ACOUSTIC BLIND SOURCE SEPARATION USING GRAPHICAL MODELS

TE-WON LEE

Qualcomm, Incorporated, San Diego, USA
tewon@qualcomm.com

We outline examples for machine learning algorithms using graphical models to represent speech signals in a systematic manner. Linear data generative models have recently gained popularity because they are able to learn efficient codes for sound signals and allow the analysis of important sound features and their characteristics to model different types of sounds, individual speech and speaker characteristics or classes of speakers. The generative model principle can be extended in time and space to handle dynamics and environmental acoustics. We present two examples for blind source separation in a graphical model. First, a method for solving the difficult problem of separating multiple sources given only a single channel observation. Second, a method for treating multi-channel observations that takes into account reverberations, sensor noise and other real environment challenges.

1 INTRODUCTION

Independent Component Analysis (ICA) [2,4,10,16,21] is now a well established method for data analysis. Its popularity is due to its simple model with a wide range of interesting theories and its applicability to many real data analysis problems. Research directions in ICA are twofold and aimed at the relaxation of strong assumptions in traditional ICA methods (no sensor noise, square mixing matrix, known number of sources, independence of sources) as well as the use of ICA methods for low level signal representation. For the former, we propose the ICA mixture model and the variational Bayesian ICA model [9] as a nonlinear extension of ICA that can handle sensor noise, estimate the number of sources, and model dependencies in the data. For the latter, ICA has been used as a tool for efficiently encoding speech signals for subsequent pattern recognition, compression and other machine learning tasks. In applying the algorithm to find a representation for speech signals, the learned speech basis functions are used for encoding speech signals for speech and speaker recognition tasks as well as the difficult problem of separating mixed sounds given only one channel. In this summary, we present these two directions in a graphical model by providing examples of source separation for one channel (simulations) and two channels (real recordings).

2 GRAPHICAL MODEL FRAMEWORK

Recently, new algorithms have been proposed to solve difficult signal processing problems. In many cases these algorithms can be described in a graphical model, which provides a general framework that allows the extension of algorithms to model new variables, parameters, signals and relationships amongst each other [18]. The learning rules and algorithms for the new model can be developed in a principled mathematical manner using tools from statistical learning theory, graphical models and signal processing. The marriage between graphical models and signal processing methods is gaining acceptance in several research disciplines and successful examples include Hidden Markov Models (HMM) for speech recognition, Independent Component Analysis (ICA) for blind signal separation, error correction codes or turbo codes in communication systems, and probabilistic algorithms in robots.

2.1 Source Models

A single audio signal can be modeled with its probabilistic representation, the time varying structure, and its decomposition into fundamental basis functions that produce an efficient coding scheme. The generative model for a single source can be extended into a multiple source observation problem. The problem is then to understand the relationship between sources and how to model their interaction with little a priori knowledge. Blind source separation is a prime example for modeling multiple sources in an environment. Furthermore, the model is realistic since audio signals do not occur isolated but are active simultaneously. Multiple source models may be given for single channel observations as well as multiple channel observations. To model the interactions with changing environments, this multiple source model needs to be further extended to include contextual changes due to the environment or non-stationary character of the sources. The model should be able to make inference about the environmental dynamics, possibly track signal sources, and understand the structure of the interacting source signals.

In its simplest form, a source can be a random variable with a fixed probability density function. A non-linear function such as the sigmoid function or the tanh-function could represent the cumulative density for the source signal. In Bell and Sejnowski [4] this non-linear function was used to separate super-Gaussian sources. This was a sufficient model because the goal was to estimate an unmixing matrix and the observation model was linear, deterministic (no sensor noise) and there were as many sources as given observation channels. There are many ways to extend this source model to include other density functions such as sub-Gaussian sources [28] and more complicated source densities that can be modeled with a mixture of Gaussians (MoG) [2].

Natural signals however are not random. To the contrary, they can have simple as well as complicated time structure. Speech signals for example are time-varying signals with correlations in time. One popular way is to model these dependencies in an HMM. The parameters of the HMM are trained on speech data and different sets of parameters can provide models for phonemes. A different approach to modeling the time structure of the source is to learn the basis functions for the signal [5]. This is a data generative model for the speech signal in which the observed speech segment can be decomposed into learned basis functions and their corresponding coefficients. The basis functions are adapted such that coefficients are statistically independent, resulting in an efficient code.

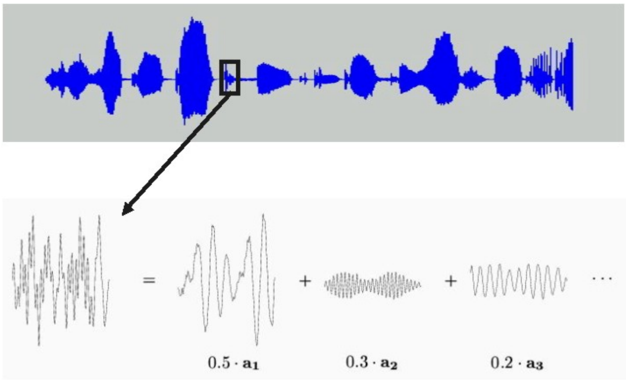

Figure 1: A speech segment is linearly decomposed into basis functions a_i and the corresponding coefficients.

In the subsequent sections, we illustrate how this data generative principle can be used to in blind separation problems in the case of one and two microphones.

3 ONE-CHANNEL SOURCE SEPARATION

The main concept behind the blind signal separation when given only a single channel recording is based on exploiting *a priori* sets of time-domain basis functions learned by ICA to the separation of mixed source signals observed in a single channel [17]. The inherent time structure of sound sources is reflected in the ICA basis functions, which encode the sources in a statistically efficient manner. We derive a learning algorithm using a maximum likelihood approach given the observed single channel data and sets of basis functions. For each time point we infer the source parameters and their contribution factors. This inference is possible due to prior knowledge of the basis functions and the associated coefficient densities. A flexible model for density estimation allows accurate modeling of the observation and our experimental results exhibit a high level of separation performance for simulated mixtures.

The single channel blind source separation can be formulated as follows:

$$y = \lambda_1 x_1 + \ldots + \lambda_i x_i + \ldots + \lambda_N x_N \qquad (1)$$

This kind of model as been studied extensively in the computational auditory scene analysis (CASA) literature [7]. The dominant approach includes the use of strong prior information about frequency clustering and the robust extraction of speech features. In the synthesis model, the observed signal y is generated by independent source signals x with different factor loadings λ. The goal is to infer the unknown source signals. This problem is highly ill conditioned and solutions can be formulated only for a constrained setting. The main idea behind our generative model approach is to make use of prior information as provided by the statistical structure of the signals of interest. The constraint is to obtain an overall efficient coding scheme, where the source signal prior information is obtained by a sparse decomposition of the signal through basis functions that have been learned. In the process of inferring the decomposition of the mixed signals, the parameters that model the linear generation of the independent source as well as the linear mixing of two sources given their basis functions and corresponding pdf structure are estimated via gradient ascent on the maximum *a posteriori* cost function. Figure 2 shows an example for separating two sources from a single observation. The details of the applied methods are described in Jang and Lee [17].

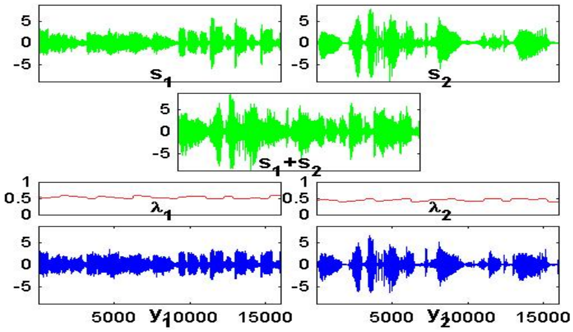

Figure 2: Single channel blind source separation. Separation results of jazz music and male speech. In vertical order: original sources (s1 and s2), mixed signal (s1+s2), and the recovered signals.

3.1 Discussion on Single Channel

The technique in [17] for single channel source separation utilizes the time-domain learned ICA basis functions. Instead of traditional prior knowledge of the sources, the statistical structures of the sources that are inherently captured by the basis and its coefficients from a training set are exploited. The algorithm recovers original sound streams through gradient-ascent adaptation steps pursuing the maximum likelihood estimate, computed by the parameters of the basis filters and the generalized Gaussian distributions of the filter coefficients. With the separation results of the real recordings as well as simulated mixtures, the proposed method is applicable to real world problems such as blind source separation, denoising, and restoration of corrupted or lost data. We are interested in including the extension of this framework to perform model comparisons to estimate the optimal set of basis functions to use given a dictionary of basis functions. This is achieved by applying a variational Bayes method [9] to compare different basis function models to select the most likely source. This method also allows us to cope with other unknown parameters such the as the number of sources. Other approaches to single channel source separation can be found in [8, 11, 22, 26] and references therein.

4 TWO-CHANNEL SOURCE SEPARATION

We consider the case where mixture signals composed of point source signals and additive background noise are recorded at different microphone locations In most practical situations the recorded microphone signals however contain a significant amount of reverberation. This phenomenon can be again modeled as a data generative model and described in the equation below

$$y_l^i = \sum_m a^i(m)x(m-l) + n^i(l) \qquad (2)$$

where y denotes the observed data, x is the hidden source, a is the mixing filter, and m is the convolution order and depends on the environment acoustics. An important distinction is made between spatially point sources and distributed background noise. Assuming little reverberation, signals originating from point sources can be viewed as identical when recorded at different microphone locations except for an amplitude factor and a delay [25]. There are many algorithms that attempt to solve this multichannel blind deconvolution problem. We outline promising approaches based on viewing this problem in a graphical model.

4.1 Microphone Array Multiple Source Models

In the case of multiple channel observations through an array of microphones, the multiple source models can be formulated as follows: $y_l^i = \sum_m a^i(m)x(m-l) + n^i(l)$, where y^i is the observed signal in channel i, $a^i(m)$ is the mixing filter and n^i is the additive noise signal.

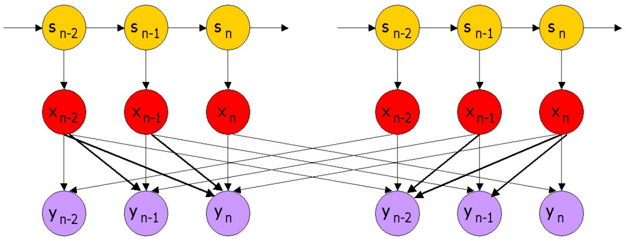

Figure 3: A generative model for representing a mixture of audio signals for two sensors. The observations can be modeled in subbands and the source models can be trained on specific audio signals such as the speech signal [3].

The directed acyclic graph (DAG) model for this mixing problem yields by Attias [3]:

$$p(y \mid x) = \prod_{ink} N(y_{in}[k] \mid \sum_{jm} H_{ij,m}[k] x_{j,n-m}[k], \lambda_i[k]) \qquad (3)$$

An EM algorithm estimates the model parameters. As the number of sources increases the E step is computationally intractable and Attias proposes to use a variational approximation to obtain the posterior distribution [3]. The benefit of solving the multichannel representation is that it not only provides with separated signals but also with the mixing filters, which provide information about the source locations with respect to each other. This additional information is useful in tracking a specific audio signal.

4.2 Separation of Real World Recordings

The separation of real world recordings poses difficult problems in many ways. Although the model in equation 2 may be sufficient, it does not take into account non-stationary issues arising from moving sources and environmental dynamics. In some cases however, the environmental setting can be controlled and the proposed solutions apply. The example below shows the separation of two voices recorded live in a conference room during a presentation. The obtained results are very encouraging and point to the right direction. The audio examples can be found on the author's website.

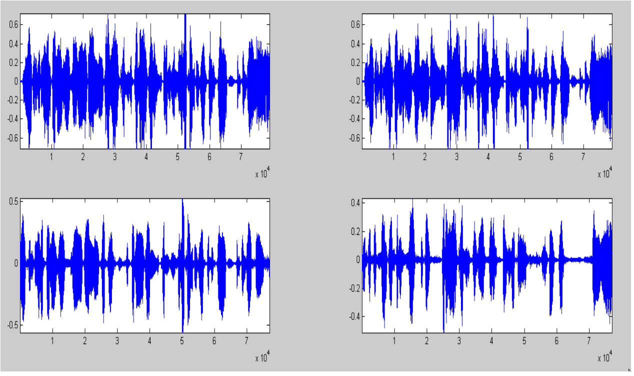

Figure 4: At the NSF workshop on speech separation Nov 2003: Real live recording of two voices (Al Bregman and Te-Won Lee) speaking simultaneously in a conference room environment. The two top plots show the time course of two microphone recordings. Since the microphones were spatially close the plots indicate only minor differences. The 2 bottom plots are the separated voice signals. The signal to noise ratio improvement was about 15dB.

4.3 Discussion on Multi Channel

There are many algorithms that attempt to solve this multichannel blind deconvolution problem. Representative work in adaptive signal processing includes [27] where higher order statistical information is used to approximate the mutual information among sensory input signals. Extensions of ICA and BSS work to convolutive mixtures include Lambert [19], Torkkola [24], and Lee et al. [20].

Traditional techniques using microphone arrays include methods for spatial filtering such as beamforming where the time delay between microphones in an array is used to steer a beam towards a sound source and therefore putting a null at the other directions. Beamforming techniques make no assumption on the sound source but assume that the geometry between source and sensors or the sound signal itself is known for the purpose of dereverberating the signal or source localization.

In contrast to beamforming techniques, ICA methods try to solve the deconvolution and automatic source localization at the same time. The main assumption is statistical independence among sources and it assumes the same number of sources as sensors. Beamforming or blind beamforming (which resembles more the ICA approach) and ICA methods make assumptions in different ways to solve similar problems. Assuming statistical independence among source is a fairly realistic assumption but it comes with several computational constraints such as the required number of sensors to be the same as the number of sources. Furthermore, no sensor noise is usually taken into account.

For practical reasons it is desirable to make stronger assumptions that elevate the sensor number, sensor noise and other constraints. A valid example is the use of a speech model. The characteristics of the speech signal can be included in many ways. In this proposal we plan to elaborate on the source model to include time structure modeled by a hidden Markov model (HMM), the observations in each state are modeled by a mixture of Gaussians in the cepstral domain. This representation is standard for modeling the dynamics of the speech signal. We believe that it will serve as a valid source model. Note that this model is used for the learning of the unmixing filters. The enhanced speech signal is obtained by filtering the observed sensor signals through the unmixing filters.

5 DISCUSSION

Relationship to other methods:

There are several research directions that are related to this research. This work relates to computational approaches for auditory scene analysis [6,8,11,12,26]. It also relates to the problem of robustly recognizing words in a realistic noisy environment [1,23]. Computational auditory scene analysis (CASA) techniques focus on techniques for grouping of frequency bands to model the auditory system [6,8,11,12,26]. The goal is to model listeners who are adept at extracting sources from mixed sounds although background noise signals can significantly overlap in time and frequency with the target speech signal.

Robust speech recognition in realistic noisy environments can be challenging when the speech signal is mixed with other acoustic sources [1,15,26]. In particular, when two speakers talk simultaneously, most speech recognition systems perform poorly.

6 CONCLUSIONS

We summarized our approaches for separating voices from mixed recordings. In the single channel case, a priori learned basis functions are used to model the temporal structure of the speech signals. A maximum likelihood approach is used to separate a voice from jazz music given only one mixed channel. In case of two microphones, the problem of separating two voices recorded by two microphones has been tackled. The mixing coefficients, time delays and reverberation coefficients are estimated using the maximum likelihood or infomax approach. The two approaches can be combined in a graphical model since both methods can be represented as data generative models where learning involves the representation of signals via the basis functions and inference involves the estimation of sources. The inference part in case of the single channel is nonlinear and linear in the two channel case.

7 ACKNOWLEDGEMENTS

The author would like to acknowledge Hagai Attias, Gil-Jin Jang and Erik Visser for valuable help and fruitful discussions.

REFERENCES

[1] Alejandro Acero, *Acoustical and Environmental Robustness for Automatic Speech Recognition*, Ph.D. Thesis, ECE Department, CMU (September 1990).

[2] H. Attias, *Independent factor analysis*. Neural Computation, vol. 11, no. 4, pp. 803-852 (1999).

[3] H. Attias, *Source separation with a sensor array using graphical models and subband filtering*. Advances in Neural Information Processing Systems 15, MIT Press, Cambridge (2003).

[4] A.J. Bell and T.J. Sejnowski, *An information-maximization approach to blind separation and blind deconvolution*, Neural Computation, vol. 7, pp. 1129-1159 (1995).

[5] A.J. Bell and T.J. Sejnowski, *Learning the higher-order structure of a natural sound*. Network: Computation in Neural Systems, vol. 7, pp. 261-266 (1996).

[6] A.S. Bregman, *Computational Auditory Scene Analysis*, MIT Press, Cambridge MA (1994).

[7] G. J. Brown and M. Cooke, *Computational auditory scene analysis*. Computer speech and language, vol. 8, pp. 297—336 (1994).

[8] de Cheveigné A, *The auditory system as a separation machine,* In Breebaart J, Houtsma AJM, Kohlrausch A, Prijs VF and Schoonhoven R (eds) Physiological and Psychophysical Bases of Auditory Function. Maastricht, The Netherlands: Shaker Publishing BV, pp. 453-460 (2001).

[9] K.-L. Chan, T-W. Lee and T.S. Sejnowski, *Variational Learning of Clusters of Undercomplete Nonsymmetric Independent Components*, Journal of Machine Learning Research, vol. 3, pp. 99-114 (2002).

[10] Pierre Comon, *Independent component analysis, A new concept?* Signal Processing, vol. 36, pp. 287–314 (1994).

[11] M. Cooke, D. Ellis, *The auditory organization of speech and other sources in listeners and computational models*, Speech Communication 35, pp. 141-177 (2001).

[12] C. Darwin, R. Carlyon, *Auditory grouping,* In: Moore, B.C.J (Ed.), The book of perception and Cognition, vol. 6, Hearing. Academic Press, New York, pp. 387-424 (1995).

[13] Daniel P. W. Ellis, *A computer implementation of psychoacoustic grouping rules*, In Proceedings of the 12th International Conference on Pattern Recognition (1994).

[14] S. Haykin, editor, *Blind Deconvolution*. Englewood Cliffs, NJ, Prentice Hall (1994).

[15] X. Huang,A. Acero, and H.-W. Hon, *Spoken Language Processing*, Prentice Hall (2001).

[16] A. Hyvärinen, J. Karhunen and E. Oja, *Independent Component Analysis*, John Wiley and Sons (2002).

[17] G.-J. Jang and T.-W. Lee, "*A Maximum Likelihood Approach to Single-channel Source Separation,*" Journal of Machine Learning Research, vol. 4, pp. 1365-1392 (December 2003).

[18] M. I. Jordan, Z. Ghahramani, and T. S. Jaakkola, *An introduction to variational methods for graphical models*, in Learning in Graphical Models (M. I. Jordan, ed.), pp. 105-161, MIT press (1998).

[19] R. Lambert, *Multichannel blind deconvolution: Fir matrix algebra and separation of multipath mixtures*, Thesis, University of Southern California, Department of Electrical Engineering (May 1996).

[20] T. Lee, A. Bell and R. Lambert, *Blind separation of convolved and delayed sources*. In Advances in Neural Information Processing Systems 9, MIT Press (1997).

[21] T-W. Lee, *Independent Component Analysis: Theory and Applications*, Kluwer Academic Publishers, Boston, ISBN: 0 7923 8261 7, (September 1998).

[22] S.T. Roweis, *One microphone source separation*. In Advances in Neural Information Processing Systems 13, (2001).

[23] R. M. Stern, A. Acero, F.-H. Liu, and Y. Ohshima, "*Signal Processing for Robust Speech Recognition*," Chapter in Speech Recognition, pp. 351-378, C.-H. Lee and F. Soong, Eds., Boston: Kluwer Academic Publishers (1996).

[24] Kari Torkkola, *Blind separation of convolved sources based on information maximization*, In IEEE Workshop on Neural Networks for Signal Processing, Kyoto, Japan (September 4-6 1996).

[25] E. Visser, M. Otsuka, T-W. Lee, "*A Spatio-Temporal Speech Enhancement Scheme for Robust Speech Recognition in Noisy Environments*" Speech Communications, vol. 41, Issues 2-3, pp 393-407 (2003).

[26] D. Wang, G. Brown, *Separation of speech from interfering sounds based on oscillatory correlation*, IEEE Transactions on Neural Networks 10 (3), 684-697 (1999).

[27] D. Yellin and E. Weinstein, *Multichannel signal separation: Methods and analysis*, IEEE Transactions on Signal Processing, 44(1):106–118 (January 1996).

[28] T.-W. Lee, M. Girolami, T. J. Sejnowski, "*Independent component analysis using an extended infomax algorithm for mixed sub-Gaussian and super-Gaussian sources,*" Neural Computation vol. 11, no. 2, pp. 409—433 (1999).