ROBUST SIMILARITY METRICS BETWEEN AUDIO SIGNALS BASED ON ASYMMETRICAL SPECTRAL ENVELOPE MATCHING

Mathieu Lagrange, Roland Badeau, Gaël Richard

Institut Telecom, Telecom ParisTech, CNRS LTCI
46, rue Barrault, 75634 PARIS Cedex 13 - FRANCE
lagrange@telecom-paristech.fr

ABSTRACT

In this paper, a new type of metric that defines the similarity between musical audio signals is proposed. Based on the spectral flatness criterion, those metrics achieve low computational cost and low sensitivity to acoustical degradations.

Validation is performed by studying the ability of the proposed metric to determine whether two audio signals have been played by the same musical instrument. For this task, proposed metrics are shown to overcome metrics based on the comparison of standard spectral features especially when the request and the records of the database are of different acoustical properties.

***Index Terms*—** Audio Similarity, Spectral Envelope Matching, Spectral Flatness

1. INTRODUCTION

As the size of personal music collections and audio broadcasting databases increases, it is becoming more and more important to be able to automatically structure large amount of audio data. For that purpose, classification schemes allow us to classify databases according to a given ontology. As far as musical databases are concerned, much attention has been taken towards genre [1], mood or musical instrument classification [2].

Complementary to this approach, there is an increase of interest towards recommendation systems that are not based on an ontology. One can alternatively consider a recommendation system that states "show me tunes that are similar to the ones I like". In this case, the similarity between musical audio signals has to be defined [3]. For this purpose, we focus in this paper on the definition of metrics that express the similarity between audio signals and that are both efficient and robust to degradations. Those degradations may come from the acoustical environment or from pre-processing methods like source separation [4, 5]. As demonstrated in this paper, we show that considering the spectral smoothness principle allows us to propose similarity metrics that are more robust to degradations than classical ones.

The remaining of the paper is organized as follows: existing approaches and applications scenarios are presented in Section 2. The proposed approach is introduced in Section 3. The degradations considered for the evaluations are presented in Section 4. Evaluation results are presented in Section 5, followed by a conclusion and a discussion about future work in Section 6.

2. MOTIVATIONS

Expressing the similarity between audio signals is of interest for many applications, especially for Music Information Retrieval (MIR) ones. If the duration corresponds to the duration of the song, it can be useful for structuring databases of musical collections. If the duration of the compared audio signals is significantly smaller than the duration of the song, the similarity matrix built from the comparison between the elements of the song can be considered to extract meaningful information concerning the musical structure of the song [6]. We will only be concerned by the latter in this paper, leaving the definition of a similarity metric among to songs for future research.

There are many ways of defining this similarity depending on the type of information that we believe to be available within the compared signals. For musical signals, timbre or orchestration, harmony and rhythm are meaningful dimensions to consider. We concentrate in this paper on the expression of timbre understood as orchestration. Simply put, the evaluation task considered in Section 5 is: given two audio signals, compute a function of the two signals whose result is high when the same musical instrument is playing in both signals and low otherwise.

2.1. Existing Approaches

State-of-the-art methods usually split this task in two subtasks. The first one involves extracting features from the audio signals and the second one aims at computing a similarity between the two features sets. The radial basis function (RBF) is usually considered for converting a distance to a similarity:

$$r_X(X_a, X_b) = e^{-||X_a - X_b||^2} \quad (1)$$

where X_a and X_b are the normalized feature vectors respectively computed from the audio signals a and b. As we are seeking feature vectors of low dimensionality that can be efficiently computed, we consider in this paper the Mel Frequency Cepstral Coefficients (MFCC)'s [7] as the representatives of this approach. We evaluated the performance of a complete set of features usually considered for music instruments classification in conjunction with features selection algorithms to obtain feature vectors of different dimensions (160, 40, 11) as in [2] using the evaluation protocol considered in this study. The experiments concluded that the MFCC are performing best.

2.2. Application Scenarios

In this section, two application scenarios are presented where the proposed approach is meaningful. The first application scenario involves comparing audio signals of homogeneous audio quality, *i.e.* the same type of degradation (if any) is applied to the records of the database and the requests. It can be considered for music structure discovery and database indexing.

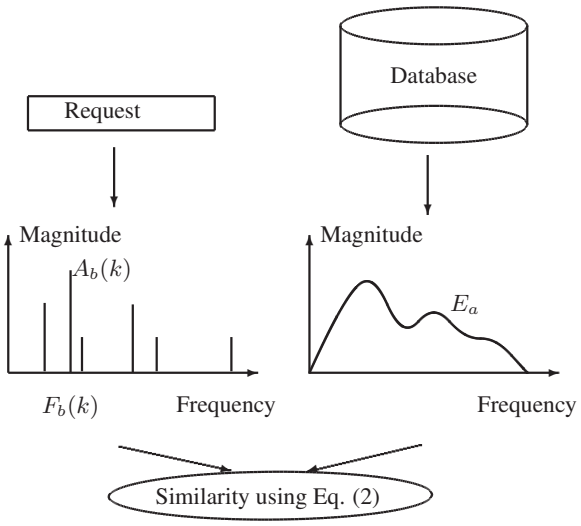

Fig. 1. *Schema of the proposed approach. The request is represented as an irregularly sampled spectral envelope (stems) whereas the records of the database are represented as a regularly sampled one (solid line).*

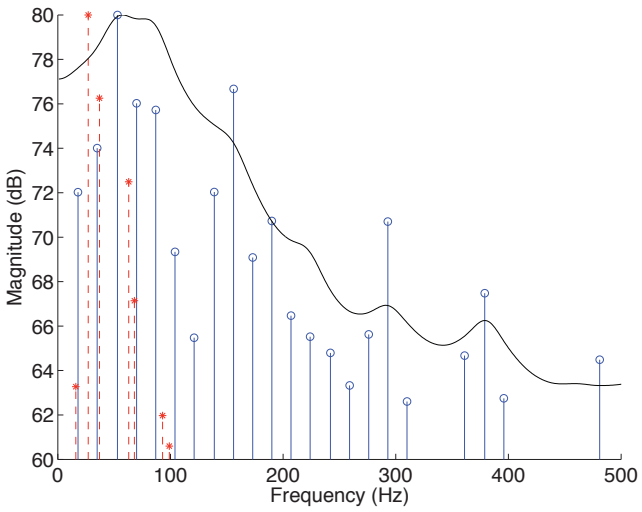

Fig. 2. *Spectral envelope (solid line) of a tone of a Double Bass and spectral peaks for a different tone of the same instrument (solid stems) and Bassoon (dashed stems).*

The second application scenario is more specific to recommendation systems as the audio quality can be different on both sides (heterogeneous). The task is then to identify which records of the database are close to the request. In this paper, the records of the database are supposed to be of good acoustical quality and the computational complexity involved in the computation of the features associated with the records is not a constraint. On contrary, on the request side, the associated features must be computed efficiently as well as the comparison with one record of the database and should be robust to degradations.

Although the two scenarios are considered in Section 5, the second one is the main motivation of the research presented in this paper and gives us the rationale of the proposed approach detailed in the next section.

3. PROPOSED APPROACH

The reference method involves the computation of the same type of features (MFCC) for both the records and the request and the comparison is done using Equation (1). Alternatively, we propose to consider an asymmetrical scheme where the features describing the records and the request are of different nature, see Figure 1.

We decided to focus on the modeling of the spectral envelope, as it is one of the major component of timbre description. On the database side, as the computation time is not a concern and the audio quality is supposed to be good, the spectral envelope is computed for a wide range of frequencies at a regular sampling rate. An attempt was made to estimate the envelope using a Linear Prediction (LP) approach but the results were not satisfying. The spectral envelope is therefore computed using a method termed "True Envelope" [8] (TE). The main interest of this iterative approach is that it results in a smooth interpolation of the observed sub-sampled spectral envelope, see Figure 2. As a consequence, the TE have good generalization properties. The order is set to 100 and the number of iterations is limited to 40.

On the contrary, the request can be noisy and the features describing the request have to be computed efficiently. Therefore, we propose to approximate the spectral envelope of the request by considering only a few peaks in the magnitude spectrum: $N = 11$ most prominent local maxima are extracted from the magnitude spectrum in order to retain the same dimensionality as the reference features (MFCC). If the request is an audio signal, the magnitude spectrum is computed using the Fast Fourier Transform (FFT). If the request is in a compressed format, it is likely that a spectral representation, usually a Modified Discrete Cosine Transform (MDCT), can be made available from the bit-stream. In this case, the spectral parameters of the peaks can be estimated directly as proposed in [9].

One can consider that the magnitude values of the peaks are the observations of the spectral envelope at an irregular sampling. Therefore, by matching those magnitudes to the magnitudes of the spectral envelope of the record at the frequencies of the peaks, one should have an estimation of whether the request and the considered record are likely to be generated by the same instrument, see Figure 2. In order to express the matching between the observations (the magnitude of the peaks) and the model (the spectral envelope of the record), we measure the spectral flatness [10] of the ratio between the observations (the peaks) with respect to the model (the spectral envelope):

$$s_{\text{ad}}(E_a, F_b, A_b) = \frac{\left(\prod_{k=0}^{N-1} |\frac{A_b(k)}{E_a(F_b(k))}|^2\right)^{\frac{1}{N}}}{\frac{1}{N}\sum_{k=0}^{N-1} |\frac{A_b(k)}{E_a(F_b(k))}|^2} \quad (2)$$

where $E_a(f)$ is the magnitude of the spectral envelope of signal a at frequency f and N is the number of peaks that describe the signal b. $F_b(k)$ and $A_b(k)$ are respectively the frequency and the amplitude of the peak k.

4. DEGRADATIONS

As the request may be of weak audio quality, we need to assess the resilience of the evaluated approaches with respect to several types

of degradations. The request signal may be degraded acoustically, for example captured using cell-phone type of devices within a noisy environment. Therefore, three types of degradations are considered: additive white noise, low-pass filtering and lossy audio encoding. For each degradation type two levels of degradations are considered:

- d_N: white noise is added to simulate channel degradation at 24 and 36 dB of signal-to-noise ratio
- d_L: low-pass filtering degradation is implemented using a FIR filter of order 100 with cutting frequency at respectively 5 kHz and 2.5 kHz
- d_E: lossy encoding degradation is done by MP3 encoding and decoding at respectively 64 and 32 Kbps.

In order to minimize the acoustical degradations coming from the environment or to focus on the musical instrument that is playing the main melody, the request may be processed by source separation algorithms as the ones proposed in [4] and [5] which respectively achieve the separation process by considering a sinusoidal modeling approach and Wiener filtering approach.

In order to evaluate the influence of those approaches over the evaluated system, we propose to simulate the degradations induced by those approaches, respectively termed d_S and d_W. As both approaches operate in the Fourier domain, the magnitude of the Short Time Fourier Transform (STFT) is first computed. Then, the effect of Wiener filtering is simulated by respectively setting 12.5 or 25 percents of the frequency bins to zero. The effect of the sinusoidal approach is simulated by respectively keeping only 25 and 12.5 percents of frequency bins who are local maxima and setting the magnitude of the remaining bins to zero. For both approaches, the degraded signal is then obtained by overlap-and-add synthesis.

5. EXPERIMENTS

In this section, several similarity metrics are evaluated by their ability to detect whether two audio signals have been played by the same instrument. We consider the SOLOS database [2] which features 505 solos recordings of mean duration 110 seconds and standard deviation 162 seconds performed by 20 different instruments for a total of 15.56 hours. Every audio signal is sampled at 44.1 kHz. The database is split into audio frames of ≈ 230 ms and the features are computed over each frame.

As the computation of the performance metrics is computationally intensive, a subset of the database is considered for a first experiment where various metrics are evaluated. The subset is built as follows. For each song available in the SOLOS database, 10 percents of the available frames are randomly selected. In a second experiment, selected metrics are next considered for exhaustive evaluation over the entire database.

For RBF-based metrics that follow Equation (1), the features are first normalized so that every feature has zero mean and unity variance. For each couple of frames, the evaluated similarity metric is computed. The *Detection Error Trade-Off* (DET) curve proposed in [11] is used to visualize the performance of the classifier corresponding to the evaluated metric at a varying classification threshold, see Figure 3. This curve represents the compromise between False Alarm Rate (FAR) and Miss Detection Rate (MDR). In order to summarize the behavior of the classifier depending on the chosen threshold, we consider two criteria. The first one called *Equal Error Rate* (EER) corresponds to a classification threshold that induces a balanced FAR and MDR represented by the crossing of the DET curve with a line that starts from $(0, 0)$ coordinates with unity slope,

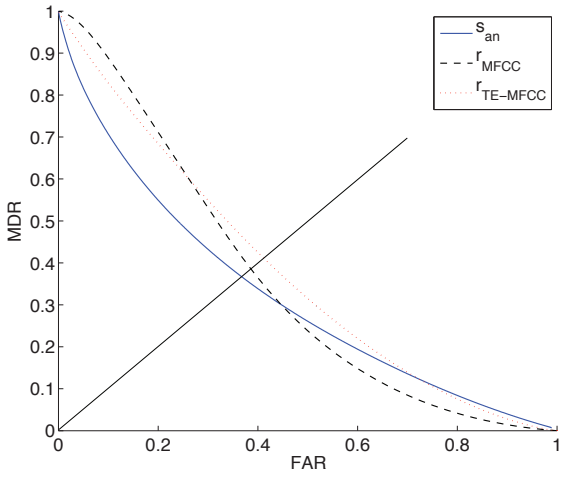

Fig. 3. *DET curves for the selected similarity metrics over the complete SOLOS database under heterogeneous mode.*

see Figure 3. The second criterion is the area under the curve. It indicates the overall behavior of the classifier. For both criteria, a lower value indicates a better performance. As a lower performance bound, a random decision would lead to an EER of ≈ 0.7 and an area of ≈ 0.5.

We consider for reference the standard RBF calculation using the MFCC and the TE-MFCC obtained by applying filter banks and the cepstrum transformation to the TE envelope. Both features consider 40 Mel sub-bands and only the first 11 Discrete Cosine Transform (DCT) coefficients (excluding the zeroth coefficient).

Several flavors of the s_{ad} defined in Equation (2) are evaluated. The metric can be made symmetrical:

$$s_{\text{sd}} = \frac{s_{\text{ad}}(E_a, F_b, A_b) + s_{\text{ad}}(E_b, F_a, A_a)}{2}. \tag{3}$$

The metric can also be normalized:

$$s_{\text{an}} = \frac{s_{\text{ad}}(E_a, F_b, A_b)}{s_{\text{ad}}(E_a, F_a, A_a) + s_{\text{ad}}(E_b, F_b, A_b) + \epsilon} \tag{4}$$

where ϵ is a small value. The symmetrical and normalized metric s_{sn} is defined similarly. As the two levels of degradations resulted in the same ranking between the evaluated metrics, only the results obtained by considering the strongest level of degradation are presented.

Those metrics are first evaluated on a reduced dataset. Table 1 shows the results in the homogeneous mode where the same degradation (if any) is applied to both the records of the database and the query. Table 2 shows the results in the heterogeneous mode, where only the request is degraded. Between the RBF-based metrics, $r_{\text{TE-MFCC}}$ is achieving better performances most of the time. As far as flatness-based metrics are concerned, s_{an} consistently provides the best performances. As a consequence, those metrics are selected together with the reference metric r_{MFCC} for the next experiment that considers the overall database.

Table 3 shows the results of the selected metrics over the entire SOLOS database. The proposed approach provides consistently better performance both in terms of EER and area under the DET curve

d	s_{sn}	s_{an}	s_{sd}	s_{ad}	r_{MFCC}	$r_{\text{TE-MFCC}}$
none	0.54	**0.52**	0.56	0.552	0.587	**0.56**
d_N	0.565	**0.563**	0.598	0.583	**0.593**	0.631
d_L	0.556	**0.541**	0.574	0.557	0.589	**0.578**
d_E	0.536	**0.516**	0.555	0.531	0.591	**0.57**
d_W	0.538	**0.518**	0.557	0.534	0.587	**0.56**
d_S	0.687	0.649	0.767	**0.641**	0.598	**0.577**

Table 1. *Performances of the evaluated similarity metrics over a subset of the SOLOS database under homogeneous degradation d. Results are expressed in terms of EER, bold values indicate best performance per type of degradation and class of similarity metric.*

d	s_{sn}	s_{an}	s_{sd}	s_{ad}	r_{MFCC}	$r_{\text{TE-MFCC}}$
none	0.54	**0.52**	0.56	0.535	0.587	**0.56**
d_N	0.567	**0.54**	0.59	0.55	**0.631**	0.789
d_L	0.62	**0.594**	0.628	0.595	**0.604**	0.615
d_E	0.553	**0.532**	0.57	0.543	0.602	**0.581**
d_W	0.553	**0.531**	0.57	0.543	0.6	**0.585**
d_S	0.578	**0.531**	0.582	0.543	**0.602**	0.617

Table 2. *Performances of the evaluated similarity metrics over a subset of the SOLOS database under heterogeneous degradation d. Results are expressed in terms of EER.*

d	s_{an}	r_{MFCC}	$r_{\text{TE-MFCC}}$
none	0.498 (**0.304**)	**0.516** (0.319)	0.525 (0.32)
d_N	**0.563 (0.36)**	0.621 (0.439)	0.851 (0.638)
d_L	**0.552 (0.351)**	0.556 (0.365)	0.589 (0.382)
d_E	**0.52 (0.325)**	0.547 (0.353)	0.576 (0.369)
d_W	**0.519 (0.324)**	0.546 (0.352)	0.582 (0.375)
d_S	**0.521 (0.328)**	0.569 (0.372)	0.622 (0.415)

Table 3. *Performances of the selected similarity metrics over the complete SOLOS database under heterogeneous degradations d. Results are expressed in terms of EER and area under DET curve (in parenthesis). Crossing of the DET curves with the straight line indicates the EER.*

by achieving a more balanced DET curve with deeper concavity, see Figure 3.

6. CONCLUSION

A new class of similarity metrics based on the asymmetrical matching of spectral envelope have been proposed. They are shown to overcome direct MFCC comparison by maintaining constraints of low dimensionality and low computational cost. This type of metrics is well suited for application scenarios where the request and the database records are of different acoustical qualities.

As future work, we plan on applying this new approach to the classification of musical instruments using the standard nearest neighbors approach (k-NN) or more dedicated approaches using classification kernels. Also, other envelope estimation methods such as the Minimum Variance Distortionless Response (MVDR) [12] as well as more realistic types of degradations such as reverberation and attenuation effects should be considered. We are also investigating the effectiveness of the proposed approach while dealing with polyphonic sounds when state of the art main melody extraction algorithms [4, 5] are considered as a front-end.

7. ACKNOWLEDGMENTS

The main author would like to thank Slim Essid for fruitful discussion. This work has been partly funded by the Quaero project within the task 6.4: "Music Search by Similarity" and the French GIP ANR DESAM under contract ANR-06-JCJC-0027-01.

8. REFERENCES

[1] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 10, no. 5, pp. 293–302, 2002.

[2] C. Joder, S. Essid, and G. Richard, "Temporal Integration for Audio Classification with Application to Musical Instrument Classification," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 17, no. 1, pp. 174–186, 2009.

[3] Jean-Julien Aucouturier and Francois Pachet, "Music similarity measures: What's the use?," in *Proceedings of ISMIR*, 2002.

[4] M. Lagrange, L. G. Martins, J. Murdoch, and G. Tzanetakis, "Normalized Cuts for Predominant Melodic Source Separation," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 16, no. 2, pp. 278–290, 2008.

[5] J.L. Durrieu, G. Richard, and B. David, "Singer melody extraction in polyphonic signals using source separation methods," in *IEEE ICASSP*, 2008, vol. 1, pp. 169–172.

[6] J. Foote, "Automatic audio segmentation using a measure of audio novelty," in *IEEE International Conference on Multimedia and Expo*, 2000, vol. 1, pp. 452–455.

[7] L. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition*, Prentice Hall Signal Processing Series, 1993.

[8] F. Villavicencio, A. Robel, and X. Rodet, "Improving LPC Spectral Envelope Extraction Of Voiced Speech By True-Envelope Estimation," in *IEEE ICASSP*, 2006, vol. 1, pp. 869–872.

[9] L. Daudet and M. Sandler, "MDCT analysis of sinusoids: exact results and applications to coding artifacts reduction ," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 12, no. 3, pp. 302–312, 2004.

[10] J. D. Johnston, "Transform coding of audio signals using perceptual noise criteria," *IEEE Journal on Selected Areas in Communications*, vol. 6, no. 2, pp. 314–323, 1988.

[11] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki, "The DET curve in Assessment of Detection Task Performance," in *Proceedings of EuroSpeech*, 1997, vol. 1, pp. 1895–1898.

[12] S. Dharanipragada, U. H. Yapanel, and B. D. Rao, "Robust feature extraction for continuous speech recognition using the mvdr spectrum estimation method," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 1, pp. 224–234, Jan. 2007.