2}(d + \frac{1}{2} d(d+1))\log(n_i + n_j) \quad (3)$$

ΔBIC is a distance between two Gaussian models which describe the same audio data with a hypothetic speaker turn. A negative value of ΔBIC indicates that the model that describes the data as a two-Gaussian process fits better than the model that describes the data as a one-Gaussian process. In the segmentation process we follow to the algorithm described in [4]. For the speaker segmentation step with BIC, we impose minimal segment duration of 1 sec.

2.4. Speech/non-speech classification

Speech/non-speech classification is useful for the speaker clustering task in order to prevent music or singing portions from being clustered as a speaker. In order to classify segments into speech and non-speech, we first classify the audio frames, labeling each individual audio frame as either 'speech' or 'non-speech'. We then determine the speech/non-speech classification of each speech segment by using a voting rule applied to the speech/non-speech labels of the audio frames contained in that segment.

For the classification of the individual audio frames three pairs of GMMs were constructed. All pairs of GMMs were trained using 3 hours of labeled data from German radio broadcasts. The process of classification is organized as three sequential maximum likelihood decisions. The first pair of GMMs has 1024/2 mixtures for speech and non-speech respectively. In the first step, audio frames containing pure speech are separated from all other audio types, namely, singing, music, speech with music, noise, telephone speech and auditorium speech. The second pair of GMMs has 512/2 mixtures for telephone speech and non-speech respectively. In the second step, telephone speech is separated from singing, noise, speech with music and auditorium speech. The third pair of GMMs has 512/2 mixtures for auditorium speech and non-speech respectively. Using a cascade of three maximum likelihood decision rules, audio frames containing pure speech, telephone speech and auditorium speech are separated from frames with the music and noise. Each frame is labeled as speech or non-speech.

2.5. Gender recognition

Gender classification is useful for speaker clustering task in order to prevent male and female audio data from being automatically merged together into the same cluster. After speech/non-speech classification, each speech frame, including telephone speech and auditorium speech, is classified as male/female speech. For gender classification one pair of GMMs was constructed. For each gender a 1024 mixture GMM was trained via EM algorithm using 45 minutes of labeled data for each class from a German radio broadcast training data. After speaker segmentation via BIC and speech/non-speech labeling, each speech segment was labeled by gender using the results of the gender classification on the frame level and a voting rule.

2.6. Clustering algorithm

Clustering is performed on audio segments that have been classified as speech and within groups of audio segments that have been classified as the same gender. At the beginning of clustering each segment within a gender class is considered as a separate cluster that is modeled by a single Gaussian. Under the standard clustering approach using BIC, ΔBIC is used to make a pair-wise comparison between audio segments. If ΔBIC is positive and maximal the segments are merged, if ΔBIC is negative, they are not. This comparison continues until there are no more pairs of the segments with a positive ΔBIC.

The baseline algorithm extends the use of the BIC for speaker clustering by constraining which pairs of segments can be considered for merging. We only permit two segments to be merged if they have the same patterns of similarity with all the other segments. We compare patterns of similarity by performing a fuzzy match on global similarity vectors. A global similarity vector represents each segment's similarity to all other segments in the audio file. Each component *j* of the vector *i* corresponds to the BIC similarity measure with the segment *j*.
The j^{th} component of the vector representing a given segment is 1 if that segment yields a positive ΔBIC and 2 if that segment yields a negative ΔBIC when compared with the segment *j*. If the j^{th} segment is a non-speech segment, the j^{th} component of all global similarity vectors is 0. We consider two global similarity vectors to represent the same similarity pattern if they constitute a fuzzy match. We define this fuzzy match between two vectors as the proportion of non-zero components of the two vectors which are either both equal to 1 or both equal to 2. Formally, a fuzzy match obtains, if the following holds:

$$l > \theta m \quad (4)$$

where l is the number non-zero components that are equal, m is the number of non-zero components and θ is a parameter that encodes fuzzy match. When value of θ is equal to 0 the second condition does not influence the process of the clustering and the result of the clustering should be the same as in the clustering based on BIC. The maximal value of θ is equal to 1.

2.7. Stop criterion

When two global similarity vectors have approximately all nonzero components equal (i.e. equation (4) holds) and when, furthermore, ΔBIC for the corresponding speech segments is positive and maximal, the two segments are merged. When no more pairs of speech segments remain whose global similarity vectors fulfill the conditions of equation (4) the process of clustering is stopped.

3. Speaker clustering based on global similarity constraints and adapted CLR

In this section we describe new clustering algorithm. In this algorithm part of the components - basic features, silence frame filtering, segmentation, speech-non classification and gender recognition are the same as in baseline algorithm. As in baseline algorithm each segment is characterized by the global similarity vector. In baseline algorithm for the local inter-cluster distance that encodes the components of global similarity vectors ΔBIC was used.

In the new suggested algorithm each model of the cluster is GMM and for local inter-cluster distance adapted CLR is suggested. We use CLR as was described in [5]. The local inter-cluster distance measure between segments c_i and c_i CLR is defined as:

$$CLR\ (c_i, c_i) = log\ \frac{L(x_i|M_j)}{L(x_i|M_{ubm})} + log\ \frac{L(x_j|M_i)}{L(x_j|M_{ubm})} \qquad (5)$$

where $L(x_i \mid M_j)$ is the average likelihood per frame of data x_i from the segment c_i given by model M_j and $L(x_i \mid M_{ubm})$ is the average likelihood per frame of data x_i from the segment c_i given by model M_{ubm}. The model M_j is MAP adapted model and M_{ubm} is universal background model (UBM). Using MAP adaptation we adapt only mean and weights. Factor τ was 26. UBM was trained on a large collection on radio broadcast audio data and had 128 diagonal Gaussian Models. We use CLR for inter-cluster similarity measure and for stop criterion. When CLR for two clusters is more than the predefined threshold these two clusters are considered as candidates for merging. For each pair of candidates for merging we also calculate global similarity measure. As was described in the section 2.6 the global similarity measure between two clusters is defined as a distance between corresponding global similarity vectors. In the described algorithm each component j of the vector i is encoded by the adapted CLR between cluster i and cluster j. When CLR is more than the threshold then corresponding component has a value of 1, otherwise 2. In the described clustering algorithm we merge pairs of clusters when their CLR is more than the threshold and is maximal in comparison with all other pairs and when their global similarity distance is also more than the predefined threshold θ. After the each iteration we adapt a new cluster model. The process of clustering is stopped when no more pairs of clusters having CLR more than threshold exist or no more pairs of the clusters having global similarity more than the threshold exist.

In section 4 the experiments with ΔBIC and CLR as inter-cluster similarity measure for global similarity vectors encoding are presented and compared.

4. Experiments and evaluation

The baseline and suggested algorithm were tested on data from four different radio programs from German radio broadcasters. The performance of the speaker clustering was measured via a mapping between reference labels and system hypotheses. We use a purity-based evaluation as was described in [6]. The test dataset consist of four 30 minute files and three 60 minute files with the total duration of 5 hours. This data includes studio speech, telephone speech, street interviews, commercials and music. The results of the clustering using standard BIC agglomerative clustering algorithm are presented in Table 1. Ref. is the number of reference clusters, Sys. is the number of clusters hypothesized by the system and acp, asp and Q are the average cluster purity, average speaker purity, and the Q-measure described in the previous section.

Table 1. *Clustering results using standard BIC agglomerative clustering algorithm.*

Data	Ref.	Sys.	asp	acp	Q
1	31	16	0.91	0.62	0.75
2	25	14	**0.75**	0.68	0.71
3	22	20	0.72	0.84	0.78
4	19	19	0.83	0.93	0.88
5	6	7	**0.95**	0.69	0.81
6	18	12	**0.76**	0.75	0.75
7	15	16	**0.93**	**0.82**	0.87

In the Table 2 the results of baseline clustering algorithm are presented. In comparison with standard BIC agglomerative clustering in the baseline algorithm, both in merging criterion and in stop criterion, additional limitation based in global similarity are used.

Table 2**.** *The results of the clustering using baseline speaker clustering algorithm.*

Data	Ref.	Sys.	asp	acp	Q
1	31	28	0.91	**0.88**	**0.90**
2	25	20	0.71	**0.80**	**0.75**
3	22	20	0.72	0.84	0.78
4	19	19	0.83	0.93	0.88
5	6	13	0.84	**0.88**	**0.86**
6	18	17	0.73	**0.81**	**0.77**
7	15	29	0.89	**0.99**	**0.94**

In both Tables 1 and 2 the results which are better are shown in bold. The comparison of the results shows that using global similarity limitation in the standard BIC clustering improves cluster purity and Q-measure.

We also conducted experiments when CLR as inter-cluster distance for both merging and stop criterion is used. The results are presented in Table 3.

Table 3. *The results of the clustering using CLR for inter-cluster distance and for stop criterion.*

Data	Ref.	Sys.	asp	acp	Q
1	31	23	**0.93**	0.74	0.83
2	25	20	0.76	0.47	0.60
3	22	14	0.90	0.75	0.82
4	19	18	0.92	0.93	0.92
5	6	22	0.73	0.78	0.75
6	18	41	0.69	0.88	0.78
7	15	24	0.94	0.96	0.95

The results of speaker clustering using suggested algorithm based on CLR and global similarity constraints both for merging and stop criterion are presented in Table 4.

Table 4. *Results of the clustering using CLR and global similarity constraints both for merging and stop criterion.*

Data	Ref.	Sys.	asp	acp	Q
1	31	25	0.91	**0.75**	0.83
2	25	28	**0.84**	**0.84**	**0.84**
3	22	14	0.90	0.75	0.82
4	19	18	0.92	0.93	0.92
5	6	32	**0.77**	**0.89**	**0.83**
6	18	48	0.69	0.88	**0.79**
7	15	24	**0.97**	**0.98**	**0.97**

For the comparison of the results presented in Tables 3 and 4 the results which are better are shown in bold. The comparison shows that using global similarity constraints improves clustering performance in cluster purity, in speaker purity and in Q-measure.

In Table 5 the average Q-measure, cluster purity and speaker purity for all data and for all tested algorithms are presented.

Table 5. *The results of comparison of tested algorithms.*

	S-BIC	G-BIC	S-CLR	G-CLR
Q-measure	0.79	0.84	0.81	**0.86**
Cluster purity	0.76	**0.88**	0.79	0.86
Speaker purity	0.84	0.80	0.85	**0.86**

S-BIC denotes standard BIC agglomerative clustering. G-BIC denotes BIC based clustering with the global similarity constraints. S-CLR denotes standard agglomerative clustering when CLR is used both for inter-cluster measure and for stop criterion. G-CLR denotes agglomerative clustering when both global similarity and CLR are used for inter-cluster measure and for stop criterion.

The experimental results show that the clustering using adapted CLR as local similarity measure in standard clustering (S-CLR) outperforms standard BIC based clustering (S-BIC). Using global similarity gives better results for BIC as local inter-cluster measure (G-BIC) and also for adapted CLR as local inter-cluster measure (G-CLR).

Global similarity constraints with CLR as local inter-cluster measure (G-CLR) outperform all described algorithms in Q-measure and in speaker purity. All described results confirm the importance of global similarity information for speaker clustering.

5. Conclusions

This paper describes a speaker clustering method that is based on the standard agglomerative clustering approach extended by global similarity features. This approach exploits the hypothesis that audio segments in the same cluster should exhibit the same similarity and dissimilarity with all other segments. The proposed global similarity is parameterized by the turning parameter θ that captures a fuzzy match between globally similar vectors. Experiments on radio broadcasts show that this approach gives an improvement in performance when compared to standard local similarity based clustering. The suggested algorithm with CLR as local similarity measure and global similarity constraints outperforms the same algorithm when ΔBIC as local similarity measure and global similarity constraints are used.

In future work we would like to use a better adapted CLR as a local measure, to test other local similarity measures and other feature sets to improve clustering robustness.

6. References

[1] Zhu X., Barras C., Meignier S. and Gauvain J.-L., "Combining speaker identification and BIC for speaker diarization", Eurospeech'2005, 2005.

[2] Ben M., Betser M., Bimbot F. and Gravier G., "Speaker diarization using bottom-up clustering based on a parameter-derived distance between adapted GMMs", Proc. ICSLP'2004, 2004.

[3] Biatov K. and Larson M., "Speaker Clustering via Bayesian Information Criterion using a Global Similarity Constraint", SPECOM'2005, 2005.

[4] Chen S. and Gopalakrishnan P., "Clustering via the Bayesian Information Criterion with the applications in speech recognition", *Proc. ICASSP'98*, 1998.

[5] Reynolds D.A., Singer E., Carson B.A., O'Leary G.C., McLaughlin and Zissman M.A., "Blind Clustering of speech utterances based on speaker and language characteristics", ICSLP'98, 1998.

[6] Solomonoff, A., Mielke A., Schmidt M. and Gish H., "Clustering speakers by their voices", ICASSP'98, vol. 2, 1998, pp.757-760.