Improved Phase Reconstruction in Single-Channel Speech Separation

Florian Mayer and Pejman Mowlaee

Signal Processing and Speech Communication Lab
Graz University of Technology, Austria
mayer@student.tugraz.at pejman.mowlaee@tugraz.at

Abstract

Conventional single-channel source separation (SCSS) algorithms are mostly focused on estimating the spectral amplitude of the underlying sources extracted from a mixture. The importance of phase information in source separation and its positive impact on improving the achievable performance is not adequately studied yet. In this work, we propose a phase estimation method to enhance the spectral phase of the underlying signals in SCSS framework. The proposed method relies on multi-pitch estimation and phase decomposition followed by applying temporal smoothing filters on the unwrapped mixture phase. We consider the combination of the proposed phase estimator with ideal binary mask and non-negative matrix factorization, as two well-known SCSS methods for separating the spectral amplitudes. Our results show that certain improvements in quality and intelligibility is achievable via replacing the mixture phase with the estimated one when reconstructing the sources.
Index Terms: Phase estimation, single-channel source separation, perceived quality, speech intelligibility.

1. Introduction

In many real-life speech applications, the clean speech signal is corrupted with some interfering sources and some background noise. While human beings perform excellent in separating sources in cocktail party situations, an ideal single-channel source separation (SCSS) performance is not yet reached using a machine. In this regard, finding new strategies to push the limited performance of the existing source separation methods is of high interest and is a quite challenging goal to reach.

Figure 1 shows the block diagram for conventional methods in SCSS where the main focus is on filtering the mixture in the spectral amplitude domain while employing the mixed phase for signal reconstruction. The phase information contributes in two ways in SCSS: i) in the spectral amplitude estimation and ii) at signal reconstruction. In [1], the importance of spectral phase difference information between the underlying sources on the resulting spectral amplitude estimator was studied. It was shown that the exact knowledge about the phase difference leads to a more accurate mixture approximation compared to the phase-averaged ones (log-max or power MMSE) in [2].

For signal reconstruction, the importance of phase was first studied in [3] where a geometry-based approach was presented and provided phase estimate up to an ambiguity in the sign of phase difference. To resolve the phase ambiguity, auxiliary group delay constraint on the sources spectral phase was used. The signals using the estimated phase were shown to outperform the ideal binary mask (IBM) [4] using the mixed phase and spectrogram inversion proposed in [5]. Later, partial phase reconstruction [6] was proposed where iterative signal reconstruction first proposed by Griffin and Lim (GL) was applied to signal components lying above a threshold in Wiener filter gain of each source. Finally, the authors in [7] proposed consistent Wiener filter by adding the short-time Fourier transform (STFT) inconsistency and solving the constrained optimization. The success of these GL-based methods relies on a large overlapping frames as more redundant information is available between STFT amplitude and phase spectra. A detailed comparison and overview on GL-based phase estimation methods can be found in [8, 9] reporting a limited performance by the methods and the restriction due to their requirement for a rather accurate spectral amplitude estimates (see [10] for an overview study).

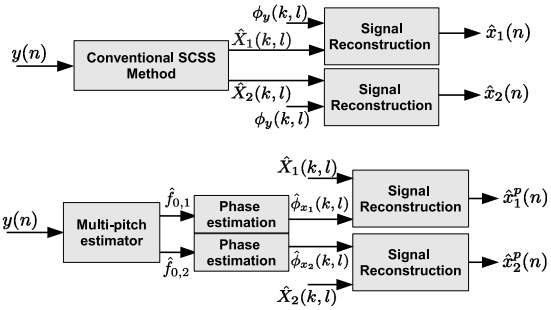

Figure 1: (Top) Conventional SCSS separating mixture in spectral amplitude versus (bottom) proposed method.

Following the success of applying phase decomposition and smoothing idea in single-channel speech enhancement in [11, 10], in this paper, we extend the idea to estimate phase of two underlying speakers in the mixture which relies on the multi-pitch estimation of the sources. Through phase decomposition, unwrapped phase for each source is obtained and temporal smoothing is applied to provide an enhanced spectral phase at signal reconstruction. As proof-of-concept, we report the effectiveness of the proposed method when combined with two SCSS methods: IBM and non-negative matrix factorization (NMF). Our results show that consistent joint improvement in separation quality and intelligibility is possible when replacing the mixture phase with the estimated one.

The rest of the paper is organized as follow. Section 2 presents an overview of previous SCSS methods. Section 3 presents the details of the proposed phase estimator, Section 4 presents the results and section 5 concludes on the work.

2. Conventional Separation Methods

2.1. Ideal Binary Mask (IBM)

Figure 1-(top) shows the block diagram for conventional SCSS. As our first example, we consider ideal binary mask method

The work was supported by Austrian Science Fund (P28070-N33).

10.21437/Interspeech.2015-416

which aims to perform an auditory scene analysis consisting of two steps: segmentation and grouping. The ideal binary mask determines the mask that provides the upper-bound separation performance for source-driven SCSS methods, also known as computational auditory scene analysis (CASA) [12] (For a recent overview see [13]). Let $y(n) = \sum_{j=1}^{2} x_j(n)$ where $x_j(n)$ denotes the underlying jth source in the mixture and n refers to time sample index. Taking the STFT, we have

$$Y(k,l)e^{j\phi_y(k,l)} = X_1(k,l)e^{j\phi_1(k,l)} + X_2(k,l)e^{j\phi_2(k,l)}, \quad (1)$$

where $Y(k,l)$ and $X_j(k,l)$ are the amplitude spectra for the mixture and the jth speaker with $j \in [1,2]$ while $\phi_y(k,l)$ is the mixture phase and $\phi_j(k,l)$ denotes the spectral phase of the jth speaker with k and l as the frequency and time index, respectively. The ideal binary mask for the first source is:

$$\hat{X}_1^{\text{ibm}}(k,l) = \begin{cases} Y(k,l) & X_1(k,l) \geq X_2(k,l) \\ 0 & \text{Otherwise} \end{cases}$$

We similarly define $\hat{X}_2^{\text{ibm}}(k,l)$ as the IBM-estimated spectral amplitude for the second source by swapping the role of the underlying sources. IBM is known to reach the highest intelligibility for the achievable separation by a SCSS method [12]. Therefore, as our first proof-of-concept, in this work, we report the effectiveness of the proposed phase estimation method (detailed in Section 3) on top of IBM. Any improvement on top of IBM is brought by the proposed phase estimation method.

2.2. Non-negative Matrix Factorization (NMF)

In contrast to source-driven SCSS methods described in previous Section, model-based SCSS methods have been of high popularity due to their top separation performance. They rely on prior knowledge in the form of pre-trained dictionaries of the spectral amplitude information, using some statistical modeling techniques including factorial hidden Markov models [14, 15], vector quantizers [16, 17], graphical models [2], Gaussian mixture models [18], non-negative matrix factorization (NMF) [19, 20, 21, 22], and deep models [23, 24]. As our second proof-of-concept in this work, we select active-set Newton algorithm for overcomplete NMF recently proposed in [19], as it outperforms other conventional source separation techniques.

The overcomplete NMF method relies on dictionaries learned for each source, characterizing the power (or magnitude) spectrum, represented as non-negative linear combinations of spectral components, selected as dictionary atom entries [19]. Let $\mathbf{x}_j$ as the time-domain signal for the jth source as observations of length F being modeled as a linear combination of basis vectors $\mathbf{b}_{m,j}$ to be selected from a dictionary $\mathbf{B}_j = [\mathbf{b}_{1,j} \cdots \mathbf{b}_{M,j}]$ where $\mathbf{B}_j$ is a matrix of size $F \times M$ with M as the number of atoms in the dictionary trained on the training set for the jth source and $m \in [1, M]$ as the atom index. The jth source estimate is then given by

$$\hat{\mathbf{x}}_j = \sum_{m=1}^{M} \mathbf{w}_j \mathbf{b}_{m,j}, \quad \text{subject to} \quad w_{m,j} \geq 0 \quad \forall m, \quad (2)$$

where $\mathbf{w}_j = [w_{1,j} \cdots w_{M,j}]$ contains M non-negative weights. In a vectorized format, the jth individual observed source is modeled as

$$\hat{\mathbf{x}}_j = \mathbf{B}_j \mathbf{w}_j. \quad (3)$$

The source decomposition is performed by minimizing a divergence measure between the weighted combination of dictionary atoms and the mixed signal power spectrum. The weights of the atoms in the active set are estimated using Newton method.

3. Proposed Phase Estimation for SCSS

3.1. Problem Definition

In both groups, the mixture phase is commonly employed at signal reconstruction stage to recover the separated time-domain signals. The choice of mixture phase for reconstruction degrades the perceived quality and introduces certain artifacts in the reconstructed signal [13, 3]. Given the mixed signal composed of several speakers, the goal in SCSS is to recover all underlying speakers in the mixture. In this work, we concentrate on mixtures composed of two speech signals. Our goal here is to estimate the phase of the underlying sources given the mixture. The proposed phase estimator requires fundamental frequency of the underlying sources provided by applying a multi-pitch estimator on the mixed signal. The estimated phase of each source is finally combined with the spectral amplitude estimates obtained from IBM and NMF (reviewed in Section 2).

3.2. Proposed Phase Estimation Method

Figure 1-(bottom) shows the configuration for the proposed phase estimator where we propose to replace the mixture phase with the estimated one at signal reconstruction stage. In the following, we present each step in the proposed method in details.

3.2.1. Phase Decomposition

Let $f_{0,j}(l)$ as the fundamental frequency for source j at the lth frame. Given the mixed signal and the estimates for the fundamental frequencies, we aim at estimating the clean instantaneous phase of the underlying sources denoted as $\hat{\phi}_{x_j}(k,l)$. We propose to decompose the mixture phase in the STFT domain to its underlying components using a harmonic model plus phase decomposition proposed in [25]. To this end, first a pitch-synchronous segmentation as: $t(l) = t(l-1) + 1/4 f_{0,j}(l-1)$ is applied to the signal $y(n)$ in order to obtain segments of length $t(l)$ given by: $y(n,l) = y(n + t(l))w(n)$ where $t(l)$ is the time instant at frame l and $w(n)$ is the analysis window of length N. We further define $\psi_j(h,l)$ as the harmonic phase of the jth speaker evaluated at hth harmonic and nth time index. The harmonic phase decomposition result is given by:

$$\psi_j(h,l) = \psi_{j,\text{lin}}(h,l) + \Psi_j(h,l), \quad (4)$$

where we define $\psi_{j,\text{lin}}(h,l)$ as the linear phase term relying only on the fundamental frequency estimate and is calculated as

$$\psi_{j,\text{lin}}(h,l) = \sum_{l'}^{l} h\omega_{0,j}(l'), \quad (5)$$

where $\omega_{0,j}(l) = 2\pi f_{0,j}(l)/f_s$ with f_s as the sampling frequency. In (4), $\Psi_j(h,l)$ is the unwrapped phase capturing the stochastic part containing the minimum phase and a phase dispersion. The proposed method only considers the spectral phase at harmonics obtained by applying a linear interpolation on the spectral phase $\phi(k,l)$ at the harmonics. Let $\Psi_j(h,l)$ as the unwrapped harmonic phase given by subtracting the linear phase part from the harmonic phase $\psi_j(h,l)$:

$$\Psi_j(h,l) = \psi_j(h,l) - \hat{\psi}_{j,\text{lin}}(h,l). \quad (6)$$

where $\hat{\psi}_{j,\text{lin}}(h,l)$ is approximated using (5), given the estimated fundamental frequency $\hat{\omega}_{0,j}(l)$.

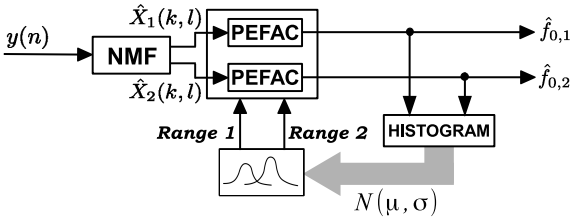

Figure 2: Block diagram for the multi-pitch Estimator.

3.2.2. Proposed Multi-Pitch Estimator

The fidelity of phase estimation for signal reconstruction in the proposed method, depends on the smoothness of the estimated fundamental frequencies for the underlying sources. Therefore, it is of high importance to have a smooth trajectory of the fundamental frequencies of the desired speaker (target/masker) with the least number of outliers from the interfering speaker. To alleviate the problem of overlapping frequency trajectories obtained from mixture, as shown in Figure 2, we propose to incorporate the histogram of the frequencies from the NMF-separated signals. The useful information regarding the frequency distribution can be obtained from the NMF-separated output signals, as they provide a more speaker-dependent statistic about $f_{0,j}$, containing less outliers of the interfering speaker. The statistical information is characterized by the mean μ_j and the variance σ_j of a Gaussian probability density function employed to fit to narrow down $f_{0,j}$ within the range determined by $f_{0,j_{\min}}$ and $f_{0,j_{\max}}$ given by:

$$f_{0,j} \in [f_{0,j_{\min}}, f_{0,j_{\max}}] \sim N(\mu_j, \sigma_j) \tag{7}$$

The fundamental frequencies together with the voicing probability, within the new frequency range, are taken into account as the new settings to re-apply the pitch estimation, which is realized using pitch estimation filter with amplitude compression (PEFAC) proposed in [26]. We found that the fundamental frequency estimate obtained by our proposed multi-pitch estimation method provides smoother trajectories compared to those in [27, 28] (detailed comparisons not shown here).

3.2.3. Temporal Smoothing of Unwrapped Phase

Inspired by our observations in speech enhancement [10, 11, 29] regarding the reduced phase variance property of the phase estimator at voiced regions, we propose to employ temporal smoothing filters on the unwrapped phase calculated for each speaker. The smoothing filters are applied across time to reduce the large circular variance of mixture phase for signal reconstruction. To this end, relying on circular statistics, we propose to calculate the circular mean for $\Psi_j(h,l)$ by averaging out the contribution by the other source. Then for the jth source phase, the smoothed unwrapped phase estimate is given by:

$$\hat{\Psi}_j(h,l) = \frac{1}{R} \sum_{l' \in R} e^{j\Psi_j(h,l')}, \tag{8}$$

where R is the smoothing filter length. The procedure is only applied to voiced frames.

3.2.4. Signal Reconstruction Using the Estimated Phase

Adding back the linear phase term using Eq. (4), we obtain the estimation of the harmonic phase $\psi_j(h,l)$. In order to combine the phase estimates with STFT spectral amplitude provided by IBM or NMF, it is required to return to the STFT domain. To this end, we replace those STFT frequency bins k, lying in the window support and adjacent to the underlying harmonic multiple $hf_{0,j}$ with the estimated harmonic phase $\psi_j(h,l)$. This method relies on the assumption that the underlying harmonics are well separated in frequency given a large enough frame length. The neighboring harmonics lying within the main-lobe width N_p of the analysis window are then modified. The enhanced spectral phase at frame l and frequency k is given by:

$$\hat{\phi}_j(\lfloor h\omega_{0,j}K \rfloor + i, l) = \hat{\psi}(h,l), \quad \forall i \in [-N_p/2, N_p/2], \tag{9}$$

with K as the DFT size. The estimated STFT phase is combined with the jth source spectral amplitude $\hat{X}_j(k,l)$ and the lth segment is given by:

$$\hat{x}_{j,l}(n) = \text{DFT}^{-1}\left\{\hat{X}_j(k,l)e^{j\hat{\phi}_j(k,l)}\right\}. \tag{10}$$

Applying overlap-add on $\hat{x}_{j,l}(n)$ gives the phase-enhanced signal $\hat{x}_j^p(n)$.

4. Results

4.1. Experimental Setup

We chose GRID corpus [30] as speech database where each utterance has a command-like structure composed of six units. To create the target and masker NMF dictionaries we randomly chose four speakers from each gender with 250 sentences per each resulting in 1000 sentences for training the gender-dependent dictionaries. For the test scenario, We chose 25 mixtures (not used in the training stage). The signal-to-signal ratio ranges from -9 to 9 (dB). In the blind separation setup, NMF-based separated signals are fed as inputs to our proposed multi-pitch estimator to extract $f_{0,j}$. Separation results are reported in terms of SDR and SIR [31], STOI [32] and PESQ [33] as they showed high correlation with subjective scores [34].

As phase estimation setup, we set the sampling frequency to 8 kHz. A Blackman window (minimizing spectral leakage [35] also reporeted in [10]) of length 32 ms was used. The temporal smoothing applied in (8), is a moving average filter with R as smoothing filter length, here chosen as 24 (ms), as we found it a reasonable trade off between a low level of circular variance of phase and preserving short-time stationarity of speech.

4.2. Proof-Of-Concept Experiment

The instantaneous phase representation in the STFT domain exhibits no harmonic pattern. Therefore, here, we consider group delay and phase variance [36]. Figure 3 shows plots to visualize the correctness of the proposed phase estimate versus the clean phase as reference. Such qualitatively evaluation informs about the effectiveness of the proposed phase estimation method in SCSS. The mixture is produced by mixing a target speaker 20 (female) saying *"bin blue at d 3 now"* and a masker speaker 17 (male) saying *"bin blue at d 8 please"* at SSR $= 0$ (dB).[1]

From spectrogram results (first row), it is observed that replacing the mixture phase with the estimated phase contributes in restoration of the harmonic phase structure available in the clean signal that were lost in the mixture. The group delay plots (middle row) further reveal the effectiveness of the proposed method in terms of recovering certain harmonic phase structure across frequencies. The circular phase variance plots (last row) demonstrate that the large phase variance in the mixture is reduced in the proposed method. The regions with successful phase recovery are highlighted by green dashed rectangles.

[1] Wave files here: http://www2.spsc.tugraz.at/people/pmowlaee/SCSS

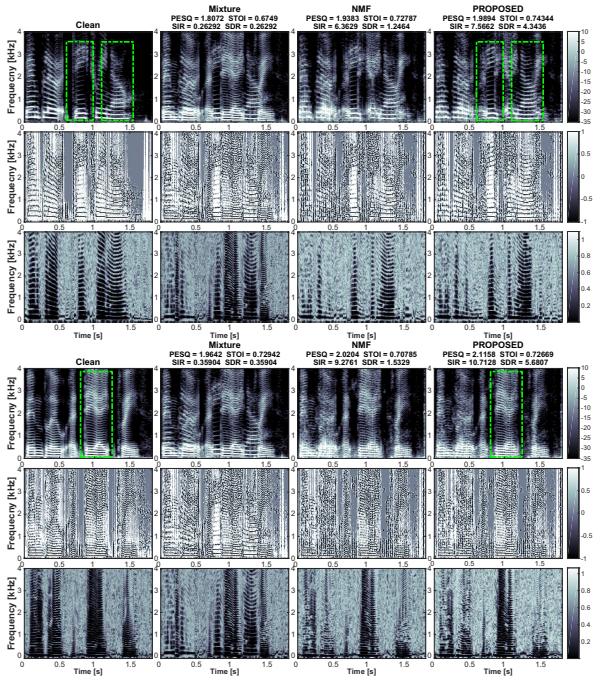

Figure 3: Proof-of-concept experiment for female target speaker (top) and male masker speaker (bottom) mixed at 0 (dB) shown as (top) spectrogram, (middle) group delay, and (bottom) phase variance. The results for clean (first), mixture (second), NMF (third) are shown for comparison, with green rectangles to highlight the regions with successfull phase restoration.

4.3. Source Separation Results

We demonstrate the effectiveness of the proposed phase estimation method when combined with IBM and NMF estimated spectral amplitudes. Any additional improvement reported here is due to phase-only enhancement, carried out by the proposed phase estimation method on top of the conventional IBM and NMF, where both use the mixture phase at signal reconstruction stage. To study the robustness of the proposed method to fundamental frequency estimation errors, we also include the results obtained for f_0-known scenario. The source separation results for IBM scenario are shown in Figure 4 categorized into target (top) and masker (bottom) reported in PESQ, STOI, SIR and SDR. The results for NMF scenario are shown in Figure 5.

Except for the target speaker in IBM scenario at -9 (dB) SSR, the proposed phase estimation method, consistently improves the speech intelligibility of both target and masker in the mixture for all SSRs. This is a crucial finding as the additional improvement is achieved on top of the conventional ideal binary mask outcome, known to have the highest intelligibility performance in SCSS [4]. For male speaker (masker) consistent improvement is feasible in most of the cases in IBM scenario, in terms of PESQ, STOI and SDR. This is an interesting observation as IBM has been reported to have a low perceived quality [13]. This observation also explains the importance of phase in SCSS. The improvement is more pronounced when f_0 is known. An exception is in terms of SIR scores for the masker at positive SSR levels where some negligible degradation in SIR scores are observed.

From the results shown for NMF scenario in Figure 5, the proposed method improves PESQ, STOI, SIR and SDR scores. In some cases, the f_0 outcome obtained from MPE shows comparable performance (not statistically significant better) than the

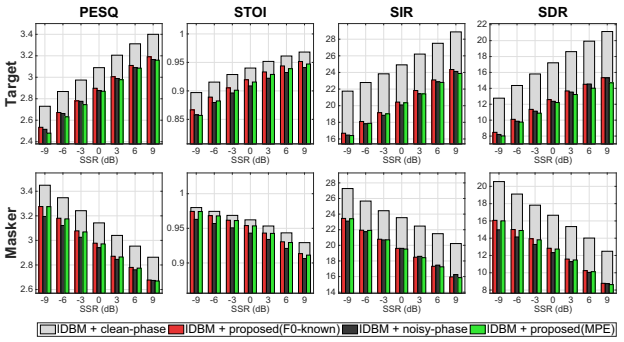

Figure 4: Ideal binary mask results: for (top) target (bottom) masker in terms of quality (PESQ), intelligibility (STOI), separation performance (SIR/SDR). The results for oracle phase and IBM (mixture phase) are reported for comparison.

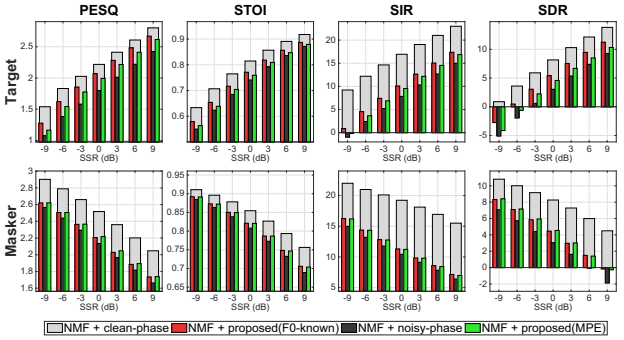

Figure 5: NMF scenario results: for (top) target (bottom) masker in terms of quality (PESQ), intelligibility (STOI), separation performance (SIR/SDR). The results for oracle phase and IBM (mixture phase) are reported for comparison.

f_0-known scenario as the proposed MPE benefits from accurate frequency boundaries. As our future work, we consider a comparative study between the proposed MPE and other benchmarks [27, 28], in terms of gross pitch error (GPE) and its consequent impact on phase estimation performance.

The comparison between f_0-oracle and f_0-estimated (using MPE) scenarios reveals the fact that the proposed phase estimation method is not that sensitive to the errors introduced by the pitch estimator. This was further justified by the low GPE results. Finally, the large improvement shown by clean phase (gray bars) justifies the importance of the phase information in terms of pushing the limited achievable performance by IBM or NMF using mixture phase.

5. Conclusion

While the conventional single-channel source separation methods rely on the mixture phase at signal reconstruction stage, in this work, we proposed to replace the mixture phase with an estimated one. The proposed phase estimation method relied on the phase decomposition principle and required the fundamental frequencies of the underlying sources obtained from a newly proposed multi-pitch estimator. Our results showed that joint improvement in perceived quality as well as speech intelligibility is feasible via replacing the mixture phase with the estimated phase. The gap between the clean and estimated phase motivates for further studies, for example by adopting the recent phase estimation methods proposed for single-channel speech enhancement [29, 10]. In this study we only included the ideal binary mask. As a future work, we report the effectiveness of our proposed phase estimator for estimated binary/ratio masks.

6. References

[1] P. Mowlaee and R. Martin, "On phase importance in parameter estimation for single-channel source separation," in *International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2012.

[2] S. J. Rennie, J. R. Hershey, and P. A. Olsen, "Graphical models for single-channel multitalker speech recognition," *IEEE Signal Processing Magazine*, vol. 27, no. 6, pp. 66–80, 2010.

[3] P. Mowlaee, R. Saiedi, and R. Martin, "Phase estimation for signal reconstruction in single-channel speech separation," in *13th Annual Conference of the International Speech Communication Association*, 2012, pp. 1–4.

[4] D. Wang and G. Brown, *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*. Wiley, 2006.

[5] D. Gunawan and D. Sen, "Iterative phase estimation for the synthesis of separated sources from single-channel mixtures," *IEEE Signal Processing Letters,*, vol. 17, no. 5, pp. 421–424, May 2010.

[6] N. Sturmel and L. Daudet, "Iterative phase reconstruction of wiener filtered signals," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, March 2012, pp. 101 – 104.

[7] J. Le Roux and E. Vincent, "Consistent Wiener filtering for audio source separation," *IEEE signal processing letters*, vol. 20, no. 3, pp. 217 – 220, 2013.

[8] P. Mowlaee and M. Watanabe, "Iterative sinusoidal-based partial phase reconstruction in single-channel source separation," in *14th Annual Conference of the International Speech Communication Association*, 2013, pp. 832–836.

[9] N. Sturmel and L. Daudet, "Signal reconstruction from stft magnitude: A state of the art," in *Proc. of the 14th International Conference on Digital Audio Effects (DAFx-11)*, 2011.

[10] P. Mowlaee and J. Kulmer, "Phase estimation in single-channel speech enhancement: Limits-potential," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 8, pp. 1–12, Aug. 2015.

[11] J. Kulmer and P. Mowlaee, "Phase estimation in single channel speech enhancement using phase decomposition," *IEEE Signal Processing Letters*, vol. 22, no. 5, pp. 598–602, May. 2015.

[12] D. Wang, "On ideal binary mask as the computational goal of auditory scene analysis," in *Speech Separation by Humans and Machines*, P. Divenyi, Ed. Kluwer, 2005.

[13] D. S. Williamson, Y. Wang, and D. Wang, "Reconstruction techniques for improving the perceptual quality of binary masked speech," *The Journal of the Acoustical Society of America*, vol. 136, no. 2, pp. 892–902, 2014.

[14] S. T. Roweis, "One microphone source separation," in *In Advances in Neural Information Processing Systems 13*. MIT Press, 2000, pp. 793–799.

[15] R. Weiss and D. Ellis, "Speech separation using speaker-adapted eigenvoice speech models," *Computer Speech and Language*, vol. 24, no. 1, pp. 16–29, 2010.

[16] P. Mowlaee, M. Christensen, and S. Jensen, "New Results on Single-Channel Speech Separation Using Sinusoidal Modeling," *IEEE Transactions on Audio, Speech, and Language Processing,*, vol. 19, no. 5, pp. 1265–1277, july 2011.

[17] P. Mowlaee, R. Saeidi, M. G. Christensen, Z. H.-. Tan, T. Kinnunen, P. Fränti, and S. Jensen, "A Joint Approach for Single-Channel Speaker Identification and Speech Separation," *IEEE Transactions on Audio, Speech, and Language Processing,*, vol. 20, no. 9, pp. 2586–2601, nov. 2012.

[18] M. H. Radfar and R. M. Dansereau, "Single-channel speech separation using soft mask filtering," *IEEE Transactions on Audio, Speech, and Language Processing,*, vol. 15, no. 8, pp. 2299–2310, Nov 2007.

[19] T. Virtanen, J. F. Gemmeke, and B. Raj, "Active-set newton algorithm for overcomplete non-negative representations of audio," *IEEE Transactions on Audio, Speech, and Language Processing,*, vol. 21, no. 11, pp. 2277–2289, Nov 2013.

[20] T. Virtanen, "Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria," *IEEE Transactions on Audio, Speech, and Language Processing,*, vol. 15, no. 3, pp. 1066–1074, March 2007.

[21] C. Joder, F. Weninger, D. Virette, and B. Schuller, "Integrating noise estimation and factorization-based speech separation: A novel hybrid approach," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2013, pp. 131–135.

[22] B. King and L. Atlas, "Single-channel source separation using complex matrix factorization," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 8, pp. 2591–2597, Nov 2011.

[23] H. Po-Sen, K. Minje, M. Hasegawa-Johnson, and P. Smaragdis, "Deep learning for monaural speech separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, May 2014, pp. 1562–1566.

[24] T. G. Kang, K. Kwon, J. W. Shin, and N. S. Kim, "Nmf-based target source separation using deep neural network," *IEEE Signal Processing Letters*, vol. 22, no. 2, pp. 229–233, Feb 2015.

[25] G. Degottex and D. Erro, "A uniform phase representation for the harmonic model in speech synthesis applications," *EURASIP, Journal on Audio, Speech, and Music Processing - Special Issue: Models of Speech - In Search of Better Representations*, 2014.

[26] S. Gonzalez and M. Brookes, "Pefac - a pitch estimation algorithm robust to high levels of noise," *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, vol. 22, no. 2, pp. 518–530, Feb. 2014. [Online]. Available: http://dx.doi.org/10.1109/TASLP.2013.2295918

[27] K. Nathwani, P. Pandit, and R. Hegde, "Group delay based methods for speaker segregation and its application in multimedia information retrieval," *IEEE Transactions on Multimedia*, vol. 15, no. 6, pp. 1326–1339, Oct 2013.

[28] M. Wohlmayr, M. Stark, and F. Pernkopf, "A probabilistic interaction model for multipitch tracking with factorial hidden Markov models," *IEEE Transactions on Audio, Speech, and Language Processing,*, vol. 19, no. 4, pp. 799–810, May 2011.

[29] P. Mowlaee and J. Kulmer, "Harmonic phase estimation in single-channel speech enhancement using phase decomposition and snr information," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 9, pp. 1521–1532, Sept 2015.

[30] M. Cooke, J. Barker, S. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition," *The Journal of the Acoustical Society of Americal*, vol. 120, p. 2421, 2006.

[31] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 4, pp. 1462 –1469, 2006.

[32] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time-frequency weighted noisy speech," *IEEE Transactions on Audio, Speech, and Language Processing,*, vol. 19, no. 7, pp. 2125–2136, 2011.

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

[34] P. Mowlaee, R. Saeidi, M. G. Christensen, and R. Martin, "Subjective and objective quality assessment of single-channel speech separation algorithms," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2012, pp. 69–72.

[35] F. Harris, "On the use of windows for harmonic analysis with the discrete fourier transform," *Proceedings of the IEEE*, vol. 66, no. 1, pp. 51–83, Jan 1978.

[36] M. Koutsogiannaki, O. Simantiraki, G. Degottex, and Y. Stylianou, "The importance of phase on voice quality assessment," in *15th Annual Conference of the International Speech Communication Association*, Singapore, September 2014.