Segmentation in singer turns with the Bayesian Information Criterion

Marwa Thlithi[1], *Thomas Pellegrini*[1], *Julien Pinquier*[1], *Régine André-Obrecht*[1]

[1] Université de Toulouse-UPS-IRIT – 118, route de Narbonne – 31062 Toulouse Cedex 9 – France

{thlithi, pellegrini, pinquier, obrecht}@irit.fr

Abstract

As part of a project on indexing ethno-musicological audio recordings, segmentation in singer turns automatically appeared to be essential. In this article, we present the problem of segmentation in singer turns of musical recordings and our first experiments in this direction by exploring a method based on the Bayesian Information Criterion (BIC), which are used in numerous works in audio segmentation, to detect singer turns. The BIC penalty coefficient was shown to vary when determining its value to achieve the best performance for each recording. In order to avoid the decision about which single value is best for all the documents, we propose to combine several segmentations obtained with different values of this parameter. This method consists of taking *a posteriori* decisions on which segment boundaries are to be kept. A gain of 7.1% in terms of F-measure was obtained compared to a standard coefficient.

Index Terms: audio segmentation, singer turns, BIC criterion.

1. Introduction

An audio document can be structured automatically by many ways according to the final objective. For example, if the goal is document indexation, we shall probably ask ourselves the questions: are we in presence of speech, music, at which moments, who is speaking, who is singing, etc.

One of the first inference steps is to clip, then to label areas or segments known as "acoustically homogeneous". In automatic speech processing, it will be a question of identifying the changes of speakers or speaker turns to know who is speaking and when, in order to facilitate eventual additional processing, such as automatic transcription [1].

In the context of the DIADEMS[1] project (Description, Indexing, Access to Ethno-musicological and Sound Documents) on indexing ethno-musicological audio documents, we studied the possibility to carry out the same structuring by identifying the changes of singers (soloists and/or choirs) within our musical recordings. We use the expression "segmentation in singer turns" by analogy with the segmentation in speaker turns. In this study, singing is taken as a broad assumption, accompanied or not by instruments, in a group or in soloist.

Figure 1 illustrates the problem of segmentation in singer turns. The "ground" truth consists in the manual annotation of the singing turns, and eventual inputs / outputs of instruments. In this paper, we present a method of segmentation in singer turns, which would precede a later step of segments clustering featuring the same singer or the same group of singers.

[1] http://www.irit.fr/recherches/SAMOVA/DIADEMS/en/welcome/

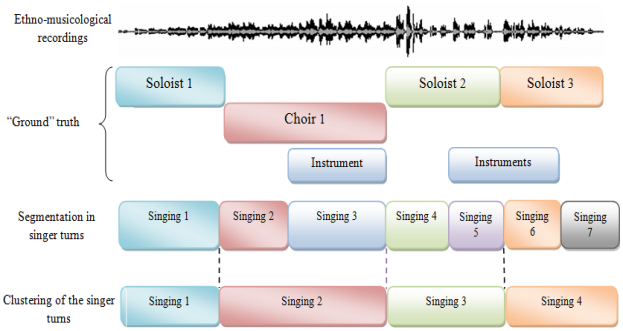

Figure 1: *Illustration of segmentation in singer turns and the step of clustering.*

For a few years our team has worked on questions related to singing, in particular on the detection of singing [2] and on the segmentation of songs in solo/choirs [3] in a musical context. This last processing is realized by means of criteria similar to those used in the detection of superposed speech. In this study, we continue to use an analogy with speech processing by adapting the method of segmentation based on the Bayesian Information Criterion (BIC), widely used in segmentation in speech turns [4]. Its application on musical recordings requires an adaptation of this criterion and its parameters. In particular, we observed that it was difficult to determine an optimal single value of the penalty coefficient for all recordings. This led us to propose a method to combine several segmentations obtained with several values of this parameter.

In section 2, we start by briefly describe the theoretical BIC Criterion and the reference algorithm, used in this study. In section 3, the application context is presented and we detail the implementation. In section 4, the problem of the adaptation of the penalty factor is discussed and the *a posteriori* consolidation with baseline results is described. Lastly, we compare the global performances obtained.

2. BIC criterion for audio segmentation

2.1. General presentation of the criterion

The Bayesian Information Criterion (BIC) is a model select criterion in a Bayesian context. For a very long time, this variant of Akaïke criterion was used in numerous application contexts [5]. These last years, BIC is at the heart of numerous works in audio segmentation [6], [7], [8], [9] and in state-of-the-art speaker diarization systems, which showed good performance.

Audio segmentation consists of dividing the audio stream into homogeneous segments by performing a hypothesis test. For each potential change point, there are two possible hypotheses: the first supposes that, on both sides of this point, the signal follows the same probabilistic model, denoted by $M_0(H0)$, the second $(H1)$ supposes that there is a change of model and it is

10.21437/Interspeech.2014-453

necessary to have two different models M_1 and M_2. In practice, the models are estimated on three analysis windows, which are used to determine if the signal is "better" represented by two distinct models or by a single model according to a threshold that is determined by an empirical method or dynamically adapted. It follows that if the analyzed signal corresponds to a sequence of *N* observations vectors (N acoustic vectors or frames) of dimension *d*, denoted by $X_0(x_1, x_2, \dots, x_N)$; a potential change point placed after the frame t induces two consecutive sub-sequences: $X_1(x_1, x_2, \dots, x_t)$ and $X_2(x_{t+1}, x_2, \dots, x_N)$. Supposing that X_0, X_1 and X_2 follow Gaussian laws given by $M_0(\mu_0, \Sigma_0)$, $M_1(\mu_1, \Sigma_1)$ and $M_2(\mu_2, \Sigma_2)$ respectively, the ΔBIC criterion at time t is given by

$$\Delta BIC(t) = R(t) - \lambda P, \quad (1)$$

where $R(t)$ is the log-likelihood ratio between the two hypothesis $(LL(H_1)/LL(H_0)$, and given by

$$R(t) = \frac{1}{2}(N \log(|\hat{\Sigma}_0|) - t \log(|\hat{\Sigma}_1|) - (N-t)\log(|\hat{\Sigma}_2|) \quad (2)$$

where $|\hat{\Sigma}_i|$ is the determinant of the matrix Σ_i, which is estimated from the sequence X_i.

P is proportional to the difference between the numbers of parameters used for each hypothesis, in the case of full covariance matrices, P has the form:

$$P = \frac{1}{2}\left(d + \frac{1}{2}d(d+1)\right)\log N \quad (3)$$

The penalty factor λ is learned so that the criterion ΔBIC is positive where the H_1 hypothesis is true, indicating a preference for two different models. Otherwise, the H_0 hypothesis is validated, indicating the preference for a single model for the window X_0.

2.2. Reference algorithm

The BIC implementation involves determining two important parameters: the size of the signal window N in which a border of segment is searched, and a penalty factor λ. As a first step, we sought to determine the value of these two parameters on a subset of our corpus described in the next section. A first version of the algorithm was based on the work of El-Khoury in which the window size was constant [10]. However, none single optimal value for all our recordings was satisfactory. We then implemented a version of the algorithm in which the size of the analysis window increases while no potential boundary is found. This method is based on studies in speech segmentation [11] and is illustrated in Figure 2.

The algorithm has two stages, involving two different temporal resolutions:

- The initial length of the analysis window is set to N_{min}. If no segment change is detected within this window, its size is increased by ΔN_{grow} until a max size N_{max} is reached. The ΔBIC values are calculated at regular intervals with sampled values of t, namely once every δ_1 observations. If no boundary is detected when N_{max} is reached, the analysis window is shifted by ΔN_{shift} observations and this step is repeated.
- If a potential boundary is detected, a window of length N_{second} is centered on this boundary and ΔBIC values are recomputed within this window with a high resolution, namely once every δ_h observations to refine the position of this boundary, with $\delta_h = \delta_1/5$.

We impose that any boundary segment cannot produces segment with a duration lower than N_{margin}, which implies that no boundary is searched between zones in $[1, N_{margin}]$ and $[N - N_{margin}, N]$.

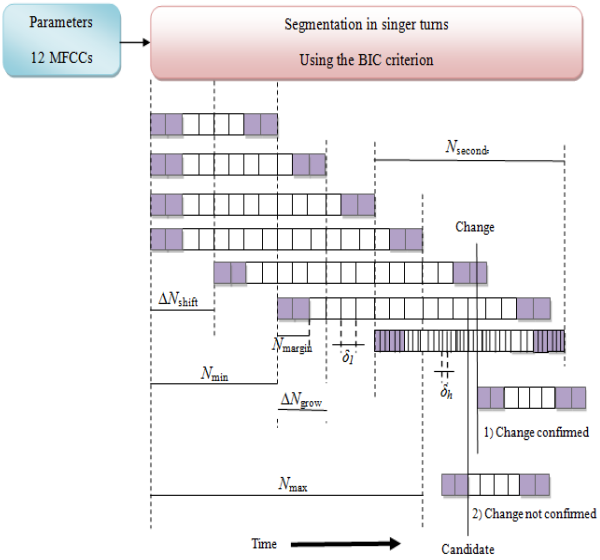

Figure 2: *Illustration of the singer turns algorithm adapted from [6].*

Segmentation is performed by a bidirectional method. Figure 3 illustrate this method .The algorithm is executed twice on each recording. A forward pass is followed by a backward pass, which acts as verification pass. F-measure is usually increased by running a backward pass.

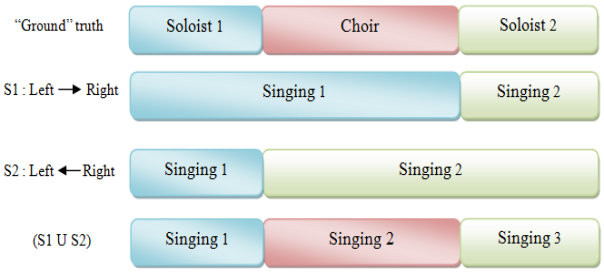

Figure 3: *Illustration of bidirectional method.*

3. Performances of the baseline system

3.1. Experiment conditions

We carried out our experiments on a corpus of recordings selected specifically for the detection of singer turns by the ethnomusicologist partners of the DIADEMS project. Examples are accessible online[1]. These records were done in several sub-Saharian countries (Congo, Gabon, Cameroon), with a variable sound quality (outdoors in general, presence of background noise and audio events other than music). They

[1] http://diadems.telemeta.org/archives/fonds/CNRSMH_DIADEMS/

mainly contain singer turns solo / choir, zones of singing voice which are alternated or overlapped with instruments or speech. The files were manually annotated in singer turns. A boundary is raised in the following situations:

- Change from a single singer (soloist) to many singers (choir) and vice versa,
- Change from a singer A to a singer B,
- Change from singing to speech and vice versa.

This corpus contains 9 music recordings of 20 minutes in total which we divided into a development corpus (DEV) and an evaluation corpus (EVAL) in the proportions 20% and 80% respectively.

3.2. Adaptation of the algorithm to singer turn segmentation

3.2.1. Acoustic features

Again by analogy between speaker turns and singer turns, we tested the acoustic parameters commonly used in speech segmentation. We used as observation vectors Mel-Frequency Cepstrum Coefficient (MFCC) with or without energy, first and second derivatives of these coefficients, Mel Spectral Coefficient (MELSPEC), PLP and RASTA-PLP. The feature extract is performing on 20 ms windows every 10 ms.

3.2.2. Adjustment of the parameters

The parameters' values of the algorithm were determined by using the development corpus:

- N_{min} : minimum size of the window, in which we search for a change boundary, is set to 0.8 s while its maximum length is N_{max} = 5 s,
- ΔN_{grow} : number of observations added to the detection window while there is no detected segment boundary and while the maximum size (N_{max}) is not reached,
- ΔN_{shift}: window shift when the maximum size is reached and no potential boundary was found, set to 0.4 s,
- N_{margin} : margin size, set to 0.7 s,
- N_{second} : size of the fine analysis window is 1.2 s.

The weak temporal resolution of the computation of the ΔBIC is $\delta_1 = 5$ (every 50 ms), while the high resolution considers all frames: $\delta_h = 1$ (10 ms of precision).

3.2.3. First Results

Performance was evaluated in terms of Recall, Precision and F-measure weighted on the files durations. The best results were obtained by limiting the observation vectors to 12 MFCC coefficients. RASTA-PLP features gave equivalent results to those obtained with MFCCs but without significant gain.

In order to define an optimal value of the penalty coefficient λ, performance was assessed by varying its value on the recordings of the development set. One standard value of λ set to 1 is used. With such a configuration, F-measure of 72.8% and 54.1% were obtained on the DEV and EVAL, respectively. These values are reported in the second line of Table 1 and in the first line of Table 2 in Section 4. Performance on EVAL corpus is much lower than on DEV corpus: an analysis of the results showed that two records, which contain very short segments of solo/choir singing, have low performance with F-measure between 30% and 40%.

4. Relevance of the penalty coefficient λ

The adjustment of the penalty factor has proved difficult; we illustrate and analyze in the first part of this section its relevance and its sensitivity to the changing of record conditions and content. In order to remedy this problem of variability, it became necessary to relax this constraint, by reasoning systematically on several values of λ and by confronting several segmentations; the new algorithm is presented in this section.

4.1. Influence of λ on the performance

The role of the coefficient λ in the BIC criterion is to penalize too complex models: in the Gaussian framework and multi models that are our framework, the larger value of λ, the more penalized the hypothesis H_1. High values of λ penalize the insertion of segment boundaries. Globally, choosing the best value of λ corresponds to finding a good compromise between precision and recall. We observed that performance varies significantly according to this factor.

However we noticed that the best value of λ varies from one record to another when we search to optimize it for each recording. For some recordings, good performance was obtained with values close to 1.2, with long obtained segments. The lowest values of λ that are close to 0.8 gave better results when shorter segments were to be found. This implies an important variability in the global performance of our system depending on λ. For purposes of comparison, we manually determined for each record in DEV and EVAL, the best value of F-measure obtained by varying the penalty coefficient and we call this artificial system, the "Oracle" system. Results are given in Table 1 and 2. On DEV and EVAL corpus, F-measure reached 89.6% and 65.2%, respectively. Comparing these results with those obtained with the standard value of λ set to 1, we find a difference in performance of about 16.8% for DEV corpus and 11.1% for EVAL corpus. This difference of 11% confirmed that fixing *a priori* the penalty coefficient to a same value for all the recordings is not optimal. This result seems to be a major difference between the singer and the speaker turn detection tasks. A single value of the penalty coefficient is usually set only once on a development set for speaker turn detection. This difference may be due to several factors. First, in the case of musical recordings, many events may be difficult singer turn detection, such as the entry or the exit of musical instruments. Second, the temporal characteristics of singing are also different from speech for which a standard syllable rate of about 4Hz is a constant. If a singer holds a long note during several seconds, a boundary may be wrongly inserted due to the fact that the corresponding segment where the long note appears is homogeneous.

4.2. Consolidated *A posteriori* Decision

In order to avoid the problem of variability and the *a priori* choice of the penalty coefficient value, we first obtain several segmentations by varying its value within the interval $[0.8 \quad 1.2]$. We tested three different steps $\{0.1, 0.05, 0.01\}$ which give 5, 9 and 41 segmentations, respectively. Second, a vote is carried out on the candidates obtained: a boundary is

validated if it was found by at least S_0 segmentations among all the segmentations. A tolerance gap of 0.5 s was used for this purpose. Therefore, we speak of a Consolidated Decision A Posteriori (DCAP) strategy. Three different values of S_0, which lie within the three intervals [1 5], [1 9] and [1 41], were determined on the DEV set. We obtained S_0 equal to 2, 3 and 15 in the three cases.

4.3. Global performance

The experimental results are presented in Table 1 and 2. The global performances obtained with a standard value of λ with a forward method only and with a bidirectional method are 71.4% and 72.8%, respectively, on the DEV corpus. A small gain of 1.4% is obtained. Therefore, the other results presented in this section are obtained by using the bidirectional method. The global performances obtained by DCAP with 5, 9 and 41 segmentations are 78.6%, 77.4% and 79.0% in F-measure, respectively, on the DEV corpus. The best global performance 61.2% was obtained with DCAP with 41 segmentations. For this reason, we used DCAP with 41 segmentations for EVAL. The gain of DCAP is around 6.2% for DEV corpus and 7.1% for EVAL corpus, compared to the results found with a single value of λ set to 1 (our baseline). This performance is still lower than the performance of the "Oracle". A possible gain of 4.0% still exists, if we consider the F-measure obtained with the "Oracle" system as the upper limit to be reached.

Table 1: *Global performance for DEV corpus.*

System	Precision	Recall	F-mesure
λ=1 - Forward only	79.7	64.7	71.4
λ=1 - Bidirectional	83.3	64.7	72.8
Oracle	92.6	86.9	89.6
DCAP - $S = 5$	74.7	82.9	78.6
DCAP - $S = 9$	75.4	79.4	77.4
DCAP - $S = 41$	82.2	75.8	79.0

Table 2: *Global performance for EVAL corpus.*

System	Precision	Recall	F-mesure
λ=1 - Bidirectional	51	57.5	54.1
Oracle	64.6	65.7	65.2
DCAP - $S = 41$	52.2	73.8	61.2

Some records on the EVAL corpus show F-measure values of about 80% and others still about 40%. Errors on these files are mostly false alarms: listening to these recordings reveals the presence of superimposed singers, rapid alternations between soloists and a choir, the presence of percussive instruments such as bells, hand claps, and background noise (speech, cries). Moreover, these recordings proved to be more difficult to manually annotate in general. In some cases of rapid alternations between singers, it is not obvious if one should insert a boundary or not. This observation would require an analysis to understand the limits of the method in terms of segmentation, and would require discussions with ethnomusicologists.

5. Conclusions and perspectives

In this article, we presented the problem of segmentation in singer turns of musical recordings, in analogy to speaker turns. The long-term objective is indexing the content of ethno-musicological musical recordings. We applied a segmentation method based on the BIC criterion. The choice of a single value for the penalty parameter of this criterion, obtained by global adjustment on a development corpus, proved unsatisfactory. Indeed, ethno-musicological recordings are very heterogeneous and have been done outdoors in presence of background noise and audio events other than music, since 1900 (old recording media). Therefore, performance varies significantly from one recording to another. It is not the case of speaker turns which recordings are homogenous because they were done in studio conditions. In order to avoid selecting a single value, we combined obtained segmentations with different values, and the final segmentation is obtained by keeping only the boundaries present in several of them. With the latter method, a gain of 7.1% in F-measure was obtained compared to a baseline system that used a single parameter value.

New versions of our segmentation approach are currently tested such as the combination of segmentations performed with other parameters (PLP, RASTA-PLP, Chromas). Larger durations of the parameter estimation window are also tested to limit the high false alarm rates observed in some cases. ΔBIC discriminative calibration and fusion of scores obtained by different segmentations is also ongoing work. We wish to confirm our results in a larger amount of data and also on different data, such as studio-quality music recordings, with more controlled acoustic conditions. In a more distant future, we will include a clustering step to label the segments.

6. Acknowledgements

This work is partly supported by a grant from the ANR (Agence Nationale de la Recherche) with reference ANR-12-CORD-0022.

7. References

[1] Anguera, X. and Bozonnet, S. and Evants, N. and Fredouille, C. and Friedland, G. and Vinyals, O. (2012). "Speaker Diarization: A Review of Recent Research". IEEE Transactions on Audio, Speech, and Language Processing, vol.20:2.

[2] Lachambre, H. and André-Obrecht, R. and Pinquier, J. (2009). "Singing voice detection in monophonic and polyphonic contexts". In Proc. European Signal Processing Conference, Glasgow, pp. 1344-1348.

[3] Le Coz, M. and André-Obrecht, R. and Pinquier, J. (2012). "Feasibility of the Detection of Choirs for Ethnomusicologic Music Indexing". In Proc. International Workshop on Content-Based Multimedia Indexing, Annecy, pp. 145-148.

[4] Zhu, X. and Barras, C. and Meignier, S. and Gauvain, J.-L. (2005). "Combining speaker identification and BIC for speaker diarization". In Proc. INTERSPEECH, Lisbon, pp. 2441-2444.

[5] Akaike, H. (1974). "A new look at the statistical model identification". IEEE Transactions on Automatic and Control, AC-19, pp. 716-723.

[6] Cettolo, M. and Vescovi, M. and Rizzi, R. (2005). "Evaluation of BIC-based algorithms for audio segmentation". In Computer Speech And Language, pp. 147-170.

[7] Siu, M.-H. and Yu, G. and Gish, H. (1991). "Segregation of speakers for speech recognition and speaker identification". In

Proc. International Conference on Acoustics, Speech, and Signal Processing, pp. 873-876.

[8] Delacourt, P. and Wellekens, C. (2000). “DISTBIC: a speaker-based segmentation for audio data indexing”. In Speech Communication, vol. 32, pp. 111-126.

[9] Chen, S. S. and Gopalakrishnan, P. S. (1998). “Speaker, Environment and Channel Change Detection and Clustering via the Bayesian Information Criterion”. In The DARPA Broadcast News Transcription and Understanding Workshop.

[10] El-Khoury, E. and Sénac, C. and Pinquier, J. (2009). “Improved speaker diarization system for meetings”. In Proc. International Conference on Acoustics, Speech, and Signal Processing, Taipei, pp. 4097-4100.

[11] André-Obrecht, R. (1998). “A new statistical approach for the Automatic Segmentation of Continuous Speech Signals”, IEEE Transactions on Audio, Speech, and Language Processing, vol. 36-1, pp. 29-40.