New Results on Single-Channel Speech Separation Using Sinusoidal Modeling

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

Published in:
I E E E Transactions on Audio, Speech and Language Processing

DOI (link to publication from Publisher):
10.1109/TASL.2010.2089520

Publication date:
2011

Document Version
Accepted author manuscript, peer reviewed version

New Results on Single-Channel Speech Separation Using Sinusoidal Modeling

Pejman Mowlaee, *Student Member, IEEE*, Mads Græsbøll Christensen, *Member, IEEE*, and Søren Holdt Jensen, *Senior Member, IEEE*

***Abstract*—We present new results on single-channel speech separation and suggest a new separation approach to improve the speech quality of separated signals from an observed mixture. The key idea is to derive a mixture estimator based on sinusoidal parameters. The proposed estimator is aimed at finding sinusoidal parameters in the form of codevectors from vector quantization (VQ) codebooks pre-trained for speakers that, when combined, best fit the observed mixed signal. The selected codevectors are then used to reconstruct the recovered signals for the speakers in the mixture. Compared to the log-max mixture estimator used in binary masks and the Wiener filtering approach, it is observed that the proposed method achieves an acceptable perceptual speech quality with less cross-talk at different signal-to-signal ratios. Moreover, the method is independent of pitch estimates and reduces the computational complexity of the separation by replacing the short-time Fourier transform (STFT) feature vectors of high dimensionality with sinusoidal feature vectors. We report separation results for the proposed method and compare them with respect to other benchmark methods. The improvements made by applying the proposed method over other methods are confirmed by employing perceptual evaluation of speech quality (PESQ) as an objective measure and a MUSHRA listening test as a subjective evaluation for both speaker-dependent and gender-dependent scenarios.**

***Index Terms*—Mask methods, mixture estimation, single-channel speech separation (SCSS), sinusoidal modeling, speaker codebook.**

I. Introduction

THERE are many speech and audio applications where the signal of interest is corrupted by highly correlated noise sources. Separating such signals from their mixture has often been considered as one of the most challenging research topics in the area of speech enhancement. An extreme case of speech enhancement, single-channel speech separation (SCSS), is often considered as one of the most difficult scenarios where a speaker signal is corrupted with other interfering speaker signals. Although there have been recent advances in speech enhancement methods [1]–[10], SCSS with high speech quality still remains as a challenge. High quality separation systems could play an integral role in offering robustness in many practical applications including speech coding, speech recognition, speaker recognition in adverse mixture scenarios, and hearing aids [11].

The main objective for an ideal speech separation system is to recover the unknown speaker signals accurately, based on their observed mixed signal recorded by one microphone. The SCSS problem is ill-conditioned since the mixing matrix is non-invertible. The problem is in principle solvable by imposing *a priori* information, e.g., about the speaker models [12]–[19]. Previous state-of-the-art SCSS systems can be divided into two groups: 1) source-driven or computational auditory scene analysis (CASA)-based method [20]–[25], and 2) model-based method [12]–[19].

The main objective in the first group is to produce the binary masks required to separate the unknown speaker signals from their mixture. The methods predominantly use estimated pitch trajectories by applying a multi-pitch estimator. According to the results reported in [22], [26], and [27], the separation quality degrades as energetic masking takes place at some overlapping time-frequency cells. Therefore, the overall separation performance is limited by the accuracy of the multi-pitch estimator especially when the relative amplitude levels of the signals differ substantially (the signal-to-signal ratio (SSR) gets either low or high). At these SSR levels, the pitch estimation accuracy is relatively lost by large gross errors [26], [28]. In addition, according to [20], the CASA-based methods are mostly able to segregate the voiced frames of the mixture and often lack perceptual quality due to a severe cross-talk problem.

The second group, model-based separation systems is based on statistical models including VQ [15]–[18], Gaussian mixture models (GMMs) [13], [19], [29], [30] and Hidden Markov models (HMMs) [12], [14], [27]. In [14], a separate HMM was applied for each speaker and a huge state space of 8000 was required in order to carefully capture every possible signal transition state. Though using HMMs enables the modeling of correlated speaker signals, according to [31], it leads to a significantly complex mixture estimation approach. MAX-VQ attempts to find two masks based on the estimated VQ codewords. According to the results reported in [17], [22], [23], [32] using such masks inevitably causes cross-talk and artifacts in the re-synthesized signals.

From a synthesis viewpoint, the methods in the second group are divided into two classes: overlap-add procedure and mask

Manuscript received December 23, 2009; revised June 15, 2010 and September 15, 2010; accepted October 04, 2010. Date of publication October 25, 2010; date of current version May 13, 2011. The work of P. Mowlaee was supported by the Marie Curie EST-SIGNAL Fellowship (http://est-signal.i3s.unice.fr) under contract MEST-CT-2005-021175. This work was accepted for presentation at ICASSP 2010. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Tim Fingscheidt.

P. Mowlaee and S. H. Jensen are with the Department of Electronic Systems, Aalborg University, 9220 Aalborg, Denmark (e-mail: pmb@es.aau.dk; shj@es.aau.dk).

M. G. Christensen is with the Department of Architecture Design and Media Technology, Aalborg University, 9220 Aalborg, Denmark (e-mail: mgc@imi.aau.dk).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASL.2010.2089520

methods. The masks to be applied are either the binary [15], [16], [22], [23] or the Wiener filter masks [13], [29], [30], [33] leading to the separation approaches of the log-max estimator [12]–[15], [34] and the Wiener filtering [13], [29], respectively. Despite the attractive appeal of using masks in speech enhancement or separation, they have problems in dealing with the energetic masking [2]. These methods suggest filtering out one of the speakers as a jammer signal thereby causing inferior performance while recovering the masked speaker signal [16], [20].

In model-based methods, difficulties arise while mapping vectors of mixed signals onto states of speaker models resulting in wrong association of the codevectors with the log-max estimator leading to the selection of poorly filtered signal vectors [16], [23]. Selecting incorrect states from the speaker models could degrade the perceptual quality of the separated signals. According to [35], the model-based approach was expected to perform better than the pitch-based methods indicating that using only the pitch information shows limited discrimination for sequential grouping. This brings forward the idea that integration of pitch and spectral envelope in [16] may not be the most efficient solution to recover both signals because accurate multi-pitch estimation from a mixture at low SSRs is still a problem [22], [26], [27], [36].

It is important to note that most of the previous separation systems achieve a rather acceptable separation quality for the underlying sources in the mixture by assuming speaker signals to have nearly the same long-term energy level, i.e., when the SSR level is around 0 dB. In practice, however, a nonzero SSR level is expected since at each frame, one speaker signal often dominates others and the energies of the sources most likely collide [1], [37], a phenomenon called energetic masking [2] that makes the signal recovery of the speakers rather difficult. Therefore, studying novel methods to improve the separation quality at different SSRs is very important.

In this paper, we present new results for SCSS by proposing a mixture estimator based on sinusoidal parameters provided by codebooks for underlying speakers in the observed speech mixture. We consider a speech mixture composed of two speakers. The proposed model-based separation method aims to find optimal sinusoidal codevectors, one from each speaker model, that when combined best describe the observed mixture segment. The speaker models pre-trained for speakers are VQ codebooks composed of sinusoidal amplitude and frequency vectors. In this paper, we focus on speaker-dependent scenario and then we relax this assumption by using gender-dependent codebooks as an intermediate scenario. Through extensive simulations and subjective evaluations, we assess the separation performance of the proposed method at different SSR levels. The separation results show that the performance of the proposed method outperforms those obtained by other previous SCSS methods.

The rest of the paper is structured as follows: In the next section, we review previous sinusoidal methods for separation. In Section III, we introduce modified unconstrained sinusoidal parameters to be employed as feature parameters. The parameter estimation procedure is presented and followed by the proposed sinusoidal mixture estimator. In Section IV, we present the experimental results to compare the separation performance of the proposed method with that of other methods. Section V presents subjective evaluations and results of our MUSHRA test to assess the perceived quality obtained by different methods. Section VI features the discussions and Section VII concludes the work.

II. Relation to Previous Sinusoidal Methods for Separation

Sinusoidal parameters have already been applied for suppressing interference from a target signal [38]–[41]. As pioneering work, [38] proposed a multi-pitch tracking approach to assign harmonics of a mixed signal to the unknown speakers. In [39] and [40] nonlinear least square method was used for detecting the pitch frequency as well as the amplitude and phase parameters. In [41], a local nonlinear least square frequency estimator was proposed and applied on harmonic or nearly harmonic musical signals. The goal in [39]–[41] was to suppress the interference signal and to recover a desired speech signal from the mixture. The method in [40] worked based on harmonic frequencies of the sources under the pre-assumption that the pitch values of sources are known, *a priori* from the signals prior to the mixing process. The idea in [41] required estimates of the fundamental frequencies of the signals obtained by the multi-pitch estimator in [28]. The proposed approach in [39] was based on either *a priori* sinusoidal frequencies or *a priori* fundamental frequencies contours. Due to the pitch dependency of the methods in [38], [40], and [41], their separation performance was limited by the accuracy of the multi-pitch estimation. More specifically, this limitation was reported as the major restriction especially for recovering the weaker speaker signal in the observed mixture [38]–[41]. Good results were reported in [39] when the frequencies of both speaker signals are obtained by peak picking of individual STFT magnitudes prior to the mixing process. However, with no *a priori* information about the speakers' pitch information, the approach achieved limited good performance for only a subset of all-voiced c-channel signals and for roughly 0-dB SSR level. As another example, the test signals used in [41] consisted of either two well-separated sinusoids or a polyphonic excerpt of music. According to [41], although the separation performance was well for signals of prominent harmonic tones, it was not yet robust enough to lead into a reliable separation performance in general case.

In contrast to previous sinusoidal methods, the proposed approach in this work is focused on separating both underlying speakers from mixture. We suggest a new pitch-independent separation approach that relies on pre-trained speaker models as codebooks composed of sinusoidal amplitude and frequency. Additionally, we employ a comprehensive database in [42] containing utterances composed of both voiced and unvoiced frames. This is comparable to sinusoidal methods in [38]–[40] applied on a limited number of all-voiced utterances. Finally, for comparison with other harmonic methods, we compare the separation performance obtained by the proposed method with respect to harmonic magnitude suppression (HMS) [24], [25] and fusion method [16] as examples for source-driven and harmonic methods, respectively.

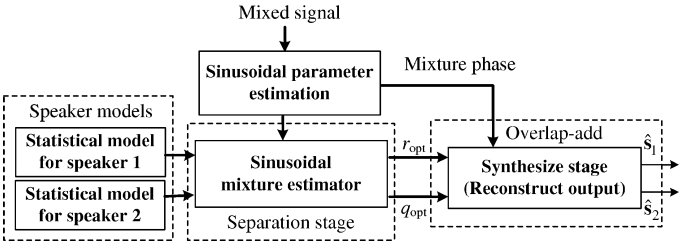

Fig. 1. Block diagram for the proposed speech separation method using sinusoidal modeling ($\hat{s}_1$ and $\hat{s}_2$ are separated signals).

III. Proposed Separation Method

We will now proceed to describe the proposed separation approach using sinusoidal modeling. Fig. 1 shows the block diagram of the proposed separation approach. The system is composed of the following blocks: sinusoidal parameter estimation, two trained speaker models, sinusoidal mixture estimator and overlap-add for signal reconstruction. In the following, we present our separation approach.

A. Sinusoidal Modeling

Before presenting the sinusoidal modeling, we will introduce some basic notation. Assume that we have a mixed signal, $\{z(n)\}_{n=0}^{N-1} = \sum_{k=1}^{K}\{s_k(n)\}_{n=0}^{N-1}$ composed of K speakers where k is the speaker index and the kth speaker signal is denoted by $\{s_k(n)\}_{n=0}^{N-1}$ with $k \in [1,K]$, n is the time sample index and N is the window length in the samples. At each frame, we represent the kth speaker signal in additive noise $e_k(n)$ as

$$s_k(n) = \sum_{i=1}^{L} A_{k,i}\cos(n\omega_{k,i} + \phi_{k,i}) + e_k(n) \qquad 0 \le n \le N-1 \quad (1)$$

where i is an index used to refer to the ith sinusoidal component characterized by the amplitude $A_{k,i}$, frequency $\omega_{k,i}$ and phase $\phi_{k,i}$, respectively. We define a parameter vector as $[\boldsymbol{\alpha},\boldsymbol{\omega},\boldsymbol{\phi}]$ of size $L \times 3$ with $\boldsymbol{\alpha} = \{A_{k,i}\}_{i=1}^{L}$, $\boldsymbol{\omega} = \{\omega_{k,i}\}_{i=1}^{L}$ and $\boldsymbol{\phi} = \{\phi_{k,i}\}_{i=1}^{L}$ denoting the kth speaker's amplitude, frequency and phase vectors, respectively, and L being the sinusoidal model order. The signal model in (1) is also used for representing observed mixed signal, $z(n)$. For the sake of simplicity and tractability, here, we focus on separating speech mixture composed of two speakers, i.e., $K = 2$ and $k \in [1,2]$.

B. Sinusoidal Parameter Estimation

We make two modifications to the unconstrained sinusoidal parameter estimator developed in [43] described as follows: 1) the spectral coefficients are translated to the Mel scale to take into account the logarithmic sensitivity of the human auditory system, and 2) at each band the spectral peak with the highest amplitude is selected [44]. These changes allow us to select the most perceptually relevant peak per band. The N-point discrete Fourier transform (DFT) vector for the ith frequency band of the kth speaker is represented by

$$\mathbf{v}_{k,i} = \begin{bmatrix} 1 & e^{j\omega_{k,i}} & \dots & e^{j\omega_{k,i}(N-1)} \end{bmatrix}^T \; i \in [1,L] \quad (2)$$

where $\omega_{k,i}$ denotes the selected peak at the ith band for the kth speaker. We define

$$\mathbf{V}_k = [\mathbf{v}_{k,1} \quad \mathbf{v}_{k,1}^* \quad \mathbf{v}_{k,2} \quad \mathbf{v}_{k,2}^* \quad \dots \quad \mathbf{v}_{k,L} \quad \mathbf{v}_{k,L}^*]^T \quad (3)$$

where $(\cdot)^*$ is the complex conjugate operator and $\mathbf{V}_k$ is a $2L \times N$ Vandermonde matrix whose rows are $\mathbf{v}_{k,i}$ defined in (2). The signal representation for the kth speaker in terms of sinusoids is given by an $N \times 1$ vector, $\hat{\mathbf{s}}_k = \mathbf{V}_k^T\mathbf{a}_k$, where

$$\mathbf{a}_k = [A_{k,1}e^{j\phi_{k,1}} A_{k,1}e^{-j\phi_{k,1}} \dots A_{k,L}e^{j\phi_{k,L}} A_{k,L}e^{-j\phi_{k,L}}]^T. \quad (4)$$

We define $S_k(\omega)$ as the complex spectrum for the kth speaker. The objective of the sinusoidal parameter estimation is to find peaks with the constraint [44]

$$\omega_{k,i} = \arg\ \max_{\omega\in\Omega_{k,i}}|S_k(\omega)|, \text{ and } A_{k,i}e^{j\phi_{k,i}} = S_k(\omega_{k,i}) \quad (5)$$

where $\Omega_{k,i}$ is a set composed of all continuous frequencies for the kth speaker within the ith band and $\arg\ \max(\cdot)$ returns the argument where $|S_k(\omega)|$ attains its maximum value.

C. Proposed Sinusoidal Mixture Estimator

In this section, we propose a mixture estimator based on the sinusoidal parametric vectors in our model-based separation approach shown in Fig. 1. Each speaker codebook is composed of a number of codevectors. The goal of a mixture estimator is to search the possible codevectors of the speaker models to find two optimal codevectors, one from each speaker model, such that when mixed, they satisfy a minimum estimation error criterion comparable to the mixed signal. These two best codevectors are denoted by $\{r_{\text{opt}}, q_{\text{opt}}\}$ in Fig. 1.

By applying the sinusoidal parameter estimator in (5) to the mixed signal, we obtain $\mathbf{z} = \mathbf{V}_z^T\mathbf{a}_z$, where $\mathbf{V}_z$ is a Vandermonde matrix composed of $2L$ frequency vectors of size $N \times 1$ as $\mathbf{v}_{z,i} = [1\ e^{j\omega_{z,i}}\ \dots e^{j\omega_{z,i}(N-1)}]^T$ defined by $\{\omega_{z,i}\}_{i=1}^{L}$, which is the set of sinusoidal frequencies obtained for the mixture at the ith band. We define $\boldsymbol{\alpha}_z = \{A_{z,i}\}_{i=1}^{L}, \boldsymbol{\omega}_z = \{\omega_{z,i}\}_{i=1}^{L}$ and $\boldsymbol{\phi}_z = \{\phi_{z,i}\}_{i=1}^{L}$ denoting, respectively, the amplitude, frequency, and phase of the ith component for the mixed signal. We derive a mixture estimator based on the sinusoidal parameters of the underlying speakers and their mixture. The key idea is to project the mixture onto its sinusoidal subspace spanned by the columns of the parametric vector $[\boldsymbol{\alpha}_z,\boldsymbol{\omega}_z,\boldsymbol{\phi}_z]$ and to find a cost function to be minimized in the mixture estimation stage. Based on (1), we define $P_k(e^{j\omega})$ as the power spectrum for the kth speaker at the ith band as [45]

$$P_k(e^{j\omega}) = \sigma_{k,i}^2 + A_{k,i}^2[\delta(\omega-\omega_{k,i}) + \delta(\omega+\omega_{k,i})] \quad (6)$$

where we assumed that $e_k(n)$ is white at each ith frequency band and $\sigma_{k,i}^2$ denotes its corresponding variance, $\{\omega_{k,i}\}_{i=1}^{L}$ is the frequency set for the peaks retained for the kth speaker signal. A similar definition holds for the mixed signal, and we define the mixture power spectrum as $P_z(e^{j\omega})$. The frequencies in $\{\omega_{z,i}\}_{i=1}^{L}$ are formed by applying (5) on the mixed sig-

nals. Considering an appropriate window denoted by $W(e^{j\omega})$ to reduce the spectral leakage, the expected value for the periodogram for each signal spectrum is $E\{\hat{P}_k(e^{j\omega})\} = P_k(e^{j\omega}) * W(e^{j\omega})$ where $\hat{P}_k(e^{j\omega})$ is the *periodogram* for the kth speaker, $E\{\cdot\}$ denotes the expectation operator and $*$ is the convolution operator. We define a cost function as the squared error between the power spectra of the mixed signal and its estimate to be sampled only at sinusoidal peaks given by $\{\omega_{z,i}\}_{i=1}^{L}$. The expected value for the mixture estimation error at the ith band is

$$\begin{aligned} &E\{\epsilon_i(e^{j\omega})\} \\ &\quad = E\{\hat{P}_z(e^{j\omega}) - \hat{P}_1(e^{j\omega}) - \hat{P}_2(e^{j\omega})\} \qquad (7) \\ &\quad = \sigma_{\epsilon,i}^2 + A_{z,i}^2\left[W\left(e^{j(\omega-\omega_{z,i})}\right) + W\left(e^{j(\omega+\omega_{z,i})}\right)\right] \\ &\qquad - \sum_{k=1}^{2} A_{k,i}^2\left[W\left(e^{j(\omega-\omega_{k,i})}\right) + W\left(e^{j(\omega+\omega_{k,i})}\right)\right]. \qquad (8) \end{aligned}$$

We define $\sigma_{\epsilon,i}^2 = \sigma_{z,i}^2 - \sigma_{1,i}^2 - \sigma_{2,i}^2$ as the variance of the error. The expected mixture estimation error in (8) is sampled at mixture sinusoidal frequencies per ith band defined by the set $\{\omega_{z,i}\}_{i=1}^{L}$. Replacing ω by $\omega_{z,i}$ in (8) and ignoring the negative part of the spectrum for real speech signals, we get

$$\epsilon_i = A_{z,i}^2 - A_{1,i}^2 W\left(e^{j(\omega_{z,i}-\omega_{1,i})}\right) - A_{2,i}^2 W\left(e^{j(\omega_{z,i}-\omega_{2,i})}\right) \qquad (9)$$

where ϵ_i captures the mixture estimation error defined between the original and the estimated mixture spectra at the ith band. $A_{1,i}$, $A_{2,i}$ and $A_{z,i}$ are the sinusoidal amplitude selected at the ith band for the first, the second, and the mixed signals, respectively. The mixture approximation error gets close to zero when the underlying speaker spectra are highly harmonic. The mixture estimation error termed as d at a given frame is $d = \sum_{i=1}^{L} |\epsilon_i|$. The distortion function in (9) only calculates the mixture estimation error at the sinusoidal peaks obtained from the mixture. The proposed mixture estimation is targeted to find the optimal indices by searching the possible codevectors in speaker one codebook $(\mathbb{C}_1)$ and speaker two codebook, $(\mathbb{C}_2)$ by solving the following minimization problem at each frame [46]

$$\begin{aligned} \{r_{\text{opt}}, q_{\text{opt}}\} = \arg\ &\min_{\mathbb{C}_1 \times \mathbb{C}_2} \\ &\times d\left(\left\{A_{z,i}, \hat{A}_{1,i}^{r}, \hat{A}_{2,i}^{q}, \hat{\mathbf{v}}_{1,i}^{r}, \hat{\mathbf{v}}_{2,i}^{q}\right\}_{i=1}^{L}\right) \qquad (10) \end{aligned}$$

where r and q are the codebook indices for speaker codebook one and two, respectively, and we define $\mathbb{C}_1 \times \mathbb{C}_2 = \{r \in \mathbb{C}_1\} \times \{q \in \mathbb{C}_2\}$ as the space formed by the union of the spaces defined by $\mathbb{C}_1$ and $\mathbb{C}_2$. In the minimization formula given by (10), $\{r_{\text{opt}}, q_{\text{opt}}\}$ addresses $\{\hat{A}_{1,i}^{r_{\text{opt}}}, \hat{\mathbf{v}}_{1,i}^{r_{\text{opt}}}, \hat{A}_{2,i}^{q_{\text{opt}}}, \hat{\mathbf{v}}_{2,i}^{q_{\text{opt}}}\}_{i=1}^{L}$ which are the optimal sinusoidal codevectors selected from codebooks $\mathbb{C}_1$ and $\mathbb{C}_2$, and $d(\cdot)$ is the 2-D cost function based on the mixture approximation error in (9). The set $\{r, q\} \in [1, M]$ can be any possible states in the speaker models with M as the codebook size. At each frame, by minimizing $d_{r,q}$ in (10), we obtain two codevectors of the speaker models, which when combined, satisfy the minimization criterion in (10). The selected codebook indices are then used to reconstruct the two separated signals by means of a weighted overlap-add (OLA) procedure as shown in Fig. 1.

It is important to note that, in this paper, we use a full search to consider all possible states during minimization of the distortion function in (10). However, it is also possible to apply some cyclic minimizer or expectation maximization (EM)-like algorithms as an approximation to solve the minimization problem more computationally efficient, which is generally sub-optimal.

Our goal here is to find the set of unknowns denoted as $\{\hat{A}_{1,i}, \hat{A}_{2,i}, \hat{\mathbf{v}}_{1,i}, \hat{\mathbf{v}}_{2,i}\}_{i=1}^{L}$ by solving the following minimization problem per band:

$$\arg\min_{\hat{A}_{1,i}^{r}, \hat{A}_{2,i}^{q}, \hat{\mathbf{v}}_{1,i}^{r}, \hat{\mathbf{v}}_{2,i}^{q}} \sum_{i=1}^{L} \left\| \left(A_{z,i}\mathbf{v}_{z,i} - \hat{A}_{1,i}^{r}\hat{\mathbf{v}}_{1,i}^{r} - \hat{A}_{2,i}^{q}\hat{\mathbf{v}}_{2,i}^{q}\right) \right\|_2^2 \qquad (11)$$

where $\hat{A}_{1,i}^{r}$ and $\hat{\mathbf{v}}_{1,i}^{r}$ are referred to the rth codevector selected from codebook $\mathbb{C}_1$, $\hat{A}_{2,i}^{q}$ and $\hat{\mathbf{v}}_{2,i}^{q}$ are referred to the qth codevector selected from codebook $\mathbb{C}_2$. By taking the Fourier transformation of the expression in (11), we get the mixture estimation in (9).

Assume that the modeling error in (9) is a zero-mean white, i.i.d. (independent and identically distributed over observations) with Gaussian noise with constant variance $\sigma_i^2 \neq 0$ at each frequency band i. Using an l_2-norm and applying band decomposition, one can show that the log-likelihood of all bands is

$$P = K - \frac{1}{2}\sum_{i=1}^{L} \frac{\|\mathbf{z}_i - \boldsymbol{\mu}_{\hat{z}_i}\|_2^2}{\sigma_i^2} \qquad (12)$$

where $\boldsymbol{\mu}_{\hat{z}_i}$ is the estimated mixed signal formed by combining the selected codewords of the speakers for the ith band and $K = -(L/2)\log(2\pi) - \sum_{i=1}^{L}\log\sigma_i$. Minimizing the likelihood of all bands using the sinusoidal estimator approximates the exact likelihood of all bands in (12). The minimization results in two sinusoids (one for each speaker) per band.

D. Training Split-VQ Codebooks on Sinusoidal Parameters

We use split-VQ codebooks composed of sinusoidal amplitude and frequency vectors as speaker models. Here, we briefly explain the split-VQ codebook generation used in our proposed separation method. The extracted sinusoidal parameters: amplitude and frequency, each of dimension L are entered to the training stage. Following [47], we apply different distance measures to produce codebooks of amplitude and frequency, respectively. For the amplitude part of the kth speaker, we apply distance measure

$$d_A = \sum_{i=1}^{L}\left(\frac{A_{k,i}}{\|\boldsymbol{\alpha}_k\|_2^2} - \frac{\hat{A}_{k,i}}{\|\hat{\boldsymbol{\alpha}}_k\|_2^2}\right)^2 \qquad (13)$$

where $\|\cdot\|_2^2$ is the l_2-norm and $\hat{\boldsymbol{\alpha}}_k = \{\hat{A}_{k,i}\}_{i=1}^{L}$ is the coded amplitude codevector, with $\hat{A}_{k,i}$ as the coded amplitude for the sinusoidal peak selected at the ith band for the kth speaker. Let M_A be the codebook size for the amplitude part of our split-VQ codebook. After establishing M_A amplitude codevectors, we select frequency vectors that are closest in terms of their related amplitude vectors. Another VQ of a lower size is performed on these frequency candidates for each amplitude codeword. To

produce frequency codevectors for the kth speaker, we apply the following distance measure

$$d_w(\mathbf{V}_k, \hat{\mathbf{V}}_k) = \sum_{i=1}^{L} w_{k,i} ||\mathbf{v}_{k,i} - \hat{\mathbf{v}}_{k,i}||_2^2 \qquad (14)$$

where $w_{k,i} = (A_{k,i}/||\boldsymbol{\alpha}_k||_2^2)$ is the energy normalized amplitude vector used for dynamic weighting of the Euclidean distance measure to make it proportional to the sinusoidal amplitude at the peak frequencies.

IV. Experimental Results

A. Separation Scenario and Database

As a proof of concept, we evaluate the performance of the proposed method in SCSS and compare it with other benchmark methods. In our implementations, we first focus on speaker-dependent scenario. Then, we relax this assumption by using gender-dependent codebooks as an intermediate scenario. The SSR is defined as the averaged ratio of the target speaker gain to the gain of the interfering signal. In our experiments, we swept the SSR level within the range $[-18, 18]$ dB. Then, the separation results are averaged at each SSR level over all pairs of test signals and quantified using PESQ [48] as objective measure and MUSHRA [49] listening test as subjective evaluation. As benchmark methods, the separation result of the proposed method is compared with other conventionally used methods: MAX-VQ [15], [23], [32], the Wiener filtering [13], [30], and STFT-VQ [17], [18]. We also compare the separation results of the proposed method with those obtained by HMS [24], [25] and fusion method [16] for both speaker-dependent and speaker-independent scenarios.

To evaluate the proposed separation algorithm, we used the database provided for SCSS in [42] consisting of 34 speakers each uttering 500 sentences. For our speaker-dependent scenario, we selected four speakers including two male (speakers 9 and 19) and two female speakers (4 and 23) from the database. We used 10 minutes of speech signals from each of the four speakers to train the speaker models. The sampling frequency was decreased from the original 25 kHz to 8 kHz. We analyzed the performance of the proposed mixture estimator for many mixture pairs to find the best values of these parameters. According to our results, throughout all experiments presented here, we used 50 sinusoidal peaks and a von Hann window of duration 32 ms with a frame-shift of 8 ms. For practical reasons, throughout the simulations presented here, the desired frequency range was set to [60, 3850] Hz at a sampling frequency of 8 kHz.

For practical reasons and according to findings reported in [47], we have opted for 11 bits for amplitude and 3 bits for frequency. For a fair comparison and consistent with the results in [17], the same codebook size was chosen for the STFT codebooks. In the experiments, we assumed that the double-talk regions in the mixture are known *a priori*. We only focus on separating the mixed regions to report the performance of different mixture estimators, which is arguably also the most difficult part. We also assumed *a priori* knowledge of speaker identities and SSR level in the observed speech mixture.

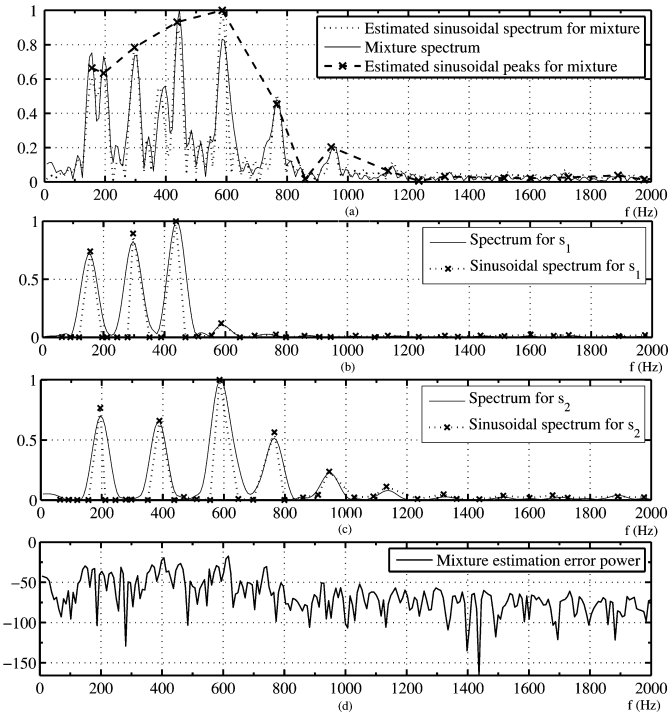

Fig. 2. Showing the magnitude spectrum for (a) the original and estimated mixture, (b) speaker one, (c) speaker two, and (d) mixture estimation error power in dB.

B. Ideal Separation Scenario

To assess the performance of our proposed mixture estimator, we consider the ideal separation scenario as was done in [32]. In an ideal separation scenario, we assume that we have access to the original underlying speakers, and from their spectral vectors, we find the optimal codevectors based on their corresponding trained speaker codebooks. We select two utterances of one male and one female and add them together at SSR $= 0$ dB to form a mixture. Fig. 2 depicts how the proposed mixture estimator works by minimizing the error at the sinusoidal peaks estimated from the mixture. The sinusoidal peaks in magnitude spectrum are shown for the original and estimated mixture in Fig. 2(a), as well as for each of the underlying single speaker signals in Fig. 2(b) and (c). From the mixture estimation error shown in Fig. 2(d), it is observed that the estimation error is reasonably low especially at sinusoidal frequencies of the mixture, explaining the high accuracy of the proposed mixture estimator.

C. Evaluating Performance for Speaker-Dependent Case

We report the separation performance of the proposed method and compare it with respect to other benchmark methods. First, we consider speaker-dependent scenario where we assume that we have *a priori* knowledge of speaker identities. To this end, we randomly selected ten sentences from the test data of each speaker in order to forming the speech mixtures. The training and test sets were disjoint. Fig. 3 shows the PESQ scores of different separation methods versus SSR. To carefully assess the gap between methods, we also included the upper-bound for the separation performance achieved by the STFT [17] and split-VQ on the sinusoidals in [47]. The performance of the proposed method was compared to previous speaker-dependent

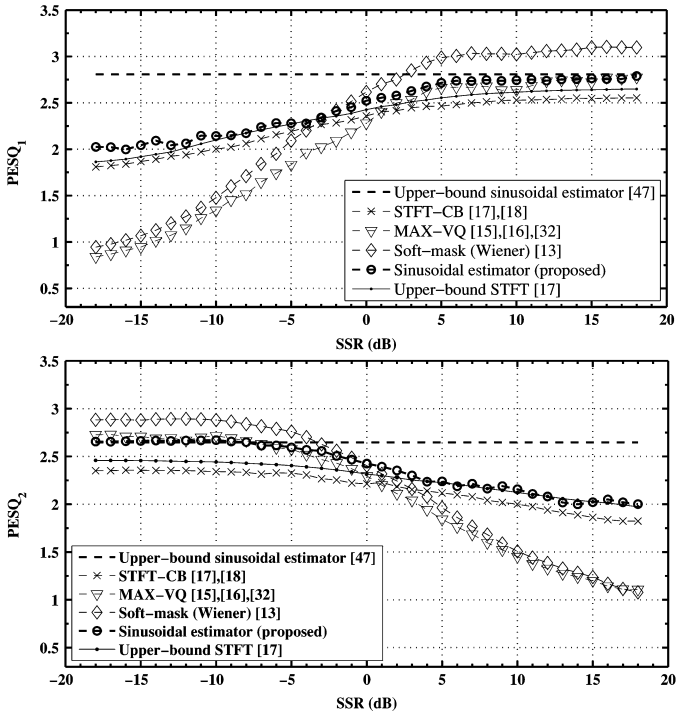

Fig. 3. Showing the separation results for speaker-dependent scenario for different methods in terms of PESQ score versus SSR levels for (a) speaker one and (b) speaker two [46].

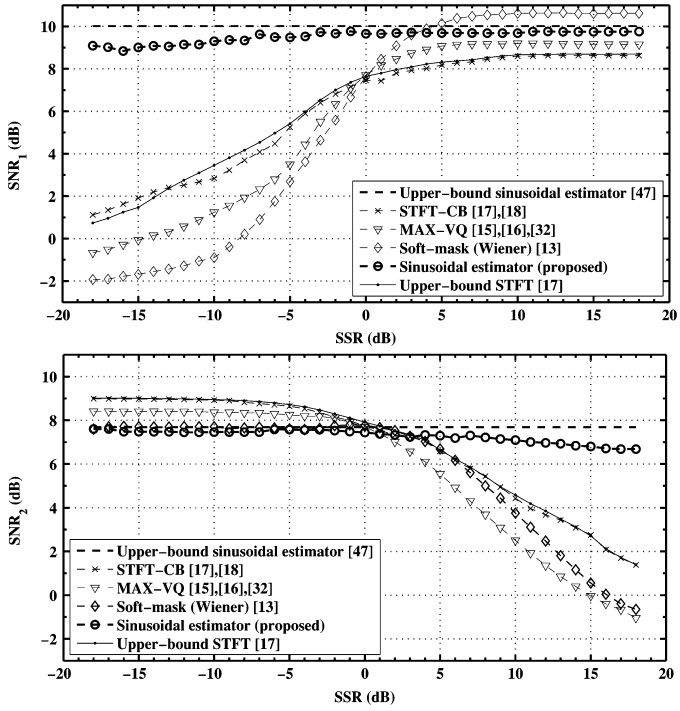

Fig. 4. Showing the separation results for gender-dependent scenario for different methods in terms of PESQ score versus SSR levels for (a) speaker one, and (b) speaker two.

methods. The methods we included in our simulations are binary mask, the Wiener filtering, and the STFT-based VQ methods. Each curve depicted in Fig. 3 is labeled with the related reference. Several results are inferred from Fig. 3: 1) according to the curves, the proposed method achieves a higher PESQ score compared to MAX-VQ and the Wiener filtering especially at low SSR levels; 2) it is observed that the proposed method achieves about 1 point improvement in PESQ score over the mask methods. The inferior performance of the mask methods can be further explained by the energetic masking effect of the dominant speaker at time-frequency cells [2], [27], [42]. The mixture estimation error observed in the mask methods is due to the fact that they originally filter out the competing speaker to recover a target signal and consequently lead to decoding errors while mapping vectors of the mixed signal into the codevectors in the codebooks of the underlying speaker in the mixture. Hence, using a log-max mixture estimator in a mask approach could result in the selection of wrong codevectors from the speaker models, and consequently, it leads to poorly filtered separated signals as reported in [23]; 3) according to Fig. 3(a) and (b), the proposed method outperforms the STFT-based approach and its upper-bound separation performance. The significant degradation in performance caused by the STFT codebook-based method (denoted by STFT-CB), as compared to the proposed approach can be observed from the gap between the PESQ curves shown in Fig. 3(a) and (b). This agrees with the recent results reported in [17] stating that compared to mask methods, performing subband transformation on the STFT features could result in improvements in the perceived speech quality of the separated signals especially at low SSR levels; 4) according to the curves shown in Fig. 3, the proposed method asymptotically reaches the upper-bound performance achieved by the split-VQ codebooks in [47].

TABLE I
SPEAKER LABELS USED FOR TRAINING THE GENDER-DEPENDENT MODELS FOR MALE AND FEMALE SPEAKERS

Male	3	5	6	9	10	12	13	14	17	19
Female	4	7	8	11	15	16	21	22	23	24

D. Separation Results for Gender-Dependent Scenario

To relax the assumption of *a priori* knowledge of speaker identities, here, we study the separation results for gender-dependent scenario. As gender-dependent models, we selected ten female and ten male speakers each producing 35 s of speech signal. We trained a male speaker model using utterance from ten speakers and a female speaker model trained on ten female speakers. These two speaker models are gender-dependent considered as an intermediate scenario between speaker-dependent and speaker-independent. The speaker labels used for training our gender-dependent models are shown in Table I. To evaluate the separation performance we formed mixtures using fifteen utterances of speakers 29, 34 as female and 30, 32 as male speakers selected as our test speakers. The separation results were then averaged over the mixture pairs at different SSR levels and speakers. Fig. 4 illustrates the separation results obtained by different methods for gender-dependent scenario. Curves demonstrate the separation performance for each speaker in terms of SNR versus SSR. To assess the gap between different methods, we also included the upper-bound separation performance. From Fig. 4, it is concluded that compared to other methods, the proposed method shows a significant improvement

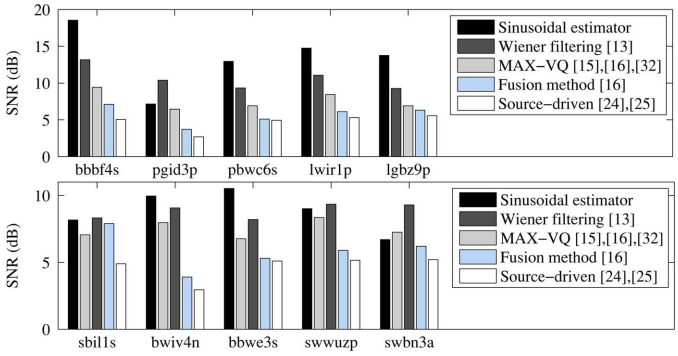

Fig. 5. Comparing the SNR results of the proposed method with MAX-VQ [15], [16], [32], Wiener filtering [13], STFT-VQ [18], source-driven [24], [25], and harmonic methods in [16] for speaker-dependent scenario.

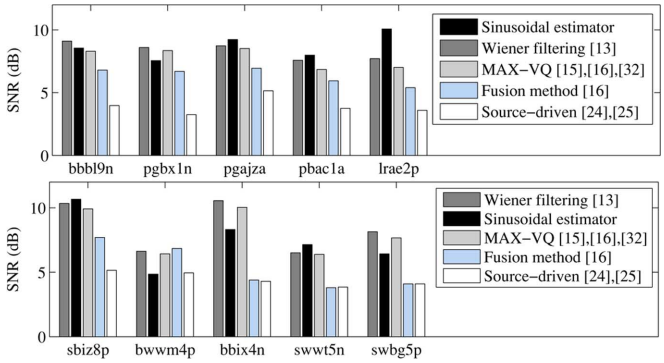

Fig. 6. Comparing the SNR results of the proposed method with MAX-VQ [15], [16], [32], Wiener filtering [13], STFT-VQ [18], source-driven [24], [25], and harmonic methods in [16] for speaker-independent scenario.

for both speaker signals, especially at extreme SSR levels (both low and high).

It is important to note that the results shown in Fig. 3 and 4 can best be interpreted separately. According to the definition of SSR, high SSR means that speaker one is dominant in the mixture while a similar interpretation goes for the second speaker but for negative values of SSR. From Fig. 3 and 4, at high SSR levels, soft mask achieves a slightly higher PESQ score compared to our method. This can be explained because of the use of masks in soft-mask method which employs information directly from the mixed signal. Since at high SSR levels, target speaker (let speaker one) is more intelligible, then mask method achieves a higher PESQ score for this speaker compared to a model-based method since the latter employs no information directly coming from the mixture, but uses pre-trained speaker spectra for signal reconstruction. This observation can be further explained by noting the fundamental difference between mask and reconstruction-based methods while synthesizing the separated signals.

E. Comparing the Separation Results With Harmonic Methods

We compare the separation performance of the proposed method in terms of SNR measure with source-driven in [24] and [25] and fusion methods in [16] both based on pitch estimates of the underlying speakers in the observed speech mixture. These two methods serve as examples for source-driven and harmonic methods, respectively. To have a fair comparison to the results reported in [16], here we select speaker 4 and 19 for the speaker-dependent scenario. In addition to the speaker-dependent scenario, here, we also consider a speaker-independent one as a more practical scenario. To train a speaker-independent codebook, we used the utterances of four speakers: 4, 7, 8, and 19. As a test, similar to [16], we selected ten speech files from the remaining 30 speakers to generate five speech mixtures. The test speakers in the speaker-independent scenario are: 2, 3, 14, 15, 16, and 22. To have a fair comparison, we used the same mixtures as described in Tables in [16] all formed at SSR $=$ 0 dB. Figs. 5 and 6 show the SNR results measured in dB per mixture described on x-axis for speaker-dependent and speaker-independent scenarios, respectively. According to the results, it is observed that the proposed approach mostly achieves a higher score compared to source-driven and fusion methods in [16]. The improvement for speaker-independent scenario is lower but the proposed method still mostly outperforms other approaches including: log-max, Wiener filtering, source-driven and fusion.

TABLE II
LABELS OF THE METHODS USED IN MUSHRA TEST

Excerpt	Separation method and SSR scenario
$BMM_{SSR=0dB}$	Binary mask at SSR=0dB
$BMM_{SSR=-18dB}$	Binary mask at SSR=-18dB
$SIN_{SSR=0dB}$	Proposed method at SSR=0dB
$SIN_{SSR=-18dB}$	Proposed method at SSR=-18dB
$FFT_{SSR=0dB}$	STFT-based VQ at SSR=0dB
$FFT_{SSR=-18dB}$	STFT-based VQ at SSR=-18dB

V. SUBJECTIVE EVALUATION

A. MUSHRA Test Setup

To assess the perceived speech quality of the separated output signals obtained by different methods, we conduct a subjective listening test by using the multi-stimulus test with hidden reference and anchors (MUSHRA test) as described in ITU-R BS.1534-1 [49]. The MUSHRA test is a double blind test for the subjective assessment of intermediate quality level benefits obtained from displaying all stimuli at the same time. This enables the subjects to carry out simultaneous comparison between the methods directly. Seven untrained listeners participated in the test (the authors not included). The excerpts used in our listening test are shown in Table II, each indicating a separated signal at a specific SSR level. The experiments are conducted for both speaker-dependent and gender-dependent scenarios. Both MAX-VQ and STFT-based VQ methods were included as benchmarks for separation methods. All the played signals were monophonic of length 2 s and sampled at 8 kHz. Many more excerpts were used in our development phase, but the excerpts shown in Table II are the ones that have been tested in our listening test. The excerpts consisted of the hidden reference denoted by HR and an anchor low-pass filtered at 2 kHz denoted by Anchor. The remaining six excerpts are the separated signals at different SSRs shown in Table II. The hidden reference shows the known quality on the scale and is used to check the consistency of the responses of a subject during the listening test. A high score is expected at this point. The anchor point is included to enable comparisons between the different listening

TABLE III
RESULTS OF THE MUSHRA LISTENING TEST FOR THE SPEAKER-DEPENDENT SCENARIO. THE MOS RESULTS OBTAINED FOR EACH CLIP AVERAGED OVER SEVEN LISTENERS ARE SHOWN FOR DIFFERENT METHODS. FOR EACH CASE, THE CONFIDENCE INTERVAL IS ALSO INCLUDED

Excerpt\Clip	1	2	3	4	5	6	7	8
$BMM_{SSR=0dB}$	47.85±15.56	39.43±13.63	26.28±12.35	34.28±17.18	35.86±11.67	30.00±12.97	45.86±16.71	23.00±12.86
$BMM_{SSR=-18dB}$	3.85±2.68	61.57±15.31	2.43±1.47	62.86±13.04	1.71±0.65	57.71±24.03	2.00±0.69	70.57±18.31
$SIN_{SSR=0dB}$	81.14±13.39	50.86±22.82	51.00±14.87	47.43±15.69	41.43±14.08	36.57±7.16	47.14±16.46	55.28±12.54
$SIN_{SSR=-18dB}$	69.28±11.66	50.57±18.48	53.14±17.58	62.71±14.76	44.86±7.37	36.28±14.31	58.00±6.07	53.14±15.55
$FFT_{SSR=0dB}$	28.14±21.19	23.00±15.01	14.43±7.79	39.57±18.14	17.86±11.32	24.28±12.41	17.14±5.64	18.28±13.63
$FFT_{SSR=-18dB}$	6.85±3.75	51.00±10.75	4.71±2.51	60.86±16.61	4.28±2.60	48.28±9.99	2.86±1.35	67.00±13.63
HR	99.14±1.59	99.14±1.94	98.28±3.88	99.86±0.32	99.86±0.32	98.71±2.02	100.00±0.00	99.57±0.97
Anchor 2 kHz	76.14±13.45	63.86±18.85	78.57±8.84	68.86±18.09	82.14±5.84	77.14±11.89	80.28±10.19	83.28±5.93

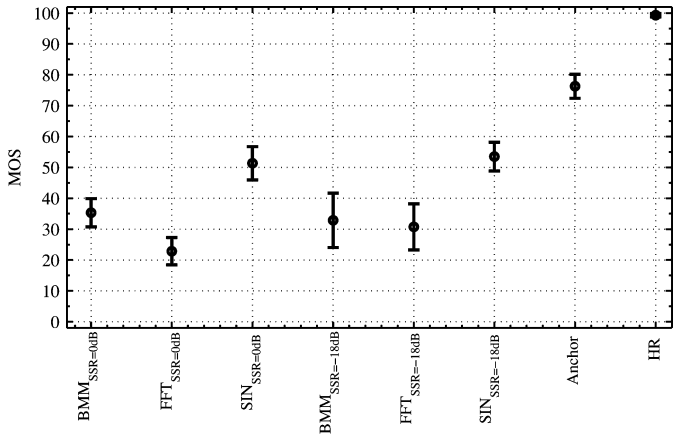

Fig. 7. Results of the MUSHRA listening test for the speaker-dependent scenario [46]. MOS scores for different separation methods over all excerpts and all listeners. Error bars indicate 95% confidence intervals.

tests since it forms a simple but well-defined modification on the reference signal. Excerpts listed in Table II were chosen and played for each subject. The listeners were asked to rank eight separated signals relative to a known reference on a scale of 0 to 100. By including different SSR levels, it is possible to assess any improvement observed in the synthesized speech quality of the proposed method compared to other methods. Further, the separation performance is evaluated for two SSRs.

B. Listening Test Results for Speaker-Dependent Scenario

We conducted the listening experiments on subjects in a silent room and a good sound quality audio, Firewire interface, was used for digital to analog conversion. Moreover, we used a high quality headphone: AKG K240 MKII. The scores obtained from different methods were averaged over all listeners and excerpts. Fig. 7 depicts the mean opinion score (MOS) for the speaker-dependent scenario. Furthermore, the performance of individual excerpts can be observed by the numbers in the first row of Table III, which shows the results obtained by each clip.[1] For each entry, the first number is the averaged value over the scores obtained by seven listeners and the second number determines the confidence interval.

The results shown in Table III are divided into two categories, i.e., target and masker speakers. Odd columns show the results for the masked signal in the mixture explaining the low scores at SSR $= -18$ dB while the even columns show the results for the target speaker in the mixture. From Fig. 7, it is observed that the proposed method scores, on average, about 20 points higher than MAX-VQ, and more than 25 points higher than the STFT-based method. According to Fig. 7, no overlap exists between the confidence intervals of the proposed method and the other methods. Therefore, it can be concluded that the proposed method achieves statistically significant improvement by consistently enhancing the performance of the perceived speech quality for both target and interference separated signals especially at low SSRs. The proposed method achieves a slightly lower quality compared to those obtained by MAX-VQ and Wiener filtering. However, as indicated by the listening experiments, some of the separated outputs achieved by MAX-VQ were found suffering from severe crosstalk. Furthermore, listeners observed that in some cases the separated signals obtained by MAX-VQ were relatively poor compared to the reference signal. They observed that these methods suffer from the cross-talk phenomenon, mostly while recovering masked signal, in which a portion of the other speaker signal exists in the separated output signal. This is mainly because a mask method applies a gain function to the mixed spectrum rather than finding a candidate from the codebooks. On the other hand, the proposed method produced artifacts in the separated signals often encountered in sinusoidal speech modeling especially in fricatives and sudden attacks [43]. However, the proposed method still outperforms the others by achieving, on average, 23 to 28 points, higher than the STFT-based method and 17 to 21 points higher than MAX-VQ. According to the listeners observations, the improvements brought about by the proposed method are perceived both as an increase in terms of speech signal quality and lower cross-talk. The tests also revealed that the separation performance of the mask methods (especially at 0 dB of SSR where their separation performance is often reported) does not necessarily produce the highest perceived quality for the separated signals. This can be observed by comparing the MOS results in Fig. 7 for $BMM_{SSR=0}$ and $SIN_{SSR=0}$ stating that the proposed method shows an advantage of 10 points in the resulting MOS compared to MAX-VQ. We also considered results shown in Table IV as the MOS results obtained by each listener averaged over eight clips defined in Table II. In gender-dependent scenario we only considered masked speaker output for subjective measurement at SSR $= -18$ dB while for speaker-dependent scenario we included both separated target speaker and masker speaker signals at SSR $= -18$ dB. By inspecting the MOS results shown in Fig. 7 along with the results of the listening experiments in Table IV, subjects often indicated that the signals related to the proposed method were close to the

[1]The mixed and separated wave files for different methods are downloadable from our webpage: http://kom.aau.dk/~pmb/IEEE_Trans.htm.

TABLE IV
RESULTS OF THE MUSHRA LISTENING TEST FOR THE SPEAKER-DEPENDENT SCENARIO. THE MOS RESULTS OBTAINED BY EACH LISTENER AVERAGED OVER EIGHT CLIPS ARE SHOWN FOR DIFFERENT METHODS. FOR EACH CASE, THE CONFIDENCE INTERVAL IS ALSO INCLUDED

Excerpt\Listener	1	2	3	4	5	6	7
$BMM_{SSR=0dB}$	51.00±9.09	33.37±13.07	39.87±9.86	24.37±10.95	26.62±11.46	48.62±17.53	23.37±6.19
$BMM_{SSR=-18dB}$	39.12±31.37	33.87±26.06	40.12±32.88	19.12±16.12	31.87±24.85	43.12±34.77	22.62±18.23
$SIN_{SSR=0dB}$	66.12±10.71	47.50±16.92	58.25±15.82	47.00±21.93	34.87±10.48	55.12±16.81	50.62±12.31
$SIN_{SSR=-18dB}$	64.00±10.08	48.12±17.71	59.75±12.49	50.50±15.58	39.00±8.43	68.50±8.96	44.62±9.79
$FFT_{SSR=0dB}$	31.50±5.53	21.12±12.09	14.37±9.16	21.62±16.18	10.12±3.99	44.87±15.72	16.25±5.59
$FFT_{SSR=-18dB}$	34.12±26.97	32.12±23.87	34.37±28.56	25.37±17.48	31±20.40	31.75±26.07	27±20.09
HR	100.00±0.00	100.00±0.00	99.87±0.00	99.12±1.64	100±0.00	96.25±3.12	100±0.00
Anchor 2 kHz	72.00±7.26	73.12±7.16	96.00±2.74	66.25±14.21	79±5.94	68.75±18.02	80±2.39

TABLE V
RESULTS OF THE MUSHRA LISTENING TEST FOR GENDER-DEPENDENT SCENARIO. THE MOS RESULTS OBTAINED FOR EACH CLIP AVERAGED OVER SEVEN LISTENERS ARE SHOWN FOR DIFFERENT METHODS. FOR EACH CASE, THE CONFIDENCE INTERVAL IS ALSO INCLUDED

Excerpt\Clip	1	2	3	4	5	6	7	8
$BMM_{SSR=0dB}$	32.57±8.99	29.71±8.56	29.14±18.10	18.86±10.03	24.86±7.25	25.43±9.78	22.86±12.00	28.14±11.79
$BMM_{SSR=-18dB}$	8.00±6.61	6.00±6.21	4.57±3.59	3.00±1.98	4.28±2.69	3.43±2.75	6.86±4.41	20.71±6.07
$SIN_{SSR=0dB}$	52.00±12.49	36.43±9.06	33.71±6.09	24.28±11.27	35.00±15.14	28.57±6.15	38.71±11.63	34.14±13.92
$SIN_{SSR=-18dB}$	35.57±13.85	33.86±13.05	32.00±5.61	17.00±6.09	41.00±12.33	28.71±7.36	38.14±10.83	33.86±10.16
$FFT_{SSR=0dB}$	34.86±8.11	28.86±13.07	28.14±9.37	16.57±3.24	32.00±13.96	27.57±9.77	29.71±10.14	19.71±11.76
$FFT_{SSR=-18dB}$	20.57±13.01	40.14±9.72	13.00±8.42	7.00±2.61	14.71±8.24	10.28±4.24	13.71±4.68	5.86±4.46
HR	97.28±3.98	97.28±3.98	97.28±3.48	97.71±3.48	97.71±4.92	97.00±3.49	99.14±1.94	99.14±1.94
Anchor 2 kHz	63.43±9.27	61.28±9.98	67.71±14.47	66.86±11.04	66.71±7.00	69.14±11.57	67.71±11.85	69.71±10.23

TABLE VI
RESULTS OF THE MUSHRA LISTENING TEST FOR THE GENDER-DEPENDENT SCENARIO. THE MOS RESULTS OBTAINED BY EACH LISTENER AVERAGED OVER EIGHT CLIPS ARE SHOWN FOR DIFFERENT METHODS. FOR EACH CASE, THE CONFIDENCE INTERVAL IS ALSO INCLUDED

Excerpt\Listener	1	2	3	4	5	6	7
$BMM_{SSR=0dB}$	19.88±6.14	26.63±5.24	42.13±10.91	18.50±7.02	18.88±4.95	21.00±7.36	38.13±9.56
$BMM_{SSR=-18dB}$	4.75±5.89	4.50±3.39	14.38±6.03	4.25±2.49	4.38±5.72	4.38±3.55	13.13±6.59
$SIN_{SSR=0dB}$	36.13±6.94	32.63±10.11	47.38±11.93	19.75±9.52	31.38±8.62	33.38±7.99	46.88±9.56
$SIN_{SSR=-18dB}$	32.63±7.35	29.88±8.68	43.88±11.41	22.00±4.20	24.63±8.44	31.50±7.58	43.13±11.82
$FFT_{SSR=0dB}$	21.25±6.95	18.75±7.57	37.13±7.93	20.75±8.11	27.25±10.36	25.88±8.38	37.50±10.86
$FFT_{SSR=-18dB}$	11.37±7.68	10.38±5.82	26.38±12.39	11.63±8.77	13.00±7.12	12.88±11.27	24.00±12.73
HR	93.25±1.09	100.00±0	100.00±0	100.00±0	100.00±0	100.00±0	91.88±4.15
Anchor 2 kHz	71.50±3.49	53.13±3.70	85.88±6.67	62.75±3.15	52.13±3.77	70.00±5.07	70.63±3.26

reference signal and showed a significant preference over other separated signals.

C. Listening Test Results for Gender-Dependent Scenario

Relaxing the *a priori* knowledge of speaker identities, we report the MOS results for the MUSHRA listening test in a gender-dependent scenario shown in Fig. 8. According to the results depicted in Fig. 8, since no overlap exists between the proposed method and the benchmark methods, it can be concluded that the proposed method can achieve statistically significant improvement compared to other methods and consistently enhances the performance of the synthesized speech quality for both target and interference separated signals. It is observed that at 0 dB of SSR the proposed method achieves greater improvement compared to other methods. From Fig. 8, it is observed that in extreme cases (low/high SSRs), the proposed method improves the perceived speech quality of the separated signals. The numbers in the first row of Table V show the results obtained by each clip for the gender-dependent scenario. We also considered results shown in Table VI as the MOS results obtained by each listener averaged over eight clips for gender-dependent scenario. It is observed that, for the gender-dependent scenario, the proposed method consistently outperforms the others in most of the cases.

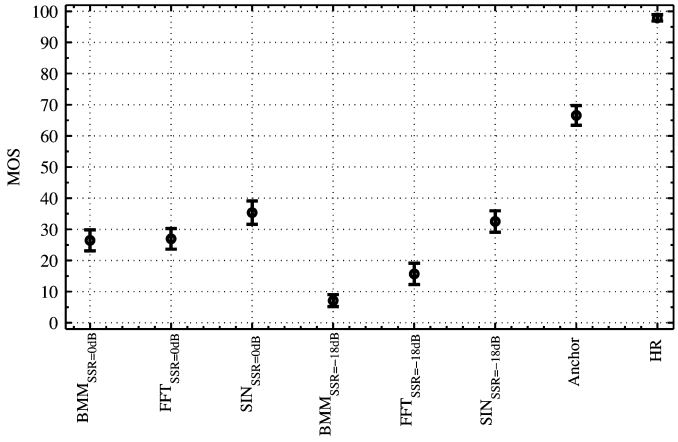

Fig. 8. Results of the MUSHRA listening test for the gender-dependent scenario. MOS scores for different separation methods over all excerpts and all listeners. Error bars indicate 95% confidence intervals.

By comparing the MOS results shown in Figs. 7 and 8 along with Tables III–VI, we observe that the proposed method achieves a higher score both in speaker-dependent and gender-dependent scenarios. The MOS results for gender-dependent scenario are lower than those obtained for speaker-dependent scenarios. At low SSR levels, both mask and STFT-based methods show inferior performance especially

in gender-dependent scenario. In contrast, the proposed method shows a shorter confidence interval both in speaker-dependent and gender-dependent scenario. By comparing the MOS results depicted in Figs. 7 and 8, it is observed that the relative difference between the methods in speaker-dependent and gender-dependent scenarios shows a remarkably similar pattern of overall performance.

VI. Discussion

In previous separation methods based on either harmonic modeling [16], [24], [25], [32], [39]–[41], or CASA [20]–[22], the speech perceived quality for separated signals was directly determined by the accuracy of the multi-pitch estimator. However, due to energetic masking [2], the pitch detection accuracy of most of the existing pitch estimators, degrades significantly, especially at low SSRs [26], [28]. Hence, the proposed method offers an attractive candidate for SCSS or similar enhancement scenarios where pitch estimation with high accuracy is either rather erroneous [26], [28] or difficult because of the energetic masking [2], [27]. In addition, it was shown in [39] that a pitch-based method is not capable of attaining the same level of enhancement compared to a system based on sinusoidal frequencies. We confirmed this by comparing the separation performance of the proposed method with source-driven in [24], [25] and fusion methods in [16]. These motivate us to present a separation strategy independent of pitch estimates, in this paper. The sinusoidal parameter estimation taken in this work leads to a high-frequency resolution at low frequencies, reflecting the pitch harmonic structure of each speaker signal and their mixture.

The proposed approach, like other well-known sinusoidal modeling methods in [39]–[41], has a major limitation in the failure to deal with unvoiced segments in a consistent manner. The methods in [39]–[41] were all suggested and limited by utterances composed of vocalic mixtures. Additionally, the sinusoidal modeling we used in this work is originally like the one described in [43] proposing that if we sample the spectrum of unvoiced speech with rate equal to 100 Hz, no perceivable degradation is observed in the synthesized speech signal at least from perceptual point of view. As future work, it is possible to consider more complex modeling of speech and jointly estimating sinusoidal model parameters and voicing states of the two underlying signals.

The study in [39] reported the problems related to the frequency resolution of the discrete Fourier transform especially when two sinusoids related to different fundamental frequencies are arbitrary close to each other. As shown in [39], [40], the solution leads to singular ill-conditioned matrix as the frequency of one speaker close to the frequency of the other speaker and the problem is only solvable if two pitch frequencies and their integer multiples are not overlapping and are well separated; a condition which is often not met when two speech sources exist in the scene. This problem is equal to extracting two unknowns (two frequencies) from single observation (mixture frequency). In order to deal with such ambiguity, [39] suggested monitoring the spacing between neighboring frequencies and using a multi-frame interpolation procedure. However, in this work, we suggest testing all possible combinations of codevectors selected from underlying speakers' codebooks. This solution guarantees leading into the minimal error in the nonlinear cost function. The work in [39] and [41] only considered enhancing the target speech while current work addresses the more challenging problem of separating both speaker signals from their observed mixture. More specifically, in [39] the interference was suppressed while changing the interference speech to noise.

The present sinusoidal mixture estimator ignores the cross-term components and phase differences which, in some situations, play a critical role and can change the position of peaks completely. This happens when the sinusoidal peaks of the underlying speakers get closer than 25 Hz. In such situations, the accuracy of the sinusoidal mixture estimator is limited but still finds the two states of the two speaker models (sinusoidal coders), which when combined, will best describe the mixture spectrum at certain frequencies (estimated from the mixture spectrum per bands).

The proposed technique uses pre-trained frequency codevectors based on peaks which makes the system more speaker-dependent. According to our simulations, the proposed method also led to good results for gender-dependent scenario which addresses an intermediate scenario. The more interesting speaker-independent scenario, most likely can be addressed by combining a speaker identification module with current separation system as reported in [50].

The present work considers the mixture scenario composed of two speaker signals. For mixtures with more than two speakers, it is possible to employ an EM-like algorithm in which for each speaker we update the signal parameters of one speaker at a time and then use these parameters in another searching scheme required for finding the optimal states of other two speakers' states. Separating mixtures of more than two speakers is an open problem and we have considered that as a potential future work.

The separation approach presented in this work neglected room reverberation and echoes as well as background noise which exist in a real recording scenario. A dereverberation approach [51], [52] together with a noise-suppression module can be integrated to each other, in order to mitigate the reverberation and background noise problem for achieving a robust speech separation system in a practical scenario. As an example, [52] proposed to suppress noise components by spectral subtraction method, followed by a dereverberation module applied to the noise-suppressed signal. In this way, it is possible to dereverberate the received echoic signal as well as to reduce background noise from the corrupted signal recorded by one microphone, and then apply our separation approach to the enhanced mixed signal.

By assuming *a priori* knowledge of double-talk regions in a given mixed signal, we apply the 2-D search only to mixed frames to find the optimal states of the underlying speaker models (codebooks). For the single-talk regions, we simply re-synthesize the single-talk speaker signals according to the corresponding speaker codebooks. It should be noted that, the quantitative performance reported in our experiments are for the entire utterances.

The proposed approach cuts the computational cost in separation by substituting STFT feature vectors with sinusoidal peaks. We conducted simulations to quantify the computational complexity of the proposed method for ten 2-s mixtures. We observed that the STFT-VQ approach, used as our benchmark, took in average 26.71 s for separating each frame while the proposed one required 5.55 s. Hence, the proposed approach leads to approximately 5 times less computation time.

The upper-bound separation results presented here confirmed recent findings in [47], where it was demonstrated that by applying the split-VQ codebooks composed of sinusoidal parameters, it is possible to achieve a better quantization performance in terms of the re-synthesized speech quality compared to the conventionally used STFT or its logarithm as the selected feature vectors. This agrees with the conclusion in [31] stating that the ultimate quality of model-based speech enhancement system is upper-bounded by the performance of the coder used. Similarly in SCSS, the selected feature type along with the statistical model determines the separation upper-bound performance. Therefore, to achieve an acceptable separation upper-bound, the selected feature type for SCSS is required to perform a high quantization performance that is in agreement with the results reported in [17], [18], and [47]. It was shown in [17] that by applying a subband perceptually weighted transform on the STFT vectors, it is possible to achieve improvements in the perceptual quality of the recovered signals especially at low SSRs. Similarly, in this work we observed that by changing STFT features with sinusoidal parameters, it is possible to achieve improvements in the separation performance.

We note that the method can also be generalized into speech enhancement in highly colored noise scenarios including babble or harmonic noise [1]–[7]. In such scenarios, the mixed signal includes less harmonics which makes the separation task rather difficult. As a future work, the proposed method is expected to be appropriately applied to speech enhancement scenarios with highly colored noise. The proposed method in this paper offers an attractive candidate similar to the weighted codebook-mapping (WCBM) in [53], as an effective tool for speech enhancement. The WCBM in [53], however, was based on harmonic plus noise model (HNM) feature parameters that require voicing estimation and pitch. In contrast, the proposed method in this research is independent of pitch estimates and benefits from the advantages inherited from modified sinusoidal features, split-VQ codebooks, and sinusoidal mixture estimator presented in this work.

VII. Conclusion

In this paper, we presented new results on single-channel speech separation and also proposed a new method based on sinusoidal parameters. In our proposed method, we suggested to use a mixture estimator in the sinusoidal domain targeted to find the optimal sinusoidal codevectors selected from speaker codebooks that, when combined, best describe the observed mixed signal in each frame. The key idea in the proposed method is to separate the signals by mapping their mixture frames onto the joint subspaces of the sources and then compute the parts that fall in each subspace. We studied the performance of the proposed method and compared its results with those obtained by previous SCSS methods. Through extensive simulations, and by comparison to other methods, it was observed that the proposed method leads to rather good re-synthesized speech quality as well as lower undesirable cross-talk for both target and interference signals. It was also concluded that minimization at sinusoidal frequencies of the mixed signal, used in the proposed mixture estimator, makes significant improvement compared to both mask approach (log-max and Wiener filtering) and STFT-based VQ approaches. To assess the improvements made by the proposed method, we used PESQ as objective measure and MUSHRA listening tests as subjective evaluation for both speaker-dependent and gender-dependent scenarios. It was observed that the proposed method achieved a higher score compared to other separation methods. In addition, it was observed that by increasing the signal-to-signal ratio, the proposed method asymptotically reaches the upper-bound separation performance (ideal separation scenario). According to the MUSHRA listening tests, the perceived speech quality of the proposed method was the highest both in speaker-dependent and gender-dependent scenarios. Finally, compared to other methods, the proposed method achieved lower cross-talk and was mostly preferred by the listeners.

References

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

[2] S. Srinivasan and D. Wang, "A model for multitalker speech perception," *J. Acoust. Soc. Amer.*, vol. 124, no. 5, pp. 3213–3224, Nov. 2008.

[3] H. Sameti, H. Sheikhzadeh, L. Deng, and R. L. Brennan, "HMM-based strategies for enhancement of speech signals embedded in nonstationary noise," *IEEE Trans. Speech Audio Process.*, vol. 6, no. 5, pp. 445–455, Sep. 1998.

[4] S. Srinivasan, J. Samuelsson, and W. B. Kleijn, "Codebook driven short-term predictor parameter estimation for speech enhancement," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 1, pp. 163–176, Jan. 2006.

[5] S. Srinivasan, J. Samuelsson, and W. B. Kleijn, "Codebook-based Bayesian speech enhancement for nonstationary environments," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 2, pp. 441–452, Feb. 2007.

[6] Y. Ephraim and H. L. V. Trees, "A signal subspace approach for speech enhancement," *IEEE Trans. Speech Audio Process.*, vol. 3, no. 4, pp. 251–266, Jul. 1995.

[7] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean square error short-time spectral amplitude estimator," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 32, no. 6, pp. 1109–1121, Dec. 1984.

[8] S. H. Jensen, P. C. Hansen, S. D. Hansen, and J. A. Sorensen, "Reduction of broad-band noise in speech by truncated QSVD," *IEEE Trans. Speech Audio Process.*, vol. 3, no. 6, pp. 439–448, Nov. 1995.

[9] P. C. Hansen and S. H. Jensen, "Prewhitening for rank-deficient noise in subspace methods for noise reduction," *IEEE Trans. Signal Process.*, vol. 53, no. 10, pp. 3718–3726, Oct. 2005.

[10] P. C. Hansen and S. H. Jensen, "Subspace-based noise reduction for speech signals via diagonal and triangular matrix decompositions: Survey and analysis," *EURASIP J. Adv. Signal Process.*, vol. 1, p. 24, Mar. 2007.

[11] D. Wang, "Time-frequency masking for speech separation and its potential for hearing aid design," *Trends in Amplification*, vol. 12, no. 4, pp. 332–353, Dec. 2008.

[12] J. R. Hershey, S. J. Rennie, P. A. Olsen, and T. T. Kristjansson, "Superhuman multi-talker speech recognition: A graphical modeling approach," *Elsevier Comput. Speech Lang.*, vol. 24, no. 1, pp. 45–66, Jan. 2010.

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

[14] S. T. Roweis, "One microphone source separation," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2000, pp. 793–799.
[15] S. T. Roweis, "Factorial models and refiltering for speech separation and denoising," in *Proc. Eurospeech*, 2003, pp. 1009–1012.
[16] M. H. Radfar, R. M. Dansereau, and A. Sayadiyan, "Monaural speech segregation based on fusion of source-driven with model-driven techniques," *Elsevier Speech Commun.*, vol. 49, no. 6, pp. 464–476, Jun. 2007.
[17] P. Mowlaee, A. Sayadiyan, and H. Sheikhzadeh, "Evaluating single channel speech separation performance in transform-domain," *J. Zhejiang Univ.-SCIENCE C, Comput. Electron.*, vol. 11, no. 3, pp. 160–174, Jan. 2010.
[18] D. P. W. Ellis and R. J. Weiss, "Model-based monaural source separation using a vector-quantized phase-vocoder representation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, May 2006, vol. 5, pp. 957–960.
[19] M. J. Reyes-Gomez, D. P. W. Ellis, and N. Jojic, "Multiband audio modeling for single-channel acoustic source separation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, May 2004, vol. 5, pp. 641–644.
[20] D. Wang and G. J. Brown, *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*. New York: Wiley-IEEE Press, 2006.
[21] Y. Shao, S. Srinivasan, Z. Jin, and D. Wang, "A computational auditory scene analysis system for speech segregation and robust speech recognition," *Elsevier Comput. Speech Lang.*, vol. 24, no. 1, pp. 77–93, 2010.
[22] G. Hu and D. Wang, "Monaural speech segregation based on pitch tracking and amplitude modulation," *IEEE Trans. Neural Netw.*, vol. 15, no. 5, pp. 1135–1150, Sep. 2004.
[23] P. Li, Y. Guan, S. Wang, B. Xu, and W. Liu, "Monaural speech separation based on MAXVQ and CASA for robust speech recognition," *Elsevier Comput. Speech Lang.*, vol. 24, no. 1, pp. 30–44, Jan. 2010.
[24] B. Hanson and D. Wong, "The harmonic magnitude suppression (HMS) technique for intelligibility enhancement in the presence of interfering speech," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Mar. 1984, pp. 65–68.
[25] J. Naylor and S. Boll, "Techniques for suppression of an interfering talker in co-channel speech," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Apr. 1987, vol. 12, pp. 205–208.
[26] D. Chazan, Y. Stettiner, and D. Malah, "Optimal multi-pitch estimation using the EM algorithm for co-channel speech separation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Apr. 1993, vol. 2, pp. 728–731.
[27] J. Barker, M. Ning, A. Coy, and M. Cooke, "Speech fragment decoding techniques for simultaneous speaker identification and speech recognition," *Elsevier Comput. Speech Lang.*, vol. 24, no. 1, pp. 94–111, 2010.
[28] M. Karjalainen and T. Tolonen, "Multi-pitch and periodicity analysis model for sound separation and auditory scene analysis," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 1999, pp. 929–932.
[29] L. Benaroya, F. Bimbot, and R. Gribonval, "Audio source separation with a single sensor," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 1, pp. 191–199, Jan. 2006.
[30] M. H. Radfar and R. M. Dansereau, "Single-channel speech separation using soft mask filtering," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 8, pp. 2299–2310, Nov. 2007.
[31] Y. Ephraim, "Statistical-model-based speech enhancement systems," *Proc. IEEE*, vol. 80, no. 10, pp. 1526–1555, Oct. 1992.
[32] 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 J. Audio, Speech, Music Process.*, vol. 1, p. 15, Mar. 2007.
[33] J. Ming, T. J. Hazen, and J. R. Glass, "Combining missing-feature theory, speech enhancement, and speaker-dependent/-independent modeling for speech separation," *Elsevier Comput. Speech Lang.*, vol. 24, no. 1, pp. 67–76, Jan. 2010.
[34] D. Burshtein and S. Gannot, "Speech enhancement using a mixture-maximum model," *IEEE Trans. Speech Audio Process.*, vol. 10, no. 6, pp. 341–351, Sep. 2002.
[35] Y. Shao and D. Wang, "Model-based sequential organization in cochannel speech," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 1, pp. 289–298, Jan. 2006.
[36] M. G. Christensen and A. Jakobsson, *Multi-Pitch Estimation*. San Rafael, CA: Morgan and Claypool, 2009, Synthesis Lectures on Speech and Audio Processing.
[37] C. J. Moore, *An Introduction to the Psychology of Hearing*. San Diego, CA: Academic, 2003.
[38] T. W. Parsons, "Separation of speech from interfering speech by means of harmonic selection," *J. Acoust. Soc. Amer.*, vol. 60, pp. 911–918, 1976.
[39] T. F. Quatieri and R. G. Danisewicz, "An approach to co-channel talker interference suppression using a sinusoidal model for speech," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 38, no. 1, pp. 56–69, Jan. 1990.
[40] F. M. Silva and L. B. Almeida, "Speech separation by means of stationary least-squares harmonic estimation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Apr. 1990, vol. 2, pp. 809–812.
[41] T. Tolonen, "Methods for separation of harmonic sound sources using sinusoidal modeling," in *Proc. 106th Audio Eng, Society Conv.*, 1999.
[42] M. Cooke, J. R. Hershey, and S. J. Rennie, "Monaural speech separation and recognition challenge," *Elsevier Comput. Speech Lang.*, vol. 24, no. 1, pp. 1–15, 2010.
[43] R. McAulay and T. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-34, no. 4, pp. 744–754, Aug. 1986.
[44] P. Mowlaee, A. Sayadiyan, and H. Sheikhzadeh, "FDMSM robust signal representation for speech mixtures and noise corrupted audio signals," *IEICE Electron. Express*, vol. 6, no. 15, pp. 1077–1083, 2009.
[45] H. Hayes, *Statistical Digital Signal Processing and Modeling*. New York: Wiley, 1996.
[46] P. Mowlaee, M. G. Christensen, and S. H. Jensen, "Improved single channel speech separation using sinusoidal modeling," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Mar. 2010, pp. 21–24.
[47] P. Mowlaee and A. Sayadiyan, "Model-based monaural sound separation by split-VQ of sinusoidal parameters," in *Proc. Eur. Signal Process. Conf.*, Aug. 2008.
[48] "Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs," ITU-T Rec. p. 862, 2001.
[49] "Method for the subjective assessment of intermediate quality level of coding systems," ITU-R BS.1534-1 2003.
[50] P. Mowlaee, R. Saeidi, Z. H. Tan, M. G. Christensen, P. Fränti, and S. H. Jensen, "Joint single-channel speech separation and speaker identification," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Mar. 2010, pp. 4430–4433.
[51] T. Yoshioka, T. Nakatani, and M. Miyoshi, "Integrated speech enhancement method using noise suppression and dereverberation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 17, no. 2, pp. 231–246, Feb. 2009.
[52] K. Kinoshita, T. Nakatani, and M. Miyoshi, "Spectral subtraction steered by multi-step forward linear prediction for single channel speech dereverberation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2006, pp. 817–820.
[53] E. Zavarehei, S. Vaseghi, and Q. Yan, "Noisy speech enhancement using harmonic-noise model and codebook-based post-processing," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 4, pp. 1194–1203, May 2007.

Pejman Mowlaee (S'07) was born in Bandar Anzali, Iran, in March 1983. He received the B.Sc. and M.Sc. degrees both with straight honors from Guilan University, Rasht, Iran, and Iran University of Science and Technology, Tehran, in 2005 and 2007, respectively. He is currently pursuing the Ph.D. degree at Aalborg University, Aalborg, Denmark.

His research interests include digital signal processing theory and methods with application to speech processing and machine learning, in particular single-channel speech separation and enhancement, and speech coding.

Mr. Mowlaee has received several awards during his academic career, to name a few: Young Researchers Award for the M.Sc. degree, Honored M.Sc. thesis in nation-wide contest between Iranian electrical engineering students, and was known as a talented student at the top class of Tehran polytechnic.

Mads Græsbøll Christensen (S'00–M'05) was born in Copenhagen, Denmark, in March 1977. He received the M.Sc. and Ph.D. degrees from Aalborg University, Aalborg, Denmark, in 2002 and 2005, respectively.

He was formerly with the Department of Electronic Systems, Aalborg University, and is currently an Associate Professor in the Department of Architecture, Design, and Media Technology. He has been a Visiting Researcher at Philips Research Labs, Ecole Nationale Supérieure des Télécommunications (ENST), University of California, Santa Barbara (UCSB), and Columbia University. He has published more than 75 papers in peer-reviewed conference proceedings and journals and is coauthor (with A. Jakobsson) of the book *Multi-Pitch Estimation* (Morgan & Claypool, 2009). His research interests include digital signal processing theory, and methods with application to speech and audio, in particular parametric analysis, modeling, and coding.

Dr. Christensen has received several awards, namely an IEEE International Conference on Acoustics, Speech, and Signal Processing Student Paper Contest Award, the Spar Nord Foundation's Research Prize awarded annually for his Ph.D. dissertation, and a Danish Independent Research Councils Young Researcher's Award. He is an Associate Editor for the IEEE SIGNAL PROCESSING LETTERS.

Søren Holdt Jensen (S'87–M'88–SM'00) received the M.Sc. degree in electrical engineering from Aalborg University, Aalborg, Denmark, in 1988 and the Ph.D. degree in signal processing from the Technical University of Denmark, Lyngby, in 1995.

Before joining the Department of Electronic Systems, Aalborg University, he was with the Telecommunications Laboratory of Telecom Denmark, Ltd., Copenhagen, Denmark; the Electronics Institute of the Technical University, Denmark; the Scientific Computing Group of Danish Computing Center for Research and Education (UNI-C), Lyngby; the Electrical Engineering Department, Katholieke Universiteit Leuven, Leuven, Belgium; and the Center for PersonKommunikation (CPK), Aalborg University. He is Full Professor and Head of the Multimedia, Information, and Signal Processing Section. He is currently heading a research team working in the area of numerical algorithms, optimization techniques, and signal processing for speech and audio processing, image and video processing, multimedia technologies, and digital communications.

Prof. Jensen was an Associate Editor for the IEEE TRANSACTIONS ON SIGNAL PROCESSING and is currently Member of the Editorial Board of *Elsevier Signal Processing* and the *EURASIP Journal on Advances in Signal Processing*. He is a recipient of the European Community Marie Curie Fellowship, former Chairman of the IEEE Denmark Section, and Founder and Chairman of the IEEE Denmark Section, Signal Processing Chapter.