RADIAL BASIS FUNCTION NETWORKS FOR CONVERSION OF SOUND SPECTRA

Carlo Drioli

Centro di Sonologia Computazionale (CSC-DEI), University of Padua, Italy
cd@csc1.unipd.it

ABSTRACT

In many high-level signal processing tasks, such as pitch shifting, voice conversion or sound synthesis, accurate spectral processing is required. Here, the use of Radial Basis Function Networks (RBFN) is proposed for the modeling of the spectral changes (or *conversions*) related to the control of important sound parameters, such as pitch or intensity. The identification of such conversion functions is based on a procedure which learns the shape of the conversion from few couples of target spectra from a data set. The generalization properties of RBFNs provides for interpolation with respect to the pitch range. In the construction of the training set, mel-cepstral encoding of the spectrum is used to catch the perceptually most relevant spectral changes. The RBFN conversion functions introduced are characterized by a perceptually-based fast training procedure, desirable interpolation properties and computational efficiency.

1. INTRODUCTION

In the field of speech and audio processing a large number of applications have been proposed up to the present which are based on the sinusoidal representation of the signal. Time and pitch scaling have been widely explored, especially in the speech processing field, and the problem of correctly reproduce the spectral characteristics has been stressed [1]. Recently a new spectral processing approach has been proposed by Stylianou et al. [2], where a conversion function was build from training samples and was used to convert the spectral features of a first speaker in the spectral features of a second speaker who uttered the same sentence.

Among the applications related to music, expressiveness processing of musical performance and sound synthesis have recently gained an increasing interest. In [3, 4] the problem of controlling the high-level musical attributes of a recorded performance by means of expressiveness models and suitable sound processing techniques is faced. In the work by Horner and Beauchamp [5], additive synthesis based on $STFT$ analysis are used as the engine for sound generation purposes, and a dynamic filter is used to gain realistic results with respect to pitch and intensity variations.

All the reported applications realizes high-level transformations by combination of simpler effects like time scaling, pitch shifting, amplitude envelope scaling, spectral processing. This work focuses on the spectral processing item, and proposes a new frequency-domain filtering model suitable for the sinusoidal representation of sound. The identification of the model parameters relies on a learning procedure based on collections of real data which represents, for example, a given musical instrument. The method is proven to be useful in preserving the spectral characteristics of sounds processed by transformations such as pitch shifting or intensity scaling.

2. SOUND ANALYSIS AND RESYNTHESIS FRAMEWORK

The investigation relies on the well known sinusoidal model of the signal (SMS) [6]. The analysis algorithm acts on windowed portions (here called *frames*) of the signal, and produces a time-varying representation as sum of sinusoids (here called *partials*). Assuming that the number of partials H is constant for all frames, for the ith frame the result of the sinusoidal modeling is a set of triples $(f_h(i), a_h(i), \phi_h(i))$ $(h = 1, \ldots, H)$ of frequency, magnitude and phase parameters describing each partial, and a residual noise component that will not be considered in this work. H is taken sufficiently high to provide the maximum needed bandwidth, and zero magnitude is assigned to the exceeding partials for the spectra with lower bandwidth.

The sinusoidal representation allows to control some of the basic sound parameters, such as pitch and intensity, by simply shifting or scaling the frequency and magnitue of the partials. However, without an accurate spectral compensation which reflects the sound characteristics, the result of a transformation performed with a constant magnitude scaling is an unrealistic sound. The proposed spectral processing method relies on learning from real data the spectral transformations which occurs when such a musical parameter changes. With this perspective, a perceptually weighted representation of spectral envelopes is introduced in the next section, so that the perceptually relevant differences are exploited in the comparison of spectral envelopes.

This work was supported by *TELECOM ITALIA* under the research contract *Cantieri Multimediali*

2.1. Representation of spectral envelopes

To move from the original sinusoidal description to a perceptual domain, the original spectral envelope is turned to the *mel-cepstrum* spectral representation, by application of the regularized discrete cepstrum method [2]: for a given sinusoidal parametrization, the magnitudes $\{a_h\}$ $(h = 1...H)$ of the partials are expressed in the log domain and the frequencies $\{f_h\}$ $(h = 1...H)$ in Hz are converted to Mel frequencies $\{\lambda_h\}$ with the formula $\lambda = mel(f) \approx 1127\log(1+f/700)$. The real mel-cepstrum parameters m_i $(i = 0,...,M)$ are finally computed by minimizing the following least squares (LS) criterion

$$\sum_{h=1}^{H}(|C(\lambda_h)| - 20\log_{10}(a_h))^2 \qquad (1)$$

with

$$|C(\lambda)| = m_0 + 2\sum_{i=1}^{M} m_i \cos(\frac{\pi\lambda i}{2B_H}) \qquad (2)$$

where M is the number of cepstral coefficients, m_0 is the frame energy, and $B_H = \min\{mel(f_H), mel(F_s/2)\}$ with F_s being the sampling frequency. The normalization factor B_H ensures that the upper limit of the band corresponds to a value of 1 on the normalized warped frequency axis. $|C(\lambda)|$, the new warped and smoothed version of the spectral envelope, is more reliable to catch the perceptually meaningful differences among spectra of different sounds. Figure 1 shows the smoothed and warped spectral envelope for a saxophone tone. The aim of this transformation is to find the perceptually relevant deviations by comparing the smoothed and warped versions of different spectral envelopes. We call now $c_h = |C(\lambda_h)| = |C(mel(f_h))|$ the hth partial magnitude (in dB) of the mel-cepstrum spectral envelope, and $\Delta\mathbf{C} = \{\Delta C_h\}$ $(h = 1,\ldots,H)$, with $\Delta C_h = (c_h^{(2)} - c_h^{(1)})$, the difference between two mel-cepstrum spectral envelopes. By comparison of two different spectral envelopes is possible to express the deviation of each partial in the multiplicative form $r_h = 10\exp[\Delta C_h/20]$, and we call *conversion pattern* the set $\{r_h\}$ $(h = 1,\cdots,H)$ generated by the comparison of two spectral envelope.

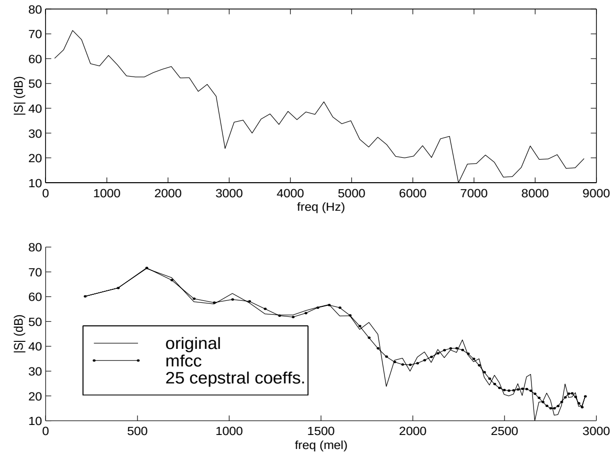

Figure 1: Sustain average spectral envelope of a saxophone note (upper figure), and frequency warped mel-cepstrum envelope (lower figure)

2.2. Spectral conversion functions

In this section, the parametric model for the conversion functions is presented as well as the parameter identification principles. The conversion is expressed in terms of deviations of magnitudes, normalized with respect to the frame energy m_0, from the normalized magnitudes of a reference spectral envelope. The reference spectral envelope can be taken from one of the tones in the data set. If the tone in the data set are notes with a classical attack-sustain-release structure, we will always consider the sustain average spectral envelopes, where the average is generally taken on a sufficient number of frames of the sustained part of the tones. Once the spectrum conversion function has been identified, the reference tone can be seen as a source for the synthesis of tones with different pitch or intensity, and correct spectral behaviour. Moreover, we are interested in keeping also the natural time-variance of the source tone, as well as its attack-sustain-release structure. To this purpose, we make the simplifying hypothesis that the conversion function identified with respect to the sustained part of notes can be used to process every frame of the source tone. In other words, the law which describes the spectral behaviour of the sustained part of a note, is assumed to well describe the behaviour in the remaining attack and release part of the note. These assumption has proven to be satisfactory in most cases, on the base of informal listening tests conducted on the processed tones.

We now make the following assumption on the structure of the conversion function:

- Due to the changing nature of the spectrum with the pitch λ_0 of the tone, the conversion function is assumed to be dependent on the pitch of the note. From the above consideration the function will then be a map $\mathcal{F} : \mathbb{R} \rightarrow \mathbb{R}^H$, where H is the maximum number of partials in the SMS representation.

- The conversion function can be decomposed in its H components $\mathcal{F}(\lambda_0) = [\mathcal{F}_1(\lambda_0), ...\mathcal{F}_H(\lambda_0)]$, and for each function $\mathcal{F}_j(\lambda_0)$ we assume the following parametric form:

$$\mathcal{F}_j(\lambda_0) = \sum_{i=1}^{U} W_{i,j} \cdot G(\lambda_0; \mathbf{q}_i) \qquad (3)$$

where $G(f;\mathbf{q}_i)$ denotes a radial basis function with parameter vector $\mathbf{q}_i$, $\mathbf{W} = \{W_{i,j}\}_{i=1\cdots U, j=1\cdots H}$ is a $U \times H$ matrix and U is the number of radial basis units used. The radial functions used as hidden units can be of various kind. Here, a cubic form $G(x;\mu) = (||x-\mu||)^3$ is used.

The conversion functions represents the behaviour of the sound spectrum in an original space whose dimension is equal in number to the number of partial used to describe the spectrum. It is quite intuitive that the number of variables involved is often redundant and should be reduced. To this purpose, singular value decomposition (*SVD*) is used.

Let $\mathbf{R}$, be the $N \times H$ matrix containing a conversion pattern in each row, one for each of the N notes in the data set (including the reference note) The singular value decomposition theorem states that $\mathbf{R}$ can be decomposed into the form $\mathbf{R}_{N\times H} = \mathbf{U}_{N\times N}\mathbf{S}_{N\times H}\mathbf{V}^T_{H\times H}$ where $\mathbf{U}$ and $\mathbf{V}$ are unitary matrices. $\mathbf{S}$ is a $N \times H$ pseudo-diagonal matrix whose non-zero elements, called singular values, are non-negative and by convention are given in decreasing order.

The singular values in matrix $\mathbf{S}$ are used to compute the rank of the decomposed matrix, which is the index of the last non-zero element. When the decomposed matrix is not square, as in our case, the rank of $\mathbf{S}$ will not be higher than the lower dimension (N), and a rank lower than N is indicated by an abrupt decrease of the magnitude between two adjacent non-zero elements in the diagonal. If we decide to use the first P components, the new set of target conversion paths will be given by $\widehat{\mathbf{R}}_{N\times H} = \widehat{\mathbf{U}}_{N\times P}\widehat{\mathbf{S}}_{P\times P}\widehat{\mathbf{V}}^T_{P\times H}$ where the unwanted columns and/or rows of the original matrices are not considered in the computation (note that for $P = N$, $\widehat{\mathbf{V}}$ is a base for the space spanned by the rows of $\mathbf{R}$ and is $\widehat{\mathbf{R}} = \mathbf{R}$). Let $\mathbf{F} = \widehat{\mathbf{U}}\widehat{\mathbf{S}}$ be the $N \times P$ new matrix which represents the spectral conversion patterns, and let $\widehat{\mathbf{V}}$ be the matrix to return to the initial conversion patterns: if we use the matrix $\mathbf{F}$ to train the RBFN, the dimensionality of the conversion function $\mathcal{F}$ is reduced from H to P with $P \leq N < H$ and the output of the RBFN will need to be multiplied by $\widehat{\mathbf{V}}$ prior to be applied to a spectral envelope.

Let now face the problem of identifying the RBFN parameters. As usually needed by the neural networks learning procedures, the original data are organized in a training set. In our case, the pitch values of the training set notes are stored in the input training vector $\mathbf{T}_{in} = [\lambda_0^{(1)}, \ldots, \lambda_0^{(N)}]$, so that each element corresponds to a row of the output matrix $\mathbf{T}_{out} = \mathbf{F}$, representing the spectral envelope conversion patterns. The centers μ of the radial basis functions are iteratively selected with the OLS algorithm [7] which places the desired number U of units (with $U \leq N$) in the positions that best explains the data. Once the radial units with centers $\mu_1, \ldots, \mu_U$ have been selected, the image of $\mathbf{T}_{in}$ through the radial basis layer can be computed as $\mathbf{G} = [\mathbf{G}_1 \cdots \mathbf{G}_U]$, $\mathbf{G}_i = [G(\lambda_0^{(1)}, \mu_i) \cdots G(\lambda_0^{(N)}, \mu_i)]^T$ ($i = 1, \ldots, U$). The problem of identifying the parameters $W_{i,j}$ of eq. (3) can thus be stated in the closed form $\mathbf{T}_{out} = \mathbf{G} * \mathbf{W}$, the LS solution of which is known to be $\mathbf{W} = \mathbf{T}_{out}\mathbf{G}^+$ with $\mathbf{G}^+$ pseudoinverse of $\mathbf{G}$.

To summarize the principal motivations why we adopted the radial basis function network model, we emphasize that the RBFNs can learn from examples, have fast training procedure, and have *generalizing* properties, meaning that if we use a training set of N tones having pitch values of $\lambda_0^{(1)} < \lambda_0^{(2)} < \cdots < \lambda_0^{(N)}$, the resulting conversion function will furnish a coherent result in the whole interval $[\lambda_0^{(1)}, \lambda_0^{(N)}]$.

3. APPLICATIONS

The method will be demonstrated in this section by using a conversion function to realize pitch transformations which preserves the spectral identity of an instrument. The procedure for the training set construction is now reviewed. From a data set of N notes we want to construct N conversion patterns (including an all-zeros pattern) comparing the sustain spectral envelope of each note with that of the note selected as *source*, whose pitch is modified each time to match the others. To this purpose, the SMS representation of the source note undergoes a modification which includes the scaling of the frequencies of partials, and optionally the interpolation of magnitudes to preserve his formant structure. This option gives the possibility to use the a-priori knowledge on the nature of sound to improve the identification process. Voice, for example, is known to be characterized by a formant structure which is, for a given vowel, approximately constant with respect to pitch variations. It is quite intuitive that, in such a case, preserving the formants can lead to a conversion pattern set with reduced magnitude range. We call *waveform preserving* the procedure where no formant preserving interpolation is performed, otherwise the procedure is called *formant preserving*. In Figure 2, the two procedures are compared with respect to a set of voiced sung notes. In Figure 3, the conversion patterns and the result of the RBF network identification is shown for a set of saxophone notes. The use of the conversion function permitted to produce pitch shifted synthetic tones whose spectral envelope reflects that of the notes in the data set, at least in the sustain part of notes. To compare the synthetic tones with the real ones, we used the spectral centroid $f_{sc} = (\sum_{h=1}^{H} f_h \cdot a_h)/(\sum_{h=1}^{H} a_h)$, which is known to be a good index of spectral similarity. Figure 4 shows the effect of the conversion function used to correct the spectral envelope of a saxophone tone pitch shifting.

The same approach seen for pitch shifting can be used to control other sound parameters implying spectral correction, like intensity. Let consider to compare couples of tones having same pitch and different intensities, say I_m the minimum and I_M the maximum intensity (no pitch shifting is now implied in the construction of the conversion pattern

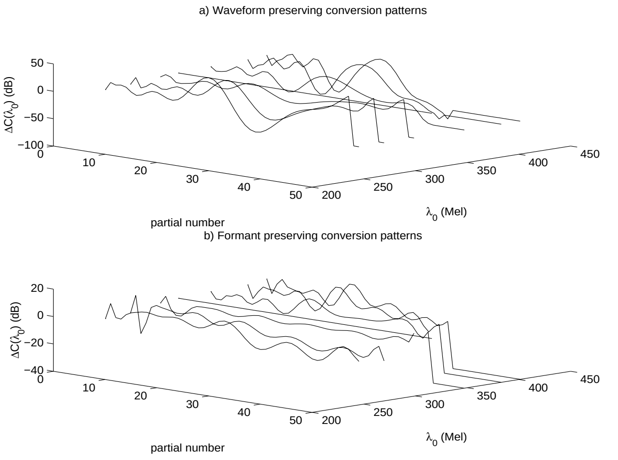

Figure 2: Conversion patterns generated from 7 voiced notes performing the same vowel: comparison between the waveform preserving procedure and the formant preserving procedure

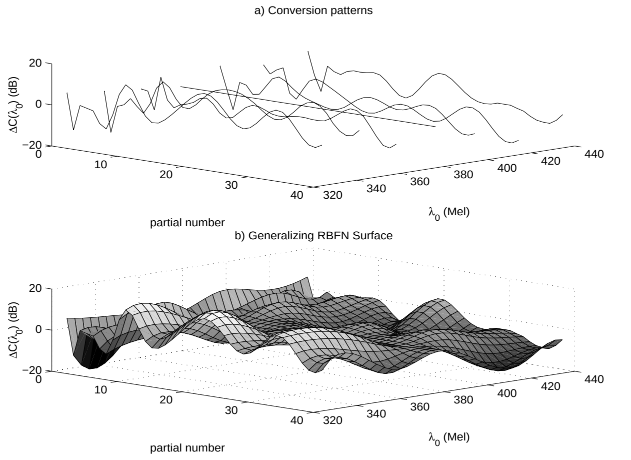

Figure 3: a) The 7 waveform preserving conversion patterns resulting from 7 sax notes. b) Interpolating surface provided by the RBFN

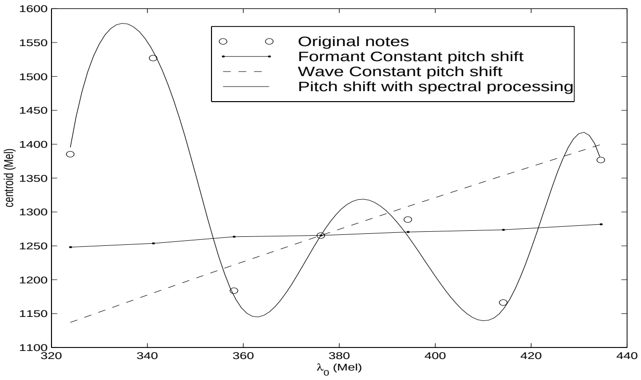

Figure 4: Spectral centroid: effect of the conversion function for pitch shifting

set). If $\mathcal{D}(\lambda)$ is the conversion function that allows to switch from I_M to I_m, the resynthesis formula that gives an intensity level $I \in [I_m, I_M]$ is $\bar{a}_h = \Delta\mathcal{D}_h(\lambda, I) \cdot a_h$, where a_h is the magnitude of the hth partial of an origin tone, λ is the pitch of that origin tone, and $\Delta\mathcal{D}_h(\lambda, I) = \mathcal{D}(\lambda) \cdot \alpha(I)$. The function $\alpha(I)$, ranging from $1/\mathcal{D}(\lambda)$, for $I = I_M$, to 1, for $I = I_m$, weights the effect of the conversion function, and can be approximated with a logarithmic function.

4. DISCUSSION AND CONCLUSIONS

A spectral processing model suitable for the sinusoidal representation of sound has been proposed. The identification procedure is characterized by a fast perceptually based learning procedure and the possibility of learning from sound examples has been stressed. Moreover, due to its low computational cost, the model is suitable for real time applications such as expressive processing or sound synthesis. The method has been applied to pitch shifting with spectral correction, and the spectral centroid of the synthesized sound has been compared with the spectral centroid of the real target sound, showing the effectiveness of this approach.

5. REFERENCES

[1] T. F. Quatieri and R. J. McAulay, “Shape invariant time-scale and pitch modification of speech,” *IEEE Transactions on Signal Processing*, vol. 40, no. 3, pp. 497–510, March 1992.

[2] Y. Stylianou, O. Cappé, and E. Moulines, “Continuous probabilistic transform for voice conversion,” *Speech and Audio Processing*, vol. 6, no. 2, pp. 131–142, March 1998.

[3] S. Canazza, G. De Poli, R. Di Federico, C. Drioli, and A. Rodá, “Symbolic and audio processing to change the expressive intention of a recorded music performance,” *Proc. of the DAFX99 Conf., Trondheim,* Dec. 1999, To be published.

[4] J. L. Arcos, R. L. de Mántaras, and Xavier Serra, “Saxex: A case-based reasoning system for generating expressive musical performances,” *Journal of New Music Research*, pp. 194–210, Sept. 1998.

[5] A. Horner and J. W. Beauchamp, “Synthesis of trumpet tones using a wavetable and a dynamic filter,” *J. Audio Eng. Soc.*, vol. 43, no. 10, pp. 799–812, October 1995.

[6] X. Serra, “Musical sound modeling with sinusoids plus noise,” *in Musical Signal Processing*, pp. 497–510, 1997, Swets and Zeitlinger.

[7] S. Chen, C. F. N. Cowen, and P. M. Grant, “Orthogonal least squares learning algorithm for radial basis function networks,” *IEEE Trans. on Neural Net.*, vol. 2, no. 2, pp. 302–309, March 1991.