Learning content-based metrics for music similarity

Sander Dieleman SANDER.DIELEMAN@UGENT.BE
Benjamin Schrauwen BENJAMIN.SCHRAUWEN@UGENT.BE
Ghent University, Sint-Pietersnieuwstraat 41, 9000 Gent, Belgium

Abstract

In this abstract, we propose a method to learn application-specific content-based metrics for music similarity using unsupervised feature learning and neighborhood components analysis. Multiple-timescale features extracted from music audio are embedded into a Euclidean metric space, so that the distance between songs reflects their similarity. We evaluated the method on the GTZAN and Magnatagatune datasets.

1. Introduction

Music similarity is an ambiguous notion. There are many different applications that require the similarity between two pieces of music to be measured, such as recommendation, automatic playlist generation, cover song detection and plagiarism detection. Each of these requires a different definition of similarity.

A common approach to constructing a music similarity metric is to project the audio data into a space where a simple distance measure (like Euclidean or cosine distance) can be used to assess similarity. This projection can be engineered or learnt from data. Slaney et al. (2008) embedded a set of engineered audio features into a Euclidean metric space with a number of different algorithms. Schlüter & Osendorfer (2011) learnt features unsupervisedly from audio data, and assessed similarity by computing the distance between songs in this feature space. Because the feature learning is unsupervised, the projection is not tailored to the task at hand.

We propose to combine both approaches: using learnt features instead of engineered ones should allow us to better capture musical structure. We can then embed these features into a Euclidean metric space with a projection that is learnt supervisedly.

Appearing in *Proceedings of the 29th International Conference on Machine Learning*, Edinburgh, Scotland, UK, 2012.
Copyright 2012 by the author(s)/owner(s).

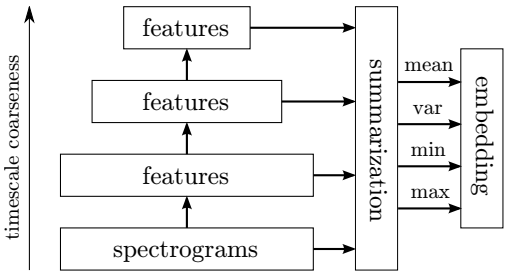

Figure 1. Schematic representation of the model.

2. Model

The proposed model consists of an unsupervised feature extraction hierarchy which operates on multiple timescales, and a linear embedding in a Euclidean metric space. Its architecture is visualized in Figure 1.

2.1. Feature extraction

To be able to extract meaningful features from audio signals, we first computed mel-spectrograms and applied PCA whitening. Short randomly sampled windows of this were used to train a sparse restricted Boltzmann machine (RBM). The learnt parameters were then used to extract features from the data convolutionally. This resulted in a first set of features which capture the local structure of the spectrograms.

Next, this first layer representation was subsampled using max-pooling. Another RBM was trained on windows of the pooled data and features were extracted in the same fashion, yielding another set of features on a coarser timescale. This process was repeated once more to get even coarser third-layer features. This multiple-stage approach has become popular recently and allows for musical structure at different timescales to be captured. We have previously used it for audio-based music classification (Dieleman et al., 2011). The features from different layers were then summarized by computing the mean, the variance, the minimum and the maximum of each feature over longer time windows of several seconds. This approach is inspired by Hamel et al. (2011). We then normalized each summarized feature to have mean zero and unit variance.

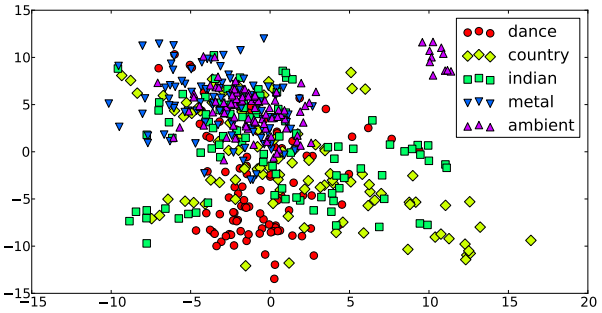

Figure 2. Visualization of a metric space learnt based on genre-related tags. Centroids for 500 excerpts are shown.

2.2. Similarity metric learning

To project the features into a low-dimensional space where similar songs are close to each other, we used Neighborhood Components Analysis (NCA) (Goldberger et al., 2004). This method is based on a version of k-nearest neighbor classification where neighbors are selected probabilistically based on their distance, and each data point inherits the class of its selected neighbor. The objective is to maximize the expected number of correctly classified points. This approach is not limited to data with class labels; it can be extended to situations where only pairs of similar data points are known, or where labels overlap. This flexibility makes it possible to learn metrics from tags or listening data. Each summarization window was used as a training or testing example.

3. Experiments and preliminary results

For our initial experiments, we used the GTZAN genre classification dataset (Tzanetakis & Cook, 2002). It contains 1,000 excerpts of songs, divided evenly over 10 genres, which makes NCA readily applicable. We also extracted three subsets from the Magnatagatune dataset (Law & von Ahn, 2009), based on instrument and genre-related tags.

We performed a random search to find good parameters for the feature extraction on GTZAN, evaluating roughly 300 trials by computing the fraction of correctly classified points on a validation set. We then learnt a metric on all datasets using these parameters.

Most parameters of the feature extraction do not seem to have much influence, including the number of units per layer, which is somewhat unexpected. Using 3 layers gives the best validation performance.

On the Magnatagatune subsets, we found that learning an instrument-based metric is difficult, indicating that our approach is suboptimal for identifying the timbres of individual instruments. A possible explanation is that the instrument each excerpt is labeled with, is not present in every summarization window, leading to mislabeled examples. Learning a genre-based metric worked well, as long as we did not use broad categories like 'rock' and 'electronic', which do not contain much similarity information. If we learn a 2D metric space, we can visualize the result (see Figure 2).

4. Future work

Evaluating the quality of the learnt metrics is challenging. We hope to use them for some of the practical applications outlined earlier, which should facilitate comparison with existing methods and make a proper evaluation possible. We would also like to evaluate our method on the Million Song Dataset (Bertin-Mahieux et al., 2011), for which a large amount of listening data was made available recently.

References

Bertin-Mahieux, Thierry, Ellis, Daniel P.W., Whitman, Brian, and Lamere, Paul. The million song dataset. In *Proceedings of the 11th International Conference on Music Information Retrieval (ISMIR)*, 2011.

Dieleman, Sander, Brakel, Philémon, and Schrauwen, Benjamin. Audio-based music classification with a pretrained convolutional network. In *Proceedings of the 11th International Conference on Music Information Retrieval (ISMIR)*, 2011.

Goldberger, Jacob, Roweis, Sam, Hinton, Geoff, and Salakhutdinov, Ruslan. Neighbourhood components analysis. In *Advances in Neural Information Processing Systems 17*, pp. 513–520, 2004.

Hamel, Philippe, Lemieux, Simon, Bengio, Yoshua, and Eck, Douglas. Temporal pooling and multiscale learning for automatic annotation and ranking of music audio. In *Proceedings of the 12th International Conference on Music Information Retrieval (ISMIR)*, 2011.

Law, Edith and von Ahn, Luis. Input-agreement: a new mechanism for collecting data using human computation games. In *Proceedings of the 27th international conference on Human factors in computing systems*, 2009.

Schlüter, Jan and Osendorfer, Christian. Music Similarity Estimation with the Mean-Covariance Restricted Boltzmann Machine. In *Proceedings of the 10th International Conference on Machine Learning and Applications (ICMLA)*, 2011.

Slaney, Malcolm, Weinberger, Kilian Q., and White, William. Learning a metric for music similarity. In *Proceedings of the 9th International Conference on Music Information Retrieval (ISMIR)*, 2008.

Tzanetakis, George and Cook, Perry. Musical genre classification of audio signals. *IEEE Transactions on Speech and Audio Processing*, 10:293–302, 2002. ISSN 1063-6676.