Unsupervised Speaker Clustering using Global Similarity and F_0 Features

Konstantin Biatov, Martha Larson

Fraunhofer Institute for Media Communication
Sankt Augustin, Germany
{konstantin.biatov, martha.larson}@imk.fraunhofer.de

Abstract

This paper investigates an unsupervised speaker clustering approach that exploits global similarity and also proposes extending the standard cepstal feature set used for speaker clustering with prosodic features, extracted from F_0. The global-similarity-based speaker clustering algorithm, initially proposed by the authors in [6], leverages the insight that audio segments within a single cluster are not only similar to one another, but also display the same patterns of similarities and differences with audio segments belonging to all other clusters. First, speaker clustering performance using the standard Bayesian Information Criterion (BIC) is compared to the performance achieved using a BIC-based algorithm incorporating global similarity. Then, both clustering techniques are tested using an extended feature set including F_0-derived features in addition to the standard cepstral features. The evaluation, which is performed on data recorded from German language radio, shows the clear benefits of using global information when performing clustering. It also demonstrates that in most cases F_0-features outperform the cepstral-only feature set both in standard BIC clustering and in the BIC global-similarity-based approach.

1 Introduction

Speaker clustering involves grouping utterances, or more generally audio segments, generated by one speaker into a single class that excludes audio of any other speaker. In the unsupervised case, the total number of speakers present in the audio is unknown and must be determined automatically. Unsupervised speaker clustering is an important technique used in applications that automatically structure spoken documents to make them available for browsing. It also serves as a pre-processing step for further audio analysis techniques, especially speech recognition used to indexing spoken documents so that they can be searched.

In recent years, much effort has been invested in researching speaker clustering [1], [2], [3], [4], [5]. Most speaker clustering approaches make use of model selection techniques. One of the dominant model selection criteria used is the Bayesian Information Criterion (BIC). Under most speaker clustering approaches, division of the audio into homogenous speaker segments is undertaken first, and the resulting segments are clustered into speaker groups in a separate second step. We use BIC as a baseline in the experiments described in this paper. The implementation of the baseline BIC clustering algorithm follows [1].

Under the baseline BIC, audio segments are clustered using BIC as a similarity comparison between segment pairs. This approach limits itself to exploiting local information about the segments and excludes important global information, namely the relationship of the two segments under consideration to all other segments. We propose a global-similarity-based speaker clustering approach in which information about the patterns of similarity and difference of a given segment with respect to the entire segment set is taken into account when making the decision about which cluster the segment should be assigned to. The proposal for global-similarity-based bottom up agglomerative clustering supplements the local BIC-criterion for merging with global information, as the same time providing a natural stop criterion. The preliminary results of this approach were published in [6].

It is intuitive that one of the major ways in which speakers' voices differ from one another is in pitch and in pitch fluctuation. For improved speaker recognition and speaker verification system different methods using F_0 features in combination with the cepstral features have proposed, for example in [7]. Addition of pitch features has yet to be fully exploited, however, for the task of unsupervised speaker clustering. In this paper the fusion $\log F_0$ and $\Delta \log F_0$ with the cepstral features for unsupervised speaker clustering task is investigated. We evaluate the extended feature set including the prosodic features both with respect to the baseline BIC clustering as well as to the global-similarity-based BIC clustering.

In Section 2, the pre-clustering steps, the clustering algorithm and the stopping criterion are presented. In Section 3, the basic feature set and also the additional F_0-based features are described. Section 4 presents the data used and the results of the experimental evaluation. In Section 5, conclusions are drawn.

2 Description of the clustering algorithms

The Bayesian Information Criterion (BIC) is a model selection criterion, meaning that is used to decide which model best represents a given set of data. The initial proposal for segmentation using BIC was published in [1], where it was defined in the general case as

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

where $\log L(X,M)$ denotes the likelihood of the data X under the model M, N is the number of points in the data, #(M) is the number of free parameters in the model and λ is a tuning parameter. In this section, both the baseline BIC clustering approach and the global-similarity-based BIC clustering approach are described.

2.1 Segmentation with BIC

The first step in both speaker clustering approaches is speaker segmentation using BIC. In order to estimate where speaker boundaries are located in the audio, the difference of the BIC (ΔBIC) between two models is used. ΔBIC is calculated for each potential position of a speaker boundary within a window of audio. The first model describes the data as drawn from two Gaussian distributions, corresponding to two different speakers; the second model describes the data as one Gaussian distribution corresponding to a single speaker.

If ΔBIC is negative the potential speaker boundary is hypothesized as an actual boundary by the system. If ΔBIC is positive, the audio is not considered to have a speaker boundary at

that position. In our implementation of BIC segmentation, we constrain the minimal segment duration to 1 sec. Before clustering, silence frames are eliminated from the audio using an experimentally determined minimal energy threshold.

2.2 Speech/non-speech classification

The second step is to separate audio segments containing speech from segments containing non-speech audio such as music. Such classification is helpful to prevent non-speech from being clustered as a speaker.

Three pairs of GMMs were constructed and used to classify the audio frames individually. The GMMs were trained via Expectation-Maximization (EM) algorithm using 3 hours of labeled data from German radio broadcasts. Classification is performed using a cascade of three maximum likelihood decisions. The first step separates pure speech from all other audio, the second step separates telephone speech from all other audio and the third step separates auditorium speech from all other audio. After the final step, the residual class contains music and other non-speech noise.

2.3 Gender recognition

After speech/non-speech classification has been carried out, gender classification is performed on the segments that have been identified as containing speech. Gender classification prevents female speakers from being clustered with male speakers and vice versa. Each speech frame is individually classified as either male of female speech. For gender classification one pair of GMMs was trained. In the training set for each class only voiced frames were included. The exclusion of unvoiced frames in gender classification is one of the ways in which the clustering algorithm presented here has been refined with respect to the algorithm previously presented in [6]. To make the voiced/unvoiced distinction, a decision threshold rule was used. If F_0 in the frame less than 60 Hz or more than 400 Hz this frame was considered as unvoiced, otherwise it was considered as voiced. A 1024 mixture GMM was trained for each gender using the EM algorithm and 45 minutes of German radio broadcast training data.

Each speech segment was assigned a gender by using a voting rule applied to the gender classes that had been determined for the component frames. The performance of GMM-based gender classifier on the segment level is presented in Table 1. The performance of the gender recognition was evaluated on the test set including 635 female and 1523 male segments.

	Classified as male	Classified as female
Male	**97.0%**	3.0%
Female	2.5%	**97.5%**

Table 1: Results of gender classification

2.4 Baseline BIC clustering algorithm

The clustering algorithm groups audio segments within the gender classes into speaker classes. At the start, each segment is assumed to be modeled by its own single Gaussian model, i.e. to represent its own cluster. Under the standard BIC clustering approach, ΔBIC is used to make a pair-wise comparison between audio segments, grouping them into clusters from the bottom up. Two segments are merged if ΔBIC is positive and maximal; if ΔBIC is negative, the segments are not merged. The process stops when there are no more pairs of the segments with a positive ΔBIC.

2.5 Global-similarity-based clustering algorithm

The global-similarity-based BIC clustering algorithm extends use of the BIC for speaker clustering by placing a constraint on which pairs of segments can be considered for merging. Two segments are merged only if they both demonstrate the same pattern of difference with all other segments. A fuzzy match performed on global similarity vectors is used to determine global patterns of difference. A global similarity vector encodes each segment's similarity/difference with all other segments that are being clustered. Each component *j* of the global similarity vector *i* corresponds to the ΔBIC between segment *i* and segment *j*. Refer to [6] for details concerning the calculation of global similarity vectors.

Two global similarity vectors are considered to display a similar pattern of global distance if they constitute a fuzzy match. A fuzzy match between two vectors is defined as the proportion of non-zero components of the two vectors which are either both equal to 1 or both equal to 2. Formally expressed, a fuzzy match obtains, if (2) holds.

$$l > \theta \; m \qquad (2)$$

where *l* is the number non-zero components that are equal between the two vectors, *m* is the number of non-zero components in the vectors and θ is a parameter that controls the fuzziness of the match. When θ is equal to 0 the global similarity condition does not influence the process of the clustering and the results are the same as the baseline BIC clustering.

If global similarity vectors are approximately equal in all of their non-zero components (i.e. equation (2) holds) and if, additionally, ΔBIC for the corresponding speech segments is positive and maximal, the two segments are merged. The process of clustering continues until no more pairs of speech segments remain whose global similarity vectors fulfill equation (2).

3 Description of the features

3.1 Basic cepstal features

The speaker clustering algorithm uses slightly difference cepstral features for each step. For speech/non-speech classification 12 mel-cepstral coefficients plus energy, Δmel-cepstral coefficients, Δenergy, ΔΔmel-cepstral coefficients and ΔΔenergy were used. For gender classification only 12 mel-cepstral coefficients were used. For speaker clustering 12 mel-cepstral coefficients plus energy, Δmel-cepstral coefficients and Δenergy were used.

3.2 F_0 features

To test the effects of incorporating pitch-based information, we extended the basic cepstral feature sets for the steps of gender classification and speaker clustering with the features log F_0 and Δlog F_0, the same prosodic features used in [7]. The feature sets used for the initial BIC segmentation of the audio and the speech/non-speech classification of the segments were not extended, since pitch is not obviously relevant to these tasks.

In order to extract F_0, we used the Edinburgh Speech Tools [9], which implement the super resolution pitch determination

algorithm [10]. The low pass filtering and peak tracking options were enabled. F_0 was extracted for overlapping windows of 0.032 sec in length with a step size of 0.010 sec. The minimum F_0 value extracted was 60 Hz, maximum F_0 value was 400 Hz. These values correspond to the minimum and maximum fundamental frequencies necessary to capture the range of the human voice.

4 Experiments

The performance of the speaker clustering algorithms was evaluated on a corpus of German-language audio data including four different news and interview programs. The radio broadcaster Deutsche Welle supplied 30 minute recordings of Funkjournal (I & II) and of Wiso (III & IV). The broadcaster Westdeutscher Rundfunk (WDR) supplied 60 minute recordings of Montalk (V & VI) and Der Tag (VII). In total, the test data amounts to five hours of radio data including studio speech, telephone speech, interviews, commercials, music, singing and artificial sounds.

The experimental conditions were evaluated by comparing system output with hand-generated reference labels. The system performance is reported using a purity-based evaluation described in [11]. The evaluation tables report for each program the number of speaker clusters in the reference labels (ref.) and the number of speaker clusters hypothesized by the system (sys.) as well as the average cluster purity (acp), average speaker purity (asp) and the Q-measure (calculated as geometric mean of asp and acp.)

In the first set of experiments, the baseline BIC clustering performance was compared to the global-similarity-based BIC clustering performance. In the second set of experiments, both clustering algorithms were tested with both the basic cepstral feature set and the feature set extended with the F_0-derived features. In all experiments for both sets the tuning parameter of the BIC had value 1.3.

4.1 Baseline vs. global similarity based BIC clustering

First, baseline speaker clustering was performed using BIC clustering together with the standard cepstral features. The results are presented in Table 2.

Data	ref.	sys.	asp	acp	Q
I	31	16	0.91	0.62	0.75
II	25	14	0.75	0.68	0.71
III	22	20	0.72	0.84	0.78
IV	19	19	0.83	0.93	0.88
V	6	7	0.95	0.69	0.81
VI	18	12	0.76	0.75	0.75
VII	15	16	0.93	0.82	0.87

Table 2: Clustering results using baseline BIC with the standard features

Then, the experiments for speaker clustering based on global similarity with the standard features were conducted. These results are presented in Table 3.

It can be seen that global clustering improves average speaker purity in many cases and average cluster purity across the board, resulting in an overall significant improvement in the Q-measure.

data	ref.	sys.	asp	acp	Q
I	31	28	0.91	0.88	0.90
II	25	20	0.71	0.80	0.75
III	22	20	0.72	0.84	0.78
IV	19	19	0.83	0.93	0.88
V	6	13	0.84	0.88	0.86
VI	18	17	0.73	0.81	0.77
VII	15	29	0.89	0.99	0.94

Table 3: Clustering results using BIC speaker clustering based on global similarity with the standard features

Apparently, incorporating global information has the concrete effect of making it possible for the system to raise the number of clusters it hypothesizes to be closer to the true number of clusters. This is reflected in the marked improvement of average cluster purity (acp). Two speech segments can be correctly assigned to two separate clusters if they display two different overall patterns of similarity/difference with other clusters. Local information is not sufficient to make this decision.

4.2 Clustering using the extended feature set

The results of the baseline BIC speaker clustering using the extended feature set including $\log F_0$ and $\Delta \log F_0$ are presented in Table 4.

Data	ref.	sys.	asp	Acp	Q
I	31	16	0.91	0.62	0.75
II	25	15	0.85	0.75	0.80
III	22	17	0.83	0.84	0.83
IV	19	19	0.83	0.93	0.88
V	6	8	0.88	0.8	0.84
VI	18	12	0.73	0.70	0.71
VII	15	16	0.95	0.80	0.87

Table 4: Clustering results using baseline BIC with the extended features

Comparing Table 4 to Table 2, we can observe that the extended feature set gives a better Q-measure for most programs.

Table 5 presents the results of global-cluster-based BIC speaker clustering using the extended feature set.

data	ref.	sys.	asp	acp	Q
I	31	26	0.92	0.85	0.89
II	25	22	0.71	0.83	0.77
III	22	17	0.83	0.84	0.83
IV	19	19	0.83	0.93	0.88
V	6	14	0.78	0.85	0.81
VI	18	18	0.73	0.86	0.79
VII	15	25	0.91	0.97	0.94

Table 5: Clustering results using speaker clustering based on global similarity with the extended features

Comparing Table 5 to Table 3, we can again observe that the extended feature set brought either improves or does not affect Q-measure for most programs. If we consider average speaker clearly (asp) we see clearer evidence for the benefits of incorporating F_0-derived features.

Table 6 summarizes the Q-measure, cluster purity and speaker purity for all experimental conditions averaged across all programs.

	Baseline BIC with standard features	Baseline BIC with extended features	Global BIC with standard features	Global BIC with extended features
Q-measure	0.79	0.81	0.84	0.84
Cluster purity	0.76	0.78	0.88	0.88
Speaker purity	0.84	0.85	0.8	0.82

Table 6: The results averaged over all programs: Q-measure, cluster purity and speaker purity

From Table 6 it can be clearly seen that integrating global information delivers significant improvement in BIC speaker clustering. This table also shows that the proposal put forth in this paper to extend the standard feature set with F_0-derived features shows potential for providing improvement in speaker clustering performance.

5 Conclusions

This paper presents a speaker clustering method that is based on global similarity vectors, which integrate information about the entirety of data to be clustered. These vectors enhance the local BIC criterion for merging and stopping in agglomerative bottom-up clustering. This approach is motivated by the idea that audio segments in the same cluster should exhibit the same pattern of similarity and dissimilarity with all other segments. The global-similarity-based clustering approach proposed here is parameterized by the turning parameter θ, which captures a fuzzy match between global similarity vectors. Evaluation of the proposed clustering algorithm on a data set composed of radio broadcasts shows that this approach gives an improvement in speaker clustering performance when compared to standard BIC clustering.

The paper also describes experiments with a new feature set that includes both mel-cepstral coefficients and prosodic features, namely $\log F_0$ and $\Delta \log F_0$. The results of the experiments show that prosodic extension of standard features by $\log F_0$ and $\Delta \log F_0$ gives slightly improvement in the performance of speaker clustering for both baseline clustering using BIC and for the algorithm based on global similarity.

6 Acknowledgements

Thank you to Deutsche Welle and WDR for providing us with the data used in these experiments. In particular, thank to the members of the project group of the AudioMining project for creating the reference labels.

7 References

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

[2] J.Ajmera J. and C. Wooters, "A robust speaker clustering algorithm", *Proc. ASRU'2003*, 2003.

[3] D. Moraru, S. Meignier, C. Fredouille, L. Besacier and J.-F. Bonastre, "The ELISA consortium approaches in broadcast news speaker segmentation during the NIST 2003 Rich Transcription Evaluation", *Proc. ICASSP'04*, 2004.

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

[5] W-H Tsai, S-S Cheng and H-M Wang, "Speaker clustering of speech utterance using a voice characteristic reference space," *Proc. ICSLP'04*, 2004.

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

[7] K. Iwano, T. Asami and S. Furui, "Noise-Robust Speaker Verification Using F0 Features," in *Proc. ICSLP'2004*, 2004.

[9] Edinburgh Speech Tools Library, *http://www.cstr.ed.ac.uk/projects/speech_tools/*

[10] Y. Medan, E. Yair and D. Chanzan. "Super resolution pitch determination of speech signals," *IEEE Transactions on Signal Processing*. Vol. 39:1, 1991

[11] A. Solomonoff, A. Mielke, M. Schmidt and H. Gish, "Clustering speakers by their voices," in *Proc. ICASSP'98*, 1998.