On the definition of the sound-types transform

Carmine Emanuele Cella
carmine.emanuele.cella@gmail.com

October 20, 2011

Abstract

There is a strong link between the short-time Fourier transform (STFT) and the theory of sound-types. The purpose of this document is to investigate such a link defining a new kind of transform called the sound-types transform (STT).

1 The theory of sound-types

1.1 A short recall

The theory of sound-types is a framework for sound analysis and synthesis designed to represent and manipulate signals at a quasi-symbolic level [1], [2]. The basic idea is to describe sounds by means of *classes of equivalence* and *probabilities*. Conceptually, the analysis is implemented with the following steps:

1. **atomize**: divide a sound in small overlapping chunks called *atoms* (this can be done by windowing or by using more complex techniques such as atomic decomposition);
2. **make classes**: compute a set of low-level features for each atom and project it onto a feature-space; apply any kind of clustering algorithm (such as GMM) to find the principal clusters of atoms in the space;

3. **compute probabilities**: apply any kind of sequential analysis (such as HMM) to estimate the probabilities that a cluster is followed by another cluster in the original signal.

While most of the details are not explicited in the description above, the core of the idea is sketched out. Following sections will give a mathematical formulation of the theory, assuming that the *atomization* is done simply by windowing the original signal. Moreover, probabilities will not be taken into consideration here since they only act in subsequent stages of the theory.

1.2 The sound-types transform

Given a signal $\overset{N}{\vec{x}}$ of length N-samples and a window $\overset{n}{\vec{h}}$ of length n-samples, it is possible to define an **atom** as a windowed chunk of the signal of length n-samples (the starting position of the chunk is not indicated here):

$$\overset{n}{\vec{a}}=\overset{n}{\vec{h}} \cdot \overset{n}{\vec{x}} \tag{1}$$

where the operator $\cdot$ is a multiplication *element-by-element*. Using an adequate hop-size t during the analysis stage (for example $t \leq n/4$), it is possible to reconstruct a *perfect*[1] version $\overset{N}{\vec{x}'}$ of the original signal with a sum of atoms as a function of time[2]:

$$\overset{N}{\vec{x}'}= \sum_{i=0}^{N/t} \overset{n}{\vec{a}}_{i \cdot t} \tag{2}$$

where N/t is the total number of atoms present in the signal $\overset{N}{\vec{x}}$. It is possible, after the computation of a set of low-level features on each atom of $\overset{n}{\vec{a}}_i$, to define a **sound-cluster** as a set of atoms that *lie* in a defined area of the feature-space (ie. that share a *similar* set of features):

$$\overset{k_r}{\vec{c}_r}= \{\overset{n}{\vec{a}}_{r,1}, \dots, \overset{n}{\vec{a}}_{r,k_r}\}. \tag{3}$$

[1] As in STFT, the reconstruction can be perfect only under special conditions (not detailed here) deriving from the type of window used and from the overlapping factor.

[2] The positions in time of the blocks of n-samples are given by an index i that counts the number of hops (ie. $i = 4 \implies 4 \cdot t$).

The content of $\overset{k_r}{\vec{c}_r}$ is given by a statistical analysis applied on the feature-space that decides the position of each sound-cluster and its belonging atoms.

A **model** $\mathcal{M}_{\overset{N}{\vec{x}}}$ of the signal $\overset{N}{\vec{x}}$ is the defined as the set of the clusters discovered on it:

$$\mathcal{M}_{\overset{N}{\vec{x}}} = \{\overset{k_1}{\vec{c}_1}, \ldots, \overset{k_r}{\vec{c}_r}\}. \tag{4}$$

The cardinality $|\mathcal{M}_{\overset{N}{\vec{x}}}|$ of the model is also called the **abstraction level** of the analisys; since the number atoms is N/t it is evident that $1 \leq |\mathcal{M}_{\overset{N}{\vec{x}}}| \leq N/t$ with higher abstraction being 1 and lower abstraction being N/t.

Each sound-cluster in the feature-space has an associate **sound-type** $\overset{n}{\vec{\tau}_r}$ in the signal-space, defined as the weighted sum of all the atoms in the sound-cluster where the weights $\overset{k_r}{\vec{\omega}_r}$ are the distances (any kind of Bregman's divergences) of each atom to the center of the cluster:

$$\overset{n}{\vec{\tau}_r} = \sum_{j=1}^{k_r} \overset{n}{\vec{a}_{r,j}} \cdot \omega_{r,j} \tag{5}$$

with $\omega_{r,j} \in \overset{k_r}{\vec{\omega}_r}$. The whole set of sound-types in the signal $\overset{N}{\vec{x}}$ is called **dictionary** and is the equivalent, in the signal-space, of the model in the feature-space:

$$\mathcal{D}_{\overset{N}{\vec{x}}} = \{\overset{n}{\vec{\tau}_1}, \ldots, \overset{n}{\vec{\tau}_r}\}. \tag{6}$$

The creation of a sound-type from a sound-cluster is also called *collapsing* and can be indicated with the symbol $\langle \overset{k_r}{\vec{c}_r} \rangle = \overset{n}{\vec{\tau}_r}$: this operation represents an interesting connection between the feature-space and the signal-space that leads to the equivalence $\langle \mathcal{M}_{\overset{N}{\vec{x}}} \rangle = \mathcal{D}_{\overset{N}{\vec{x}}}$.

It is possible to define a function Ψ that maps an atom to its corresponding sound-type as:

$$\Psi_{\overset{n}{\vec{a}_i}} : \overset{n}{\vec{a}_i} \to \langle \overset{k_r}{\vec{c}_r} \rangle. \tag{7}$$

For a complete decomposition of the signal, it is also useful to define a function Θ that returns the original time position of each atom:

$$\Theta_{\overset{n}{\vec{a}_i}} : \overset{n}{\vec{a}_i} \rightarrow i. \tag{8}$$

It is now possible to define the **sound-types decomposition** $\overset{N}{\vec{x}''}$ of a signal by *replacing* each atom of equation 2 with the corresponding sound-type defined throught Ψ, in the right time position given by Θ:

$$\overset{N}{\vec{x}''} = \sum_{i=0}^{N/t} \overset{n}{\vec{\tau}_{r,p}} \tag{9}$$

where $p = \Theta_{\overset{n}{\vec{a}_i}}$. Finally, it is possible to define a function of time and frequency by multiplying the sound-types in a given dictionary with complex sinusoids:

$$\boxed{\overset{N}{\vec{\Phi}}_{\overset{n}{\vec{k}}} = \sum_{i=0}^{N/t} \overset{n}{\vec{\tau}_{r,p}} \cdot e^{-j\cdot\frac{2\cdot\pi}{n}\cdot\overset{n}{\vec{k}}}} \tag{10}$$

where $\overset{n}{\vec{k}} = \{f_1, \ldots, f_n\}$ is a vector of frequencies. Equation 10 is called the forward **sound-types transform** (STT); the inverse transform can recreate the sound-types decomposition and is given by:

$$\overset{N}{\vec{x}''} = \frac{1}{n}\sum_{i=0}^{N/t} \overset{n}{\vec{\Phi}}_{i\cdot t,\overset{n}{\vec{k}}} \cdot e^{j\cdot\frac{2\cdot\pi}{n}\cdot\overset{n}{\vec{k}}}. \tag{11}$$

As the next section will show, equation 10 is connected to STFT.

2 STT and STFT

The usual way to mathematically define the discrete short-time Fourier transform $\overset{N}{\vec{X}}_{\overset{n}{\vec{k}}}$ of a signal $\overset{N}{\vec{x}}$ of length N-samples taken n at a time while hopping by t-samples, is a function of both time and frequency:

$$\overset{N}{\vec{X}}_{\overset{n}{\vec{k}}} = \sum_{i=0}^{N/t} \overset{n}{\vec{h}} \cdot \overset{n}{\vec{x}}_{i\cdot t} \cdot e^{-j\cdot\frac{2\cdot\pi}{n}\cdot\overset{n}{\vec{k}}} \tag{12}$$

where $\overset{n}{\vec{h}}$ is a window of length n-samples [3] and $\overset{n}{\vec{k}}$ is as above. A general resynthesis equation is then given by:

$$\overset{N}{\vec{x}} = \frac{1}{n}\sum_{i=0}^{N/t} \overset{n}{\vec{X}}_{i\cdot t,\overset{n}{\vec{k}}} \cdot e^{j\cdot\frac{2\cdot\pi}{n}\cdot\overset{n}{\vec{k}}}. \tag{13}$$

Clearly, equations 10 and 12 have a strong resemblance. As observed in the previous section, the abstraction level of a model can be at most equal to the number of atoms (N/t) in the original signal. The extreme case for $|\mathcal{M}| = N/t$ is interesting: for that abstraction level, each sound-cluster is a singleton made of a single atom and consequently each sound-type reduces to that single atom scaled in amplitude:

$$|\mathcal{M}| = N/t \implies \overset{1}{\vec{c}}_r = \{\overset{n}{\vec{a}}_1\} \implies \overset{n}{\vec{\tau}}_r = \overset{n}{\vec{a}}_r \cdot \omega_{r,1}. \tag{14}$$

For equation 1, an atom is defined simply a windowed chunk of the original signal. Not considering the amplitude scaling factor, this also makes the sound-types decomposition $\vec{x}''$ equivalent to the simple decomposition $\vec{x}'$, leading to the important consequence that STT is a **generalization** of STFT:

$$\boxed{\overset{n}{\vec{\tau}}_r = \overset{n}{\vec{a}}_r = \overset{n}{\vec{h}} \cdot \overset{n}{\vec{x}} \implies \sum_{i=0}^{N/t} \overset{n}{\vec{\tau}}_{r,p} \cdot e^{-j\cdot\frac{2\cdot\pi}{n}\cdot\overset{n}{\vec{k}}} = \sum_{i=0}^{N/t} \overset{n}{\vec{h}} \cdot \overset{n}{\vec{x}}_{i\cdot t} \cdot e^{-j\cdot\frac{2\cdot\pi}{n}\cdot\overset{n}{\vec{k}}}} \tag{15}$$

with p defined as above. This property also holds for the inverse transform case but the prove will not be given here. The abstraction level of a model is directly connected to the *goodness* of the representation: the higher the abstraction (closer to 1) the more compact the representation. On the contrary, the quality of the synthesis given by the inverse transform degrades with high abstractions and increases with low abstractions becoming a perfect reconstruction for $|\mathcal{M}| = N/t$ as proved above.

3 Conclusions

Previous sections showed how the sound-types transform is a general case of the short-time Fourier transform. The *abstraction level* of this transform controls the compactness of the representation, making possible new operations on signals such as selective trasformations on sound-types.

References

[1] C. E. Cella, **Towards a Symbolic Approach to Sound Analysis**, MCM 2009, Yale University - New Haven (CT), Springer.

[2] C. E. Cella, **Sound-types: a new framework for sound analysis and synthesis**, ICMC 2011, Huddersfield (UK).

[3] A. Oppenheim and R. W. Shafer, **Discrete-time signal processing** - 3rd edition, Prentice Hall, 2010.