Speech Separation in Multi-Channel by ICA and in Monaural by Deep Recurrent Neutral Networks

Wei Zhou **Yufei Zhang**

Dept. of Electrical and Computer Engineering, University of Rochester, Rochester, NY, USA

{wzhou17, yzh242}@rochester.edu

ABSTRACT

Speech separation is an important topic with increasingly broad applications. Plenty of methods have been proposed to solve this problem in various challenging situations. In this paper, we first investigate the influence of the number of channels on the performance of the Independent Component Analysis (ICA), a widely used method dealing with multichannel sound source separation. Then, for monaural sound source condition, we report the Independent Subspace Analysis (ISA), a model removing the channel limitation in ICA, and finally the Deep Recurrent Neutral Networks (DRNN) is presented. The DRNN model is improved by using denoising autoencoder to solve the problem. Our best separation results achieve SNR to 13.93 on given datasets.

1. INTRODUCTION

Source separation aims to recover one or more source signals of interests from a mixture of signals. An important topic is to obtain clean speech signals mixed with non-stationary noises. An effective solution towards this question can facilitate human-human or human-machine communication in unfavorable acoustic environments, such as enhancing the accuracy of automatic speech recognition (ASR) [1].

Source separation tasks can be divided into two categories according to the number of channels. First, multi-channel separation, the system has more than one microphone to record audio. Under this condition, the number of channels is greater, equal and less to the number of sound sources are defined as overdetermined, determined and under-determined, respectively. For under-determined condition, Degenerate Unmixed Estimation Technique was presented by Yilmaz and Rickard [2] and works well with anechoic environments. Nonetheless, it would fail if the sources overlap too much. While for overdetermined source separation, Cardoso and Souloumiac proposed the Beamforming model [3]. It is simple and robust, but needs to know the direction of the target source.

ICA [4-5], meanwhile, is a model most commonly used in overdetermined and determined conditions. It achieves elegant results. And many ICA-based techniques have been developed and proposed [6-8]. However, ICA cannot be directly used for the separation of monaural time domain signals.

ISA removes the above limitation. Based on ICA model, ISA has been proposed to apply in monaural sound source separation, for example, by Casey and Westner [7] and Orife [9]. Also, a sound recognition system based on ISA has been adopted in the MPEG-7 standardization framework [10]. What's more, statistical model-based spectral subtraction [11] in speech denoising, sparse and low-rank model [12] are presented. However, they can't deal with drums sound whose acoustic components may lie in sparse subspaces instead of being low rank. Without sparse subspaces and low rank prior assumption, Non-negative Matrix Factorization (NMF) [13] is used to factorize time-frequency spectral representations. Nevertheless, in real-world scenarios, since signals might not always follow Gaussian distributions, linear models, including NMF model, are not robust enough to model the complicated relationship between separated and mixture signals.

Recently, deep learning techniques attracts lots of attention. Dealing with the nonlinear relationship between mixture signals and separated sources, deep neural networks (DNN) model is proposed [14]. Meanwhile, considering the continuity characteristic of audio signal, DRNN model is presented [15-17], which achieve good results in dealing with sequential data. Moreover, the application of autoencoder in speech enhancement shows good results [18-19].

In this paper, we introduce the basic theory of ICA, ISA and DRNN model in Section 2. In Section 3, we illustrate the improvements on solving the DRNN model. And Section 4 presents the experimental setting and corresponding results using the NOIZEUS dataset [20], a noisy speech corpus for evaluation of speech enhancement algorithms. We conclude the paper in Section 5.

2. SOURCE SEPARATION MODEL

2.1 Multi-channel

2.1.1 ICA model

Under multi-channel source separation condition, the relationship between recorded time signals and source sources can be expressed as

$$x_j(t)=a_{j1}s_1(t)+a_{j2}s_2(t)+\boxtimes+a_{jn}s_{1n}(t) \qquad (1)$$

where $x_j(t)$ is the j^{th} observation and $s_n(t)$ is the n^{th} sound sources. The recorded signals are the mixture of sound sources in different ways (depend on the position of the microphones). If we knew the parameters a_{jn}, we could solve the linear equation in (1) by classical methods, and finally get separated sound sources.

ICA tries to solve this problem by using the statistical properties of the sound sources $s_n(t)$. It assumes the sources, at each time instant *t*, are *statistically independent*. Though the assumption is unrealistic in many cases, it doesn't need to be exactly true in practice. To explain the model clearly, we denote $\boldsymbol{X}$ as the matrix whose row vector $\boldsymbol{x}_j$ are the samples of mixtures $x_j(t)$, $\boldsymbol{S}$ as the matrix whose row vector $\boldsymbol{s}_n$ are the samples of $s_n(t)$, and denote **A** as the mixing matrix with elements a_{jn}. Thus, the mixing model can be written as

$$\boldsymbol{X}=\mathbf{A}\boldsymbol{S} \tag{2}$$

The mixing matrix **A** is unknown. We estimate both **A** and $\boldsymbol{S}$ only by using the observations $\boldsymbol{X}$. If matrix **A** can be estimated successfully, we can compute its inverse, say **W**, and obtain the independent component simply by

$$\boldsymbol{S}=\mathbf{W}\boldsymbol{X} \tag{3}$$

In ICA model, the independent components are restricted and assumed as having *non-Gaussian distribution*. And in most of classical statistical theory, random variables, here is the recorded signals which is mixed by the sound sources "randomly", are assumed to have Gaussian distributions. Meanwhile, the Central Limit Theorem, a classical result in probability theory, indicates that the distribution of a sum of independent random variables tends toward a Gaussian distribution, under certain conditions. Thus, the recorded signals usually have distributions that are closer to Gaussian than any of the original sound sources.

The problem of finding **A** becomes *maximizing the non-Gaussianity* of $\mathbf{w}^{\mathbf{T}}\boldsymbol{x}$, where $\mathbf{w}^{\mathbf{T}}$ denotes a row vector of **W**. In this paper, we use *negentropy* as a quantitative measure of non-Gaussianity of a random variable. The *entropy* of a discrete random variable $\boldsymbol{Y}$ can be defined as

$$H(Y)=-\sum_i P(Y=a_i)\log P(Y=a_i) \tag{4}$$

where the a_i are the possible values of *Y*. Meanwhile, the differential entropy *H* of a random vector $\boldsymbol{y}$ with density $f(\boldsymbol{y})$ can be defined as (Cover & Thomas, 1991; Papoulis, 1991):

$$H(\boldsymbol{y})=-\int f(\boldsymbol{y})\log f(\boldsymbol{y})\mathrm{d}\boldsymbol{y} \tag{5}$$

A fundamental result of information theory is that *a Gaussian variable has the largest entropy among all random variables of equal variance*. Therefore, the negentropy *J* can be expressed by entropy as

$$J(\boldsymbol{y})=H(\boldsymbol{y}_{\text{gauss}})-\mathrm{H}(\boldsymbol{y}) \tag{6}$$

Where $\boldsymbol{y}_{\text{gauss}}$ is a Gaussian random variable of the same covariance matrix as $\boldsymbol{y}$. The estimation of negentropy is difficult, we use approximation

$$J(\boldsymbol{y})\propto[E\{G(\boldsymbol{y})\}-E\{G(\boldsymbol{y}_{\text{gauss}})\}]^{\mathbf{2}} \tag{7}$$

where the function *G* is some non-quadratic function, *E* denotes the expected value of its arguments (Hyvärinen, 1998b).

2.1.2 FastICA Algorithm

The basic principles of ICA have been introduced in preceding section. To maximizing the contrast function in Eq. (7), FastICA, an efficient algorithm, is presented in this section. Before using FastICA, *centering* and *whitening* are preprocessing techniques to simplify the estimation and make it better conditioned.

Centering is making **x** a zero-mean variable, i.e. subtract its mean vector $\mathbf{m}=E\{\mathbf{x}\}$. After estimating matrix **A**, we need to add the mean vector of **s** back to the centered estimates of **s**. The mean vector is given by $\mathbf{A}^{-1}\mathbf{m}$. Meanwhile, whitening is making the components of **x** uncorrelated and their variances equal unity. One popular method for whitening is to use the eigenvalue decomposition (EVD) of the covariance matrix $E\{\mathbf{x}\mathbf{x}^{\mathrm{T}}\}=\mathbf{EDE}^{\mathrm{T}}$, where **E** is the orthogonal matrix of eigenvectors of $E\{\mathbf{x}\mathbf{x}^{\mathrm{T}}\}$ and **D** is the diagonal matrix of its eigenvalues, $\mathbf{D}=\mathrm{diag}(d_1, d_2, \ldots, d_n)$. Whitening signals can now be done by

$$\tilde{\mathbf{x}}=\mathbf{E}\mathbf{D}^{-1/2}\mathbf{E}^{\mathbf{T}}\mathbf{x} \tag{8}$$

where $\mathbf{D}=\mathrm{diag}(d_1^{-1/2}, d_2^{-1/2}\ldots, d_n^{-1/2})$.

Begin with estimating one weight vector $\mathbf{w}^{\mathrm{T}}\mathbf{x}$, i.e. calculating one independent sound sources, the FastICA algorithem is organized as:

1. Choose an initial (e.g. random) weight vector **w**.
2. Let $\mathbf{w}^{+}=E\{\mathbf{x}g(\mathbf{w}^{\mathbf{T}}\mathbf{x})\}-E\{g'(\mathbf{w}^{\mathbf{T}}\mathbf{x})\}\mathbf{w}$.
3. Let $\mathbf{w}=\mathbf{w}^{+}/\|\mathbf{w}^{+}\|$.
4. If not converged, go back to 2.

Above process is based on a fixed-point iteration scheme for finding a maximum of the non-Gaussianity. And *g* ($g(u)=u^3$ in this paper's algorithm) is the derivative of the function *G* used in Eq. (7). The convergence means that the dot-product of the old and new values of **w** equal to 1. We don't cover the concrete deduction of FastICA in this paper, the details can be get from the paper by A. Hyvärinen and E. Oja [6]. Then, applying above process on every weight vectors $\mathbf{w}_i$. To prevent the vectors from converging to the same maxima, we need *decorrelate* the outputs $\mathbf{w}_1^{\mathrm{T}}\mathbf{x}, \ldots, \mathbf{w}_n^{\mathrm{T}}\mathbf{x}$ after every iteration. The decorrelation method in this paper is

$$\mathbf{w}_p=\mathbf{w}_p-\sum_{j=1}^{p-1}(\mathbf{w}_p^{\mathrm{T}}\mathbf{w}_j)\,\mathbf{w}_j \tag{9}$$

In sum, applying FastICA on a set of mixing time signals, we can get the separated sound sources under determined and overdetermined conditions.

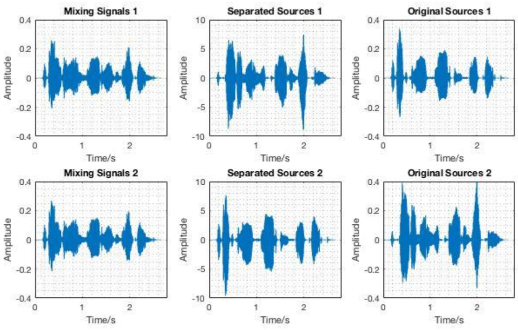

Figure 1. The results of applying FastICA on two mixing time signals, which are composited by two male speeches in random ratio, respectively.

The two original sound sources in Fig. 1 are two speeches spoken by different male, respectively. And the curves in Figure. 1 are the samples of the signals in time domain. We can find that the curves of separated sources are similar with the original sources, except different in the magnitude of amplitude. In next section, we will present how to develop ICA model in monaural conditions.

2.2 Monaural

2.2.1 ISA

Based on the concept of reducing redundancy in time-frequency representations of signals, ISA represents sound sources as low dimensional subspaces in the time-frequency plane. It extends ICA by identifying independent multicomponent source subspaces as input vectors.

First, the audio data is mapped to spectrogram by Short Time Fourier Transform (STFT). The computed spectrogram is an n by m matrix stored magnitude information $\boldsymbol{X}$ and phase information $\boldsymbol{\Phi}$. To extracted subspaces from the spectrogram, the transposed spectrogram $\boldsymbol{X}^T$ is calculated by *Singular Value Decomposition* to get the eigenvalue decomposition of the covariance matrix of $\boldsymbol{X}^T$. The deductions are given by

$$\boldsymbol{X}^T=\boldsymbol{U}\cdot\boldsymbol{D}\cdot\boldsymbol{V}^T \tag{10}$$

where the computed diagonal matrix $\boldsymbol{D}$ stores a set of singular values in decreasing order, two orthogonal matrices $\boldsymbol{U} = (\boldsymbol{u}_1, \ldots, \boldsymbol{u}_m)$ and $\mathrm{V} = (\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n)$ equal to the eigenvectors of $\boldsymbol{XX}^T$ and $\boldsymbol{X}^T\boldsymbol{X}$, respectively. The singular basis vectors (e.g. $\boldsymbol{u}_m$ and $\boldsymbol{v}_n$) are linearly independent.

The singular values in $\boldsymbol{D}$ represent the standard deviations of the principal components of $\boldsymbol{X}$. These standard deviations are proportional to the amount of information contained in the corresponding principal components. Therefore, the subspace of the $\boldsymbol{X}$ can be obtained by

$$\overline{\boldsymbol{X}}=\overline{\boldsymbol{D}}\cdot\boldsymbol{V}^{\mathbf{T}}\cdot\boldsymbol{X} \tag{11}$$

where $\overline{\boldsymbol{D}}$ is a submatrix containing the upper d rows of $\boldsymbol{D}$. As we can see from the Eq. (11), the number of d determines the amount of information to keep remains. However, there is a trade-off between the amount of information to retain and the reconcilability of the resulting features. Usually, a threshold ϕ is set for the estimation of d as

$$\frac{1}{\sum_{i=1}^{m}\sigma_i}\sum_{i=1}^{d}\sigma_i \geq \phi \tag{12}$$

where σ_i is the i^{th} singular value of matrix $\boldsymbol{D}$. Based on their experiments, *M. A. Casey and A. Westner*'s [7] indicates that the value of ϕ is 0.7 when extracts string quartet.

Then, the reduced rank spectrogram $\overline{\boldsymbol{X}}$ can be interpreted as an observation matrix, where each column is regarded as realizations of a single observation. By performing FastICA algorithm, as discussed in Section 2.1, we can get the mixing matrix $\boldsymbol{A}$. The pseudo-inverse $\boldsymbol{A}^{-1}$ represents the unmixed matrix, correspondingly. Thus, the independent temporal amplitude envelops $\boldsymbol{E}$ can be computed as

$$\boldsymbol{E} = \boldsymbol{A}^{-1}\cdot\overline{\boldsymbol{X}} \tag{13}$$

The estimation of the independent frequency weights $\boldsymbol{F}$ is achieved by

$$\boldsymbol{F}^{-1} = \boldsymbol{A}^{-1}\cdot\boldsymbol{T} \tag{14}$$

Finally, the independent sound sources' spectrograms are computed by multiplying one column of $\boldsymbol{F}$ with the corresponding row in $\boldsymbol{E}$,

$$\boldsymbol{S}_c = \boldsymbol{F}_{u,c}\cdot\ \boldsymbol{E}_{c,v} \tag{15}$$

where u=1, …, n, v=1, …, m and c=1, …, d. Combined with mixing signals' phase information $\boldsymbol{\Phi}$ directly, the time signals of the separated sources are gained by inverse STFT of each spectrograms $\boldsymbol{S}_c$.

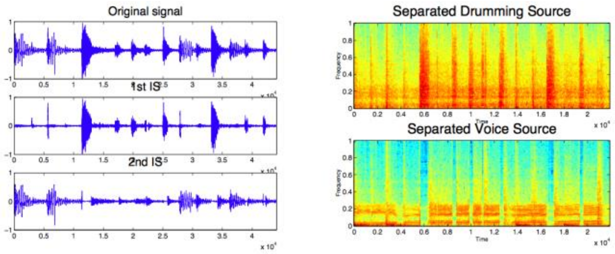

Figure 2. The time and frequency domain results of applying non-stationary ISA model on voice-drum mixing signals presented by S. Dubnov [21].

The discussion above has the assumption that the frequency basis functions are invariant which means that no pitch or frequency structure changes are possible for the separated sound sources. However, such assumption is too week to deal with the practice. To solve the separation of sound sources with non-stationary spectrogram. The mixing signals' spectrogram is decomposed into blocks of frames, each block having a unique subspace decomposition. Then, independent components which belongs to same source but in different block are identified and group together depending on the similarity of probability density functions. Traditional clustering methods are presented by Casey and Westner [7] and Dubnov [15].

2.2.2 DRNN

In this part, we try to use deep neural networks to solve the speech separation problem in spectral domain. Assume $x(t)$ is the training input at time t, and $y'_1(t)$, $y'_2(t)$ are the prediction output of the network, corresponding to the first and second channels, respectively. Meanwhile, in an RNN model, the output of the n^{th} hidden layer can be expressed as $\boldsymbol{h}^{n}=\mathrm{f}(\mathbf{W}^{n}\cdot\boldsymbol{h}^{n-1}(x(t))+b^{n}+\mathbf{U}^{n}\cdot\boldsymbol{h}^{n-1}(x(t-1)))$, where $\mathbf{W}$ and $\mathbf{U}$ are weights and $\boldsymbol{b}$ is the bias. In a DNN model, the weight matrix $\mathbf{U}$ equals to zero.

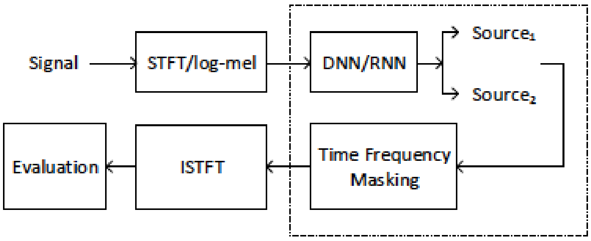

Figure 3. Framework of DRNN

In our project, we simplified this problem. We turned $x(t)$ to the mixture of clean voice and noise, denoted as $x(t) = x_{\text{clean}}(t) + x_{\text{noise}}(t)$). Correspondingly, $y'_1(t)$ means the prediction of clean voice, and $y'_2(t)$ means the prediction of noise. Usually, the noise is not required, so we only need to use the neural network to predict the output of clean voice.

The input of the neural network is the magnitude spectrogram of $x(t)$. The outputs of the neural network are the magnitude spectrograms of $y'_1(t)$ and $y'_2(t)$ and we only retain the isolated voice signal($y'_1(t)$). Then a frequency mask is implemented to the outcome and then we get the estimated spectrogram of $y'_1(t)$.

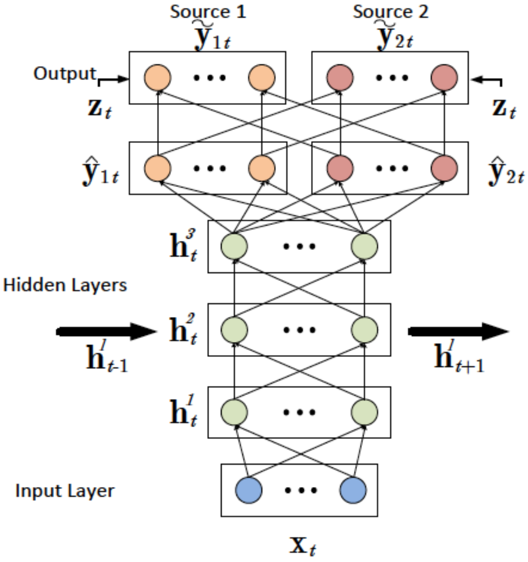

Figure 4. Architecture of RNN model

3. IMPROVEMENT ON DRNN

As we have turned the input into the mixture of clean voice and noise. We can use denoising autoencoder to help us solve the problem more effectively.

In this part, we turn some of the original data values to 0 to corrupt the data as the input $x'(t)$. The output of the network $y(t)$ is the reconstruction of the original signal $x(t)$. On the contrary, the loss function is calculated by comparing the output values $y(t)$ with the original input $x(t)$, not the corrupted output $x'(t)$.

4. EVALUATION

4.1 Dataset and Evaluation Measures

The noisy speech corpus NOIZEUS, used in our project evaluation, was developed to facilitate comparison of speech enhancement algorithms. The noisy database contains 30 IEEE sentences (produced by three male and three female speakers) corrupted by eight different real-world noises at different SNRs. The noise was taken from the AURORA database and includes suburban train noise, babble, car, exhibition hall, restaurant, street, airport and train-station noise. The sentences were originally sampled at 25 kHz and down-sampled to 8 kHz. The IRS filter is independently applied to the clean and noise signals. A noise segment of the same length as the speech signal is randomly cut out of the noise recordings, appropriately scaled to reach the desired SNR level and finally added to the filtered clean speech signal.

The inputs of the proposed ICA algorithms are the mixing of the clean voice in the database. Two voices are mixed in a random way (e.g. two signals in time domain are multiplied by a random matrix). Changing the number of mixing inputs, we investigate the performance of ICA on speech voice separation. Moreover, the proposed DRNN method is evaluated by the noisy speech directly.

4.2 Results

SAR	M_1-M_1	F_1-F_1	M_1-M_2	F_1-F_2	M_1-F_1
2	41.6266	39.3103	67.1854	39.3073	55.2952
3	41.5690	39.2454	67.1785	39.3059	55.6942
4	41.5636	38.1494	67.1784	39.3059	55.6942

Table 1. The average SAR evaluation of FastICA Algorithm.

In Table. 1, the rows represent the number of channels that is 2, 3 and 4, respectively. M_i and F_i (i=1, 2) represents the speaker. The table indicates that the separation results of two human voices. There are five clean speeches for each person in the datasets. Instead of providing all three valuation criterions SAR (*Sources to Artifacts Ratio*), SDR (*Source to Distortion Ratio*) and SIR (*Source to Interferences Ratio*), the results in Table. 1 only presents SAR because the results have same SAR and SDR value and SIR is positive infinity for all data. Therefore, we can know that every separation has very high SAR, which means the separation results are significantly good. Moreover, the number of mixing inputs has little influence on the separation results. In addition, as changing the voice spoken by different speaker, there is no obvious indication that the timbre will affect the separation result. However, our evaluation need more diversity speech separation data, such as daily conversation.

While for the ISA model, we try to build corresponding algorithm as indicated in [7]. However, our algorithms didn't achieve good separation results for lack of effective clustering methods. So we use the experiment results presented by Virtanen to give an overview of this method [22]. Virtanen implemented it on pitched instrument and drum sound separation. The test signals contain several sound production mechanisms, variety of spectra, and also modulations, such as vibrato. Mixture signals were generated by choosing a random number of pitched instrument sources and a random number of drum sources. And each source was scaled to obtain a random total energy between 0 and 20dB. The total number of test mixtures was 300. The results indicate that about 3.6dB overall signal to noise ratio can be achieved by ISA model. While ISA does provide an effective means of separating sound mixtures in monaural conditions, it also have some limitations. Combining *Principle Component Reduction* (e.g. the process of building subspaces in the preceding discussion) and ICA, ISA not only makes use of the properties of each method but also retains the problems associated with each method. First is ICA's indeterminacy with regards to source ordering and scaling. In addition, the variance-based nature of PCA inherently biases the analysis towards sources of high amplitude, which can make it difficult to recover sources of low amplitude. What's more, the sound with clear frequency characteristics, such as the harmonic structure, will achieve better separation results.

At last, we apply the purposed DRNN model on the database. Ten noisy speech were tested and the average signal to distortion ration, source to interference ration and source to noise ration are 7.40, 12.51 and 7.56, respectively. Table 2 shows 3 evaluation outcomes of ten noisy speech. The SNR of these 3 examples are 3dB, 5dB and 5dB respectively.

Objects	SDR	SIR	SAR
1	4.19	2.89	8.95
2	7.11	12.81	7.40
3	7.27	11.51	7.41

Table 2. SDR, SIR and SAR evaluation of DRNN

5. CONCLUSION

In this paper, we introduce two traditional methods, ICA and ISA for multi-channel and monaural condition speech separation, respectively. The ICA model shows significant clear speech separation. Meanwhile, the number of input mixing, that is the number of channel in some degree, won't influence the separation results clearly. And the differences in the timbre of speech won't affect the separation results as well. For ISA model, we introduce the basic theory and give a brief view on the pitched instruments and drum separation performance. Finally, we investigate the state of the art deep learning model, the DRNN, on noisy speech separation. It achieves better results than other linear models. However, the DRNN method requires many training data to train to extract the features ahead of time. Combining the methods presented in this paper, we can solve the cocktail problem in some specific situations.

6. REFERENCES

[1] O. Vinyals, S. V. Ravuri, and D. Povey, "Revisiting recurrent neural networks for robust ASR," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 4085–4088, 2012.

[2] Ö. Yilmaz and S. Rickard, "Blind separation of Speech Mixtures via Time-Frequency Masking," *IEEE Transactions On Signal Processing*, vol. 52, no. 7, 2004.

[3] J. F. Cardoso and A. Souloumiac, "Blind beamforming for non-Gaussian signals," *IEE Proceedings-F*, vol. 140, no. 6, 1993.

[4] P. Comon, "Independent component analysis, a new concept?" *Signal Process.*, vol. 36, no. 3, pp. 287–314, 1994.

[5] P. Smaragdis, "Blind separation of convolved mixtures in the frequency domain," *Neurocomputing*, vol. 22, pp. 21–34, 1998.

[6] A. Hyvärinen and E. Oja, "Independent component analysis: algorithms and applications," *Neural Networks*, vol 13, pp 411-430, 2000.

[7] 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.

[8] J. C. Brown and P. Smaragdis, "Independent component analysis for automatic note extraction from musical trills," *J. Acoust. Soc. Amer.*, vol. 115, pp. 2295–2306, May 2004.

[9] I. Orife, "A rhythm analysis and decomposition tool based on independent subspace analysis," *Master's thesis*, Dartmouth College, Hanover, NH, 2001.

[10] M. A. Casey, "MPEG-7 sound-recognition tools," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 11, no. 6, pp. 737–747, Jun. 2001.

[11] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 32, no. 6, pp. 1109–1121, Dec. 1984.

[12] Y. H. Yang, “Low-rank representation of both singing voice and music accompaniment via learned dictionaries,” in *Proceedings of the 14th International Society for Music Information Retrieval Conference (ISMIR)*, 2013.

[13] D. D. Lee and H. S. Seung, “Learning the parts of objects by non-negative matrix factorization,” *Nature*, vol 401, no 6775, pp 788-789, 1999.

[14] Y. Wang, “Supervised speech separation using deep neural networks,” *Doctoral thesis*, 2015

[15] P.-S. Huang, M. Kim, M. Hasegawa-Johnson, and P. Smaragdis, “Deep learning for monaural speech separation,” *in Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), pp. 1562–1566, 2014.

[16] P.-S. Huang and M. Kim and M. Hasegawa-Johnson and P. Smaragdis, “Singing-voice separation from monaural recordings using deep recurrent neural networks,” *in Proceedings of the 15th International Society for Music Information Retrieval* (*ISMIR*), 2014.

[17] P.-S. Huang and M. Kim and M. Hasegawa-Johnson and P. Smaragdis, “Joint optimization of masks and deep neural networks for monaural source separation,” *Ieee/Acm Transactions On Audio, Speech, And Language Processing*, vol. 23, no. 12, 2015.

[18] X. Lu, Y. Tsao, S. Matsuda and Chiori Hori, “Speech enhancement based on deep denoising autoencoder,” *Interspeech*, 2013.

[19] Y. Hu and P. C. Loizou, “Subjective comparison and evaluation of speech enhancement algorithms,” *Speech Communication*, vol 49, pp 588-601, 2007.

[20] P. G. Shivakumar and P. Georgiou, “Perception optimized deep denoising autoencoder for speech enhancement,” *Interspeech*, 2016.

[21] S. Dubnov, “Extracting sound objects by independent subspace analysis,” *ASE 22nd International Conference on Virtual, Synthetic and Entertainment Audio*, 2002.

[22] T. Virtanen, “Monaural sound source separation by non-negative matrix factorization with temporal continuity and sparseness criterion,” *IEEE Transactions On Audio, Speech, And Language Processing*, vol. 15, no. 3, 2007.