Time-Frequency Masks for Monaural Speech Separation: A Comparative Review

Belhedi Wiem[#1], Ben Messaoud Mohamed anouar[#2], Bouzid Aicha[#3]
[#]*Electric Department, National School of Engineers*
Le Belvédère BP.37, 1002 Tunis, Tunisia
[1]bel.hedi.wiem@gmail.com
[2]anouar.benmessaoud@yahoo.fr
[3]bouzidacha@yahoo.fr

***Abstract*— In this paper we present a comparative analysis of different time-frequency (T-F) masking techniques used for single channel speech separation (SCSS). We survey T-F masking concept and compare different types of masks in different criteria. The comparison is conduct theoretically by mathematical study and numerically by objective and subjective assessment. Also, we study the effect of the masking techniques on the perceptual quality of speech and their ability to separate a target speech from monaural mixing.**

***Keywords*— Time-Frequency Masking, Wiener Mask, Binary Mask, DNN-based Mask, Soft Mask, Sinusoidal Mask, Ideal Binary Mask, Speech Separation.**

I. INTRODUCTION

Previous studies who treat speech separation have divided approaches on three essential steps which namely are: Computational auditory scene analysis (CASA) based-approaches, Subspace decomposition (SD) based-approaches and Model-based approaches.

A CASA-based approach aims to determine the discriminative characteristics of the observed mixture in order to separate the desired speaker. CASA aims to extract psycho-acoustic indices from a mixture. In the CASA process, the input mixture is first decomposed into time-frequency cells in order to determine the regions dominated by the target. However, as the pitch is a key tool in the CASA process, this makes it more suitable for voiced regions of speech.

The main goal of subspace decomposition approaches is to determine an adequate base to re-represent a composite signal. The main goal of subspace decomposition approaches is to determine an adequate base to re-represent a composite signal. The mixture is then projected into new subspace in which undesired components are illuminated.

In model-based approaches a number of parameters are determined in order to establish a viable model that best describes the speech. To do so, deep statistical computations and fractional research through leant dictionaries are operated. Thus Model-based approaches lead to significantly complex mixture estimator algorithms thus they are difficult to implement in real time systems.

Most algorithms that deal with single channel speech separation are also based on masking. Therefore, this paper focuses on studying different types of masks and their effect on speech quality. The rest of the paper is structured as follows. The next section gives a general overview about time-frequency masking concept. Section 3 reviews different mask types existing in literature. Section 4 is devoted to discussion. Finally, section 5 concludes the paper.

II. TIME-FREQUENCY MASKING CONCEPT

The masking technique is built on the fact that a dominant speaker (also known as desired speaker) can, in specific bands, hide a less "powerful" speaker. Then the roles of speakers can be reversed between desired and masker, it depends on the frequency band to be treated. Masking types could be classified into two broad categories: informational masks [1] and energetic masks [2]. The main differences between these two categories is that in the first, the decision of maintaining or eliminating a signal is energy. In addition, energetic masking is when a signal is dominated by a stronger one which makes it non audible. However in the informational masking case, both speakers are audible but we cannot distinguish one from the other. As pioneering work, Time-Frequency masking technique was first introduced by Wang in [3] as major computational goal of CASA. Time-frequency masks have been of great effectiveness in various speech processing researches in order to suppress unwanted energy. The unwanted energy could be either a concurrent talker, in the monaural speech separation case, or a noise which could be present in the original signal or an artifact.

978-1-5090-4712-3/16/$31.00 ©2016 IEEE

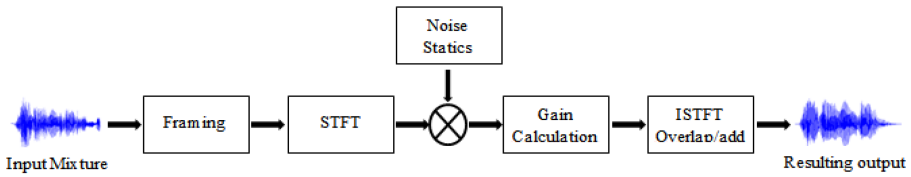

Fig. 1. Wiener filtering framework

In literature, various approaches are either based on masking or use it as a complimentary tool. It has been widely used in speech enhancement [4], speech separation [5][6] as well as denoising [7].

III. Overview and Comparison

In this section, an overview of several T-F masking techniques is given. The methods included in our study are Wiener mask [8,9], Binary mask[17,18] and its optimal version which is namely Ideal Binary mask [3], Soft mask [6] Sinusoidal mask [15] and DNN-based mask [20][21].

A. *Wiener Mask*

Wiener filter is a widely used tool in speech processing. It has been much used either as the principal tool or added to other tools as complementary device. In [8], Wiener filter has been extended and used for wavelet denoising for image. It estimates wavelet shrinkage as a means to design Wiener filter that operates in wavelet-domain. It improves the mean square error (MSE) of thresholding by estimating each wavelet coefficient using the Wiener filter. In [9] Wiener filter has been used for speech enhancement in one of the most difficult cases which is real environment. In this approach Wiener filtering and spectral subtraction cooperate, in multi-channel case, to deal with low and high frequencies, respectively. It allows the reduction of musical noise effectively. Wiener filtering has also proved its effectiveness in speech separation. In this context Benaroya et al. developed a source separation approach based on Wiener filter, hidden Markov model (HMM) and Gaussian mixture models (GMM) [10]. This approach is a smooth adaptive Wiener filter which runs mainly on two stages, the first involves training of the model parameters and the second is a re-estimation and separation phase using HMM and GMM. It gives better performance compared to standard Wiener filter.

Here we give a brief description of Wiener filter operating principle. Let the input mixture y(n) be as shown in following equation:

$$s(n) = x(n) + y(n) \tag{1}$$

such as x(n) is the target speech and y(n) is either a concurrent speaker or noise . The problem that arises is how to recover x(n) from y(n). One solution is to filter y(n) such that the output $\hat{x}(n)$ is as close as possible of x(n). One can measure the quality of the estimation defined by:

$$e(n) = x(n) - \hat{x}(n) \tag{2}$$

Obviously, the more e(n), the lower the estimation will be good. Therefore we look for a filter that minimizes the error. It is convenient to seek to minimize $e^2(n)$ because it is a quadratic function easily differentiable. Furthermore, since the interesting signals are random, the cost function to be minimized is MSE defined by:

$$\zeta(n) = E(e^2(n)) \tag{3}$$

The optimal Wiener filter corresponds to the filter that minimizes the minimal square error (MSE). It is based on a statistical approach. Unlike ordinary filters that work for a well-defined frequency, the optimal Wiener filter performs differently. It requires information about spectral properties about the original signal and about the noise and searches for the output that is close to the original signal. Thus the criterion used is signal to noise ratio (SNR) that can be expressed as:

$$SNR = \frac{P_x(\omega)}{P_y(\omega)} \tag{4}$$

Where P_x and P_y are power spectral density of clean and noisy speech, respectively.

The Wiener filter can be interpreted as T-F mask where T-F cells of the mask represent the SNR. Wiener mask can be expressed as:

$$M_{Wiener} = \begin{cases} 1-\dfrac{1}{SNR} & \text{if } 1-\dfrac{1}{SNR} \succ 0 \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

The Figure1 illustrates the principle of Wiener filtering.

B. *Binary Mask and Ideal Binary Mask*

Binary masking is the most common way to retrieve the target speaker by comparing the two subspaces [17, 18]. In fact, it was of great effectiveness in CASA-based approaches as well as in subspace decomposition approaches to decompose the input mixture into low rank and sparse subspaces.

Theoretically, when prior information about the target and the intrusion is given, the optimal binary mask can be determined. It is known as the ideal binary mask (IBM) [5] in which speech separation becomes a binary classification problem. To estimate the IBM, the spectrum of both mixture (masker) and target are first decomposed in spectral domain using either a short-time Fourier transform (STFT) or a Gammatone bank filter. Energy is calculated along the time domain. Then the IBM is obtained by comparing the SNR at each T-F unit. As denoted in the equation below, when the SNR within a T-F

unit overcomes the local criterion (LC), the unit is assigned the value of 1. Otherwise, the value attributed to the unit is 0.

$$M_{IBM}(t,k)=\begin{cases}1 \text{ if SNR(t,k)>LC}\\ otherwise\end{cases} \quad (6)$$

such as t is the time index and k is the frequency index. LC is typically chosen to be 0 dB.
In the synthesis phase, T-F matrix is applied with the binary value to the original mixture as shown in the next equation.

$$\hat{S}(t,k)=S(t,k)*M_{IBM}(t,k) \quad (7)$$

IBM separation leads to large speech intelligibly improvements; Improvement for stationary noise is about 7 dB in normal hearing listener [12] and about 9 dB for hearing-impaired [13].

C. Soft Mask

Soft mask identifies the contribution of speaker *x* or *y* to the mixture by the use of probability calculation. In fact, unlike hard masking, deciding whether a speech component belongs to the speaker *x* or *y* is not binary. It is conditioned and assigned by a probability-based estimation.
Estimating soft mask can make the use of statistical calculations or fuzzy logic.
Calculating the probability of maintaining or rejecting a T-F cell of the mixture is equivalent to find weighting factors in which frames were mixed. We retain the same speech model of Equation 1.
Thus the probability that the segment of s belongs to x, in the frame *i*, is the probability that x is greater than y in the same frame. Hence soft mask can be expressed by the following equation:

$$M_{soft}=\begin{cases}P(\mathrm{x}_i=\mathrm{s}_i\,|\,\mathrm{s}) \text{ if } \mathrm{x}\rangle y\\ 0 \text{ otherwise}\end{cases} \quad (8)$$

D. Sinusoidal Mask

Sinusoidal mask is a binary mask which is projected in the sinusoidal space and improved with the Wiener filter. To estimate this mask we must first extract, from the input mixing, the constituents of the sinusoidal subspace and that amplitude, frequency and phase. The general idea is to train codebooks of sinusoidal amplitude-frequency for speaker-dependent scenario, and decode the codewords (states) via comparing the mixture approximation with the observation, and apply the codebook to reconstruct the spectral amplitude. The sinusoidal model used is first introduced in [14] such as:

$$s(t)=\alpha_1 x(t)+\alpha_2 y(t)=\sum_{l=1}^{L(t)} A_l(t)\exp(j\varphi_l(t)) \quad (9)$$

where α_1 and α_2 are the gains of target concurrent speakers, $A_l(t)$ and $\varphi_l(t)$ are the amplitude and phase of the l^{th} sine wave in the frequency $\omega_l(t)$.
The major difference between the binary mask and the sinusoidal mask is that the second one relies on amplitudes for decisions; it selects the peak of the highest amplitude per band. The sinusoidal mask is established in each frequency ω_l component as follow:

$$M_{\mathrm{Sin}}(\omega_l)=\begin{cases}1 \text{ if } A_{1,l}\geq A_{2,l}\\ 0 \text{ otherwise}\end{cases} \quad (10)$$

In order to reach a better separation performance, Wiener filter is used as constrained optimization. We define $g1(\omega_l)$ as the sinusoidal gain function. The challenge of speech separation is resumed to find $g_1(\omega_l)$ that solves the constrained problem of minimizing the cross-talk without causing speech distortion. As detailed in [15], $g_1(\omega_l)$ can be expressed as follow:

$$g_1(\omega_l)=\frac{\alpha_1 X(\omega_l)}{\sqrt{\alpha_1 X^2(\omega_l)+\alpha_2 Y^2(\omega_l)}} \quad (11)$$

E. DNN-based Mask

With the Deep Neural Network (DNN) speech separation is turned from target estimation to a classification problem [20-22]. In fact, DNN-based masks are a novel type of approaches that apply deep learning models to predict a T-F mask for one of the sources. Generally, incorporating T-F context in classification takes place over two DNN- training stages; a convolutional neural DNN is employed to learn the ideal binary mask for a two-speaker speech separation problem. In fact, selected speech segments are used as training set to build a model used for the test set. From the model a probabilistic predictions of the ideal binary mask are obtained.
For each T-F element, of each source, the mean prediction is computed and applied a confidence threshold (α); two different masks were designed; a binary mask for the male speaker M^{male}_{DNN} and a binary mask for the female speaker M^{female}_{DNN} which are respectively:

$$M^{male}_{DNN}(t,f)=\begin{cases}1 \text{ if } \frac{1}{T}\sum_{i=0}^{T} S_{t+i,f}>\alpha\\ 0 \text{ otherwise}\end{cases} \quad (12)$$

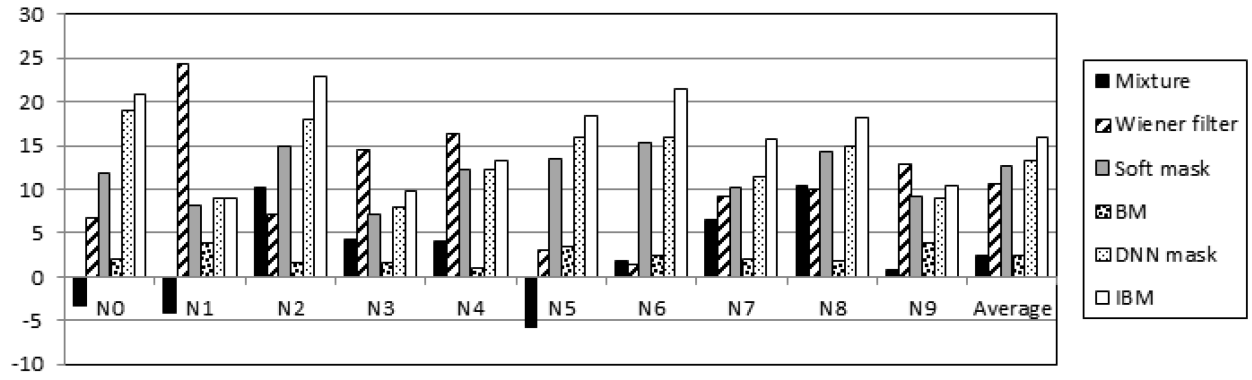

Fig. 2. SNR results of T-F approaches. Mixtures are composed of speech and ten different types of intrusion (N0:1khz /N1: random noise /N2: noise bursts /N3: "cocktail party" /N4:rock music /N5: siren /N6: trill telephone /N7: female speaker /N8: male speaker /N9: female speaker

and

$$M_{DNN}^{female}(t,f)=\begin{cases}1 \text{ if } \frac{1}{T}\sum_{i=0}^{T} S_{t+i,f} < (1-\alpha) \\ 0 \text{ otherwise}\end{cases} \quad (13)$$

Such as T is window size, t is the time index, f is the frequency bin and S is the mixture spectrogram.

IV. Discussion

The average SNR [18] values are given in figure 2. From this figure, it is observed that IBM mask give better results in almost all cases and the DNN-mask approach the limits of the IBM. In contrast, binary mask that gives a limited separation performance in most of the test cases. It is also shown that soft mask performs better than Wiener mask for the 1Khz tone (N0), the noise bursts (N2), the siren (N5), the trill telephone (N6), the female speaker (N7) and the male speaker (N8) and it performs worst for the random noise (N1), the cocktail party noise (N3), the rock music and the female speaker (N9).

As SNR does not distinguish amplification distortion and attenuation distortion, the separation performance for each technique is reported in terms PESQ [19] in order to conduct a meaningful comparison. Averaged PESQ are given in figure 3. The PESQ results are in harmony with those of SNR.

Based on the previous study, we can notice that Wiener filter has fixed frequency response at all frequencies and also needs an estimation of the power spectral density of clean and noisy speech prior to filtering. This was overcome by the optimal Wiener filter which requires information about the noise as well as original signal and it tends to minimize the MSE to find the output that is more close to the original signal. The fact of being adaptive makes the optimal Wiener filter very robust in noisy mixture but still enable to achieve SCSS. Also, providing informations about the noise is not always the case in real-time applications.

IBM in the speech separation problem is transformed into a binary classification problem and it greatly improves speech intelligibility; for stationary noise improvement is more than 7 dB for normal hearing listeners and more than 9 dB for hearing-impaired listeners. Furthermore, improvement is even more significant for modulated noise than stationary noise [16]. In addition, to determine the IBM, the source signals must be available which could not be the case in most of real-time application. The results provided by this approach are therefore ideal data treated as a reference to evaluate separation and enhancement approaches. Binary mask is the generalized form of IBM and it is used when no prior information about the target and the noise are given. Compared to IBM, binary mask decreases speech intelligibility. In fact, decision of rejecting or maintaining a speech component cannot be done with certainty in most of the cases. Hence, perfect separation can be achieved if the desired signal is known which is not possible in real-time applications. However, soft masks attribute a probability that varies between 0 and 1 instead of a hard decision. Nevertheless soft masking is computationally expensive because it is mainly based on complex statistic calculations. Performing masks in the spectral domain have three major deficiencies due to the use of STFT; the first one is that some portion of the weaker speaker are relatively masked by the other speaker which causes speech distortion in desired signal. The second problem is that these masks are generally enabled to recover both speakers at the same time. The third problem is cross-talk and it happens when some segments of the concurrent speaker are still audible. Hence the importance of sinusoidal masks that operate in the sinusoidal subspace and rely on the amplitudes for making decision of maintaining or suppressing a speech portion. This strategy has shown a great effectiveness in improving speech separation performance.

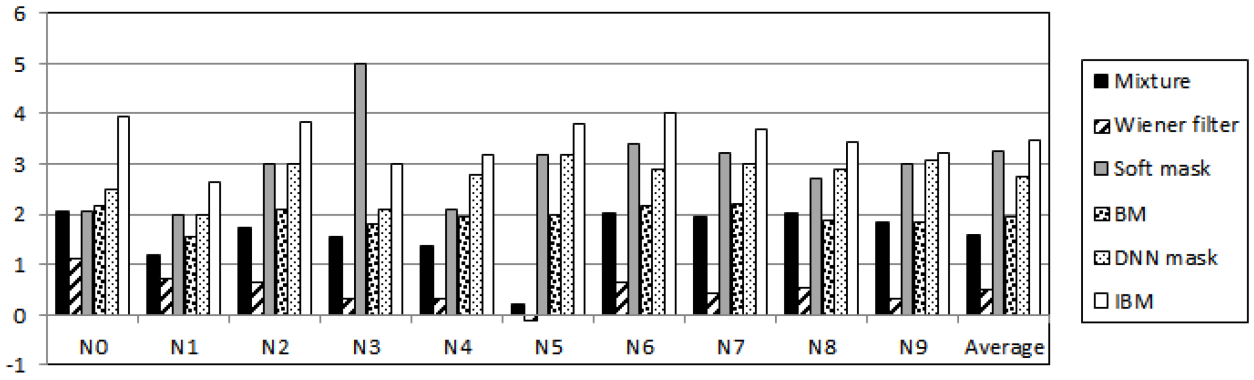

Fig. 3. PESQ results of studied approaches. Voiced speech is segregated from a mixture of speech and ten different intrusions as indicated in figure 2

From objective and objective evaluation, given in figure 2 and figure 3, we conclude that DNN-mask can approach the limits of the IBM separation performance. As DNN are inspired by brain functionalities and by the complex neural circuits, it is possible that the neural signal processing might be employed in the human auditory system. However, deep structure is hard to train from random initializations and unpredicted speech signals. This might be obstacle to the use in real-time applications.

Figure 4 illustrates the output spectrograms of each previous studied techniques tested on a noisy mixture as well as double speaker mixture of the Cooke database [11].

V. Conclusion

This paper presents a comparative review of different techniques of masking to reduce noise or for speech separation. Among the methods discussed, those based on Wiener filter, on binary masking, ideal binary masking, soft masking, DNN-mask and sinusoidal masking. This comparison was mainly conduct in terms objective and subjective measurements. The SNR and PESQ results show that T-F masking techniques can be remarkably successful at single channel speech separation. IBM gives the best speech separation results if certain knowledge about the target and noise are met. These requirements may not be available in real environment. Also DNN-based mask is capable of approaching the IBM separation performance. The only weakness of this kind of mask is that it requires a training phase which is of high computational complexity cost. Also, deep structure is hard to train from random initializations and unpredicted speech signals.

However with soft and sinusoidal masks, requirements are becoming ever less but still need trained dictionaries or statistical calculations which are computationally expensive as well. This may be an inconvenience in their implementation in real time applications. Binary masks have proven their effectiveness when preceded by subspace decomposition. The main advantage of binary mask is that it can be implemented very easily. Finally, Wiener mask performs better in the case of a noisy mixture and performs less in the case of double talk mixtures.

References

[1] Divenyi, P.: Speech Separation by Humans and Machines. Springer, (2004).

[2] Srinivasan . S., Wang, B.: A model for multitalker speech perception. The Journal of the Acoustical Society of America, vol. 124, no. 5, pp. 3213–3224 (2008).

[3] Wang, D.: On ideal binary mask as the computational goal of auditory scene analysis," in Speech Separation by Humans and Machines, edited by P. Divenyi (Kluwer Academic, Norwell, MA), pp. 181–197 (2005).

[4] Jeong, S. Y., Jeong, J. H., Oh, K. C.: Dominant speech enhancement based on SNR-adaptive soft mask filtering. IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, pp. 1317-1320 (2009).

[5] Belhedi Wiem and Ben Messaoud Mohamed anouar and Bouzid Aicha. Single channel speech separation based on PCA and Fuzzy logic. Neural, Parallel & Scientific Computations v24, pp 489-504, 2016

[6] Radfar, M. H., Dansereau, R. M.: Single-channel speech separation using soft mask filtering. IEEE Transactions on Audio, Speech, and Language Processing, vol.15(8), pp. 2299-2310 (2007).

[7] Yu-jing, T., Hong-wei, Z., He, L., De-sheng, W.: 0 Notice of Retraction A new algorithm of the wavelet packet speech denoising based on masking perception model. Seventh International Conference on Natural Computation (ICNC), vol. 1, pp. 33-37(2011).

[8] Ghael, S. P., Sayeed, A. M., Baraniuk, R. G. : Improved wavelet denoising via empirical Wiener filtering. International Society for Optics and Photonics , In Optical Science, Engineering and Instrumentation, pp. 389-399 (1997).

[9] Meyer, J., Simmer, K. U.: Multi-channel speech enhancement in a car environment using Wiener filtering and spectral subtraction. IEEE International Conference on Acoustics, Speech, and Signal Processing, 1997. ICASSP-97, vol. 2, pp. 1167-1170 (1997).

[10] Benaroya, L., Bimbot, F.: Wiener based source separation with HMM/GMM using a single sensor. In Proc. ICA, pp. 957-96 (. (200).

[11] M.P. Cooke, J. Barker, S.P. Cunningham, and X. Shao.: An audiovisual corpus for speech perception and automatic speech recognition. The Journal of the Acoustical Society of America, vol. 120, no. 5, pp.2421–2424(2006).

[12] Brungart, D. S.: Informational and energetic masking effects in the perception of two simultaneous talkers. The Journal of the Acoustical Society of America, vol. 109(3), pp. 1101-1109 (2001).

[13] Anzalone, M. C., Calandruccio, L., Doherty, K. A., Carney, L. H. (2006). Determination of the potential benefit of time-frequency gain manipulation. Ear and hearing, vol. 27(5), pp.480 (2006).

[14] McAulay, R. J., & Quatieri, T. F.: Speech analysis/synthesis based on a sinusoidal representation. IEEE Transactions on Acoustics, Speech and Signal Processing, vol. 34(4), pp. 744-754 (1986).

[15] Mowlaee, P., Christensen, M. G., Jensen, S. H.: Sinusoidal masks for single channel speech separation. 2010 IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP), pp. 4262-4265 (2010).

[16] Li, N., & Loizou, P. C.: Factors influencing intelligibility of ideal binary-masked speech: Implications for noise reduction. The Journal of the Acoustical Society of America, vol. 123(3), pp. 1673-168 (2008)2.

[17] Liu, L., Ding, Z., Li, W., Wang, L., Liao, Q.: Speech enhancement via low-rank matrix decomposition and image based masking. 9th International Symposium on Chinese Spoken Language Processing (ISCSLP), pp. 389-392 (2014).

[18] Loizou, P.: Speech Enhancement: Theory and Practice, CRC Press, Boca Raton (2007).

[19] Rix. A.W., Beerends,. J. G., Hollier, M. P., Hekstra, A. P.: Perceptual evaluation of speech quality (PESQ)-a new method for speech qualityassessment of telephone networks and codecs. IEEE International Conference on Acoustics, Speech and Signal Processing, vol. 2, pp. 749–752 (2001).

[20] Simpson, A. J. (2015). Probabilistic Binary-Mask Cocktail-Party Source Separation in a Convolutional Deep Neural Network. *arXiv preprint arXiv:1503.06962*.

[21] Simpson, A. J. (2015). Deep Transform: Cocktail Party Source Separation via Probabilistic Re-Synthesis. arXiv preprint arXiv:1503.06046.

[22] Huang, P. S., Kim, M., Hasegawa-Johnson, M., & Smaragdis, P. (2014, May). Deep learning for monaural speech separation. In 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 1562-1566). IEEE.