10ème **C**ongrès **F**rançais d'**A**coustique

Lyon, 12-16 Avril 2010

Spectral similarity measure invariant to pitch shifting and amplitude scaling

Romain Hennequin[1], Roland Badeau[1], Bertrand David[1]

[1] Institut TELECOM, TELECOM ParisTech, CNRS LTCI

46, rue Barrault - 75634 Paris Cedex 13 - France

<prénom>.<nom>@telecom-paristech.fr

This paper presents a statistical model aiming at quantitatively evaluate the spectral similarity between two sounds. The measurement of similarity is a central issue in the field of Music Information Retrieval as several popular applications rely on comparisons between sound objects as for instance musical sequence seeking in a big database or automatic transcription. To take musicological considerations into account, the measure is intended to be invariant to pitch shifting and to amplitude scaling.
The main idea of the method is to compare a target spectrum to a reference spectrum using the reference to drive a statistical model, the target being an outcome of the model. The likelihood of the target spectrum is then derived in order to measure the similarity between both spectra. To be able to compare sounds of unequal intensity and pitch, the reference spectrum is made tunable in term of transposition and rescaling. Transposition and scaling parameters maximizing the likelihood are selected and the values are kept to compute the similarity measure. Thanks to a joint model, the measure is then made symmetrical. The measure is used to assess the similarity between two simple sounds (*i.e.* single isolated notes). Experimental results illustrate the usefulness of the approach: Applications of the method to classification and multipitch estimation are presented.

1 Introduction

Assessing the similarity between the spectra of two sounds has been a concern in speech and music processing, since it can be used in the context of pitch estimation [11], prosody tracking [12] or, perhaps in a broader sense, for automatic information retrieval and recognition. From these different fields, spectral distances, similarity measures or so-called divergences have appeared, relying on psychoacoustics properties or not, as for instance the log-spectral distance, the Itakura-Saito divergence [4] or the Kullback-Leibler divergence. These similarity measures are found useful for minimizing the distortion between an original signal and its coded version [5, 16] or to reduce spectral discontinuities [10]. Most of these measures are derived from Linear Prediction Coefficients [8] or AutoRegressive modeling of the signal [9]. On the other hand, a spectral distance or divergence is employed to define the cost function in the algorithm of Non-Negative-Matrix factorization (NMF, [7, 13]). Kullback-Leibler and Itakura Saito divergence are common examples of such criteria [6], when an application of the NMF to the spectrogram of an audio excerpt is targeted.

In many cases, the spectral similarity measure is designed to compare timbral characteristics of sounds. From this standpoint, desirable properties are the scale invariance (the spectra of sounds recorded at dissimilar levels should be considered as close) and pitch shift invariance (we want to compare sounds independently from their fundamental frequency). To cope with the scale invariance issue, modified versions of the log-spectral distance and the Itakura-Saito divergence have been proposed in [5]. The fact that a human listener normally exposed to music is able to recognize an instrument (or a kind of instrument) without absolute pitch reference [3] leads to many works in the field of Music Information Retrieval (MIR) to derive pitch-independent features, like for instance Mel frequency cepstral coefficients [15]. It is also not uncommon to plot the data spectrum or time-frequency representation along log-frequency axes [11, 12, 14] where the transposition becomes a simple shift. In this paper, one important goal is to define a measure which describes the similarity between two spectra up to a transposition, which allows in particular the comparison between different notes produced by a single instrument.

The main idea in this paper for comparing a target spectrum B to a reference spectrum A, is to use a scaled and transposed version of A to drive a statistical model for which B is an outcome. The likelihood of B, maximized over the different versions of A can then be used to measure the similarity $S(A, B)$ between both spectra. In the course of the method, best scaling and transposition factors are estimated and can be considered sometimes as an interesting by-product (for instance in pitch estimation problems). Since the roles of A and B in the definition above are not symmetrical the obtained measure is not symmetrical either and this leads to sometimes surprising results. An effort has then been undertaken to make the model symmetrical. A simple example of the behavior of our measure with

real audio spectra is presented and we introduce a simple application of the estimation of the best transformation parameters in multipitch estimation.

In the next section the statistical model is presented and a first similarity measure is derived. In section 2.3 a technique to obtain a symmetrical measure is described and examples on real world audio signals are given in section 3.

2 Statistical model

In this section, the model is described and subsequently a similarity measure is derived.

2.1 Spectral model

Let $S(f)$ be the Discrete Time Fourier Transform (DTFT) of a sound snapshot, f denoting the normalized frequency ($f \in]-\frac{1}{2}, \frac{1}{2}[$). As it is commonly stated (see for instance [1]), the values of the Fourier transform are assumed to be independent complex random variables with circular Gaussian probability density function (pdf):

$$S(f) \sim \mathcal{N}_{\mathbf{C}}(0, \sigma_f^2) \text{ for } f \in]-\frac{1}{2}, \frac{1}{2}[\tag{1}$$

The independence assumption is only asymptotically true, but is largely used when dealing for instance with short time spectra.

The first important idea of this work is to form a "data driven model" by expressing the preceding density with the help of a reference spectrum $S_r(f)$. This reference spectrum could be either a Fourier transform computed from real data or a synthetic pattern, which the target spectrum is intended to be compared with. This reference data is included in the model as follows:

$$\sigma_f = A\Phi_\theta(S_r, f) \tag{2}$$

where A is a scale parameter and $\Phi_\theta(S_r, f)$ is a parametric functional applied to the reference spectrum S_r; θ being a scalar or vector parameter of the functional. Then $\sigma_f = A\Phi_\theta(S_r, f)$ is obtained from a transformation Φ_θ of the whole spectrum S_r and the scaling by a factor A. This leads to a "data driven model" expressed as:

$$S(f) \sim \mathcal{N}_{\mathbf{C}}(0, A^2|\Phi_\theta(S_r, f)|^2) \text{ for } f \in]-\frac{1}{2}, \frac{1}{2}[\tag{3}$$

Thus, we obtain the pdf of S:

$$p(S(f)) = \frac{1}{\pi A^2 |\Phi_\theta(S_r, f)|^2} e^{-\frac{|S(f)|^2}{A^2|\Phi_\theta(S_r,f)|^2}} \text{ for } f \in]-\frac{1}{2}, \frac{1}{2}[\tag{4}$$

Φ_θ can describe a broad range of possible spectral modifications. The particular case of a simple transposition (an homothety on the frequency-axis of the spectrum) will be addressed in the next sections.

2.2 Asymmetric comparison of spectra

From the statistics of S, Log-Likelihood function (LL) of S with respect to A and θ is derived. The LL is large when the data fit with the model variance all along the frequency axis.

Then, we derive the best parameter θ and the best scale factor A by maximizing the likelihood of S according to the parameters. The LL of S is equal to:

$$L(S|A^2, \theta) = C - \int_{-\frac{1}{2}}^{\frac{1}{2}} \left\{ \log(A^2|\Phi_\theta(S_r, f)|^2) + \frac{|S(f)|^2}{A^2|\Phi_\theta(S_r, f)|^2} \right\} df \tag{5}$$

where C is a constant independent of A and θ. For all θ, the scale factor which maximizes the LL is given by:

$$A^2(\theta) = \int_{-\frac{1}{2}}^{\frac{1}{2}} \frac{|S(f)|^2}{|\Phi_\theta(S_r, f)|^2} df \tag{6}$$

Equation (6) is obtained by calculating the root of the derivative of the LL with respect to A^2.

By substituting $A^2(\theta)$ in the expression of the LL (5), we obtain a function $L_{A^2}(\theta)$. This function cannot generally be analytically maximized. Thus we simply use a griding strategy to find the best θ.

Particular case of the transposition: A particular case of transformation Φ_θ that will be considered in the following is the pitch transposition. The transposed version of the spectrum is obtained by stretching the frequency axis. λ is the dilatation factor (here $\theta = \{\lambda\}$). Then, the transformation is:

$$\Phi_\lambda(S_r, f) = S_r(\lambda f) \tag{7}$$

The transposition can be expressed in semitone: $12\log_2(\lambda)$.

We can notice that for $\lambda > 1$, λf can exceed $\frac{1}{2}$, thus equation (3) does not make sense since $S_r(\lambda f)$ is not defined. This issue is addressed at the end of this section.

Figure 1 shows the function $L_{A^2}(\lambda)$ for $S = S_r$: we can observe a strong peak for $12\log_2(\lambda) = 0$ semitones *i.e.* when there is no transposition. Figure 2 shows this function where S and S_r are two different notes played by a piano (respectively a $F\#3$ and a $D\#3$): we can observe a strong peak for $12\log_2(\lambda) = -3$ semitones *i.e.* for a transposition of a minor third down. We can also notice a secondary peak at $12\log_2(\lambda) = 9$ semitones which corresponds to an octave above the primary peak.

While computing the function $L_{A^2}(\lambda)$, a particular care is required to calculate the transposed spectrum $S(\lambda f)$. Actually when $\lambda \leq 1$, the transposed spectrum $S(\lambda f)$ can be calculated on the full frequency range ($f \in]-\frac{1}{2}, \frac{1}{2}[$), but when $\lambda > 1$, bins with frequencies $f > \frac{1}{2\lambda}$ need to be extrapolated. Thus, to properly compute a value of $L_{A^2}(\lambda)$ that makes sense and can be compared to the value computed from other spectra, we reduce the frequency band according to the maximum value λ_{max} of the transposition factor λ. Then we compute the LL for $f \in]-\frac{1}{2\lambda_{max}}, \frac{1}{2\lambda_{max}}[$.

2.3 Measure of similarity

From the model presented in the previous section, we can derive a measure of similarity between two spectra:

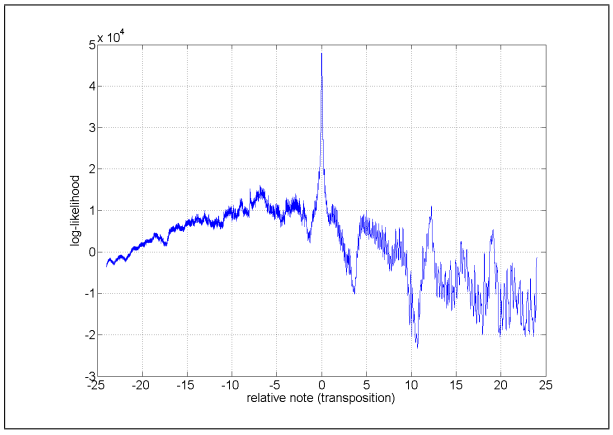

Figure 1: $L_{A^2}(\lambda)$ as a function of $12\log_2(\lambda)$ (pitch transposition in semitones) where $S = S_r$ is a recorded piano note ($D\#3$) played alone.

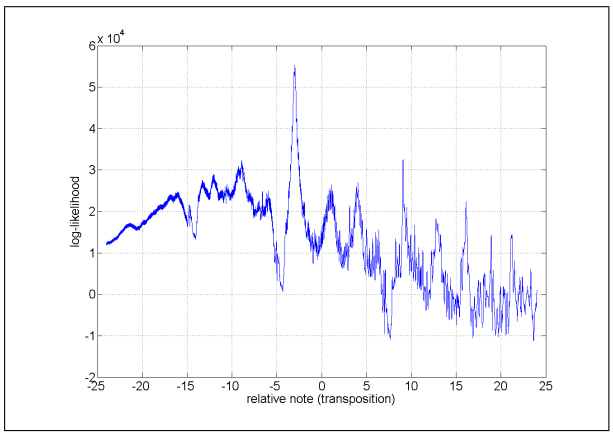

Figure 2: $L_{A^2}(\lambda)$ as a function of $12\log_2(\lambda)$ (pitch transposition in semitones) where S is a $F\#3$ piano note and S_r is a $D\#3$ recorded piano note.

the value of the maximum of LL gives a good idea of the similarity between the spectrum S and the reference spectrum S_r up to a transposition. However this value needs to be normalized in order to enable comparison between values of the measure for various spectra. Thus we add a term to the LL in equation (5) in order that its value for $S = S_r$, $A^2 = 1$ and $\lambda = 1$ is equal to zero.

Our measure is defined as:

$$\mu(S||S_r) = \max_{\lambda} \int -d(|S(f)|^2, A^2(\lambda)|S_r(\lambda f)|^2)df \quad (8)$$

where $d(x, y)$ is defined by:

$$d(x, y) = \frac{x}{y} - \log\frac{x}{y} - 1 \quad (9)$$

Values taken by the measure μ are negative or zero. The more the spectra are different (according to μ), the more the value of the measure is negative. The null value occurs only if S can be derived from S_r by a scaling and a transposition. This never happens with spectra computed from real audio data.

We can notice that the function to maximize in (8) is the Itakura-Saito divergence between $|S(f)|^2$ and $A^2(\lambda)|S_r(\lambda f)|^2$.

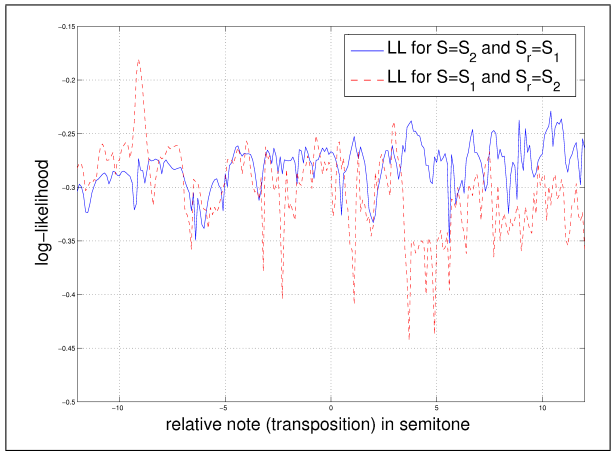

Figure 3: $L_{A^2}(\lambda)$ as a function of $12\log_2(\lambda)$ (pitch transposition in semitones) - dashed red plot: for $S = S_1$ (spectrum of 100ms frame of a piano $A\#2$) and $S_r = S_2$ (spectrum of 100ms frame of a piano $C\#2$). - plain blue plot: for $S = S_2$ and $S_r = S_1$.

2.4 Asymmetric comparison issue

In the model presented in section 2.2, there is a clear asymmetry: first, the role played by S and the role played by S_r in the definition of the model are asymmetric, which results in the asymmetry of d in equation (9), second, the way LL is computed from the data (see 2.2) clearly shows an asymmetry. This asymmetry can lead to different estimations for the best transposition and scaling parameters. An example of such a difference is shown in figure 3: the dashed red plot $L_{A^2}(\lambda)$ has a significant maximum for $12\log_2(\lambda) \approx -9$ (transposition of a major sixth), but the plain blue plot, for which the spectra were inverted in the computation of $L_{A^2}(\lambda)$, has no significant maximum at $12\log_2(\lambda) \approx 9$. Thus, $L_{A^2}(\lambda)$ does not exhibit a clear maximum in the second case. This undesirable effect results from the fact that, since we reduce the frequency band while computing $L_{A^2}(\lambda)$ (see last paragraph of section 2.2), at the location of the expected maximum, we do not compute $L_{A^2}(\lambda)$ from the same spectra when $S = S_1$ and $S_r = S_2$ as we do when $S = S_2$ and $S_r = S_1$.

This asymmetry is then an objectionable characteristic of the designed measure. Notably when spectra S and S_r are both harmonic, we expect that the location of the maximum of $L_{A^2}(\lambda)$ corresponds to the actual difference of pitch between S and S_r so that the value of the measure is meaningful.

2.5 Symmetric comparison of spectra

As shown in section 2.4, the main drawback of the method of comparison presented in 2.2 is that there is no symmetry: the function $L_{A^2}(\lambda)$ obtained while estimating $S = S_1$ with $S_r = S_2$ can differ a lot from the function obtained while estimating $S = S_2$ with $S_r = S_1$. To overcome this drawback, we propose a symmetrized version of the model. We symmetrize the previous model by simultaneously considering that S_1 is an observation of a random variable parameterized by the scaled and transposed spectrum S_2 and S_2 is an observation of a random variable parameterized by the

scaled and transposed spectrum S_1. Thus, we jointly assume:

$$\begin{cases} S_1(f) \sim \mathcal{N}_{\mathbf{C}}\left(0, A^2\,|S_2(\lambda f)|^2\right) \\ S_2(f) \sim \mathcal{N}_{\mathbf{C}}\left(0, A^{-2}\left|S_1(\frac{f}{\lambda})\right|^2\right) \end{cases} \tag{10}$$

We still suppose the values of the DTFT independent of one another and moreover that S_1 and S_2 are independent of each other. Then, we can calculate the joint LL of S_1 and S_2:

$$\begin{aligned} L^s\left(\begin{pmatrix} S_1(f) \\ S_2(f) \end{pmatrix} | A, \lambda\right) = \\ C' - \int \Big\{ \log|S_1(\frac{f}{\lambda})|^2 + \log|S_2(\lambda f)|^2 \\ + \frac{|S_1(f)|^2}{A^2|S_2(\lambda f)|^2} + \frac{A^2|S_2(f)|^2}{|S_1(\frac{f}{\lambda})|^2} \Big\}\, df \end{aligned} \tag{11}$$

By calculating the roots of the partial derivative of L^s with respect to A^2, we obtain the best scale factor for each λ:

$$A^2(\lambda) = \sqrt{\frac{\int \frac{|S_1(f)|^2}{|S_2(\lambda f)|^2} df}{\int \frac{|S_2(f)|^2}{|S_1(f/\lambda)|^2} df}} \tag{12}$$

As in section 2.2, we obtain a function ${L^s}_{A^2}(\lambda)$ of λ. As in section 2.3, we can derive a similarity measure (which will be symmetric) from this function, by normalizing it and selecting its maximum. Thus we obtain the new measure:

$$\begin{aligned} \mu_s(S_1, S_2) = \max_{\lambda} \int \big\{ d(|S_1(f)|^2, A^2(\lambda)|S_2(\lambda f)|^2) \\ + d(|S_2(f)|^2, A^{-2}(\lambda)|S_1(\frac{f}{\lambda})|^2) \Big\}\, df \end{aligned} \tag{13}$$

where $d(x, y)$ is defined by (9).

3 Preliminary experiments

In this section, we introduce two preliminary experiments: the first one is a simple experiment of sound classification and the second one provides a good representation for multipitch estimation. These two examples are quite simple, meant for illustrating and do not have the ambition to compete with state-of-the-art methods.

3.1 Comparison of spectra with μ_s

In this section, an example of comparison of spectra obtained with the symmetric measure presented in section 2.5 is outlined.

The symmetric measure μ_s defined in 2.5 is used to compare spectra of single notes played by different instruments. Results are gathered in a similarity matrix (see figure 4). The inputs of the matrix are made of five classes (which more or less correspond each to an instrument) of five elements (which are 1s long spectra of different notes). Notes played are located within the range $F3/F4$ for each instrument. Classes respectively gather spectra of:

1. single notes of an oboe
2. single notes of a trumpet
3. single notes of a piano played *forte*
4. single notes of a piano played *piano*
5. Gaussian white noise

Spectra are computed from real recorded sounds of single notes (except for white noise which is synthesized).

In figure 4, black corresponds to high values of the measure (values near 0), and white corresponds to low values (strongly negative values). We can see that according to the measure μ_s, spectra of different notes played by the same instrument are very similar. We can also notice that spectra of piano notes played at different dynamics (piano and forte) are quite similar.

To visualize more clearly, a simple Multidimensional Scaling (see [17]) in 2 dimensions was computed from the similarity matrix (using the mdscale function of Matlab). Results are shown in figure 5: spectra of each instrument seem to be accurately grouped together.

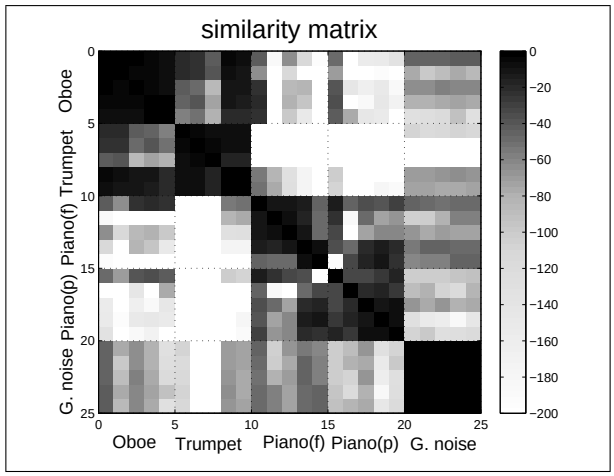

Figure 4: Similarity matrix computed with the measure μ_s.

3.2 Multipitch estimation with *Log-Likelihoodogram*

In this section, the model presented in section 2 is used for pitch estimation. This approach is based on a relative estimation of pitch in relation to a reference harmonic spectrum. The Short Time Fourier Transform (STFT) is computed from the signal to analyze. Thus we obtain for each frame a short term spectrum. The function ${L^s}_{A^2}(\lambda)$ (see 2.5) is then computed between the short term spectrum corresponding to each frame and a reference harmonic spectrum. Then, the location of significant local maxima of ${L^s}_{A^2}(\lambda)$ should give the relative pitches of the frame (relative to the pitch of the reference

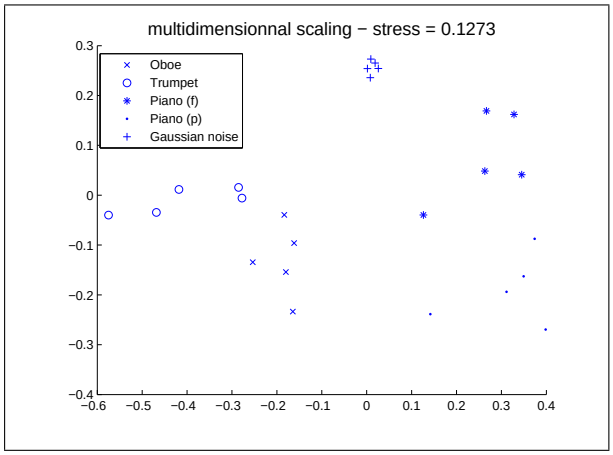

Figure 5: Multidimensional Scaling of spectra computed from the similarity matrix in figure 4 .

spectrum). We call *Log-Likelihoodogram* the representation of the function $L^s_{A^2}(\lambda)$ for each frame (which is a time/relative pitch representation).

Figure 6 shows a *loglikelihoodogram* computed from the STFT of a short synthetic piano extract with frames of 100ms and overlap of 75%. The x-axis corresponds to the number of the frame, and the y-axis corresponds to the transposition (in semitones) between the spectrum computed from the current frame and the reference spectrum. The reference harmonic spectrum was taken within the extract as the first note of the extract is played alone. We can clearly see significant local maxima (in black) at the right time/pitch position (white rectangles). However there are some other significant local maxima located in places where no notes were played: most of them correspond to fifths or octaves of effectively played notes.

This simple system is not very robust but is quite promising, and interesting since in opposition to most of the state of the art multipitch estimation systems (which are based on absolute pitch estimation), this one is based on relative pitch estimation. This is an approach more similar to the musician analyzing technique since most of the musicians do not have the absolute pitch.

4 Conclusion

In this paper we proposed a new way of comparing spectra based on a statistical model with preliminary applications. This method can be particularly useful to compare harmonic spectra of different pitches and different global amplitudes, since it allows a scaling and a transposition. We saw that the measure seems relevant to compare single note spectra of various instruments and that the computation of the log-likelihood can give a good representation for pitch estimation.

Future improvements can include a better modeling of the transposition than a simple dilation of the frequency axis. Moreover the statistical model of spectra is quite simple and could be improved to better fit harmonic spectra (it could include sinusoidal modeling).

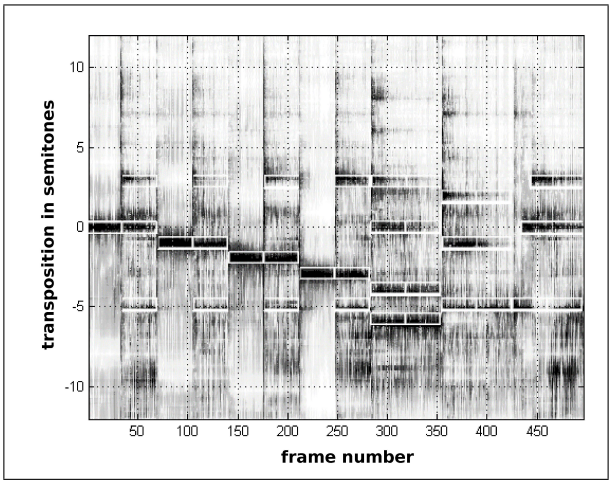

Figure 6: *Log-Likelihoodogram* obtained from a short piano extract. The reference short term spectrum was taken from the first note of the extract (which is played alone).

5 Aknowledgement

The research leading to this paper was supported by the Quaero Programme, funded by OSEO, French State agency for innovation.

References

[1] Y. Ephraim, D. Malah, "Speech Enhancement Using a Minimum Mean Square Error Short-Time Spectral Amplitude Estimator", *IEEE Transactions on acoustics, speech and signal processing*, 32 (6), 1109–1121 (1984).

[2] W. A. Pearlman, R. M. Gray, "Source Coding of the Discrete Fourier Transform", *IEEE Transactions on information theory*, 24 (6), 683–692 (1978).

[3] J. Profita, T.G. Bidder, "Perfect Pitch," *American Journal of Medical Genetics*, 29 (4), 763–771 (1988).

[4] F. Itakura, S. Saito, "Analysis synthesis telephony based on the maximum likelihood method", *Proceedings of the 6th International Congress on Acoustics*, C-17–C-20 (1968).

[5] R. M. Gray, A. Buzo, A. H. Gray, "Distortion Measures for Speech Processing", *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 28 (4), 367–376 (1980).

[6] C. Févotte, N. Bertin, J.L. Durrieu, "Nonnegative matrix factorization with the Itakura-Saito divergence. With application to music analysis", *Neural Computation* 21(3), 793–830 (2009).

[7] D. D. Lee, H. S. Seung, "Learning the parts of objects with nonnegative matrix factorization", *Nature*, 401, 788–791 (1999).

[8] G.E. Kopec, M.A. Bush, "An LPC-based spectral similarity measure for speech recognition in the

presence of co-channel speech interference", *International Conference on Acoustics, Speech, and Signal Processing*, 270–273 (1989).

[9] K. Drouiche, P. Gomez, A. Alvarez, R. Martinez, V. Rodellar, V. Nieto, "A spectral distance measure for speech detection in noise and speech segmentation", *11th IEEE Signal Processing Workshop on Statistical Signal Processing*, 500–503 (2001).

[10] E. Klabbers, R. Veldhuis, "Reducing Audible Spectral Discontinuities", *IEEE Transactions on Speech and Signal Processing*, 9 (1), 39–51 (2001).

[11] J. C. Brown, "Musical fundamental frequency tracking using a pattern recognition method", *Journal of the Acoustical Society of America*, 92 (3), 1394–1402 (1992).

[12] C. Wang, S. Sene, "Robust pitch tracking for prosodic modeling un telephone speech", *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1343–1346 (2000).

[13] P. Smaragdis, J. C. Brown, "Non-Negative Matrix Factorization for Polyphonic Music Transcription", *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics on Music Information Retrieval*, 177–180 (2003).

[14] P. Smaragdis, B. Raj, M.V. Shashanka, "Sparse and shift-invariant feature extraction from non-negative data", *Proceedings of the IEEE International Conference on Audio and Speech Signal Processing*, 2069–2072 (2008).

[15] G. Peeters, "A Large Set of Audio Features for Sound Description (similarity and classification) in the CUIDADO project", *Technical report* (2004).

[16] M. Schroeder, B. Atal, "Code-excited linear prediction(CELP): High-quality speech at very low bit rates", *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 937–940 (1985).

[17] I. Borg, P. Groenen, *Modern Multidimensional Scaling: theory and applications*, Springer-Verlag New York (2005).