Audio Source Separation with Matching Pursuit and Content-Adaptive Dictionaries (MP-CAD)

Conference Paper · November 2007

DOI: 10.1109/ASPAA.2007.4393000 · Source: IEEE Xplore

3 authors, including:

C.-C. Jay Kuo
University of Southern California
1,273 PUBLICATIONS **22,724** CITATIONS

SEE PROFILE

AUDIO SOURCE SEPARATION WITH MATCHING PURSUIT AND CONTENT-ADAPTIVE DICTIONARIES (MP-CAD)

Namgook Cho, Yu Shiu and C.-C. Jay Kuo

Department of Electrical Engineering and Signal and Image Processing Institute
University of Southern California, Los Angeles, CA 90089-2564
namgookc@usc.edu, yshiu@usc.edu, cckuo@sipi.usc.edu

ABSTRACT

A single-channel audio source separation algorithm based on the matching pursuit (MP) technique with content-adaptive dictionaries (CAD) is proposed in this work. The proposed MP-CAD algorithm uses content-dependent atoms that capture inherent characteristics of audio signals effectively. As compared with previous methods based on spectral decomposition and clustering in the time-frequency domain, the MP-CAD algorithm projects the time-domain audio signals onto a subspace spanned by content-adaptive atoms efficiently for their concise representation and separation. The effectiveness of the MP-CAD algorithm in audio signal approximation and single-channel source separation is demonstrated by computer simulation.

1. INTRODUCTION

Speech signals are often perceived against the background of other audio sounds with different characteristics in a real world situation. Humans are able to listen and concentrate on each individual sound from complex acoustic mixtures even under a noisy environment. Audio source separation, which aims to estimate individual sources, is one of the emerging research topics in recent years due to its potential applications to audio signal processing, *e.g.*, automatic music transcription and speech recognition.

We focus on the separation of speech and background music sounds from mixed single-channel recordings in this work. Analysis of audio signals from multiple sources using single-channel observation is a challenging problem. In contrast with multi-channel audio source separation problem, there is no spatial cue between channels to be exploited in the single-channel setting. Furthermore, since a musical signal tends to have a broader spectrum, it often overlaps with the speech signal in the time-frequency domain and it is more difficult to impose the sparsity assumption, *i.e.* only one signal is present [1, 2].

Methods proposed for single-channel source separation can be roughly classified into two categories: parametric analysis (or model-based inference) and non-parametric analysis. Model-based inference methods adopt a parametric model for each source under separation and its model parameters are estimated based on observed mixture signals [1, 3]. The main difficulty with the parametric approach is that it is not easy to find a proper model for a wide range of signals. For example, the optimal state space for music sources with the hidden Markov model (HMM) is much larger than that for the speech source due to their wider frequency range and dynamic range [2]. The sinusoidal model is also limited in its applicability [1]. Non-parametric methods are typically performed in the time-frequency domain by finding decomposition according to transform-coefficient magnitudes. After that, a clustering scheme [1, 4, 5] is needed to complete the signal separation. However, two signals could overlap significantly in some region of the time-frequency domain, and their separation in these overlapping regions could be difficult.

In this work, we decompose a given mixed signal using the matching pursuit (MP) technique [6] and attempt to maximize the sparsity of multiple signals by seeking a compact representation of each source audio with content-adaptive dictionaries (CAD). The proposed MP-CAD algorithm learns the essential representation of signals so as to produce a new dictionary containing content-adaptive atoms. For audio separation, the mixed signal is projected onto each CAD to extract the desired signal accordingly.

The rest of this paper is organized as follows. Signal decomposition using the MP technique is described in Sec. 2. The generation of content-adaptive dictionaries is discussed in Sec. 3. The proposed MP-CAD algorithm for audio source is presented in Sec. 4. The experimental setup and the corresponding results are shown in Sec. 5. Finally, concluding remarks and future research directions are given in Sec. 6.

2. SIGNAL DECOMPOSITION USING MATCHING PURSUIT AND MULTIPLE DICTIONARIES

To analyze an observation vector, $\underline{x}$, in the time-frequency domain, it is often to decompose its magnitude or power spectrum vector $\underline{X}$ into a weighted sum of functions $\underline{\Phi}_k$ as

$$\underline{X} = \sum_k \Lambda_k \cdot \underline{\Phi}_k, \tag{1}$$

where Λ_k is the gain of $\underline{\Phi}_k$. In previously proposed non-parametric methods [1, 4, 5], the spectrogram magnitude of the mixture signal is decomposed into time-varying gains and basis spectra derived from independent component analysis (ICA) or nonnegative matrix factorization (NMF). Then, the audio separation task is accomplished by clustering functions $\underline{\Phi}_k$ and their gains into disjoint sets corresponding to different original sources as

$$\hat{\underline{Y}}_m = \sum_{j \in \Gamma_m} \Lambda_j \cdot \underline{\Phi}_j, \tag{2}$$

where Γ_m is the index set of representative functions for source m, and $\Gamma_m \cap \Gamma_n = \emptyset$, $m \neq n$.

There are several challenges associated with the above approach. First, the clustering process is a complicated task by itself [1, 2]. Second, it is difficult to find a good set of functions

978-1-4244-1619-6/07/$25.00 ©2007 IEEE

$\underline{\Phi}_k$ that is truly disjoint with respect to all underlying source signals. Several methods have been proposed for this problem yet with limited success. For example, independent subspace analysis (ISA) [4] groups gain series that exhibit the highest dependencies together by calculating similarities of basis functions with some statistical distance measure. Instrument-specific features were employed in [5]. Virtanen [1] used original signals before mixing as the reference for clustering, which appears to be too restrictive in real world applications. Third, since phase-invariant features [4, 5] and the non-negativity constraint [1] are used in some methods, the synthesis formula in (2) have to be modified by compensating the phase information, which can be obtained from the source spectrogram [1, 4, 5].

In this work, we consider an even more generic framework by allowing $\underline{\Phi}_k$ to be an atom taken from a dictionary. Then, the decomposition as shown in (1) corresponds to the matching pursuit representation in the time domain [6]. In addition, based on the observation in time-frequency plane, different types of sources have different characteristics; *e.g.*, the speech signal contains irregular frequency co-modulation, frequent upward/downward sweeps and non-harmonic components, whereas the music signal exhibits relatively well-structured harmonicity and spectral continuity of partials. This observation motivates us to allow multiple dictionaries to be used in (1). Here, we do not assume the availability of the original audio signals but the audio signal types. For example, if we know that a mixture is obtained from speech and piano sounds, then we can adopt two content-adaptive dictionaries tailored to speech and piano sounds. This concept can be described in mathematical terms below.

Suppose that one type of audio source signals can be represented by a subspace as depicted in Fig. 1 (a). We may choose functions in the subspace as atoms, and all atoms of that subspace form a content-adaptive dictionary. Given a mixture consisting of audio source types s (speech) and m (music), we can express it as

$$\underline{x} = \underline{s} + \underline{m},$$

where $\underline{s} \in S_s$ and $\underline{m} \in S_m$, and subspaces S_s and S_m are subsets of the universal audio space denoted by U. If atoms of a specific subspace are known a priori, we can extract the desired audio content by projecting the mixture onto the subspace as

$$\hat{\underline{s}} = P_s \underline{x}, \qquad \text{and} \qquad \hat{\underline{m}} = P_m \underline{x}, \tag{3}$$

where P_s and P_m represent the orthonormal projections onto subspaces S_s and S_m, respectively.

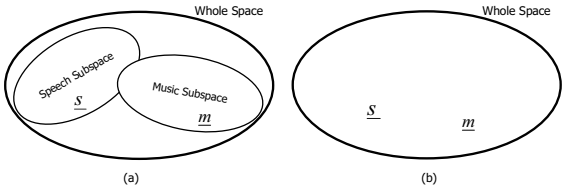


Figure 1: Signal decomposition using (a) content-adaptive dictionaries and (b) one content-independent dictionary.

For comparison, all the previous work is done based on one single dictionary (or basis functions) as illustrated in Fig. 1 (b). The use of CAD allows multiple projections *i.e.*, P_s and P_m. This gives us more flexibility in choosing a proper dictionary for each audio source separately without the burden of clustering demanded at a later stage. Besides, it is easy to reduce the overlapping of atoms from two different dictionaries, *i.e.*, enhancing the desired "sparsity" property between mixed signals. Finally, thanks to the time and frequency translation invariant Gabor dictionary [6], all the computation in the proposed MP-CAD algorithm can be performed in the time domain.

3. CONTENT-ADAPTIVE DICTIONARIES

We propose to learn the structure of a signal type from its latent basic components. For example, individual notes are basic components in pitched sounds of a specific musical instrument, and the learning of musical notes can be accomplished by a matching pursuit mechanism and a grouping procedure as depicted in Fig. 2 and detailed below.

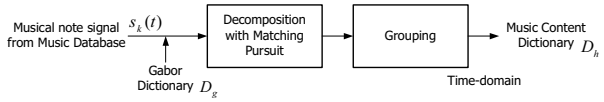


Figure 2: Learning atoms and the corresponding dictionary from musical notes.

We consider a dictionary consisting of Gabor atoms only and use g_{γ_m} to denote a Gabor atom at scale $s_m > 0$, time position u_m and frequency $\xi_m, m = 1, \cdots, M$. [6]. After initialization by setting $R_0(t) = s_k(t)$, the training signal of the *k*th musical note, denoted by $s_k(t)$, can be decomposed into a linear combination of M Gabor atoms chosen among Gabor dictionary $D_g = \{g_\gamma\}_{\gamma \in \Gamma_g}$, indicated by index set Γ_g plus residual term $R_M(t)$ as

$$s_k(t) = \sum_{m=0}^{M-1} \langle R_m, g_{\gamma_m} \rangle \cdot g_{\gamma_m}(t) + R_M(t), \tag{4}$$

where g_{γ_m} is chosen to maximize the correlation with residual $R_m(t)$ at the *m*th step, *i.e.*,

$$g_{\gamma_m} = \arg \max_{g_\gamma \in D_g} |\langle R_m, g_\gamma \rangle|. \tag{5}$$

After decomposition of the *k*th musical note, we obtain a set of Gabor atoms chosen from the decomposition, which can be represented by

$$D_k = \{g_{\gamma_m}, \gamma_m = (s_m, u_m, \xi_m) \in \Gamma_k\}, \quad \Gamma_k \subset \Gamma_g \tag{6}$$

for some index set Γ_k. The set of selected Gabor atoms is regrouped to create a subspace to represent the strong harmonic content of the musical note signal effectively. The *k*th new atom $h_k(t)$ and the corresponding subspace S_{h_k} obtained by the grouping procedure are represented by

$$h_k(t) = \sum_{\gamma_m \in \Gamma_k} c_m \cdot g_{\gamma_m}(t), \; S_{h_k} = span\{g_{\gamma_m}, \gamma_m \in \Gamma_k\}, \tag{7}$$

where c_m is a normalization constant, *i.e.*, $\|h_k(t)\|^2 = 1$ and the music subspace can be represented by $S_{music} = \cup_k S_{h_k}$. Note that one atom can be constructed for one musical note.

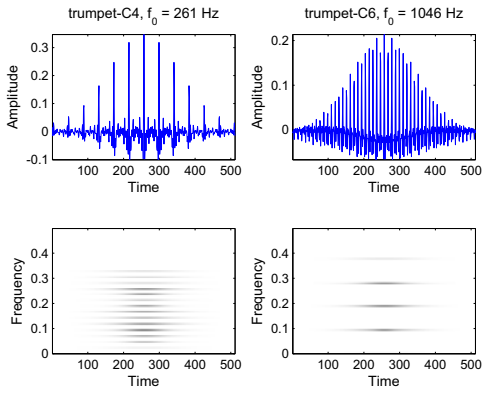


Figure 3: Examples of atoms in the content-adaptive dictionary that correspond to notes C4 and C6 of the trumpet, where f_0 is the fundamental frequency of notes.

Then, atoms h_k, $1 \leq k \leq K$, form a content-adaptive dictionary, denoted by D_h, where K is the number of notes learned from the music database. It is obvious that the size of dictionary D_h is much smaller than that of Gabor dictionary D_g. Since the dictionary size has an important impact on complexity, the complexity can be reduced by the use of CAD as well. Fig. 3 shows two atoms in the dictionary D_h, which are learned from note signals of the trumpet. We see that the time-frequency representation of these two atoms has strong harmonic components, which often exist in musical signals.

It is worthwhile to point out that a similar approach was proposed by Gribonval and Bacry [7] in creating the so-called harmonic dictionary. However, their approach and goal are different from ours. Their harmonic atoms are synthesized from linear combinations of Gabor atoms to achieve a better approximation of musical signals. Their harmonic dictionary is an extended Gabor dictionary by adding synthesized harmonic atoms to existing Gabor atoms. Thus, the size of the resultant harmonic dictionary is actually larger than the traditional Gabor dictionary.

4. AUDIO SOURCE SEPARATION BY MP-CAD

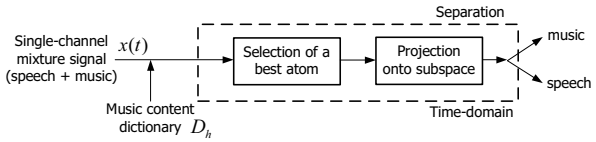


Figure 4: Audio separation by projecting a single-channel mixture onto a subspace spanned by atoms in a content-adaptive dictionary.

After obtaining the CAD, audio separation can be achieved by the projection of mixture $x(t)$ onto the subspace spanned by atoms from dictionary D_h as illustrated in Fig. 4. That is, after initialization by setting $R_0(t) = x(t)$, we can compute at the mth step as

$$P_{h_k} R_m(t) = \langle R_m, h_k \rangle \cdot h_k(t), \tag{8}$$

where

$$h_k = \arg_{h_\gamma \in D_h} (|\langle R_m, h_\gamma \rangle|^2 > \eta_e), \tag{9}$$

and η_e is a pre-defined threshold based on an energy criterion. The new residual can be computed as $R_{m+1}(t) = R_m(t) - \langle R_m, h_k \rangle \cdot h_k(t)$. Note that musical notes can be easily identified by (9), which yield large projection values.

Finally, the desired signal is reconstructed as a weighted sum of atoms chosen from D_h by

$$\hat{m}_j(t) \simeq \sum_k \sum_m \langle R_m, h_k \rangle \cdot h_k(t), \quad j = 1, \cdots, J, \tag{10}$$

where J is the number of non-overlapping window frames and h_k's are identified atoms in frame j. When $x(t)$ consists of speech and music signals only, we can get the speech signal simply via $\hat{s}(t) = x(t) - \hat{m}(t)$.

It is worthwhile to point out that all the computation in the proposed MP-CAD algorithm is actually performed in the time domain. Therefore, unlike the spectral decomposition methods described in Sec. 2, no re-synthesis procedure is needed in the proposed MP-CAD algorithm.

5. EXPERIMENTAL RESULTS

Experiments were conducted to evaluate the performance of the proposed MP-CAD algorithm. We first created a music content dictionary as discussed in Sec. 3 using real Gabor atoms as [8]

$$g_{s,u,\xi,\phi}(t) = K_{s,u,\xi,\phi} \cdot g(\frac{t-u}{s}) \cdot \cos(2\pi\xi(t-u) + \phi), \tag{11}$$

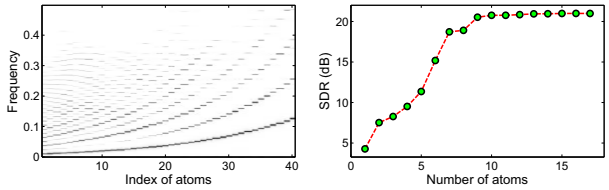
where parameter s is the scale, u is the position, ξ represents the frequency and ϕ is the phase,

$$g(t) = \frac{1}{\sqrt{s}} \cdot e^{-\pi t^2}, \tag{12}$$

and normalizing constant $K_{s,u,\xi,\phi}$ is chosen such that $\|g_{s,u,\xi,\phi}\|^2 = 1$. We used a Gabor dictionary built on atoms of length 92.8 msec, where scales are dyadic and the phase is set to zero in (11), and adopted 800 different frequencies uniformly spread over the interval of normalized frequencies, $[0, 0.5]$. Thus, the overall size of the Gabor dictionary is $|D_g| = 8000$. The time-shift parameters in atoms are considered in the computation of the correlation with the current residual signal.

Instrument sounds from RWC Music Database can be used to build music dictionaries. Since we would like to separate the mixture of speech and clarinet sounds in our experiment, a total of 40 note signals of the clarinet sound from the music database was used as the training data to generate the CAD as discussed in Sec. 3. The harmonic components of atoms in the resulting CAD are illustrated in Fig. 5 (a). We observe an excellent harmonic structure of musical instrument sounds. Test signals were chosen from several excerpts of real audio signals, *e.g.*, recordings of solo musical instruments sounds or speech signals. All sounds in the experiments were downsampled to 11,025 Hz. To measure the quality of extracted audio with respect to the original one, we used the source-to-distortion ratio (SDR) as proposed in [9]. To evaluate the capability of approximating musical sounds, a real clarinet audio signal was approximated using atoms from the CAD obtained above and the SDR values are shown in Fig. 5 (b). Note that the real audio signal of clarinet consists in nine different notes. As observed from Fig. 5 (b), after nine different atoms are chosen from

the CAD with respect to large projection values in (9), SDRs of the resynthesized signal become saturated around 20 dB. In other words, it shows that nine atoms are enough to capture most of the energy of the original clarinet signal.



(a) (b)

Figure 5: (a) Spectral properties of clarinet atoms (where each column represents an atom in clarinet's CAD), and (b) the approximation capability (in terms of SDRs) of the set of cumulative atoms.

In the audio source separation experiment, the proposed MP-CAD algorithm is compared with recently published algorithms for performance benchmarking. Two algorithms were used to factorize the magnitude spectrogram of the mixture signal; namely, ISA [4] and NMF [10, 1]. NMF was tested with the two different cost functions, which minimize the Euclidean distance (NMFEUC) or the divergence (NMFDIV) between two non-negative matrices. After that, separation was performed by clustering basis spectra $\underline{\Phi}_k$ using the *k*-means algorithm with the symmetric Kullback-Leibler divergence as a distance measure. The results obtained with these methods are shown in Table 1. The proposed MP-CAD outperforms the other two algorithms by a significant margin. The poor separation performance in two benchmark methods using spectral factorization is probably due to poor clustering. NMFDIV produces slightly better results than NMFEUC and ISA.

Table 1: Simulation results, where C, M and F stand for the clarinet music, the male and female speech, respectively.

algorithm	mixture	reconstructed signals in SDR (dB)	
		speech, $\hat{s}(t)$	music, $\hat{m}(t)$
ISA	M + C	-0.47	-1.99
NMFEUC	M + C	-0.49	0.13
NMFDIV	M + C	1.18	0.20
MP-CAD	M + C	6.34	8.91
MP-CAD	F + C	6.03	7.78

The energy-based threshold η_e, as described in Sec. 4 works reasonably well. However, for vowel speech sounds of larger energy in a non-music region, it might select an atom from the dictionary D_h incorrectly due to large projection values in (9). The reason is that only the music dictionary was used for the separation of speech and musical sounds currently. One possible solution is to obtain a CAD for speech signals as well, which will be our future work. Here, to have a quick fix, we use the speech basis spectra to improve the separation performance. The results are shown in Table 2, where a mixture of male speech and clarinet sound was used for comparison. We do see a significant improvement by using the speech information in the separation process.

Table 2: Comparison results.

dictionary	reconstructed signals in SDR (dB)	
	speech $\hat{s}(t)$	music $\hat{m}(t)$
only music	6.34	8.91
speech + music	10.28	10.87

6. CONCLUSION AND FUTURE WORK

An algorithm for single-channel audio source separation was presented by using the matching pursuit technique with content-adaptive dictionaries. The proposed MP-CAD algorithm has demonstrated good performance in its approximation and source separation capability. However, results presented in this work is still preliminary, and further work is needed to understand the strength and limitation of this algorithm. Future research directions include the search of dictionaries associated with various effects on audio signals such as reverberation, inharmonicity, and irregular pitch sweeps.

7. REFERENCES

[1] T. Virtanen, "Sound source separation in monaural music signals," in *Ph.D. Dissertation, Tampere Univ. Technol.*, Tampere, Finland, 2006.

[2] E. Vincent, M. G. Jafari, S. A. Abdallah, M. D. Plumbley, and M. E. Davies, "Blind audio source separation," *Technical Report, Queen Mary University of London*, 2005.

[3] S. Roweis, "One microphone source separation," *in Proc. Neural Inf. Proc. Syst. (NIPS)*, vol. 13, pp. 793–799, 2000.

[4] M. A. Casey and A. Westner, "Separation of mixed audio sources by independent subspace analysis," in *Proc. Int. Comp. Music Conf.*, Berlin, Germany, 2000, pp. 154–161.

[5] C. Uhle, C. Dittmar, and T. Sporer, "Extraction of drum tracks from polyphonic music using independent subspace analysis," in *Proc. 4th Int. Symp. Independent Compon. Anal. Blind Signal Separation*, 2003, pp. 843–848.

[6] S. G. Mallat and Z. Zhang, "Matching pursuit with time-frequency dictionaries," *IEEE Trans. Signal Process.*, vol. 41, pp. 3397–3415, Dec. 1993.

[7] R. Gribonval and E. Bacry, "Harmonic decomposition of audio signals with matching pursuit," *IEEE Trans. Signal Process.*, vol. 51, pp. 101–111, Jan. 2003.

[8] P. Jost, P. Vandergheynst, and P. Frossard, "Tree-based pursuit: Algorithm and properties," *IEEE Trans. Signal Process.*, vol. 54, pp. 4685–4697, Dec. 2006.

[9] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," *IEEE Trans. Audio, Speech, and Lang. Process.*, vol. 14, pp. 1462–1469, Jul. 2006.

[10] D. D. Lee and H. S. Seung, "Algorithms for nonnegative matrix factorization," in *Neural Inf. Proc. Syst.*, Denver, CO, 2001, pp. 556–562.