Literature Study and Evaluation of the Various Techniques for Multi-source Separation to Improve Speech Recognition

Jonathan A. F. Blixt
jonbli@kth.se
Lee Chun Yat
cylee@kth.se

June 11, 2018

Abstract

The focus of this paper is on the various state-of-the-art techniques being used for multiple-source separation, with the aim of improving speech and speaker recognition in the classic Cocktail Party problem. Speech recognition in multiple-source problems focuses on accurately distinguishing the utterances coming from the various sources; whereas the speaker recognition problem attempts to assign the separated utterances to the corresponding sources. Solving these issues would prove to be particularly interesting for applications in transcribing meetings and conversations in a multiple-speaker scenario, where recognizing both the content of the speech and the speaker producing that speech are of similar importance. Some of the techniques that we will be discussing in this paper includes methods for both single and multiple channel conditions, such as Geometric Source Separation (GSS), Missing Features, Bi-directional Long Short-term Memory (BLSTM) and Supervised Non-negative Matrix Factorization (SNMV).

1 Introduction

The Cocktail Party problem is defined as "our ability to listen to, and follow, one speaker in the presence of others", by British cognitive scientist Colin Cherry in 1957. The main issue revolves around separating overlapping signals to their component parts to be recognized as separate coherent speech. There are already well-documented methods to perform stationary noise separation such as spectral subtraction and noisy modelling. However, non-stationary noise is more difficult to model statistically, making it a much bigger challenge to separate. The problem of non-stationary noise is further exacerbated when the interference signal is similar to the signal of interest, as in the case of another

speaker speaking at the same time. This paper shall attempt to review the latest techniques designed to address this problem. The well established method of independent component analysis (ICA) will only be mentioned here due to the fact that there is exhaustive previous research within the area. Instead we will shed some light on more resent approaches calling for further investigation. For clarity, the techniques described will be segmented into two main types, namely single-channel and multi-channel methods. Multi-channel techniques such as sound source localization and beamforming aim to determine the number of speakers speaking at each time interval and their positions, and to use that information to separate the signals. On the other hand, single-channel techniques learn the characteristic features of the sound and use that to separate the signals.

2 Geometric source separation (GSS)

2.1 Sound source localization in general

Localizing a sound source is greatly facilitated by using multiple sensors often referred to as a *microphone array*. This is because a microphone array enables measurement of sound waves from the same source, but at different distances causing differences in amplitude and time delay, among other phenomena. Using binary omnidirectional audition it is possible to locate a speaker in one dimension, e.i. the angular direction in the plane of the microphones, in combination with active behaviour to distinguish between front and rear. Adding one microphone in another plane gives you the complete 3D direction of the speaker but the distance is still ambiguous. If several speakers in the same direction, one could have even more microphones to estimate the distance. Some work has been done to show that the distance can be estimated using particle filter and a circular array of 8 microphones [1].

If we assume the impinging sound waves at different microphones, but from the same source, to be parallel the time delay of arrival (TDOA) τ is given by

$$\tau = \frac{d \sin \theta}{c}, \tag{1}$$

where d is the distance between the microphones, θ the angle between the symmetry line of the microphones and the source and c the speed of sound. This delay corresponds to the factor $e^{-j\omega\tau}$ in frequency domain, where ω is the frequency of the emitted signal.

The localization problem is now reduced to figuring out what parts of the separate signals from each microphone correspond to the same sound wave from the common source. This is hampered by aliasing, repeats in the source signal, indirect repeats caused by reverberation and noise. The most common way of tackling this issue of estimating the time delay is using the cross-correlation between the signals [2]. The cross-correlation between signal a(t) and b(t) is given by

$$R_{ab}(\tau) = E[a(t)b(t+\tau)]. \tag{2}$$

But because of disturbances just described we may only make an estimation

$$\hat{R}_{ab}(t,\tau) = \frac{1}{T}\int_{t-\frac{T}{2}}^{t+\frac{T}{2}} r_a(u) r_b(u+\tau) du, \tag{3}$$

where r_a and r_b are the measurements that approximate the sound signal of the source. In modern applications, however, the calculations are performed in frequency domain

$$R_{ab}(\tau) = \int S_{ab}(\omega) e^{j\omega\tau} d\omega. \tag{4}$$

using the cross-spectrum S_{ab}. The true delay τ_d corresponding to the distance d between the microphones is then found as

$$\tau_d = \arg\max_{\tau} R_{ab}(\tau), \tag{5}$$

i.e. the match is found by maximizing the correlation between the signals. Finally, the *angle of incidence* θ is found as

$$\theta = cos^{-1}(\frac{c\tau_d}{d}). \tag{6}$$

There are, however, many ways of computing the cross-spectrum depending on what physical phenomena are taken into consideration. As an example, one could consider multiple but mutually uncorrelated sources, including noise. Allowing a single reflection we arrive at an expression for the auto-spectrum at microphone a, namely

$$S_{aa} = \sum_{i=1}^{n} [S_{s_i s_i}(\omega) + S_{s_i r_i}(\omega) + S_{r_i s_i}(\omega) + S_{r_i r_i}(\omega)] + S_{n_a n_a}(\omega) \tag{7}$$

where n is the number of sources, $S_{s_i s_i}$ the auto-spectrum due to the source itself, $S_{r_i r_i}$ correspondingly for the reflection and the cross terms are complex conjugates. The last term is due to the noise. Keeping microphone a as the reference the auto-spectrum for microphone b becomes

$$\begin{aligned} S_{bb} = &\sum_{i=1}^{n} [S_{s_i s_i}(\omega) + \alpha_i S_{s_i r_i}(\omega) e^{j\omega(\tau_{s_i} - \tau_{r_i})} + \\ &+ \alpha_i S_{r_i s_i}(\omega) e^{-j\omega(\tau_{r_i} - \tau_{s_i})} + S_{r_i r_i}(\omega)] + \alpha_i S_{n_b n_b}(\omega) \end{aligned} \tag{8}$$

where τ_{s_i} and τ_{r_i} represent the delay of the direct sound s_i and reflected sound r_i at microphone b compared to microphone a. The a_i correspond to the attenuation at mic b compared to a. This results in the cross spectrum

$$\begin{aligned} S_{ab} = &\sum_{i=1}^{n} S_{s_i s_i}(\omega) e^{-j\omega(\tau_{s_i}} + \alpha_i S_{s_i r_i}(\omega) e^{-j\omega\tau_{r_i}} + \\ &+ S_{r_i s_i}(\omega) e^{-j\omega\tau_{s_i}} + \alpha_i^2 S_{r_i r_i}(\omega) e^{-j\omega\tau_{r_i}} \end{aligned} \tag{9}$$

	Delay and Sum	MVDR	MUSIC
1 speaker	0.7035	0.1012	0.0851
2 speakers	0.4992	0.4990	0.1903

Table 1: RMSE in angle for different localizers and number of sources [5].

where the noise terms has vanished due to the assumption of uncorrelation and infinitely long averaging.

The output of the GSS still contains background noise and also interference. One drawback of using several microphones is that the multiple channels exhibits spectral leakage between one another. This issue can be tackled by postfiltering. It has been shown, especially when decomposing the noise into both stationary and transient components for each source, that the performance is improved by using a postfilter [3].

2.2 Beamforming

If the direction of the sound source is known one may use the technique of *beamforming* which is a means of emitting, or as in this case receiving, signals directionally [4]. More precisely, the signals from the different microphones are combined in such a way as to amplify or attenuate signals depending on the direction of arrival. The simplest beamformer (Delay and Sum) delays each signal corresponding to the individual distance to the source and then adds them up. This distance can be estimated using any sound source localization technique. In fact, beamforming itself can be used for source localization by scanning all direction by shifting the delays and seek maxima in output energy. This simple technique, called *steered beamformer*, combined with reliability weighted phase transform (RWPHAT) has shown an accuracy better than 1 degree and only 10% RMS error for the distance in the context of conferance calls [1]. The RWPHAT helps reducing the influence of noise and reverberation by giving the corresponding frequency bins less weight.

By adding a constraint to minimize the power of angles different from the source angle we get the more advanced beamformer called Minimum Variance Distortionless Respons (MVDR).

Another possible solution is to use Eigen Value Decomposition to work in separate signal and noise subspaces. This method is called Multiple Signal Classification (MUSIC) and has been shown to outperform both Delay and Sum and MVDR in measures of RMSE of the estimated direction [5]. One drawback, however, is that the number of sources need to be specifically stated for MUSIC to work.

3 Missing features

By using probabilistic to estimate the reliability for measurement in the time-frequency plane one may set a threshold as to what data should be considered

too unreliable and ignored as noise. For conventional missing feature estimation Hidden Markov Models (HMM) are used. The output emission probabilities are then modified to only keep the reliable distributions. It is suggested that the HMMs are trained on clean data to recognize true speech. The discrete missing feature mask is computed as follows without any explicit modeling of noise. First a noise surpression post-filter is applied. Then, for each mel-frequency band, that feature is considered reliable if the ratio of the post-filter output energy over the input energy is greater than a chosen threshold. One could also use a continues mask instead of an absolute threshold. It has been shown that using postfilter can reduce the relative recognision error rate by 24% alone, and 42% when combined with a missing feature mask on average for a humanoid robot using source separation [3].

4 Bi-directional Long Short-term Memory

We will now explore a few of the latest techniques used for monaural speech separation. Neural networks have a great potential in today's society where more and more data is becoming available. Features not found by GMMs in the can be discovered by these networks. As pointed out in earlier sections of this paper, we are faced with a challenging problem of separating multiple signals of interest from one another, which involves repeatedly focusing on one particular signal at a time and treating the remaining signals as non-stationary noise. The main challenge would be that the non-stationary noises would also be in a similar time-frequency band and share similar characteristics with the signal of interest, such as their pitch. These constraints would render techniques such as spectral subtraction and noise adaptation models less effective in dealing with non-stationary noise segregation. This motivates us to look into a structure based on recurrent neural networks (RNN) known as a bi-directional long short-term memory (BLSTM) network. [6]

The BLSTM unit architecture is essentially the same as that of the common long short-term memory (LSTM) unit architecture, consisting of a memory cell, a forget gate, an input gate and an output gate which are denoted by c_t, f_t, i_t and i_t respectively, as shown in Figure 1a. In a conventional RNN, the output of each hidden layer h_t is usually calculated by the equation

$$h_t = f(W_{xh}x_t + W_{hh}h_{t-1} + b_h), \tag{10}$$

where W are the weights, b is the bias vector and x_t is the input vector at time t. Given its ability to retain some information from the past, the RNN is able to perform learning on time sequential data, as in speech recognition. However, the limited temporal range of conventional RNNs diminishes its ability to effectively perform temporal context driven regression and classification over long utterances with mixed signals from multiple speakers. This is due to the vanishing gradient problem that affects neural networks using gradient-based learning algorithms, including RNNs, whereby the gradient-based updates become so small as training progresses that the weights almost stop changing and training stops prematurely.

An LSTM network is able to go around this issue by using the backpropagation through time (BPTT) update rule that does not face a vanishing gradient. The LSTM also often uses the logistic sigmoid function σ as its activation function. The hidden layer output h_t for an LSTM can then be calculated using these equations

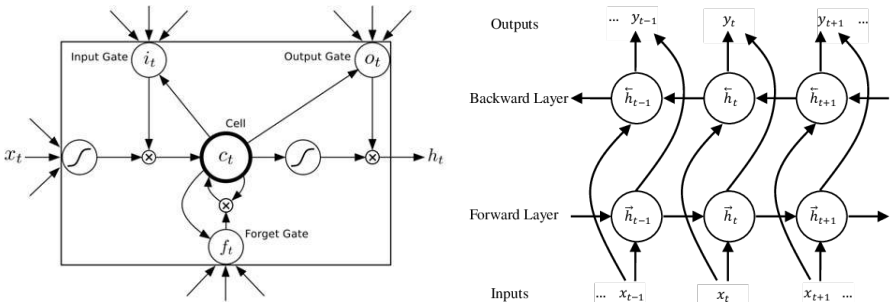
$$i_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1} + W_{ci}c_{t-1} + b_i), \tag{11}$$

$$o_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1} + W_{co}c_t + b_o), \tag{12}$$

$$f_t = \sigma(W_{xf}x_t + W_{hf}h_{t-1} + W_{cf}c_{t-1} + b_f), \tag{13}$$

$$c_t = f_t c_{t-1} + i_t \cdot tanh(W_{xc}x_t + W_{hc}h_{t-1} + +b_c), \quad h_t = o_t \cdot tanh(c_t). \tag{14}$$

The main difference between BLSTM and conventional unidirectional LSTM networks is that normal LSTM networks only take into account past context, whereas BLSTM also takes into account future context. This is being done by each BLSTM hidden layer having both a forward and backward layer going in opposite directions from each other, while also being simultaneously connected to both the input and output layers, hence the name bi-directional LSTM. An illustration of a single BLSTM hidden layer is shown in Figure 1b.



(a) LSTM Unit Cell Diagram [7] (b) BLSTM Hidden Layer [8]

Figure 1: LSTM and BLSTM Architecture

There are a few experiments [6, 9] conducted that make use of a BLSTM network with 2 such hidden layers, as in Figure 2. Although those experiments were mainly for non-stationary noises in general and are not particularly pertaining to multi-speaker speech separation tasks, the methods are transferable [10]. With just 2 BLSTM hidden layers in [6], they were able to perform dynamic wind noise reduction off-line, where the network was trained on a mix of clean and noisy speech. The BLSTM network achieved better source-to-distortion ratio (SDR) than the LSTM and conventional deep feedforward networks across nearly all signal-to-noise ratios (SNR). Admittedly, while the BLSTM network

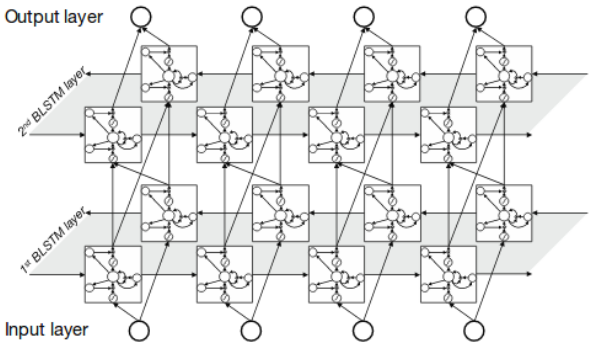


Figure 2: Bi-directional long short-term memory network, 2 hidden layers [6]

shows promise in non-stationary noise separation, it would still be unable to address the Cocktail Party problem as a stand alone solution. These networks would have to be implemented in conjunction with other feature enhancements or clustering methods before effective speech separation is possible.

5 Supervised Sparse Non-negative Matrix Factorization

The sparse non-negative matrix factorization (SNMF) technique is a popular method for single-channel speech separation, which functions by producing sparse representations of the mixed signals through estimated dictionaries of the different speakers. In a supervised SNMF, these dictionaries are obtained from the various original separated speech signals, which is then used to approximate the signal of interest from the mixed signal [11]. Consider a non-negative matrix V of dimension $F \times N$ of the mixed speech signal, which we then factorize approximately into two separate non-negative matrices W and H of dimensions $F \times k$ and $k \times N$ respectively

$$V \approx WH, \tag{15}$$

where k represents the number of hidden or latent features that is pre-defind for the SNMF to find, W represents the dictionary and H denotes the activation matrix as depicted in fig. 3. The non-negative nature of the matrices mean that only additive operations are engaged to reconstruct V from the 2 lower level matrices, forcing the dictionary W to search for the most basic building blocks comprised in the spectrogram[12]. These building blocks would then be representative of the most fundamental sounds making up speech from the respective sources. In supervised SNMF, the first step is to learn the dictionary W based on the non-mixed speech data by optimizing W and H to approximate V using a certain optimization function. Popular optimization approaches include

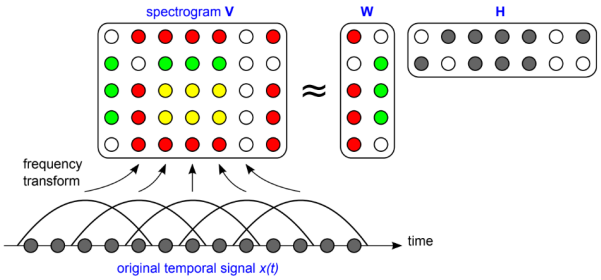


Figure 3: Decomposition of non-negative matrix V into non-negative matrices W and H in NMF analysis [12]

the Itakura-Saito distance, the generalized Kullback-Leibler divergence and the Euclidean distance.[13] After we get a fixed dictionary W, the optimization function focuses on optimizing H during testing. The ultimate aim is to get a set of dictionaries and activation coefficients for each of the different sources, used in conjunction to recreate the separate utterances from the mixed signal.

6 Conclusion

The standard sound source localization method described in section 2.1 using time delay TDOA is very useful for separating sources far away from the microphone array and with diverse directions. If the sound sources are too close to the microphone array some assumptions cannot be made demanding a more complex model and a higher computational cost. There is also the problem of spectral leakage between the channels, which can be aided by a multichannel and multisource postfilter. The use of missing features also drastically improves the recognition rate when severe interference, like multiple speakers, are present. If the sound sources are located to near each other one has to rely on the characteristics of the emitted sound solely, which may be solved using the single-channel methods described in sections 4 and 5. The BLSTM network is capable of performing non-stationary noise separation effectively, which could then be extended to speaker separation. Its main advantage lies in the powerful temporal context dependency, but a drawback would be that it must be given the number of speakers beforehand. Similarly to supervised SNMF, given that the dictionary needs to first be trained, this technique also requires prior knowledge of the speakers before it can achieve effective speech separation. Although SNMF has the advantage of not requiring as much data as BLSTM the training must be performed on a supervised individial level. At this point, the multi-channel methods such as beamforming come in to solve the issue of estimating the number of speakers. This could then be an inspiration for a hybrid system of both single and multi-channel techniques that could potentially address the challenging Cocktail Party problem.

6.1 Incorporation of feedback

Jonathan Blixt: First I want to mention that my partner will have his grade converted to simply Past and Fail for his education in Singapore, so it has mainly been me who incorporated all the feedback.

The feedback I got in the peer review was simply to add comparisons between methods and some brief history. I have therefor mentioned the independent component analysis method and why we chose to focus on other methods in this report. I have also now compared methods using additional microphones to show that although we run into the issue of spectral leakage between channels we can counteract this effect by postfiltering, although this method of course adds computational cost. I have also added comparison between the already mentioned methods and motivated the use of neural networks, especially the BLSTM, in this data rich society. While SNMF requires less data it has a drawback of having to be trained individually, supervised, for each speaker. I also heard from my partner that he got feedback conserning the conclusion so I have added conclusions drawn from missing feature masks. I have tried to be a bit more concise in the remainder of the text to make room for all the improvement while not exceeding the already hit limmit of 8 pages. Missing references to the figures have also been added.

Thank you and have a great summer!

References

[1] J.-M. Valin, F. Michaud, and J. Rouat. Robust 3d localization and tracking of sound sources using beamforming and particle filtering. volume 4, pages IV–IV, USA, 2006. IEEE.

[2] Brent C. Kirkwood. Acoustic source localization using time-delay estimation. *Technical University of Denmark*, August 2003.

[3] J.-M. Valin, S. Yamamoto, J. Rouat, F. Michaud, K. Nakadai, and H.G. Okuno. Robust recognition of simultaneous speech by a mobile robot. *Robotics, IEEE Transactions on*, 23(4):742–752, August 2007.

[4] Angelica Kjellson, Jun Yu, and Leif Nilsson. Sound source localization and beamforming for teleconferencing solutions, 2014.

[5] Ramin Anushiravani. Sound source localization with microphon arrays. *Department of Electrical and Computer Engineering, University of Illinois*, July 2016.

[6] Jinkyu Lee, Keulbit Kim, Turaj Shabestary, and Hong-Goo Kang. Deep bi-directional long short-term memory based speech enhancement for wind noise reduction. pages 41–45. IEEE, 2017.

[7] Yu-Hui Qu, Hua Yu, Xiu-Jun Gong, Jia-Hui Xu, and Hong-Shun Lee. On the prediction of dna-binding proteins only from primary sequences: A deep

learning approach.(research article)(author abstract). *PLoS ONE*, 12(12), December 2017.

[8] Amr Mousa and Björn Schuller. Deep bidirectional long short-term memory recurrent neural networks for grapheme-to-phoneme conversion utilizing complex many-to-many alignments, September 2016.

[9] Martin Wollmer, Felix Zixing Zhang, Bjorn Weninger, Gerhard Schuller, and Gerhard Rigoll. Feature enhancement by bidirectional lstm networks for conversational speech recognition in highly non-stationary noise. pages 6822–6826. IEEE, May 2013.

[10] John R. Hershey, Zhuo Chen, Jonathan Le Roux, and Shinji Watanabe. Deep clustering: Discriminative embeddings for segmentation and separation. August 2015.

[11] Tuan Pham, Yuan-Shan Lee, Yu-An Chen, and Jia-Ching Wang. A review on speech separation using nmf and its extensions. pages 26–29. IEEE, December 2015.

[12] Alexey Ozerov Cédric Févotte, Emmanuel Vincent. Single-channel audio source separation with nmf: divergences, constraints and algorithms, November 2017.

[13] Weninger F.J. Hershey J.R Le Roux, J. Sparse nmf – half-baked or well done?, March 2015.