

ON COMPUTING MORPHOLOGICAL SIMILARITY OF AUDIO SIGNALS

Martin Gasser

Austrian Research Institute
for Artificial Intelligence
martin.gasser@ofai.at

Arthur Flexer

Austrian Research Institute
for Artificial Intelligence
arthur.flexer@ofai.at

Thomas Grill

Austrian Research Institute
for Artificial Intelligence
thomas.grill@ofai.at

ABSTRACT

Most methods to compute content-based similarity between audio samples are based on descriptors representing the spectral envelope or the texture of the audio signal only. This paper describes an approach based on (i) the extraction of spectro-temporal profiles from audio and (ii) non-linear alignment of the profiles to calculate a distance measure.

1. INTRODUCTION

Many real-world applications in the field of audio similarity would greatly benefit from an approach that explicitly models the temporal evolution of certain aspects of the signal and derives similarity values from this high-level description of the signal. Apart from being able to calculate similarities between signals that would be totally indistinguishable under a *bag-of-frames*-type [1, 2, 3] approach, such a method would also facilitate the classification of audio signals according to *morphological* descriptions.

Schaeffer [4, 5] proposed the description of *sound objects* according to a set of criteria, which include descriptions of the sound matter, the sound shape, and variation criteria. The notion of such sound-shapes appears in everyday applications e.g. with so-called *up-* and *downlifting* sounds as used in fields of sound-design and jingle production. Recently, some of those criteria have been implemented [6, 7, 8]. These works focus on classification of audio signals based on simple categories like “ascending” or “descending” based on standard audio features (e.g. loudness, spectral centroid, pitch).

The contribution of this paper is two-fold: First, we show some deficiencies of a standard descriptor (the spectral centroid [9]) for modeling spectral evolution over time. In particular, the spectral centroid is very sensitive to background noise if the noise level exceeds the level of the signal, whereas humans can easily spot spectral movement, even in the presence of strong (stationary) noise. We present a noise-robust and efficient approach to model spectro-temporal evolution based on tracking window-to-window cross-correlations of Constant-Q magnitude spectra over time.

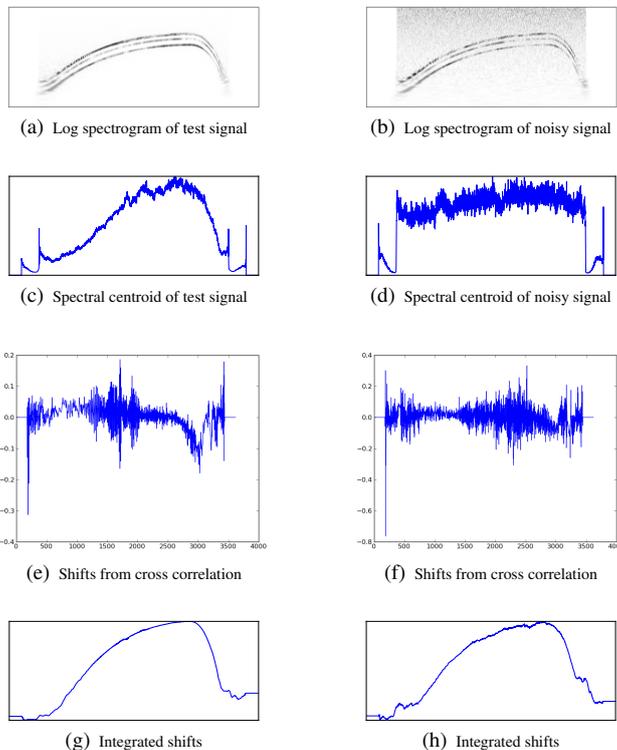


Figure 1. Profiles calculated from a clean test signal and a noisy signal ($SNR_{db} = 1$)

We also propose to calculate similarities between spectral evolution trajectories reconstructed from the aforementioned descriptor by using *Dynamic Time Warping* [16] and briefly evaluate our approach on a set of synthetic audio samples.

2. FEATURE EXTRACTION

A natural candidate for capturing the spectral evolution of a signal is the *spectral centroid* [9]. The spectral centroid is calculated as a weighted mean of the frequencies present in the signal. Thus, the lower the signal-to-noise ratio of a signal is, the less meaningful is the spectral centroid value. Figure 1(d) shows a typical result of a spectral centroid calculation on a noisy signal.

Another approach to capturing the spectral movement of signals is F0-tracking [10]. Since we want to be able to capture spectral evolutions of highly inharmonic sounds as well, F0-tracking approaches have been rejected in the first place.

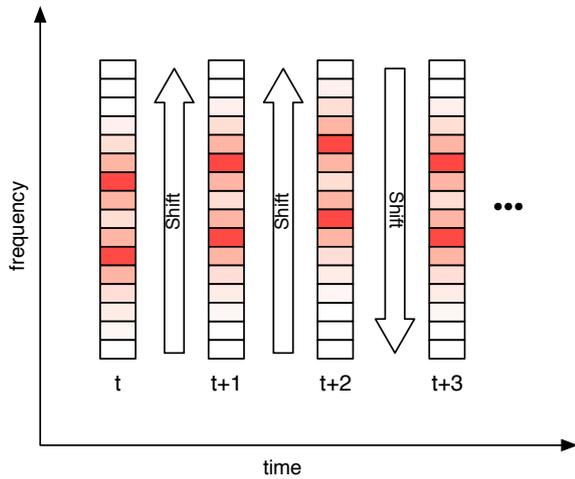


Figure 2. Shifted log-scaled magnitude spectra

Instead of trying to find instantaneous descriptors like F0 or the spectral centroid, we take a different route: Based on the observation that successive logarithmically scaled magnitude spectra calculated from a coherent spectral movement are cross-correlated (see figure 2), we cross-correlated Constant-Q magnitude spectra, from the cross-correlations we derive the hypothetical shift values to optimally align two magnitude spectra, and finally we calculate the final trajectory by cumulative summing of the shift values. Figure 1 demonstrates the weakness of spectral centroid under noisy conditions (Fig. 1(c) and Fig. 1(d)) and how we derive the final spectral profile from the shift values (Fig. 1(e)–Fig. 1(h)).

To calculate the logarithmically scaled short-time spectra of the input signal, we apply a Constant-Q transform [11] (CQT) and use the absolute value of the result.

For the Constant-Q analysis, we use a frequency resolution of 32 bins per octave, with a minimum frequency of 50 Hz and a maximum frequency corresponding to the Nyquist frequency (the sampling rate of the signals being 44.1 kHz). The hop size for the analysis is set to 1.451 ms and the time domain basis functions of the CQT are tapered with Hann windows. The window size resulting from the aforementioned parameters is 371.5 ms. We use the CQT algorithm as described by Brown and Puckette [12] and discard 98% of the coefficients in the spectral kernel SK by thresholding it with $0.01 * \max(SK)$, taking advantage of sparse matrix multiplications.

2.1 Mathematical background

According to the correlation theorem [13], the cross-correlation r of two signals can be calculated efficiently in the Fourier domain as:

$$\mathbf{G}_a = \mathcal{F}\{g_a\} \quad , \quad \mathbf{G}_b = \mathcal{F}\{g_b\}$$

$$R = \frac{\mathbf{G}_a \mathbf{G}_b^*}{|\mathbf{G}_a \mathbf{G}_b^*|} \quad , \quad r = \mathcal{F}^{-1}\{R\}$$

where \mathbf{G}_a and \mathbf{G}_b are the Fourier transforms of the sig-

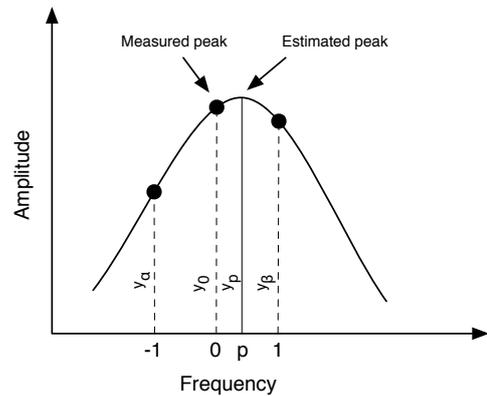


Figure 3. Fitting a parabola to the correlation function

nals g_a and g_b , respectively, and \mathbf{G}_b^* denotes the complex conjugate of \mathbf{G}_b .

By peak-picking the cross-correlation function, the shift factors can be recovered.

One problem that we encountered was that the frame-to-frame shift factors were potentially smaller than the resolution of the spectral analysis. Therefore, the resulting shift factors were much too coarse and in many cases, the correlation function always peaked at lag 0. Our solution to the problem was to use correlation interpolation [14] by fitting a parabola to the peak and the two surrounding points in the cross-correlation function (see figure 3).

$$p = \frac{1}{2} \frac{y_\alpha - y_\beta}{y_\alpha - 2y_0 + y_\beta} \quad (1)$$

By adding the value p to the index of the peak index of the cross correlation function, the final shifting factor for a pair of magnitude spectra is computed.

3. SIMILARITY COMPUTATION

We assume that similar audio signals show a similar spectral evolution in time. In order to measure the similarity between two signals, we (i) align a low dimensional representation of the spectral evolution of the signals and we (ii) derive a similarity measure from the quality of the optimal alignment.

A well-researched method that solves the problem of non-linear alignment of time series is Dynamic Time Warping (DTW) [15, 16]. DTW is a dynamic programming algorithm, that is, it calculates a matrix of partial optimal solutions to sub-problems and finds the optimal solution of the problem in a back-tracking manner. The result of DTW is a list of matchings $(a_i, b_j)_k$, where a match in time step k relates the elements a_i and b_j from the time series a and b (a_i and b_j are the elements at positions i and j in the time series a and b , respectively). By summing and normalizing the distances between matched elements, a distance measure for the time series is derived.

Figure 4 shows a simple example of DTW (the first column shows the distance matrix and the optimal path, the second column shows the resulting point-to-point align-

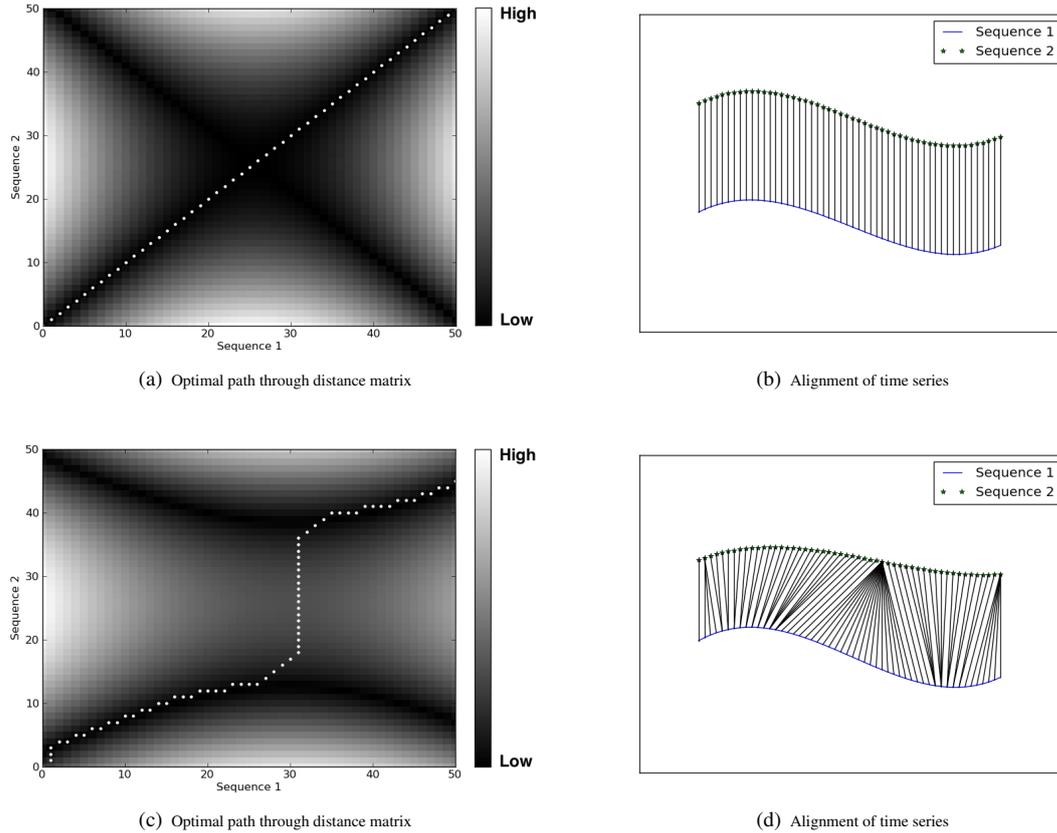


Figure 4. DTW example

ment). Whereas in figure 4(a)/figure 4(b) the two time series are identical, figure 4(c)/figure 4(d) shows two slightly different series, yielding a non-zero distance.

In order to be independent of translations along the y axis, we used a variant of DTW called Derivative Dynamic Time Warping [17], that is, instead of directly matching values of time series, we matched their first order differences.

Our algorithm for calculating morphological similarities can be summed up as follows:

1. Calculate the Constant-Q magnitude spectrum of the signals
2. For each pair of spectral frames, calculate the cross-correlation, pick the peaks and derive the shift value
3. Integrate shifts (compute cumulative sums)
4. To account for different signal lengths and to reduce computation time, resample the integrals of the spectral evolution trajectories to length n (we used $n = 100$)
5. Align the trajectories with DTW and compute the distances

4. EVALUATION

To evaluate our approach, we generated short audio samples consisting of a sinusoidal oscillator following a set of

pitch envelopes $\{e_i(x) | i \in [0, 17]\}$ modeling up-down and down-up movements.

$$t_\alpha(x) = \begin{cases} \sin(\frac{x}{\alpha} \cdot \frac{\pi}{2}), & x \leq \alpha \\ \sin(\frac{\pi}{2} + \frac{x-\alpha}{1-\alpha} \cdot \frac{\pi}{2}), & x > \alpha \end{cases} \quad (2)$$

$$e_i(x) = \begin{cases} t_{(i+1)/10}, & 0 \leq i < 9 \\ 1 - t_{(i-8)/10}, & 9 \leq i \leq 17 \end{cases} \quad (3)$$

See figure 5 for a plot of the resulting 18 functions.

As a proof of concept, we calculated DTW distances directly on the pitch envelopes without synthesizing audio and extracting the trajectories in the first place. Figure 6 demonstrates the theoretical applicability of the DTW approach to our problem: For each pair of pitch envelopes, the DTW distance is mapped to a color (dark corresponds to low, light to high values). As can be seen, the distances vary correspondingly to the choice of the α parameter in equation 2.

From the profiles in figure 5, we generated short audio samples by applying a bank of resonators with time-varying filter frequencies to an excitation signal consisting of white noise. The resonator frequencies were set to $(env * 1.1, env * 1.6, env * 2.1, env * 2.2)$, where env is the value of a prototype envelope. We also generated two sets of signals with added white noise ($SNR_{db} = 6$ and 1, respectively) and calculated their distances to the clean signals. Figures 7(a)- 7(c) show that the distance measure is

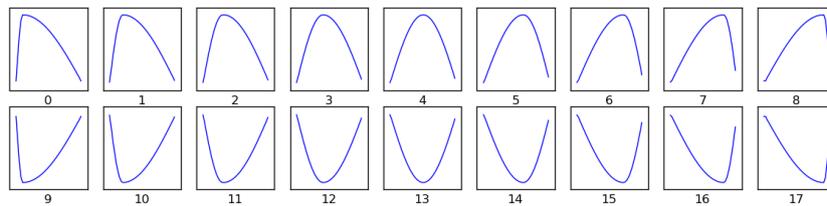


Figure 5. Pitch envelopes used for the evaluation

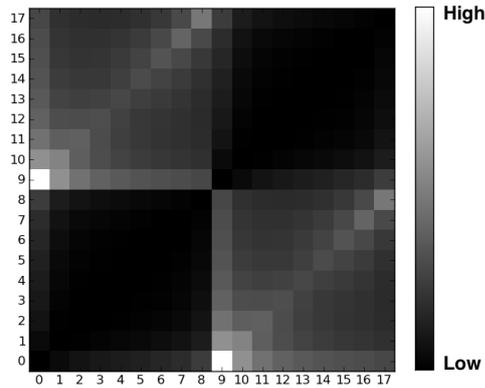


Figure 6. DTW distance matrix calculated from the 18 prototypes

very robust, even under very noisy conditions.

5. CONCLUSIONS AND FUTURE WORK

We have presented an approach to calculating similarity between audio samples based on morphological criteria. We model the spectro-temporal evolution of signals by calculating the pairwise interpolated cross-correlation between successive log-scaled short-time magnitude spectra and integrating the resulting shift values to a trajectory. To calculate similarities, trajectories are aligned with Derivative Dynamic Time Warping and a distance value is derived from the resulting DTW path.

In this paper, we have only considered spectral evolution of signals; we plan on using the same methodology to calculate similarities from loudness functions as well.

We also think that the method has potential for interactive retrieval of audio samples, since it should be straightforward for users to directly draw the desired spectral evolution trajectory. We will investigate this possibility in future research.

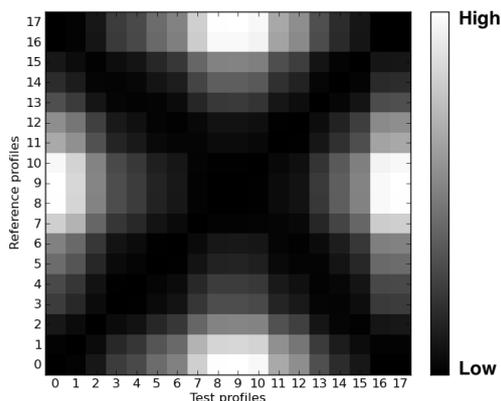
6. ACKNOWLEDGMENTS

This research is supported by the Austrian Science Fund (FWF, P21247, and Z159) and the Vienna Science and Technology Fund (WWTF, project “Audiominer”)

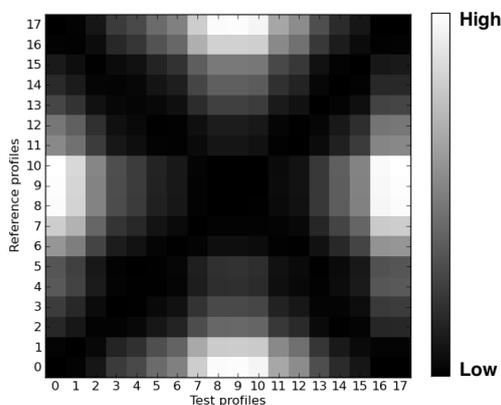
7. REFERENCES

- [1] M. Mandel and D. Ellis, “Song-level features and support vector machines for music classification,” in *Proceedings of the 6th International Conference on Music Information Retrieval*, London, United Kingdom, 2005.
- [2] E. Pampalk, “Computational models of music similarity and their application in music information retrieval,” Ph.D. dissertation, Vienna University of Technology, Vienna, Austria, March 2006. [Online]. Available: <http://www.ofai.at/~elias.pampalk/publications/pampalk06thesis.pdf>
- [3] J.-J. Aucouturier, B. Defreville, and F. Pachet, “The bag-of-frames approach to audio pattern recognition: A sufficient model for urban soundscapes but not for polyphonic music,” *Journal of the Acoustical Society of America*, vol. 122, no. 2, pp. 881–891, 2007.
- [4] P. Schaeffer, *Traité des objets musicaux*. Editions du Seuil, Paris, 1966.
- [5] M. Chion, *Guide des objets sonores, Pierre Schaeffer et la recherche musicale*. Ina-GRM/Buchet-Chastel, Paris, 1983.
- [6] J. Ricard and P. Herrera, “Morphological sound description computational model and usability evaluation,” in *AES 116th Convention*, 2004. [Online]. Available: <files/publications/AES116-jricard.pdf>
- [7] J. Bloit, N. Rasamimanana, and F. Bevilacqua, “Modeling and segmentation of audio descriptor profiles with segmental models,” *Pattern Recogn. Lett.*, vol. 31, pp. 1507–1513, September 2010. [Online]. Available: <http://dx.doi.org/10.1016/j.patrec.2009.11.003>
- [8] G. Peeters and E. Deruty, “Sound indexing using morphological description,” *IEEE Transactions on Audio, Speech & Language Processing*, vol. 18, no. 3, pp. 675–687, 2010.
- [9] G. Peeters, “A large set of audio features for sound description (similarity and classification) in the CUIDADO project,” IRCAM, Tech. Rep., 2004.
- [10] A. de Cheveigne and H. Kawahara, “Yin, a fundamental frequency estimator for speech and music,” *The Journal of the Acoustical Society of America*, vol. 111, no. 4, pp. 1917–1930, 2002. [Online]. Available: <http://link.aip.org/link/?JAS/111/1917/1>
- [11] J. C. Brown, “Calculation of a constant q spectral transform,” *Journal of the Acoustical Society of America*, vol. 89, no. 1, pp. 425–434, 1991.
- [12] J. Brown and M. Puckette, “An efficient algorithm for the calculation of a constant q transform,” *Journal of the Acoustical Society of America*, vol. 92, no. 5, pp. 2698–2701, 1992.
- [13] A. V. Oppenheim and R. W. Schaffer, *Discrete-Time Signal Processing*, 3rd ed. Upper Saddle River, NJ, USA: Prentice Hall Press, 2009.

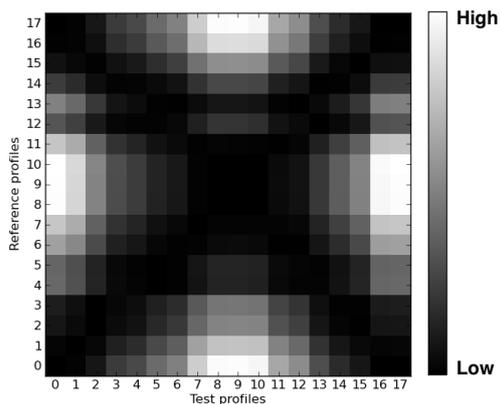
- [14] Q. Tian and M. N. Huhns, "Algorithms for subpixel registration," *Computer Vision, Graphics, and Image Processing*, vol. 35, no. 2, pp. 220–233, 1986.
- [15] L. Rabiner and B.-H. Juang, *Fundamentals of speech recognition*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 1993.
- [16] D. J. Berndt and J. Clifford, "Finding patterns in time series: a dynamic programming approach," pp. 229–248, 1996. [Online]. Available: <http://portal.acm.org/citation.cfm?id=257938.257961>
- [17] E. J. Keogh and M. J. Pazzani, "Derivative Dynamic Time Warping," in *In First SIAM International Conference on Data Mining (SDM'2001)*, 2001. [Online]. Available: <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.23.6686>



(a) Reference profiles to themselves



(b) Reference profiles to noisy version ($SNR_{db} = 6$)



(c) Reference profiles to noisy version ($SNR_{db} = 1$)

Figure 7. Distance matrices of reference profiles to noisy versions