COMBINING AUDIO SIMILARITY MEASURES USING GENERALIZATIONS OF LOGICAL CONNECTIVES

Klaas Bosteels, Etienne E. Kerre
Fuzziness and Uncertainty Modelling Research Group
Department of Applied Mathematics and Computer Science
Ghent University, Krijgslaan 281 (S9), B-9000 Gent, Belgium

ABSTRACT

A triangular operator is an aggregation operator that can be understood as a generalization of a logical connective. Although weighted means are very often used for combining audio similarity measures, theoretical considerations suggest that it might be better to use a triangular operator instead. With the MIREX 2007 submissions described in this paper, we investigate if triangular operators really are suitable for combining audio similarity measures.

1 INTRODUCTION

Generally speaking, aggregation is the task of intelligently combining several values to one value. An aggregation operator is a mathematical object that can be used for this task. In the past few years, these mathematical objects have been studied extensively by many researchers in the fuzzy set community, which has led to a wide plethora of fuzzy aggregation operators. Since audio similarity measures can be modelled by binary fuzzy relations, we can use such a fuzzy aggregation operator to combine them. In this paper, we investigate the usage of a specific class of fuzzy aggregation operators for combining audio similarity measures, namely, the class of triangular operators.

2 AUDIO SIMILARITY MEASURES

A fuzzy set A in a universe U is a $U \to [0,1]$ mapping that associates with each element $u \in U$ a degree of membership $A(u)$. The higher $A(u)$, the more u is a member of A. In particular, u belongs fully to A when $A(u) = 1$, and $A(u) = 0$ implies that u is not at all an element of A. A (binary) fuzzy relation R on U is a fuzzy set in $U \times U$. For a fuzzy relation M on the set of all possible audio fragments, we can interpret the membership degree $M(a,b)$ of a pair of audio fragments (a,b) as the degree to which a and b are similar. The higher $M(a,b)$, the more a and b are considered similar. Formally defining audio similarity measures in this way has the advantage that the extensive range of fuzzy aggregation operators available in the literature can then be used for combining different measures. In the remainder of this section, we describe two audio similarity measures that are both defined as a fuzzy relation. Combining these measures makes perfect sense because the first one is timbre-related, while the second one is related to the rhythm.

2.1 Timbre-related

Mel-frequency cepstral coefficients (MFCCs) are a short-time spectral decomposition of an audio signal that conveys the general frequency characteristics important to human hearing. In [3], Mandel and Ellis proposed a simple but effective MFCCs-based approach to determine the similarity between two audio fragments. They use a single Gaussian to model the distribution of the MFCCs computed for subsequent segments of an audio fragment, and they calculate the Kullback-Leibler divergence between two distributions to determine the similarity between the corresponding audio fragments. In [6], the distances obtained in this way were rescaled in order to improve the results when combining them with other information: $d' = -\exp(-d/450)$, with d the symmetric Kullback-Leibler divergence. Since $d' \in [-1, 0[$, we can interpret $-d'$ as a membership degree of a fuzzy relation on the set of all audio fragments. The fuzzy relation SG obtained by interpreting all rescaled distances in this way is, to some extent, related to the perceived timbre, i.e., $SG(a,b)$ is usually high when the timbres of a and b are similar.

2.2 Rhythm-related

The fluctuation pattern (FP) [6, 7] of an audio signal describes the loudness fluctuations for each frequency band. By taking the median of the FPs computed for subsequent segments of an audio fragment, we obtain a single FP that represents this fragment. The similarity between two audio fragments can then be determined by interpreting the corresponding FPs as vectors and calculating the Euclidean distance between these vectors [6]. In this paper, however, we use the cosine similarity measure instead of the Euclidean distance because the experimental observations in [1] indicate that it is more suitable for this task. Moreover, the cosine similarity measure has the advantage that it generates values that can directly be interpreted as membership degrees of a fuzzy relation on the set of all audio fragments. We use the notation FP for this fuzzy relation, which is, to some extent, related to the perceived

rhythm, i.e., $FP(a,b)$ is usually high when a and b are rhythmically similar.

3 COMBINED AUDIO SIMILARITY MEASURES

In the context of fuzzy set theory, an aggregation operator of arity $n \in \mathbb{N}$ is an increasing $[0,1]^n \to [0,1]$ mapping $\mathcal{H}$ such that $\mathcal{H}(0,0,\ldots,0) = 0$ and $\mathcal{H}(1,1,\ldots,1) = 1$. For this paper, we restrict ourselves to binary aggregation operators, i.e., aggregation operators of arity 2. The pointwise extension $\overline{\mathcal{H}}$ of a binary aggregation operator $\mathcal{H}$ can be used to combine two fuzzy sets A and B in a universe U: $\overline{\mathcal{H}}(A,B)(u) = \mathcal{H}(A(u),B(u))$, for all $u \in U$. In particular, we can use the pointwise extension of a binary aggregation operator $\mathcal{H}$ to combine SG and FP into a single audio similarity measure $\overline{\mathcal{H}}(SG,FP)$.

3.1 Using weighted means

For $\lambda \in [0,1]$, the convex linear combinations $(1-\lambda)\cdot x + \lambda \cdot y$ of two values $x, y \in [0,1]$ give rise to the following operator: $K_\lambda(x,y) = (1-\lambda)\cdot x + \lambda \cdot y$, for all $x, y \in [0,1]$. This operator is a binary aggregation operator. In fact, it is equivalent with the binary form of the well-known weighted (arithmetic) mean. Convex linear combinations were used to combine audio similarity measures in [6]. However, the author made the following remark: "There is a conceptual problem with the linear combination: A human listener does not compute a weighted sum of the similarities with respect to different aspects. In contrast, a single aspect which is similar is sufficient to consider pieces to be similar". By using a triangular operator instead of a weighted mean, we can get rid of this problem.

3.2 Using triangular operators

An associative and commutative binary aggregation operator $\mathcal{T}$ is called a triangular norm (t-norm) [4] if is satisfies $\mathcal{T}(x,1) = x$ for all $x \in [0,1]$. Since $\mathcal{T}(0,0) = \mathcal{T}(0,1) = \mathcal{T}(1,0) = 0$ and $\mathcal{T}(1,1) = 1$ hold for such a mapping $\mathcal{T}$, a t-norm can be understood as a generalization of the logical conjunction from the two-valued set $\{0,1\}$ to the whole unit interval $[0,1]$. Consequently, t-norms can be used to generalize intersection from ordinary to fuzzy sets, which explains why $\overline{\mathcal{T}}(A,B)$ is usually written as $A \cap_{\mathcal{T}} B$. The logical disjunction can be generalized to $[0,1]$ by means of a triangular conorm (t-conorm), i.e., an associative and commutative binary aggregation operator $\mathcal{S}$ that satisfies $\mathcal{S}(x,0) = x$ for all $x \in [0,1]$. Usually, the notation $A \cup_{\mathcal{S}} B$ is used instead of $\overline{\mathcal{S}}(A,B)$.

T-norms and t-conorms are collectively referred to as triangular operators (t-operators). Since the inequalities $\mathcal{T}(x,y) \leq K_\lambda(x,y)$ and $K_\lambda(x,y) \leq \mathcal{S}(x,y)$ hold for all $x, y, \lambda \in [0,1]$ when $\mathcal{T}$ is a t-norm and $\mathcal{S}$ is a t-conorm, all t-operators return either lower or higher values than weighted means. The minimum T_{M} and the maximum S_{M} are the most popular t-operators, but these operators are noninteractive [2], i.e., a modification of x or y does not necessarily imply an alteration of $T_{\mathrm{M}}(x,y)$ or $S_{\mathrm{M}}(x,y)$. We avoid this problem by using the algebraic product T_{P} and the probabilistic sum S_{P}. These interactive t-operators are defined as $T_{\mathrm{P}}(x,y) = x \cdot y$ and $S_{\mathrm{P}}(x,y) = x + y - x \cdot y$, for all $x, y \in [0,1]$.

Our second MIREX submission (BK2) implements the combined audio similarity measure $\overline{S}_{\mathrm{P}}(SG,FP)$. This measure considers two audio fragments similar when they are timbrally *or* rhythmically similar. Hence, similarity of a single aspect is sufficient to consider two fragments similar, and thus this combined measure solves the abovementioned conceptual problem. Considering two audio fragments similar when they are similar with respect to all aspects, i.e., when they are both timbrally *and* rhythmically similar, is an alternative approach that can also be regarded closer to human reasoning than computing a weighted sum. This is precisely what the combined audio similarity measure $\overline{T}_{\mathrm{P}}(SG,FP)$ does. Our first submission (BK1) is an implementation of this measure.

4 DISCUSSION OF RESULTS

The algorithms submitted to the MIREX 2007 Audio Music Similarity and Retrieval task were run on a collection of 7000 songs. For each of the 10 considered genres, 10 songs were randomly selected as queries, and the 5 most similar songs were determined for each query (after filtering out the songs by the same artist). Then, for each query, the returned results (candidates) from all submissions were grouped and evaluated by human graders. A single grader evaluated each individual set of query/candidate pairs by assigning a score on a scale from 0 to 10 (with a resolution of 0.1) to each pair. A global score was then computed for each query/algorithm pair by taking the mean of the similarity ratings associated with it. Figure 1 was obtained by applying a Friedman test [5] to these global scores.

First of all, we notice that BK2 performs significantly worse than the other submissions. This suggests that humans do not necessarily consider two songs similar when they are similar with respect to a single aspect. We should, however, also take some practical issues into account here. For instance, BK2 is very sensitive to false positives generated by one of its submeasures. When either SG or

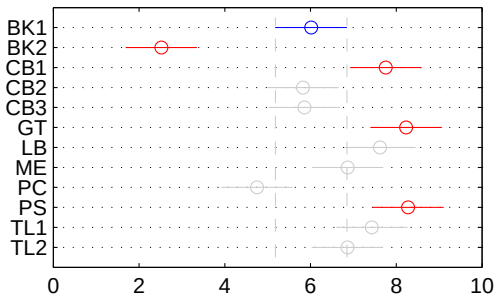

Figure 1. Evaluation results using the Friedman test. The circles mark the mean ranks, and the lines represent the significance boundaries.

FP generates a false positive, e.g., two songs are considered rhythmically similar by FP while their rhythms really are very different, then BK2 will always consider the two songs similar, independent of the value returned by the other submeasure. Hence, it is quite likely that BK2 generated a substantial amount of false positives, which is an important problem because false positives strongly influence the conducted evaluation method. Another practical issue is that SG and FP are only to some extent related to the perceived timbre and rhythm, respectively. A generalized disjunction might perform better when the measures that it combines are more closely related to a particular aspect of human perception.

Furthermore, Figure 1 indicates that only three submissions perform significantly better than BK1. Since BK1 is a very simple audio similarity measure that can still be enhanced and optimized in many ways, this is a rather good result. Hence, a generalized conjunction appears to be quite suitable for combining the considered audio similarity measures.

5 CONCLUSIONS AND FUTURE WORK

Inspired by the conceptual problem mentioned in [6], we investigated the usage of triangular operators, i.e., generalized logical connectives, for combining audio similarity measures by submitting two simple combined measures to the MIREX 2007 Audio Music Similarity and Retrieval task. Our first submission combines the audio similarity measures SG and FP using a generalized conjunction, while the second one uses a generalized disjunction to combine these measures. Although both approaches are appealing and intuitive from a conceptual point of view, the one based on a generalized conjunction proved to perform significantly better in practice.

Both our submissions can be enhanced in many ways. For instance, it might be possible to improve the performance by considering more than two submeasures, or by using weighted versions of the triangular operators [9]. Moreover, optimizations similar to the ones described in [8] could also lead to significant improvements. Since there is so much room for enhancement and optimization, the fact that only three of the submitted algorithms perform significantly better than our conjunction-based submission suggests that it might be worthwhile to further investigate the usage of generalized conjunctions for combining audio similarity measures.

6 REFERENCES

[1] K. Bosteels and E.E. Kerre. Fuzzy audio similarity measures based on spectrum histograms and fluctuation patterns. In *Proceedings of the International Conference on Multimedia and Ubiquitous Engineering*, Seoul, Korea, 2007.

[2] D. Dubois and H. Prade. *Fuzzy sets and systems: Theory and applications*. Academic Press, 1980.

[3] M. Mandel and D. Ellis. Song-level features and support vector machines for music classification. In *Proceedings of the International Symposium on Music Information Retrieval*, London, UK, 2005.

[4] R. Mesiar. Triangular norms – An overview. In B. Reusch and K.-H. Temme, editors, *Computational intelligence in theory and practice*, pages 35–54. Physica-Verlag, 2001.

[5] E. Pampalk. Audio-based music similarity and retrieval: Combining a spectral similarity model with information extracted from fluctuation patterns. Implementation submitted to the 3rd annual Music Information Retrieval Evaluation eXchange, 2006.

[6] E. Pampalk. *Computational models of music similarity and their application in music information retrieval*. PhD thesis, Vienna University of Technology, 2006.

[7] E. Pampalk, A. Rauber, and D. Merkl. Content-based organization and visualization of music archives. In *Proceedings of the ACM International Conference on Multimedia*, Juan les Pins, France, 2002.

[8] T. Pohle and D. Schnitzer. Striving for an improved audio similarity measure. Implementation submitted to the 4th annual Music Information Retrieval Evaluation eXchange, 2007.

[9] R. Yager. Weighted triangular norms using generating functions. *International Journal of Intelligent Systems*, 19:217–231, 2004.