Available online at www.sciencedirect.com
ScienceDirect
Speech Communication 53 (2011) 643–657
SPEECH COMMUNICATION
www.elsevier.com/locate/specom

Auditory-inspired sparse representation of audio signals

Ramin Pichevar *, Hossein Najaf-Zadeh, Louis Thibault, Hassan Lahdili

Communications Research Centre, 3701 Carling Ave., Ottawa, Canada

Available online 8 October 2010

Abstract

This article deals with the generation of auditory-inspired spectro-temporal features aimed at audio coding. To do so, we first generate sparse audio representations we call spikegrams, using projections on gammatone/gammachirp kernels that generate neural spikes. Unlike Fourier-based representations, these representations are powerful at identifying auditory events, such as onsets, offsets, transients, and harmonic structures. We show that the introduction of adaptiveness in the selection of gammachirp kernels enhances the compression rate compared to the case where the kernels are non-adaptive. We also integrate a masking model that helps reduce bitrate without loss of perceptible audio quality. We finally propose a method to extract frequent audio objects (patterns) in the aforementioned sparse representations. The extracted frequency-domain patterns (audio objects) help us address spikes (audio events) collectively rather than individually. When audio compression is needed, the different patterns are stored in a small codebook that can be used to efficiently encode audio materials in a lossless way. The approach is applied to different audio signals and results are discussed and compared. This work is a first step towards the design of a high-quality auditory-inspired "object-based" audio coder.

Keywords: Sparse representations; Masking; Quantization; Temporal data mining; Episode discovery; Audio coding; Matching pursuit; Auditory pattern recognition

1. Introduction

Non-stationary and time-relative structures such as transients, timing relations among acoustic events, and harmonic periodicities provide important cues for different types of audio processing techniques including audio coding, speech recognition, audio localization and auditory scene analysis. Obtaining these cues is a difficult task. The most important reason why it is so difficult is that most approaches to signal representation/analysis are block-based, i.e., the signal is processed piecewise in a series of discrete blocks. Therefore, transients and non-stationary periodicities in the signal can be temporally smeared across blocks (Smith and Lewicki, 2005). Moreover, large changes in the representation of an acoustic event can occur depending on the arbitrary alignment of the processing blocks with events in the signal. Signal analysis techniques such as windowing or the choice of the transform can reduce these effects, but it would be preferable if the representation was insensitive to signal shifts. Shift-invariance alone, however, is not a sufficient constraint on designing a general sound processing algorithm. A desirable representation should capture the underlying 2D-time-frequency structures, so that they are more directly observable and well represented at low bit rates (Smith and Lewicki, 2005). These structures must be easily extractable as audio objects for further processing in coding, recognition, etc.

The aim of this article is to propose an auditory-inspired coding scheme, which includes many characteristics of the auditory pathway such as sparse coding, masking, audio object extraction, and recognition (see Fig. 1). More specifically, we introduce an adaptive approach to the extraction of sparse codes and show by objective and subjective tests that the adaptive approach outperforms the classical matching pursuit approach (as described in Smith and Lewicki, 2005). We also show that the addition of a masking model to the classical MP can enhance the

* Corresponding author.
E-mail address: Ramin.Pichevar@usherbrooke.ca (R. Pichevar).

0167-6393/$ - see front matter
doi:10.1016/j.specom.2010.09.008

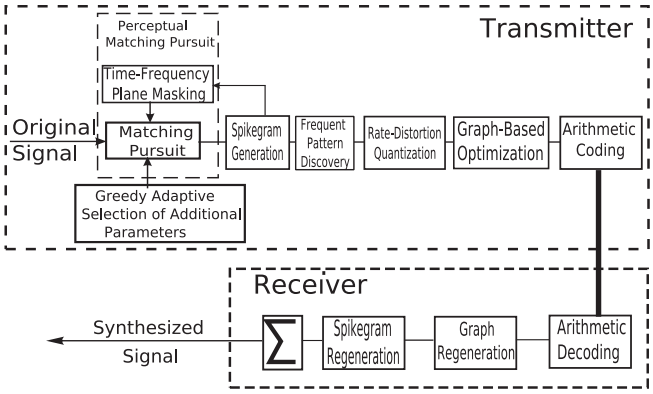

Fig. 1. Block diagram of the universal bio-inspired audio coder.

perceived quality of the reconstructed signal. We call this new technique perceptual matching pursuit (PMP). Finally, we propose an audio object extraction algorithm based on temporal episode discovery (Patnaik et al., 2008). The audio object extraction helps us address sparse codes collectively rather than individually and consequently save in coding bitrate.

In the next section we will give a brief survey of different coding schemes to justify our choices for our proposed approach.

2. Coding schemes

In this section we compare different coding approaches and will justify our choice of using sparse overcomplete codes. For this purpose, we briefly compare three approaches: block-based coding, filterbank coding, and overcomplete representations.

2.1. Block-based coding

Most of the signal representations used in speech and audio coding are block based (i.e., DCT, MDCT, FFT). In the block-based coding scheme, the signal is processed piecewise in a series of discrete blocks, causing temporally smeared transients and non-stationary periodicities. Moreover, large changes in the representation of an acoustic event can occur depending on the arbitrary alignment of the processing blocks with events in the signal. Signal analysis techniques such as windowing or the choice of the transform can reduce these effects, but it would be preferable if the representation was insensitive to signal shifts.

2.2. Filterbank-based shift-invariant coding

In the filterbank design paradigm, the signal is continuously applied to the filters of the filterbank and its convolution with the impulse responses are computed. Therefore, the outputs of these filters are shift invariant. This representation does not have the drawbacks of block-based coding mentioned above, such as time variance. However, filterbank analysis is not sufficient for designing a general sound processing algorithm. Another important aspect not taken into account in this paradigm is coding efficiency or, equivalently the ability of the representation to capture underlying structures in the signal. A desirable code/representation should reduce the information redundancy from the raw signal so that the underlying structures are more directly observable. However, convolutational representations (i.e., filterbank design) increase the dimensionality of the input signal.

2.3. Overcomplete shift-invariant representations

In an overcomplete representation, the number of basis vectors (kernels) is greater than the real dimensionality (number of non-zero eigenvalues in the covariance matrix of the signal) of the input. The approach consists of matching the best kernels to different acoustic cues using different convergence criteria such as the residual energy. However, the minimization of the energy of the residual (error) signal is not sufficient to get an overcomplete representation of an input signal. Other constraints such as sparseness must be considered in order to have a unique solution (Graham and Field, 2006). Overcomplete representations have been advocated because they have greater robustness in the presence of noise (Graham and Field, 2006). They are also a way to maximize information transfer, when different regions/objects of the underlying signal have strong correlations (Graham and Field, 2006). In other terms, the peakiness of values can be exploited efficiently in entropy coding. In order to find the "best matching kernels", matching pursuit is used.

2.4. Generating overcomplete representations with matching pursuit (MP)

In mathematical notations, the signal $x(t)$ can be decomposed into the overcomplete kernels as follow:

$$x(t) = \sum_{m=1}^{M} \sum_{i=1}^{n_m} a_i^m g_m\left(t - \tau_i^m\right) + r_x(t), \tag{1}$$

where τ_i^m and a_i^m are the temporal position and amplitude of the ith instance of the kernel g_m, respectively. The notation n_m indicates the number of instances of g_m, which need not be the same across kernels. In addition, the kernels are not restricted in form or length.

In order to find adequate τ_i^m, a_i^m, and g_m matching pursuit can be used (Mallat and Zhang, 1993). In this technique the signal $x(t)$ is decomposed over a set of kernels so as to capture the structure of the signal. The approach consists of iteratively approximating the input signal with successive orthogonal projections onto some basis. The signal can be decomposed into

$$x(t) =< x(t), \quad g_m > g_m + r_x(t), \tag{2}$$