Towards Home-Made Dictionaries for Musical Feature Extraction
La Cour-Harbo, Anders

Publication date: 2003

Citation for published version (APA):
TOWARDS HOME-MADE DICTIONARIES FOR MUSICAL FEATURE EXTRACTION

Anders la Cour-Harbo

Aalborg University
Department of Control Engineering
Fredrik Bajers Vej 7C
9220 Aalborg East, Denmark
alc@control.auc.dk

ABSTRACT

The majority of musical feature extraction applications are based on the Fourier transform in various disguises. This is despite the fact that this transform is subject to a series of restrictions, which admittedly ease the computation and interpretation of transform coefficients, but also imposes arguably unnecessary limitations on the ability of the transform to extract and identify features. However, replacing the nicely structured dictionary of the Fourier transform (or indeed other nice transform such as the wavelet transform) with a home-made dictionary is a dangerous task, since even the most basic properties are easily lost.

1. INTRODUCTION

The extraction of features from music signal, and indeed many types of signals, often starts with a transformation of the signal. The purpose is to rearrange the energy in the signal such that various features of interest are concentrated in few samples. A large class of transforms of particular interest in feature extraction is the linear transforms, because they can be interpreted as a correlation between a signal and a dictionary of atoms. The choice of dictionary is usually guided by some knowledge of the signals to be transformed and by the properties associated with each dictionary. The shape and structure of the atoms in the dictionary determine which shapes and structures in the signal the transform will ‘look for’, i.e. which features can be extracted by choosing only a few samples in the transformed signal, and the interpretation of the atoms is therefore crucial in understanding how the transform responds to various properties of the signal.

The by far most widely used linear transform for feature extraction in music is the Fourier transform. Although there are a number of good reasons for using this particular transform (in particular the conception of music as a linear combination of individual frequencies is a good reason) there are a number of indications that the Fourier transform (and indeed also the more recent wavelet transform) is inadequate for extracting high level and detailed information from music.

Firstly, in the majority of the more advanced Fourier transform based feature extraction applications reported in the literature (such as classification of notes and harmonics [1, 2], identification of genre [3, 4] and instruments [5, 6], automated transcription [7, 8], beat and rhythm detection [9, 10], to mention just a few) the success rate is lower than any slightly trained listener is capable of. Secondly, the trigonometric dictionary corresponding to the Fourier transform fulfills a set of rather restrictive and in the context of musical analysis arguably unnecessary conditions that introduces a series of limitations on the use of Fourier coefficients for feature extraction (see for instance [11]).

Relaxing just some of the conditions introduces a significant freedom in the design of the dictionary, a freedom which can be used to tailor a dictionary to a certain signal type, such as music. However, even slightly loosened conditions come at a surprisingly high cost in the form of more complicated interpretations, confusion in coefficient order, numerical instability, and loss of fast implementations. Indeed, the mathematical as well as practical challenges in applying home-made dictionaries is far from trivial.

2. FREEDOM AND LIMITATIONS

In theory one has absolute freedom when choosing atoms for the dictionary. This is a quite appealing fact in the sense that it provides the freedom to design atoms which resembles features of interest, which is known to appear in the signal. The corresponding sample in the transformed signal is then an indication of to what extent that particular feature is present in the signal. Nonetheless virtually no-one exploits this freedom, but rather stick to transforms where the atoms are generated by some method independent of the signals to be transformed. In many cases the dictionary is an orthogonal (or orthogonal-like) set, and in some cases the atoms are merely dilated and translated versions of one another.

The major reasons for choosing a priori given and rather restricted transforms in favor of the freedom to design signal-specific dictionaries is that some transforms have a set of properties which anyone would be very reluctant to abandon. A list of these properties is given below. All of the well-known transforms (Fourier, wavelet, Gabor, etc.) posses most (often all) of these properties. A home-made ‘arbitrary’ dictionary does not necessarily posses any of these properties!

Orthogonality In an orthogonal dictionary all atoms are independent in the sense that changing one coefficient in the transformed signal is equivalent to altering the original signal (by addition) with exactly the corresponding atom. This makes the calculations and interpretation of transformed signal easy, and allows for a simple reconstruction based on the same dictionary.

Uniqueness When their is a one-to-one correspondence between the original and the transformed signal (this implies that
the dictionary is a basis) the transformed signal is unique. 
Thus, the original signal is represented by only one particu-
lar set of coefficients. If the dictionary contains more atoms 
than necessary for representing any signal the dictionary is 
redundant.

**Mother atom** A dictionary can be generated by simple alterations, 
like translation, scaling, and dilation, of a single atom. In 
that case the interpretation of every atom is closely related 
to the interpretation of the mother atom, and thus, the whole 
dictionary has in some sense a uniform interpretation.

**Greedy** When the best \( m \)-term approximation of a signal is given 
by the \( m \) largest transform coefficients the dictionary is call-
ed greedy (see [15, 16] for a precise definition). Thus, in 
a greedy dictionary the most dominating features always 
correspond to the largest coefficients.

Note that an orthogonal transform requires the signal as well as 
the atoms to exist in a Hilbert space, while greediness is defined in 
a Banach space. Note also that orthogonality implies uniqueness 
and greediness.

It is important to realize that the main reasons for sticking to 
transforms with these properties is algorithmic and computational 
as well as a fairly simple interpretation of transform coefficients. 
When it comes to targeting specific, significant features in the sig-
nals one can only hope (or have a qualified believe) that an a priori 
given set of atoms will perform reasonably well.

For instance, a wavelet basis is (in \( L_2 \) norm) an orthogonal 
dictionary based on a mother atom (it also possesses the other 
properties). As a consequence the computation of transform co-
efficients is easy, and the interpretation of the transform result is 
straight forward. However, orthogonality combined with a mother 
atom introduces significant restrictions to the freedom of choosing 
atoms for the dictionary. Once we choose the first atom, we auto-
matically exclude a rather large set of other atoms. We can only 
hope that none of these other atoms resemble important features in 
the signal.

The obvious question at this point is why the vast majority of 
feature extraction results are based on these transforms instead of 
transforms tailored to particular purposes. The short answer is that 
it is very difficult (if not impossible) to design a fast transform ac-
cording to some arbitrary dictionary, and the interpretation of the 
coefficients is by no means straight forward. When each single 
coefficient is potentially affect by all features (non-orthogonality) 
and more than one coefficient corresponds to each feature (non-
uniqueness) and each atom is constructed according to it’s own 
rule (no mother atom) and the most dominating features are not 
necessarily represented by the largest coefficients (not greedy dic-
tionary), it does become difficult to extract as well as exploit the 
transform coefficients.

### 3. REDUNDANT DICTIONARIES

In feature extraction applications the most common reason for aban-
donning orthogonality is the need or desire for a redundant dic-
tionary. It may be that in order to target all variations of all interesting 
features it is essential to introduce more atoms than is strictly nec-
necessary for representing the signal in the transform domain. Or one 
might be interested in having multiple ways of representing the 
same signal and therefore purposely introduce more atoms (as is 
the case with the Gabor dictionary). In most cases a redundant dic-
tionary is needed for targeting all features, often simply due to the 

### 4. EXISTING HOME-MADE DICTIONARIES

Despite the many challenges there do exist quite a few results on 
applications of home-made dictionaries ranging from slight relax-
tations of one or two of the above mentioned properties to com-
plete abandonment of most of the properties, with a tendency to 
the former rather than the latter! The general impression of these 
results is that virtually any small step towards an ‘arbitrary’ home-
made dictionary is accompanied by a major increase in interpreta-
tional complexity and post-transform processing, and, thus, com-
putational load. This is exemplified by the fact that while the best 
representation in a orthogonal dictionary is readily available from 
the coefficients, it is an NP-hard problem to obtain the optimal 
representation in a union of two orthogonal dictionaries.

Existing results are also characterized by the lack of a unified 
theory for arbitrary or slightly structured dictionaries (as such a 
theory is still in its infancy) and as a consequence there is a large 
variety in the methods and terminology applied in different pub-
llications. A few examples of home-made-dictionary applications 
are: Merging of different well-structured bases with basis pursuit 
and matching pursuit (see above), adaption of an existing orthog-
onal construction to a particular signal [21], using a very large set 
of features for generating a dictionary (sparse component analysis) 
[22]. There also exist several musical applications, for instance 
chirp atoms [23], matching pursuit [24], and tone model design of 
atoms [25].
To illustrate the points made in the first part of this paper we here present a simple home-made dictionary. The signal to be analyzed is a part of a trumpet fanfare played by a single trumpet, and consists of three different tones played in rapid succession. The waveform of the sampled (11025 Hz) signal is shown in Figure 1a along with markings of the which note is played. The time-frequency distribution of energy is shown by means of a smoothed pseudo Wigner-Ville distribution. While this method seems appealing it often turns out to be surprisingly inefficient. The reason is that the features we are looking for are identified not by an exact waveform, but rather a particular structure of the waveform. Therefore we want the atom to respond to this structure rather than a particular waveform. Consequently, we need to reproduce the structure of the feature in order to be able to disregard the small (or not so small!) differences between waveforms corresponding to the same feature.

Depending on what sort of differences one can expect various methods can be applied in an attempt to make such differences transparent to the atom. In the present case the main difference is low amplitude noise, phase shift, and small variations in frequency (too small to be noticed in the time-frequency plane in Figure 1b, but large enough to cause a significant discrepancy between a waveform-replicated atom and (other) occurrences of that feature).

To reproduce the structure of the three features in the present signal, three vectors have been designed such that they have approximately the same frequency content as the three features (thus the time-frequency plane in Figure 1b). While this approach relies on the Fourier transform as a design tool other transforms, such as the wavelet transform, might just as well have been applied, since the choice of transform for the purpose of reproducing a certain structure is governed by the ability of the transform to analyze as well as synthesize in a nice and easy fashion rather than producing coefficients with a specific interpretation.

The length of the designed vectors is 100 entries, i.e. long enough to capture sufficiently low frequencies. A matrix of size $300 \times 100$ is now constructed with the first vector inserted in the first 100 rows and starting on the diagonal with wrap around. The following 100 rows are filled in the same fashion with the second vector, and so on. In this fashion the unknown phase is captured. The entire signal is then transformed by applying this matrix to consecutive parts of the signal (each part being 100 samples), that is, no overlap, and finally, the average of the absolute value of the first 100 samples (1), the second (2) and third 100 samples (3) of the transform coefficients are computed. The resulting signals are shown in Figure 1c.

There are two apparent properties of the three curves: Firstly, large coefficients in the transform does not necessarily indicate the presence of the corresponding feature. For instance the response from the second atom (2) is larger for feature (1) than for its ‘own’ feature (2). Secondly, it is nonetheless easy to tell which feature is present at what point in time since the total response (all three curves) differs significantly between features as well as vary only a little for different instances of the the same feature.

This second property allows some simple non-linear method to map the three dimensional transform output to the set {1, 2, 3}, i.e. the three notes. In this simple case a one-hidden layer neural network with two perceptrons would suffice (the challenge resembles the classical XOR problem, see [27]).

There are a number of obvious improvements, such as exploiting the chirp-like structure in some of the notes (as evident in the
time-frequency plot) and applying the transform in a more subtle way. However, the construction suffice for the present example.

6. DISCUSSION

The purpose of this paper is to bring attention to the potential as well as the challenges of home-made dictionaries. In a musical feature extraction application based on a linear transform one has a choice of dictionary ranging from the classical, well-structured, restricted dictionaries such as Fourier and wavelet to ‘arbitrary’, home-made dictionaries. It was argued that the degree of structure in a dictionary is quite important because 1) a ‘too high’ degree of structure imposes unnecessary restrictions on the choice of atoms, 2) a ‘too low’ degree of structure means loss of very useful computational and interpretational properties, and 3) even a slight reduction in the degree of structure comes at a high cost. The union of orthogonal bases is currently being investigated by several people and is at present perhaps the most interesting way of constructing less restricted dictionaries.

The simple example of identifying notes in a trumpet fanfare demonstrates one of many ways of constructing a feature-based transform. The simplicity of the example is deceptive, though, as more extensive sound examples would require a significantly larger effort, and indeed the purpose is only to illustrate some of the points made in the preceding discussions of home-made dictionaries.

7. ACKNOWLEDGEMENT

The author would like to thank Lasse Borup for references and comments on the theoretical aspects of home-made dictionaries.

8. REFERENCES