REVIEW OF SINGLE CHANNEL SOURCE SEPARATION TECHNIQUES

Kedar Patki
University of Rochester
Dept. of Electrical and Computer Engineering
kedar.patki@rochester.edu

ABSTRACT

The paper reviews the problem of single channel audio source separation and methods from recent research literature to solve the problem. I attempt to provide a review of the basic and advanced approaches, the assumptions behind each model, the pre-processing of the input and outline the involved algorithms. The paper focusses on statistical approaches to signal processing and does not include CASA methods, those based on information such as pitch/ periodicity, continuity and other cues from human auditory system. It also makes comparisons between methods based on both subjective and objective evidence as provided in the referenced papers.

1. INTRODUCTION

The general single channel source separation problem can be framed like this: Estimate the signals $s_1(t), s_2(t)..s_r(t)$ given only a signal $y(t)$, which follows

$$y(t) = s_1(t) + s_2(t) + s_3(t) + ... + s_r(t) \quad (1)$$

$y(t)$ is called the 'mixture' signal. In practical cases, t is a discrete quantity. The length of the available mixture signal as well as each source is finite, say N. Then, the goal of source separation is to find N values per source. i.e. total of $N * r$ values, given only N values of mixture $y(t)$.

The problem of source separation is researched extensively. To date, no machine has been built to solve this problem in a general way. Many techniques have been published in signal processing and machine learning literature, each with its own assumptions of the source signals and mixture signal.

Source separation is called 'Blind' when no information about the source signals or the process of their mixing process is presented in the problem. On the other hand, "Non-blind" source separation has some information about the sources prior to separation available.

The following sections give an overview of five techniques developed in the last decade. Section 2 presents the technique of source separation based on Bayesian statistical inference. Section 3 presents the NMF technique and Sparse NMF method. Section 4 presents the Empirical Mode Decompostion and 2-Dimensional NMF based method. Section 5 presents the Hilbert Spectrum Subspace Decomposition method. Section 6 presents the Bark-Scale Wavelet Decomposition method. Section 7 presents a general comparision between the presented methods.

2. BAYESIAN SINGLE CHANNEL SOURCE SEPARATION

Authors Thomas Beierholm, Brian Dam Pedersen and Ole Winthert [3] present a bayesian model for single channel speech separation using factorized source priors in transform domain. The priors are trained on speech samples, followed by separation of mixture signal. Mixing coefficients are estimated using Maximum Likelihood estimation.

2.1 Basic Assumptions

The method is presented for separation of two speaker sources. In the assumed model, both speakers share the same basis filters. Discrete Cosine Transform transform is used to make basis filters. The source priors are assumed to be a mixture of gaussians, one for each band in the DCT domain. Such an assumption is made for low computational burden and mathematical simplicity.

Further, the model assumes that the sources are independent of each other and independent over time.

2.2 Signal Representation

The signals should be represented in the transform domain using a transform that must be a) linear b) invertible and c) representing independent features.

Choice of transform is DCT as there are several advantages

1) Coefficients are real valued

2) Coefficients are robust to noise

3) No information loss in transformation from time domain to DCT domain

4) Has a decorrelating affect on coefficients.

Discrete Fourier Transform is not a good choice because it has complex coefficients and the sources in the transform domain are not independent.

2.3 Training of Priors and Separation

Training is done by transforming the speech samples to DCT and creating histograms. The assumed model for sources is mixture of gaussians (MoG). Estimation of source coefficients is done using posterior mean estimator and the estimation of mixing coefficients is done by maximum likelihood estimation. The Expectation-Maximization algorithm is used.

3. SPARSE NON-NEGATIVE MATRIX FACTORIZATION

The Non-negative matrix factorization (NMF) method is a basic technique to source separation and two of the techniques presented in this paper use the NMF as base.

Non-negative matrix factorization is a factorization as shown below:

$$V = D * H \tag{2}$$

Here, all three matrices are required to have non-negative elements.

When applied to audio, a spectrogram V can be factored into a matrix D of dictionary elements, which are different patterns in frequency domain and a matrix H containing weights which modify the dictionary elements over time. The matrix H is also known as an 'encoding'.

Also, we assume that this spectrogram is a sum of source spectrograms, each individually decomposable into dictionary elements and encoding matrices. The full dictionary is the concatenation of individual dictionaries and encoding is also a concatenation.

3.1 Training and Separation

In the training step, the NMF learning algorithm is run on audio samples from individual sources and dictionary elements are learnt for each source. The algorithm is initialized with random matrices which change over several iterations of the learning algorithm to produce an approximate factorization.

Concatenation of learned dictionaries is used for initializing the NMF algorithm, and the encoding matrix is learnt over iterations while the dictionary is kept constant. The resulting encoding is a concatenation of encodings of estimated source spectrograms.

Learning the non-negative factors of the given spectrogram matrix is essentially an optimization problem. Traditional NMF has several choices of the objective functions to optimize. These are called divergence rules. The most common divergence rules used are the Frobenius norm [2] and Kullback - Liebler divergence rule [4]

Sparsity is enforced on matrix H so that V is separated into sources if D is diverse enough. The Sparse NMF learning algorithm modifies the Frobenius norm and adds a sparseness criteria term into it.

Statistically, this optimization is equivalent to computing MAP estimate given a Gaussian LIkelihood function and a one-sided exponential prior distribution over H.

4. SEPARATION USING EMPIRICAL MODE DECOMPOSITION AND 2-D SPARSE NMF

Gao *et al* [5] introduce a unique way to source separation, based on decomposing the mixture into a series of oscillatory components called Intrinsic Mode Functions (IMFs), which in turn are used by a variably tuned two-dimensional sparse non-negative matrix factorization algorithm to achieve separation. This method is an unsupervised separation method.

4.1 Empirical Mode Decomposition

Empirical Mode Decomposition (EMD) is an analytical tool used to analyze non-stationary non-linear time-series. It decomposes a signal into simple oscillatory functions called IMFs. An IMF is defined as a function that follows two rules: (1) The number of extrema and the number of zero-crossings must either be equal or differ at most by one, and (2) at any point, the mean value of the envelope defined by the local maxima and the envelope defined by the local minima is zero. Instead of a rigid sinusoidal variation obtained through the DFT, an IMF can have time-varying amplitude and time-varying frequency. The first IMF of a signal contains the highest frequencies of oscillation. The subsequent IMFs contain oscillations of decreasing frequencies.

IMFs are advantageous because each IMF consists of a sub-band of frequencies where the degree of mixing is reduced [5], and the contribution to the IMF from the original sources is skewed. Therefore, the IMF spectrograms are used as observations for the subsequent NMF based decomposition.

4.2 General Algorithm

A general case of the sparse NMF, called the 2-Dimensional Sparse NMF is used in this approach. Two dimensional sparse NMF offers two advantages [9] over conventional sparse NMF (1) The 2D-NMF considers the relative position of each spectrum thereby incorporating the temporal information, (2) the NMF does not model notes but rather unique events only. Thus, if two notes are always played simultaneously they will be modeled as one component. On the other hand, a major disadvantage of 2D sparse NMF is the lack of generalized criterion for sparseness. This is overcome by introducing a sparseness term, and so the modified approach is called variable-regularized 2D sparse NMF. This modification improves accuracy in resolving speactral bases and temporal codes. This is covered in much detail in [9].

The overall algorithm is as follows: First, the mixture signal is decomposed into IMFs using EMD, followed by STFT of each IMF. These STFT representations are input to the variable 2D sparse NMF algorithm which essentially derives an overcomplete set of basis vectors. These bases are inverse STFT transformed before their grouping into sources. The grouping is done using the k-means clustering algorithm that uses the Kullback-Leibler divergence as a proximity rule.

5. SEPARATION BY HILBERT SPECTRUM SUBSPACE DECOMPOSITION

Hirose and Molla [4] introduce a method to separate sources from a single channel signal using the Hilbert spectrum and Independent Subspace Analysis (ISA). The Hilbert spectrum is used in place of the STFT for time-frequency representation. This method does not require prior knowledge of the sources and hence is an unsuperivsed method.

5.1 Signal Representation

The STFT-based time-frequency representation includes a remarkable amount of cross-spectral terms due to the harmonic assumption and the window-overlapping between successive time frames [11]. Independent Subspace Analysis [8], which represents a signal as a sum of individual source subspaces, is difficult to apply with an STFT representation because of its cross-spectral nature. The authors propose the Hilbert Spectrum (HS) as a TF representation. The Hilbert Spectrum is the application of the Hilbert-Huang Transform (HHT) to the IMFs obtained by Empirical Mode Decomposition of (EMD) of the mixture signal. The authors demonstrate the lack of cross-spectral energy terms in the Hilbert Spectrum of a signal, which leads to independent basis vectors to separate sources, and claim that the Hilbert Spectrum is a better alternative to the STFT.

5.2 General Algorithm

The general algorithm for this method is as follows: First, the mixed signal is decomposed into IMFs using EMD. Then, the Hibert Spectrum of the mixed signal is obtained from the IMFs using the HHT. Next, the coherence (spectral projection) vectors between the mixture and individual IMF components are computed, followed by deriving spectral independent bases from the set of coherence vectors by applying PCA and ICA. Then, the bases are grouped into source subsets using Kullback-Leibler divergence based K-means clustering. Then, the mixture Hilbert spectrum is projected on to the pseudoinverse nth subset of basis vectors to get the corresponding subset of temporal bases. The nth independent source subspace is derived as product of the nth subset spectral independent bases and corresponding temporal bases. Finally, time domain signals of each source are obtained by applying the inverse transformations.

6. SOURCE SEPARATION USING BARK-SCALE WAVELET PACKET DECOMPOSTION

Litvin and Cohen [6] present a unique input representation based on the Bark scale Wavelet Packet Decomposition (BS-WPD). The method is a supervised source separation technique, based on a statistical model of source signals. The source models are trained separately on individual source training data. Separation is done using MAP estimation.

6.1 Signal Representation

The Wavelet Packet Decomposition is a type of the Discrete Wavelet Transform (DWT), where the basic component resulting from the decomposition is a tiny waveform along with information about its position and frequency. The Bark scale is a psychoacoustical scale that basically enumerates indescernible frequency ranges.

For this approach, the wavelet packet decomposition is modified to become shift-invariant using a special mapping prior to DWT, so that the signal space achieves some redundancy, thereby increasing the training data [6, 9]. The wavelet decomposition uses the Bark scale critical bands for filtering.

Compared with the STFT, wavelet packet analysis produces significantly less sub-bands, with approximately the same frequency resolution in low frequencies. Frequency resolution at higher frequency range is sacrificed, in accordance with human auditory system which also has a coarser resolution in high frequency range. Reduction in the number of sub-bands results in smaller dimension of data that is used in training and separation stages [9].

6.2 Training and Separation

A simple additive mixture model is assumed. Posterior Mean is used to estimate the sources in the wavelet domain. The sources are assumed to be a mixture of gaussians (GMM). The training of the GMM models for each source is performed using Expectation Maximization (EM) algorithm and K-means algorithm.

7. COMPARISON AND PERFORMANCE OF SOURCE SEPARATION METHODS

While it is generally difficult to compare all of the different source separation methods with each other because the authors seldom test the implementations of their methods on standardized datasets, some authors explicitly compare two or more competing methods with their own. A summary of such comparisons for the methods so far are given below.

7.1 NMF and Bark-Scale Wavelet Packet Decomposition

The authors [6] found that a combination of using Bark-Scale Wavelet Packet Decomposition along with sparse NMF technique and training on phoneme level audio signals results in huge computational savings over traditional NMF and general speech signals from individual speakers.

7.2 ICA and Bark-Scale Wavelet Packet Decomposition

The Bark-Scale Wavelet Packet Decomposition (BS-WPD) approach [9] shows comparable performance with ICA, proposed by Jang and Lee [10]. In the wavelet method, the achieved separation is less in comparison to that of ICA. However, the reconstructed signals also have minimal artifacts compared with ICA. On the other hand, ICA has slightly better separation performance as well as audible artifacts.

The BS-WPD approach works well if the sources are better separated in frequency.

7.3 ICA method versus Hiilbert Spectrum based method versus EMD and sparse 2D NMF based method

The TIMIT (for speech) and RWC (for music) databases were used for training and separation of the three models by [5]. The Hibert Spectrum method does not require any training. In general, the Undetermined ICA method performed lowest, followed by Hilbert Spectrum method for speech-speech as well as speech-music mixtures. The ICA method performed slightly better than the Hilbert Spectrum method in the music-music mixtures. In all mixtures, the EMD-Sparse 2D NMF performed the best.

7.4 NMF versus EMD and sparse 2D NMF based method

The NMF methods are unable to determine the number of sources and this number usually has to be provided manually. One the other hand, the EMD preprocessing has the advantage that the number of derived IMFs exactly correspond to the number of sources and thus does not require manual input.

The EMD based 2D-NMF method and NMF method with temporal comtinuity and spaseness criteria were tested on the same datasets with the same types of mixtures (music-music, speech-speech and music-speech) and the EMD based method was found to be overall better in terms of the ISNR.

8. CONCLUSION

In this paper, I attempted to cover some state-of-the-art methods for single channel source separation. All methods exploit the underlying statistics of the data, while assuming certain probabilistic models of the source signals. These methods chose input representations that fit their model assumptions for e.g. independence of TF bins criteria. The individual methods achieve successful separation on data used by respective authors. In some cases, the data used are not naturally occuring signals or signals which one would hope to fit the type that the method aims to decompose well. This necessitates a standardized way to compare different methods. To overcome the difficulty of objective comparison, one standard dataset - SISEC - has been used for evaluating results.

Another motivating problem in this area is that a truly general source separation algorithm, one which could be applied to any kind of data, is yet to be developed. One approach that comes close is by Ozerov *et al* [12] that creates a unified approach tothe problem by consolidating all of the known statistical techniques - NMF, GMM, HMM - into one, and also creating a general way to accept prior information about sources. It employs a highly generalized expectation-maximization algorithm, based on MAP estimation, for training and separation stages. By its inclusive nature, this algorithm performs well on most types of data [12].

One key elment missing from these approaches is the use of psychoacoustical cues. As indicated in the introduction section, such methods using human auditory information are separately studied under the label CASA - Computational Auditory Scene Analysis. I believe that incorporating psychoacoustical information into the statistical approaches, along with improvements in signal processing techniques, would greatly improve the quality of separation in the future.

9. REFERENCES

[1] Mikkel N. Schmidt, Rasmus K. Olsson: “Single-Channel Speech Separation using Sparse Non-Negative Matrix Factorization”

[2] Paris Smaragdis, Judith C. Brown: “Non-Negative Matrix Factorization for Polyphonic Music Transcription,” *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2003.

[3] Thomas Beierholm, Brian Dam Pedersen, Ole Winthert: “Low Complexity Bayesian Single-Channel Source Separation,” *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2004.

[4] K. I. Molla and K. Hirose: “Single-Mixture Audio Source Separation by Subspace Decomposition of Hilbert Spectrum,” *IEEE Transactions on Audio, Speech and Language Processing*, 2004.

[5] Bin Gao, W. L. Woo and S. S. Dlay: “Single-Channel Source Separation Using EMD-Subband Variable Regularized Sparse Features,” *IEEE Transactions on Audio, Speech, and Language Processing*, 2011.

[6] Yevgeni Litvin and Israel Cohen: “Source Separation using Bark-Scale Wavelet Packet Decompostion,” *IEEE International Workshop on Machine Learning for Signal Processing*, 2009.

[7] N. E. Huang et al: “The empirical mode decomposition and hilbert spectrum for nonlinear and non-stationary time series analysis, *Proc. R. Soc. London A, vol. 454, pp. 903995*, 1998.

[8] M. A. Casey and A.Westner: “Separation of mixed audio sources by independent subspace analysis, *Proc. Int. Comput. Music Conf.*, 2000.

[9] Israel Cohen: ”Enhancement of Speech Using Bark-Scaled Wavelet Packet Decomposition” *7th European Conference on Speech Communication and Technology*, 2001.

[10] Gil-Jin Jang and Te-Won Lee: “A Maximum Likelihood Approach to Single-channel Source Separation”

[11] K.I. Molla, K. Hirose and N. Minematsu: On the empirical mode decomposition in time-frequency representation of audio signals considering disjoint orthogonality, *Proc. IEEE-EURASIP Int. Workshop Nonlinear Signal Image Process*, 2005.

[12] A. Ozerov, E. Vincent and F. Bimbot: A General Flexible Framework for the Handling of Prior Informaion in Audio Source Separation, *IEEE Transactions on Audio, Speech and Language Processing*, 2012.