Sinusoidal masks for single channel speech separation

Mowlaee, Pejman; Christensen, Mads Græsbøll; Jensen, Søren Holdt

Published in:
I E E E International Conference on Acoustics, Speech and Signal Processing. Proceedings

DOI (link to publication from Publisher):
10.1109/ICASSP.2010.5495679

Publication date:
2010

Document Version
Accepted author manuscript, peer reviewed version

SINUSOIDAL MASKS FOR SINGLE CHANNEL SPEECH SEPARATION

Pejman Mowlaee[1]*, Mads Græsbøll Christensen*[2]*, and Søren Holdt Jensen*[1]

[1]Dept. of Electronic Systems, Aalborg University, Aalborg, Denmark
[2]Dept. of Media Technology, Aalborg University, Aalborg,Denmark
emails: {pmb,shj}@es.aau.dk, mgc@imi.aau.dk

ABSTRACT

In this paper we present a new approach for binary and soft masks used in single-channel speech separation. We present a novel approach called the sinusoidal mask (binary mask and Wiener filter) in a sinusoidal space. Theoretical analysis is presented for the proposed method, and we show that the proposed method is able to minimize the target speech distortion while suppressing the crosstalk to a predetermined threshold. It is observed that compared to the STFT-based masks, the proposed sinusoidal masks improve the separation performance in terms of objective measures (SSNR and PESQ) and are mostly preferred by listeners.

Index Terms— Mask-based method, mixture estimator, sinusoidal mask, single-channel speech separation.

1. INTRODUCTION

Speech signal processing in adverse environments has widely been studied during recent years. Many solutions have been proposed to improve the performance of speech enhancement systems under highly colored noise scenarios [1, 2]. In general, when the interfering noise is non-stationary, the overall performance of the enhanced signal is corrupted by undesired artifacts, speech distortion or residual noise in the background called musical noise [2]. In this regard, there is a crucial need to develop an efficient speech enhancement approach to minimize the residual noise while keeping the quality of the enhanced speech unchanged. We here focus on single-channel speech separation (SCSS).

Mask-based methods have predominantly been applied in many speech enhancement [2] and separation [3]. The key idea behind any mask method is to estimate two masks and apply them to the mixture spectrogram to recover the speaker signals. The mask-based methods are generally categorized into two groups: binary [3–5], and Wiener filter [6], [7]. Binary mask was applied as MAX-VQ [5]) which employs log-max mixture estimator [1] to find two binary masks extracted from the speaker codebooks and then apply them on the mixture. In [8] the MAX-VQ system was applied as a model-based where the codewords are provided by taking the mean value among the vectors trained by a clean speech dataset. According to [8], the separation stage leads to errors while estimating masks for the underlying speakers and the re-synthesis speech quality was reported relatively low because of crosstalk caused by the interfering signal [8].

The greatest asset of mask-based methods lies in its simplicity and the fact that all that is required, is an estimate of the masks time-frequency pattern. Although the use of a mask-based approach is often recommended in speech enhancement [2], it is not yet optimal for SCSS paradigm. The performance of the mask-based methods is influenced by the non-stationarity behavior of speech segments. It is of high interest to incorporate a model of non-stationary speech into the binary mask or Wiener filtering frameworks. The main concern in mask-based method is attributed to the energetic masking occurring at frames where one speaker signal dominates the other. In such a case, the speaker signals energies collide at mixture time-frequency cells and make the signal recovery rather difficult. The mask-based methods explicitly suggest to filter out one of the speaker as a jammer signal which contradicts with the objective of an ideal separation system targeted to recover both signals.

In this paper, we present a new mask-based method for speech enhancement in general and in particular for SCSS. The proposed sinusoidal mask are constructed by using sinusoidal parameters extracted from the speaker models. It balances a tradeoff between the crosstalk suppression and the target speech distortion. Extensive simulation results are conducted to evaluate the speech separation performance for the proposed sinusoidal masks and compare them with those obtained by predominantly used STFT masks and VQ-based methods with STFT feature. The results show that the proposed masks could achieve a higher performance in terms of Perceptual Evaluation of Speech Quality (PESQ) as objective measure and are mostly preferred according to the informal listening experiments. The rest of the paper is organized as follows. In Section II the problem formulation for the mask-based SCSS is reviewed. The proposed method is presented in Section III. Section IV describes the separation algorithm. Section V presents the simulation results. Section VI concludes on the work.

2. MASK-BASED SPEECH SEPARATION

We now briefly review the key idea behind mask-based methods for SCSS. The main objective here is to design two masks, either binary or Wiener filter based, to be applied to the mixture spectrogram. The filtered time-frequency representations are then used to recover the individual speaker signals. Note that the binary mask aims at retaining the dominant time-frequency cells in a mixture spectrogram. This is implemented by removing the interference-dominant units. Such masking approaches are mostly unable to recover both target and masked signals at the same time [4], [5]. On the other hand, the Wiener filter weights each time-frequency cell of the mixture spectrum by taking a soft-decision according the *a priori* SNR [2]. There are two deficiencies for STFT masks; 1) some portions of the weaker speaker signal (often of high importance) is relatively masked by the other speaker (causing speech distortion in target signal), and 2) in some parts of the recovered speech signal (target) some portion of the interfering speaker signal is still audible (called cross-talk). This is similar to musical noise in speech enhancement but introduces a more severe effect for the listeners. Furthermore, the Wiener filter in the STFT domain is not able to recover both speaker signals with

a high quality (especially when one of them is dominant). Hence, we aim at generalizing the STFT-based masks to sinusoidal space to improve the separation performance.

3. PROPOSED SINUSOIDAL MASKS

In this section, we present the sinusoidal masks aimed at recovering the underlying speaker signals $\mathbf{s}_1$ and $\mathbf{s}_2$ according to the mixture $z = \alpha_1\mathbf{s}_1 + \alpha_2\mathbf{s}_2$ where α_1 and α_2 are the gains.

3.1. Sinusoidal Feature Parameters

According to the sinusoidal model of speech signals, each frame of the signal can be represented as a $N \times 1$ time vector as

$$\mathbf{s} = \mathbf{V}^T\mathbf{a} \quad , \tag{1}$$

where $\mathbf{V} = [\mathbf{v}_1 \dots \mathbf{v}_M]^T$ is a Vandermonde matrix of $M \times N$ whose rows are $\mathbf{v}_k = [1 \quad e^{j\omega_k} \quad \dots \quad e^{j\omega_k(N-1)}]^T$ with $k \in [1, M]$ as the sinusoidal frequency vector of dimension $N \times 1$, ω_k indicates the frequency of the kth selected peak, N is the time window length in samples, M is the order, and $\mathbf{a} = [a_1 \dots a_M]^T$ is a $M \times 1$ complex sinusoidal amplitude vector whose components are defined as $a_k = A_k e^{j\phi_k}$. The sinusoidal model used here is [9]; however sinusoidal parameter estimation is a bit different which is described in [10], [11]. We simply select the peak of the highest amplitude per Mel scale band and is characterized with triple $M \times 1$ vectors of amplitude, frequency and phase of the selected peaks. The decision rule of taking the highest peak per band is similar to maximum approximation used as a minimum mean square error (MMSE) mixture estimator for log amplitude spectra [1].

3.2. Sinusoidal Binary mask

We now consider the SCSS problem in a frame and define k as the frequency bin index. We incorporate the selected sinusoidal peaks within the bands to establish a sinusoidal binary mask defined as

$$H_1(\omega_k) = \begin{cases} 1 & \text{if} \quad A_{1,k} \geq A_{2,k} \\ 0 & \text{if} \quad A_{1,k} < A_{2,k} \end{cases} \quad , \tag{2}$$

where ω_k denotes the kth frequency component. The hard decision making in (2) can be summarized as an on-off keying (OOK) between two states C_1 the class of $A_1(k)$ and C_2 in favor of $A_2(k)$. A similar definition goes for $H_2(\omega_k)$ as complement of $H_1(\omega_k)$. The decision rule is similar to the ideal binary mask which compares the gain ratio of each time-frequency cell to the 0 dB local SNR [3].

3.3. Sinusoidal Wiener Filter as a Constrained Optimization

Speech enhancement with negligible perceived distortion is of high interest. In order to achieve an ideal separation performance we need to satisfy two requirements [2]; 1) it is required to guarantee minimal speech distortion of the target signal, and 2) the separated signals are required to have no portions of the other speaker signal. Without loss of generality, we assume in the following that $\mathbf{s}_1$ is the target and the other speaker is the interfering signal. We aim to find the kth frequency bin of the sinusoidal gain function as $g_1(\omega_k)$ that solves a constrained minimization problem by keeping the cross-talk of the other speaker below a predefined threshold and minimizing the target speech distortion. We define $\varepsilon(\omega_k)$ as the separation error for the target signal in the kth frequency bin as

$$\varepsilon(\omega_k) \quad = \quad \underbrace{(g_1(\omega_k) - 1)S_1(\omega_k)}_{\varepsilon_{s_1}(\omega_k)} + \underbrace{g_1(\omega_k)S_2(\omega_k)}_{\varepsilon_{s_2}(\omega_k)} \quad , \tag{3}$$

where $\varepsilon_{s_1}(\omega_k)$ is the speech distortion term for target speaker while $\varepsilon_{s_2}(\omega_k)$ is the the crosstalk term of the interfering speaker. We define $\mathbf{S}_i$ as a $N \times 1$ vector containing the spectral components of the ith underlying speakers defined as $\mathbf{S}_i = \mathcal{F}\{\mathbf{s}_i\} = [S_i(\omega_1) \cdots S_i(\omega_K)]^T$ where K is the number of frequency points used in calculating the DFT. The speech distortion energy for the target signal is calculated as $\varepsilon_{s_1}^2 = E\{\boldsymbol{\varepsilon}_{s_1}^H(\omega)\boldsymbol{\varepsilon}_{s_1}(\omega)\}$ and the cross-talk energy of the other speaker is $\varepsilon_{s_2}^2 = E\{\boldsymbol{\varepsilon}_{s_2}^H(\omega)\boldsymbol{\varepsilon}_{s_2}(\omega)\}$. We consider an optimization problem addressed as below

$$\min_{g_1,\mu} \quad \varepsilon_{s_1}^2 \quad \text{s.t.} \quad \varepsilon_{s_2}^2 \leq \delta \quad . \tag{4}$$

We define $\mathbf{G}_1$ as a $N \times N$ diagonal matrix with entries of $g_1(\omega_k)$ on its diagonal. The periodogram estimation of the PSD for the ith speaker is denoted by $\mathbf{P}_{s_is_i}$ defined as the Fourier transform of the autocorrelation function, $\mathbf{R}_{s_is_i}$ which is Toeplitz. Then, the power spectrum components are the diagonal elements of $\mathcal{F}^H R_{s_is_i}\mathcal{F}$ where $\mathcal{F}$ is the N-point Fourier transform matrix [2] and $\mathbf{P}_{s_is_i} = \text{diag}(P_{s_is_i}(\omega_1), \cdots, P_{s_is_i}(\omega_K))$. By using the Lagrangian multiplier method, we are required to solve the following constrained optimization in sinusoidal domain as

$$\mathbf{L} = (\mathbf{G}_1 - \mathbf{I})\mathbf{P}_1(\mathbf{G}_1 - \mathbf{I}) + \mu\mathbf{G}_1\mathbf{P}_2\mathbf{G}_1 \quad , \tag{5}$$

where $\mathbf{L}$ is a diagonal matrix whose (k,k)th element is given by the lagrangian of $\mathcal{L}(g_1(\omega_k), \mu)$ calculated at the kth frequency bin, and μ is the Lagrange multiplier as a parameter to trade off crosstalk suppression against speech distortion. Setting $\frac{\partial\mathcal{L}}{\partial\mathbf{g}_1} = \mathbf{0}$ with $\mathbf{0}$ as a $N \times 1$ zero vector, we obtain

$$(g_1(\omega_k) - 1)P_{s_1s_1}(\omega_k) + \mu g_1(\omega_k)P_{s_2s_2}(\omega_k) = 0 \quad . \tag{6}$$

The kth component of the sinusoidal Wiener gain is

$$g_1(\omega_k) = \frac{P_{s_1s_1}(\omega_k)}{P_{s_1s_1}(\omega_k) + \mu P_{s_2s_2}(\omega_k)} \quad . \tag{7}$$

Since we have no access to speakers' PSD, we replace them by the squared spectral vectors in discrete frequency domain and we obtain

$$g_1(\omega_k) = \frac{\xi_k}{\xi_k + \mu} \quad , \tag{8}$$

where we define $\xi_k = \frac{P_{s_1s_1}(\omega_k)}{P_{s_2s_2}(\omega_k)}$ as the *a priori* SSR computed at sinusoidal frequency peaks. The idea is to make the noise imperceptible by a proper choice of μ. In this paper we assume that the SSR level is known *a priori* and we set $\mu = (\frac{\alpha_2}{\alpha_1})^2$ which is agreement with the relevant discussion in chapter 6 of [2] where μ was such chosen to minimize the speech distortion in speech dominated frames while reducing the residual noise in noise dominated frames. Replacing μ into (8) and taking square root from (8) we have

$$g_1(\omega_k) = \frac{\alpha_1 S_1(\omega_k)}{\sqrt{\alpha_1 S_1^2(\omega_k) + \alpha_2 S_2^2(\omega_k)}} \quad , \tag{9}$$

which is similar to parametric Wiener filter in [2] and we call it sinusoidal Wiener mask. The proposed masks: sinusoidal binary in (2) and Wiener mask in (8) are used to recover the signals.

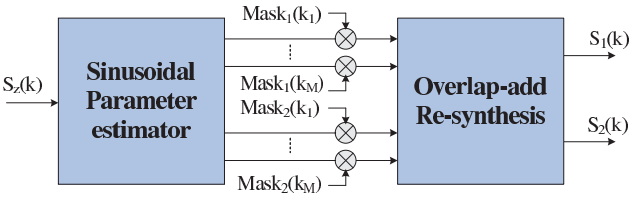

Fig. 1. SCSS system based on sinusoidal masks. The speaker signals are recovered by multiplying the mixture spectrum using a mask.

4. SEPARATION ALGORITHM

The key part of a separation algorithm is to find the optimal states of the source models of the speakers in the mixture. In this section we present the idea on how to find these states. These states refer to a codeword each composed of sinusoidal amplitude and frequency vectors denoted by $\mathbf{a} = [a_1 \dots a_M]^T$ and $\mathbf{f} = [f_1 \dots f_M]^T$, respectively. These codewords found by mixture estimation will be used to produce the sinusoidal masks. The codebooks are designed by using split-VQ of the sinusoidal parameters [10]. In the separation stage, two estimators are used. As our first estimator we use the optimum mixture estimation in [12]. In [12], it was demonstrated that under the uniformity assumption of mixture phase, the optimal estimator for SCSS, $S_{z,\text{opt}}(k)$ in the MMSE sense is

$$S_{z,\text{opt}}(\omega_k) = (S_1(\omega_k) + S_2(\omega_k)) \frac{E(\gamma_k)}{\pi} \quad , \tag{10}$$

with $\gamma_k = \frac{4S_1(\omega_k)S_2(\omega_k)}{S_1^2(\omega_k)+S_2^2(\omega_k)}$ and $E(\cdot)$ is the complete Elliptic integral of the second kind given by

$$E(\gamma_k) = \pi \left[1 - \frac{\gamma_k^2}{4} - \left(\frac{1 \times 3}{2 \times 4} \right)^2 \left(\frac{\gamma_k^4}{3} \right) - \dots \right] \quad . \tag{11}$$

As our second method, according to the optimum mixture estimator in (10) we replace the sinusoidal masks in (9) and we obtain

$$S_{z,m}(\omega_k) = \frac{1}{\pi} S_w(\omega_k)(g_1(\omega_k) + g_2(\omega_k))E(\gamma_k), \tag{12}$$

where $S_{z,m}(\omega_k)$ denotes the mask-based estimated mixture at the kth frequency bin and we define $S_w(\omega_k) = \sqrt{S_1^2(\omega_k) + S_2^2(\omega_k)}$ as the Wiener filter mixture estimation. To include SSR levels other than 0 dB in (12), we can consider the gain values α_1 and α_2. The sinusoidal mixture estimation is accomplished by searching for the optimal states of the composite sources by minimizing $\sum_{k=1}^{M} |S_z(\omega_k) - \hat{S}_z(\omega_k)|^2$, where $\hat{S}_z(\omega_k)$ can be replaced by either $S_{z,\text{opt}}(\omega_k)$ in (10) or $S_{z,m}(\omega_k)$ in (12). The solution of this minimization problem gives two states in the split-VQ codebooks to be used produce the masks in (9). To re-synthesize the separated outputs the mixture phase ϕ_z is used. Using the sinusoidal binary mask the kth frequency bin of the refiltered spectrum is

$$S_i(\omega_k) = S_z(\omega_k) g_i(\omega_k) \quad i \in \{1, 2\} \quad , \tag{13}$$

where $S_z(\omega_k)$ is the mixture power spectrum, $S_i(\omega_k)$ is the recovered spectrum for the ith speaker signal and $g_i(\omega_k)$ is either sinusoidal binary mask or sinusoidal Wiener mask given by (2) and (8), respectively. By using IDFT along with the mixture phase we get recovered time signals of each speaker. Fig. 1 shows the block diagram describing the SCSS based on the sinusoidal mask. In Fig. 1, $\{k_1, \dots, k_M\}$ indicate the frequency bins of sinusoidal peaks defined in Section II and M is the sinusoidal model order.

According to the suppression rule of Ephraim and Malah in [13], the proposed sinusoidal Wiener filter can be expressed as

$$g_1(\omega_k) = \sqrt{\frac{\xi_k}{\xi_k + 1} \frac{S_1^2(\omega_k) + S_2^2(\omega_k)}{S_z^2(\omega_k)}} = \sqrt{\frac{\xi_k}{\xi_k + 1} \left(\frac{1 + \nu_k}{\zeta_k} \right)} \tag{14}$$

Similar to [13], we define $\zeta_k = \frac{S_z^2(\omega_k)}{S_2^2(\omega_k)}$ as the *a posteriori* SSR and $\nu_k = \zeta_k \frac{\xi_k}{\xi_k + 1}$ as the instantaneous SSR. Then the proposed mask given by (14) is similar to Ephraim and Malah suppression rule already given in [13].

5. SIMULATION RESULTS

To assess the separation performance, we use the comprehensive database in [14] consisting of 34 speakers each containing 500 utterances. The sampling rate is decreased to 8 kHz from the original 25 kHz. Ten minutes of the speech signals of each speaker was used to produce the split-VQ and STFT codebooks with a codebook size of 2048. Twenty utterances are chosen from speakers 9 and 23 as test signals to evaluate the separation algorithms in a speaker-dependent scenario. The mixed signal is generated by adding the signals at different SSR. The separation performance for each method is reported in terms of PESQ [15] and segmental SNR (SSNR) [2]. The methods included in our simulations are the sinusoidal binary mask, sinusoidal Wiener filter and their STFT counterparts. As our benchmark methods, we applied algorithms similar to [4], [6], [16]. We also include the upper-bound performance for both STFT [17] and split-VQ [10] determining the highest performance obtainable by using the same source model if no mixture estimation error occurs. We used window length of 32 ms along with a frame shift of 8 ms. The codebook size for STFT and split-VQ was 2048. The number of sinusoidals used in our simulations is 50 and the number of DFT points in the STFT-based methods is 1024.

Fig. 2 shows the averaged PESQ scores for the separated signals obtained from their mixture[1]. From Fig. 2(a) it is observed that the optimal mixture estimator in sinusoidal space given in (10) is very close to the sinusoidal masks approximation in (12). Furthermore, by increasing the SSR level, both curves asymptotically attain the same performance, which is determined by the split-VQ upper-bound quantization performance [10]. Fig. 2(a) illustrates the PESQ curves obtained by masking approaches in STFT and sinusoidal domain. It is observed that applying the sinusoidal Wiener mask to the mixture improves, the separation performance compared to the STFT-based methods. This is also validated by listening to the separated signals at different SSRs for different genders. The improvement introduced by sinusoidal wiener filter over the STFT-based mask is rather significant at low SSR. Fig. 2(b) shows the separation performance for the second speaker signal in terms of SSNR in dB, and it is observed that employing the proposed sinusoidal masks, either binary or Wiener filter, cause improvements in the separation performance compared to the STFT binary mask [3–5] or Wiener masks in [3], [6]. This improvement is significant at low SSRs. The curves in Fig. 2(b) show that the performance obtained by the optimal mixture estimator in (10) is very close to the operational upper-bound determined by the STFT VQ. It is also observed that the proposed sinusoidal masks outperform the results obtained by both optimal estimator in STFT domain and the STFT-based masks. The

[1]The mixed and separated signals of different methods are downloadable from webpage at http://kom.aau.dk/∼pmb/IEEE_ICASSP2.htm

listening tests revealed that the re-synthesized signal quality is significantly improved compared to those obtained by the STFT based methods. From curves shown in Fig. 2(b) we observe that by employing the optimal estimator in (10) we reach to the operational upper-bound performance (where we assumed that the correct indices are known *a priori*). The results presented here are in accordance with our recent results in [17] where we showed that the model-based speech separation in transform domain results in improvements over the mask-based methods especially at low SSR.

According to the listening results, as SSR level decreases the STFT-based masks mostly lead to inferior performance. In contrast, the proposed sinusoidal masks achieve a superior performance and introduce significant improvement especially at low SSRs. This could be explained by the fact that the proposed sinusoidal mask minimizes the mixture estimation error at sinusoidal peaks of the mixture making a tradeoff between less crosstalk and small speech distortion. The proposed masks retain the highest peaks per bands and exclude other peaks mostly caused by main-lobe windowing or low-frequency modulation effect. This strategy would exclude those peaks vulnerable to be masked by the other speaker signal. Therefore the method is expected to result in lower crosstalk compared to the STFT masks.

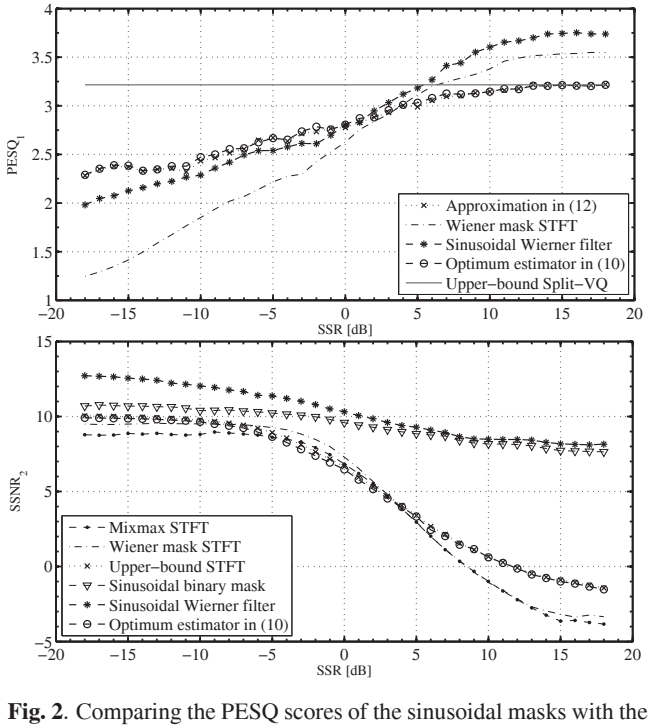

Fig. 2. Comparing the PESQ scores of the sinusoidal masks with the STFT masks and VQ-STFT versus SSR$\in [-18, 18]$.

6. CONCLUSION

In this paper, we proposed a new sinusoidal version for both binary mask and Wiener filter and compared their performance with their STFT counterparts. It was observed that the proposed sinusoidal masks could result in a significant improvement in the re-synthesized speech quality for both the recovered signals. We presented a framework to minimize the signal distortion while keeping the crosstalk below a predefined threshold. It was demonstrated that by the proposed approach, it is possible to reach the optimal performance for SCSS in a MMSE sense. From the simulation results, It was observed that, compared to the STFT masks, sinusoidal masks improved the separation performance in terms of SSNR and PESQ and were mostly preferred by informal listening tests.

We focused on speech separation scenario. As a future work, it is highly desirable to evaluate the proposed masks in other noisy environments including babble noise, car noise and other noise types. It is expected that the proposed method results in improvements compared to the STFT masks.

7. REFERENCES

[1] D. Burshtein and S. Gannot, "Speech enhancement using a mixture-maximum model," *IEEE Trans. Audio, Speech, and Lang. Process.*, vol. 10, no. 6, pp. 341–351, Sep. 2002.

[2] P. Loizou, *Speech Enhancement: Theory and Practice*, CRC Press, Boca Raton, 2007.

[3] D. Wang and G. J. Brown, *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*, Wiley-IEEE Press, 2006.

[4] M. H. Radfar, R. M. Dansereau, and A. Sayadiyan, "A maximum likelihood estimation of vocal-tract-related filter characteristics for single channel speech separation," *EURASIP Journal on Audio, Speech, and Music Processing*, pp. 84–186, March. 2007.

[5] S. Roweis, "Factorial models and refiltering for speech separation and denoising," *European Conference on Speech Communication and Technology*, pp. 1009–1012, 2003.

[6] A. M. Reddy and B. Raj, "Soft mask methods for single-channel speaker separation," *IEEE Trans. Audio, Speech, and Lang. Process.*, vol. 15, no. 6, pp. 1766–1776, Aug. 2007.

[7] M. H. Radfar and R. M. Dansereau, "Single-channel speech separation using soft mask filtering," *IEEE Trans. Audio, Speech, and Lang. Process.*, vol. 15, no. 8, pp. 2299–2310, Nov. 2007.

[8] P. Li, Y. Guan, S. Wang, B. Xu, and W. Liu, "Monaural speech separation based on MAXVQ and CASA for robust speech recognition," *Computer Speech and Language*, vol. 24, no. 1, pp. 30–44, Jan. 2010.

[9] R. McAulay and T. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Trans. Audio, Speech, and Lang. Process.*, vol. 34, no. 4, pp. 744–754, Aug. 1986.

[10] P. Mowlaee and Sayadiyan, "Model-based monaural sound separation by split-VQ of sinusoidal parameters," in *European Signal Processing Conference*, Aug. 2008.

[11] P. Mowlaee, A. Sayadiyan, and H. Sheikhzadeh, "FDMSM robust signal representation for speech mixtures and noise corrupted audio signals," *IEICE Electronics Express*, vol. 6, no. 15, pp. 1077–1083, 2009.

[12] P. Mowlaee, A. Sayadiyan, and M. Sheikhan, "Optimum mixture estimator for single-channel speech separation," *IEEE International Symposium on Telecommunications (IST)*, pp. 543–547, Aug. 2008.

[13] Y. Ephraim, "Statistical-model-based speech enhancement systems," *Proceedings of the IEEE*, vol. 80, no. 10, pp. 1526–1555, Oct. 1992.

[14] M. P. Cooke, J. Barker, S. P. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition," *The Journal of the Acoustical Society of America*, vol. 120, no. 5, pp. 2421–2424, 2006.

[15] 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," *IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 2, pp. 749–752, Aug. 2001.

[16] D. P. W. Ellis and R. J. Weiss, "Model-based monaural source separation using a vector-quantized phase-vocoder representation," *IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 5, pp. 957–960, May 2006.

[17] P. Mowlaee, A. Sayadiyan, and H. Sheikhzadeh, "Evaluating single-channel separation performance in transform domain," *Journal of Zhejiang University Science-A, Engineering Springer-Verlag, in press*, 2009.