Filterbank coefficients selection for segmentation in singer turns

Marwa Thlithi, Julien Pinquier, Thomas Pellegrini, Régine André-Obrecht

Filterbank coefficients selection for segmentation in singer turns

Marwa Thlithi
LIUM
Université du Maine
Le Mans, France
Marwa.Thlithi@univ-lemans.fr

Julien Pinquier, Thomas Pellegrini, Régine André-Obrecht
IRIT
Université de Toulouse – UPS
Toulouse, France
{pinquier, pellegrini, obrecht}@irit.fr

***Abstract*—Audio segmentation is often the first step of audio indexing systems. It provides segments supposed to be acoustically homogeneous. In this paper, we report our recent experiments on segmenting music recordings into singer turns, by analogy with speaker turns in speech processing. We compare several acoustic features for this task: FilterBANK coefficients (FBANK), and Mel frequency cepstral coefficients (MFCC). FBANK features were shown to outperform MFCC on a "clean" singing corpus. We describe a coefficient selection method that allowed further improvement on this corpus. A 75.8% F-measure was obtained with FBANK features selected with this method, corresponding to a 30.6% absolute gain compared to MFCC. On another corpus comprised of ethno-musicological recordings, both feature types showed a similar performance of about 60%. This corpus presents an increased difficulty due to the presence of instruments overlapped with singing and to a lower recording audio quality.**

I. Introduction

A music audio document can be structured automatically by many ways according to the final objective. For example, if the goal is singing voice detection, we shall probably ask ourselves the question: are we in presence of singing or not? Some studies report methods for singing voice detection [1, 2]. In [2], the proposed method of singing voice detection consists, firstly, in distinguishing monophonies from polyphonies by using the short term mean and variance of a confidence indicator which are modeled with bivariate Weibull distributions. The parameters of these distributions are estimated with the moment method. Secondly, the detection of singing voice in a monophonic context is performed by detecting the presence of vibrato, which is an oscillation of the fundamental frequency between 4 and 8 Hz, on the pitch. In a polyphonic context, a frequency tracking on the whole spectrogram is carried out, and then searching the vibrato on each frequency tracks is performed.

In the context of music document indexing, some studies report methods for automatic detection and tracking of target singers [4] and singer identification [3]. For example, the work detailed in [3] propose an hybrid singer identifier system for automated singer recognition which uses multiple features extracted from both vocal and non-vocal music segments to improve the system's effectiveness, and a probabilistic model based on mixture models and logistic regression for singer characteristics.

In this context, we asked ourselves the questions: who is singing and when? Segmentation in singer turns consists in detecting changes of singers (soloists and/or choirs) to determine who is singing and when [5]. Figure 1 illustrates the task. The "ground" truth consists of a manual annotation in singing turns, and eventual entry/exit of instruments.

In the context of the ANR DIADEMS[1] project (Description, Indexing, Access to ethno-musicological and Sound Documents) on indexing ethno-musicological audio documents, a system of segmentation in singer turns [5] and a system of choirs detection were developed [6]. Our segmentation system was inspired by speaker turn segmentation systems. Almost of all these systems use MFCC for the parameterization stage and the Bayesian Information Criterion (BIC) for the segmentation stage [7-11].

In previous work [5], we applied segmentation in singer turns system on ethno-musicological recordings, which present a variable sound quality. In the current study, to further validate our segmentation method, we applied it on studio-quality music recordings with more controlled acoustic conditions. The songs we chose contain singing only. Thus, applying the segmentation method on these recordings, we were expecting better performance on this clean corpus but it was not the case. This led us to test other acoustic features than MFCC.

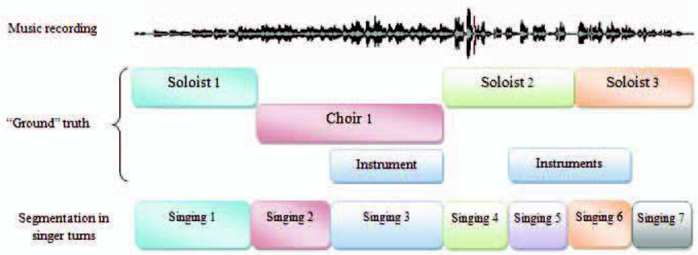

Fig. 1. Illustration of segmentation in singer turns

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

978-1-4673-8695-1/16/$31.00 © 2016 European Union

In this paper, we report performance with several types of acoustic features and we present a new parameterization method, which consists in selecting the coefficients which contain most of the spectral information.

The paper is organized as follows. In the next section, we start by briefly describing our segmentation system. In section 3, the two corpora used in this study are presented. In section 4, the acoustic features and the method implemented to select feature coefficients are detailed. Lastly, performance in terms of F-measure are compared and discussed.

II. Segmentation in singer turns

Segmentation in singer turns consists in segmenting musical recordings, and then to label areas or segments known as "acoustically homogeneous", our final goal is to obtain segments comprised of the singing of a single group of singers (soloist or choir).

Our method is based on the Bayesian Information Criterion (BIC), which is a model select criterion in a Bayesian context and a variant of Akaïke criterion. These last years, BIC is at the heart of numerous works in audio segmentation [7-11] and in state-of-the-art speaker diarization systems, which showed good performance. The application of this criterion in audio segmentation consists of considering two hypotheses test for each potential change point: the first (H_0) supposes that, on both sides of this point, the signal follows the same probabilistic model, denoted by M_0, the second (H_1) supposes that there is a change of model and it is necessary to have two different models M_1 and M_2. The ΔBIC which is the difference between the models of these two hypotheses is calculated in order to decide of the existence of a change point. At time t, the ΔBIC is given by:

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

where $R(t)$ is the log-likelihood ratio between the two hypothesis ($LL(H_1)/LL(H_0)$) and P is proportional to the difference between the numbers of parameters used for each hypothesis. 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.

The application of this criterion on music recordings required an adaptation of two parameters: the size of the signal window, in which a border of segment is searched, and the penalty factor. The adaptation of the window analysis size was solved by implementing a version of the algorithm in which the window size increases while no potential boundary is found [5, 7]. A more detailed description of the BIC and the sequence of the used algorithm can be found in [5].

For the penalty parameter, we observed that no single value was optimal for all the recordings [5]. This led us to propose the Consolidated *A Posteriori* Decision (DCAP) method which is illustrated in Figure 2. First, this method consists in combining several segmentations (M segmentations) obtained with several values of this parameter. Each new obtained segment, whose duration is lower than a certain threshold (which is the tolerance used in the evaluation of our segmentation system) is replaced with a border located at the middle. Second, a vote is carried out on the candidates obtained from all these segmentations: a boundary is validated if it was found by at least S_0 segmentations among all the segmentations. S_0 is determined on a development set.

III. Audio material

We used two different corpora with two different sound qualities. The first one is called "clean corpus" and the second one, the "DIADEMS corpus". Table I represents the duration of the development (DEV) and evaluation (EVAL) of each corpus used in this work. All the audio material comprises 16-bit 16 kHz mono files.

A. The "Clean" corpus

The recordings of this corpus were done in controlled acoustic conditions. They were taken from a few music albums which contain singing without instruments such as the song called "Mayingo" from the "Lambarena Bach to Africa" album, the vocal tracks of the "Sloop John B" song by the Beach Boys and "Marions les roses", a pop song by the "Malicorne" band. These recordings mainly contain singer turns (solo/choir), zones of singing voice which are alternated.

This corpus is comprised of 11 minutes of singing, which was divided into a development corpus (DEV) and an evaluation corpus (EVAL) in the proportions 27% and 73%, respectively. The DEV subset is composed of 10 groups of singers and it is used to set the features selection method and the parameter of our DCAP method. The EVAL subset is composed of 13 groups of signers and it is used to evaluate our system and the different feature types.

B. The DIADEMS corpus

The "DIADEMS" corpus was provided by the ethnomusicologist partners of the DIADEMS project. Examples are accessible online[2] from a platform called Telemeta.

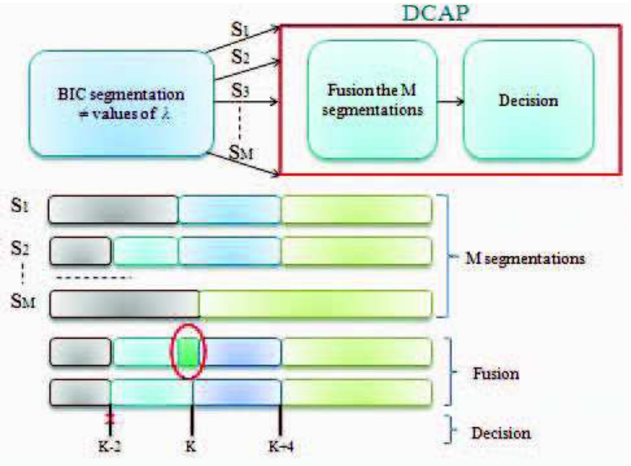

Fig. 2. Illustration of DCAP method

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

TABLE I. DURATION OF THE DEV AND EVAL OF EACH CORPUS

Corpus	DEV	EVAL
"Clean"	3 minutes	8 minutes
DIADEMS	4 minutes	16 minutes

This corpus is comprised of music recordings with a variable sound quality (outdoors in general, presence of background noise and audio events other than music).

Most recordings were done between 1940 and 1980 in several sub-Saharian countries (Congo, Gabon and Cameroon). They mainly contain singer turns solo / choir, zones of singing voice which are alternated or overlapped with instruments or speech. We divided this corpus into a DEV and an EVAL set in the proportions 20% and 80%, respectively. DEV and EVAL of this corpus are composed of 14 and 41 groups of singers, respectively.

C. Manual annotation

In order to evaluate our segmentation system, we manually annotated both corpora in terms of singer turns. A segment boundary is inserted in the following situations:

- Change from a group of *i* singers G_i (i=1…N) to another group of *j* singers G_j (j=1… N'):

$$G_i \neq G_j \; \forall i,j$$

- Change from singing to no-singing (silence, instruments, speech, etc.) and vice versa.

IV. ACOUSTIC FEATURES

A. MFCC and FBANK

Acoustic features play a major role in audio segmentation. Mel Frequency Cepstral Coefficients (MFCC) are commonly used in speech segmentation [12, 13]. The MFCC feature extraction process is based on the filterbank approach which is applied for modification of the magnitude spectrum. The modification of the magnitude spectrum consists in integrating spectral energies by a set of band-limited triangular filter weighting functions. Filters are equally spaced along the Mel scale, which is defined in "(2)". They are linearly spaced with equal bandwidth under 1000 Hz. Then, they are logarithmically spaced until 8000 Hz.

$$Mel(f)=2595log_{10}(1+f/700) \qquad (2)$$

A logarithm of the energies is finally taken to compute the FBANK coefficients. Figure 3 presents the different computing steps of FBANK coefficients.

To obtain the MFCC, a projection of these FBANK to cosine bases is performed. In this work, features are extracted on 20 ms windows with a hop size of 10 ms.

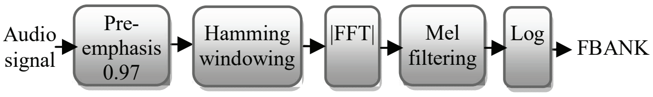

Fig. 3. Computing steps of FBANK coefficients

B. Selection of feature type

By analogy with speaker turns, we tested acoustic parameters commonly used in speech segmentation to detect singer turns on the "clean" recordings.

We tested several types of features with different configurations: MFCC with or without energy, first and second derivatives, Perceptual Linear Prediction (PLP), RASTA-PLP, FBANK and musical features Chromas. We report performances obtained with MFCC, Chromas and FBANK only since the other features gave lower performance. Table II illustrates the results of MFCC, Chromas and FBANK on the DEV subset of the "clean" corpus. Using 12 MFCC, 12 Chromas and 24 FBANK, the best F-measure values were 59.5%, 66.8% and 78.7%, respectively. Using 24 FBANK, precision increases of 1.9 and 1.5 times more than with 12 MFCC and 12 Chromas respectively. Indeed, the number of false alarms decreases sharply for all files in the corpus. Using 12 MFCC, the system produces a lot of false alarms: detection of long notes. The high number of false alarms obtained when we use Chromas is due to the fact that these parameters mainly follow the melody.

C. Selection of FBANK coefficients

We further improved the FBANK performance by limiting the coefficient range. We also found that selecting the FBANK coefficients based on their variance was helpful. The final selected FBANK coefficients are the result of the combination of these two selection steps.

1) Selection of the highest FBANK coefficient

Varying the number of FBANK coefficients, we noticed that using more than 12 or 13 FBANK coefficients degraded performance. For this reason, we decided to use the 12 first coefficients of FBANK only. This gave a performance of 87.4%. Absolute gains of 8.7%, 20.6% and 27.9% were obtained compared to the results found with 24 FBANK, 12 Chromas and 12 MFCC, respectively.

2) Variance-based selection

The large performance gain obtained with FBANK compared to MFCC and Chromas led us to consider their characteristics by examining their variance in order to find the most informative and relevant coefficients. We noticed that some FBANK coefficients have much larger variance values than others.

TABLE II. PERFORMANCE ON THE DEV "CLEAN" CORPUS

Features	Precision(%)	Recall (%)	F-measure (%)
12 MFCC	47.3	80.0	59.5
12 Chromas	59.4	76.3	66.8
24 FBANK	88.1	71.1	78.7
12 FBANK	91.5	84.0	87.4
Selected FBANK	82.6	94.3	88.1

Figures 4 and 5 show the variances of FBANK coefficients computed on two different songs of the DEV subset. By observing these histograms, we decided to develop a method to remove the FBANK coefficients with the lowest variance values. The method involves examining the first coefficient variance:

- If it is the smallest, all the coefficients starting from the second are kept. This is the case for the example presented in Figure 4 (only the first coefficient is removed).
- Otherwise all the coefficients with a variance higher than the first one are kept. This is illustrated in Figure 5 (the third and fourth coefficients are removed).

Using this method, the best performance obtained on the DEV is 88.1% as reported in the selected FBANK row of Table II. This corresponds to absolute gains of 9.4% and 0.7% compared to the results found with 24 and 12 FBANK, respectively.

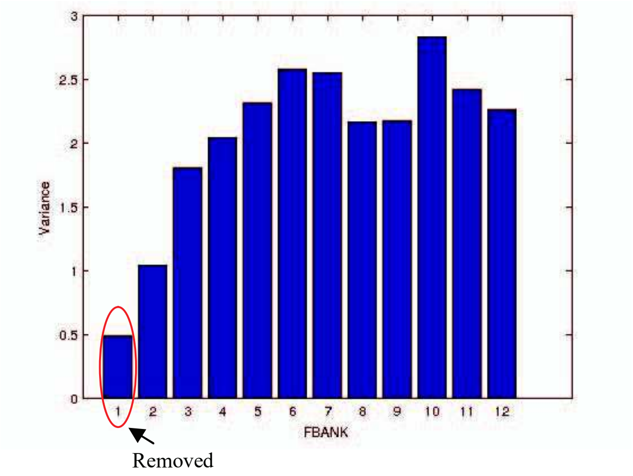

Fig. 4. Variance of the FBANK of an example of 38 seconds on the DEV "clean" corpus

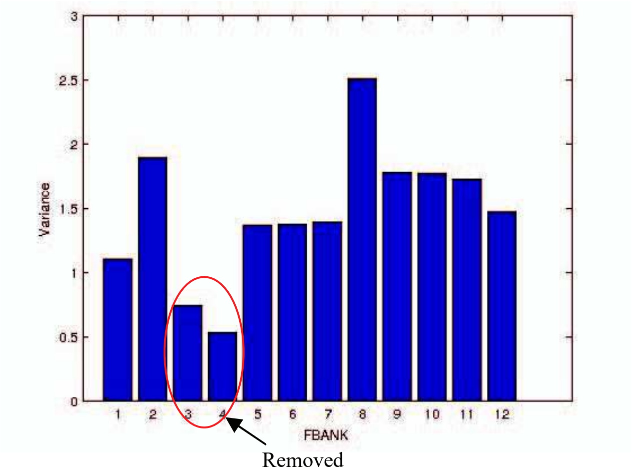

Fig. 5. Variance of the FBANK of an example of 30 seconds on the DEV "clean" corpus

The gain of the strategy in terms of F-measure is not very important compared to the results found with 12 FBANK, but we prefer this method of parameterization because it improves the recall of approximately 10%. The precision has declined somewhat but it stills compliant and can improve during the clustering step.

V. EXPERIMENTS

A. Results on the "Cleancorpus"

Table III presents the results of the DCAP method with 12 MFCC and selected FBANK coefficients on the EVAL corpus. We set the S_0 parameter of the DCAP method on the DEV corpus, which was used for the selection of features. We use for this method 161 segmentations, which are obtained by varying the penalty coefficient value λ within the interval [2.0, 10.0] with a 0.05 step. We obtained S_0 equal to 14 and 71 with 12 MFCC and with the selected FBANK coefficients, respectively. We used diagonal covariance matrices for the BIC probabilistic model and a tolerance gap of 0.5 s.

We observed that the performance obtained with the selected FBANK coefficients is always better than the one found with 12 MFCC. The absolute gain is around 30.6%. This confirms the results obtained on the DEV corpus. As reported in Table III, this large difference in performance may be due to the fact that with MFCC, the system produces 2.2 times more false alarms than with FBANK. We observed that using MFCC tends to over-segment by detecting long notes instead of singer turns.

To analyze these results more deeply, we can distinguish two cases. Figure 6 illustrates different singer turn situations encountered in our music recordings. As it can be seen, these situations can be grouped into two cases. The first one corresponds to alternates between singing and silence. The second one contains alternates between groups of singers. A group of singers can be composed of one (soloist) or several singers (choir). We were expecting to observe performance differences between detecting silence-singing or singing-singing transitions. Nevertheless, when we distinguish these 2 cases, the performance with our best system is very similar in both situations: about 75%. These results show that our system allows segmenting both alternate cases equally well.

We noted that the soloist-choir transition situations are easier to detect than soloist-soloist and choir-choir transition situations. This can be explained by the fact that passing from soloist to choir and vice versa, the number of sources increases and therefore detection becomes easier.

TABLE III. PERFORMANCE ON THE EVAL "CLEAN" CORPUS

Features	Precision	Recall	F-measure
12 MFCC	32.2	75.8	45.2
Selected FBANK	71.7	80.5	75.8

Group 1
Group 2
Silence
Group 1
Group 3

Fig. 6. Cases encountered in singer turns

B. *Results on the "DIADEMS corpus"*

In order to further validate our feature selection, we tested the approach on the "DIADEMS corpus" with the same system configuration as the one used in our previous study [5]: full covariance matrices and 41 segmentations for the DCAP method. These segmentations were obtained by varying λ within the interval [0.8, 1.2] with a step of 0.01. We used the DEV subset to determine the S_0 parameter of the DCAP method. We obtained S_0 equal to 15 and 16 with 12 MFCC and the selected FBANK coefficients, respectively. The same tolerance gap of 0.5 s was used.

Table IV presents the results found with 12 MFCC and selected FBANK coefficients on the EVAL corpus subset. MFCC slightly outperformed selected FBANK coefficients by 3.4%. Recalls obtained with these two features are similar: about 73%. Nevertheless, the FBANK precision is lower than the MFCC one. For recordings that contain percussive instruments such as bells and hand claps, FBANK tend to over-segment by detecting these claps or bells. For the recordings which contain singing only, performance is the same with both features.

One can note that the performance on this corpus is lower than the one obtained on the "clean" corpus, which may be due to the increased difficulty caused by the presence of instruments and the variable audio quality of the DIADEMS recordings. Indeed, the recordings of the "clean" corpus contain singing only, whereas the DIADEMS recordings are very heterogeneous. Some DIADEMS recordings which contain singing only, show a performance of 80% and others about 40%. Errors on these files are mostly false alarms: listening to these recordings reveals the presence of percussive instruments, background noise, superimposed singers and rapid alternates between soloists and choir. Moreover, these recordings proved to be more difficult to annotate manually in general.

In some cases of rapid alternates between singers, it is not obvious if a boundary should be inserted or not. This observation would require an analysis to understand the limits of the method in terms of acoustic features and also in terms of segmentation, for variable audio quality recordings.

VI. Conclusions and Perspectives

In this paper, we presented the problem of segmentation in singer turns and discussed the importance of the type of acoustic features used for this task.

TABLE IV. PERFORMANCE ON THE DIADEMSEVAL CORPUS

Features	Precision	Recall	F-measure
12 MFCC	52.2	73.7	61.2
Selected FBANK	47.5	73.4	57.8

We started by using MFCC as it is standard in speech processing but performance revealed poor. We have also tested musical parameters Chromas but performance stills poor. FBANK was shown to outperform MFCC and Chromas. This led us to consider the characteristics of FBANK by implementing a method to select the coefficients, which are the most informative and relevant for our task. With the selected FBANK coefficients, an absolute gain of 30.6% in F-measure was obtained compared to our baseline performance obtained with 12 MFCC. This result was achieved on a "clean" corpus, comprised of songs with singing only (no instruments). On a corpus with more heterogeneous ethno-musicological recordings, FBANK coefficients showed slightly worse performance than MFCC.

As future work, our segmentation system will be followed by a clustering step in order to build a complete singer diarization system, similar to a speaker diarization system. This clustering step is expected to decrease the number of false alarms. We will also consider the possibility to add a preprocessing step to detect singing before performing segmentation.

Acknowledgment

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

References

[1] W. Chou and L. Gu,"Robust singing detection in speech/music discriminator design," in *Proc.International Conference on Acoustics, Speech, and Signal Processing*, Salt Lake City, UT, 2001, pp. 865-868.

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

[3] J. Shen, B. Cui, J. Shepherd and K. L. Tan,"Towards efficient automated singer identification in large music databases," in *Proc. The 29th Annual International ACM SIGIR conference on Research and Development in Information Retrieval*, Washington, USA, 2006, pp. 59-66.

[4] W. H. Tsai and H. M. Wang,"'Automatic detection and tracking of target singer in multi-singer music recordings,'" in *Proc. International Conference on Acoustics*, *Speech, and Signal Processing*, Montreal, QC, Canada, 2004, vol. 4, pp. 221-224.

[5] M. Thlithi, T. Pellegrini, J. Pinquier and R. André-Obrecht,"Segmentation in singer turns with the Bayesian Information Criterion,'" in *Proc.International Speech Communication Association*, Singapore, 2014, pp. 1988-1992.

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

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

[8] M.-H. Siu, G.Yu and H. Gish, "Segregation of speakers for speech recognition and speaker identification," in *Proc.International Conference on Acoustics, Speech, and Signal Processing*, Toronto, Canada, 1991, pp. 873-876.

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

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

[11] X. Anguera Miro, “Robust speaker diarization for meetings,” *PhD Thesis*, 2006.

[12] H.P. Combrinck and E.C. Botha,“On the Mel-scaled Cepstrum,” in *Proc. The Seventh Annual South African Workshop on Pattern Recognition,* University of Pretoria, Pretoria, 1996.

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