Aalborg Universitet

Improved single-channel speech separation using sinusoidal modeling

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.2010.5496263

Publication date:
2010

Document Version
Accepted author manuscript, peer reviewed version

Link to publication from Aalborg University

Citation for published version (APA):
Mowlaee, P., Christensen, M. G., & Jensen, S. H. (2010). Improved single-channel speech separation using sinusoidal modeling. *I E E E International Conference on Acoustics, Speech and Signal Processing. Proceedings*, *2010*, 21-24. https://doi.org/10.1109/ICASSP.2010.5496263

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal -

Take down policy
If you believe that this document breaches copyright please contact us at vbn@aub.aau.dk providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from vbn.aau.dk on: December 04, 2025

IMPROVED SINGLE-CHANNEL SPEECH SEPARATION USING SINUSOIDAL MODELING

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

We present a novel single-channel separation approach to improve the separation performance while recovering the signals from a mixture. The key idea in this research is to employ a mixture estimator based on unconstrained modified sinusoidal parameters. Compared to the mixmax (binary mask) and Wiener filter (softmask) approaches, the proposed approach works independently of pitch estimates. Furthermore, it is observed that it can achieve acceptable perceptual speech quality with less cross-talk at different signal-to-signal ratios while bringing down the complexity by replacing STFT with sinusoidal parameters. Improvements made by the proposed approach are demonstrated by employing PESQ as our objective measure and MUSHRA listening test as our subjective evaluation.

***Index Terms*—** Mixture estimation, single-channel speech separation, mask-based methods, speaker codebook.

1. INTRODUCTION

Although there have been recent advances in many speech enhancement methods [1], single-channel speech separation (SCSS) systems with high quality are still of great importance and remain as an unsolved problem. Ideal separation systems are targeted to provide accurate estimations for both sources from their mixture. In this aspect having a high quality separation system would play an integral part offering robustness to many practical applications including speech recognition and speaker identification from mixtures of signals.

Previous single-channel speech separation systems are mainly divided into two categories: source driven [2], [3] and model-based methods [4], [5]. Most methods in either group are often required to estimate two masks at each frame and applying them to the given mixture to recover the unknown sources [3], [6–8]. The mask to be applied could be either binary (hard decision) [3], [6], [7] or soft mask [8] leading to MAX-VQ system (with log-max mixture approximation) [6], [7] and Wiener filter (soft masks) [8], respectively. Most of the previous separation systems led to rather satisfying performance for both sources mostly at signal-to-signal ratio (SSR) around 0 dB [4–8]. However, it is often expected that the SSR level vary from 0 dB since the underlying speakers in the mixture often mask each other as time evolves. As a consequence, the SSR level can vary in frames [1] making signal recovery of speakers difficult. One reason for this problem is the fact that usually at a frame level one speaker signal dominates the other and the energies of sources collide at a time-frequency cell. The mask-based methods explicitly suggest to filter out one of the speakers to recover the target speaker. This would degrade the performance of the signal recovery for the masked speaker. Further, using masks inevitably causes cross-talk and artifacts in the separated signals as reported in [3]. From these aspects, there is a strong motivation in finding novel methods to recover both signals at different SSR levels. According to the results in [3] the Computationally Auditory Scene Analysis (CASA) often lacks enough perceptual quality due to severe cross-talk problems in the separated output signals. The separation performance of CASA-based methods are mainly determined by multi-pitch estimation accuracy. Further, according to the simulation results given in [9], the pitch estimation shows large gross errors especially at low SSR levels because of energetic masking. In this aspect, integration of pitch as proposed in [6] may not be the best solution at low SSR levels, since extracting pitch frequencies from a mixture is both challenging and difficult [9]. This, as a consequence, causes errors in mixture estimation stage which is targeted to find the pair of states of composite sources of the speakers that best fit the given mixture. These indices are then sent to the reconstruction stage, therefore any mixture estimation error would degrade the perceptual quality of the synthesized outputs. Compared to the mask-based methods, a model-based system is able to achieve a rather acceptable separation quality for known speakers at SSR of 0 dB. Model based systems are mostly based on statistical models including vector quantization (VQ) [4–6], Gaussian mixture models (GMM) [8] and Hidden markov models (HMM) [7]. As the most representative method of this group, the MAX-VQ separation system tries to produce two masks based on the estimated VQ states [2], [6], [7] and integrate them with the log-max approximation as its mixture estimation. According to the results reported in [2], [5], [6] using these estimated masks provides re-synthesis signals often corrupted with undesirable cross-talk effects. Furthermore, based on the analysis recently given in [10], we showed that log-max approximation in [6], [7] and Wiener filter [8] are both biased mixture estimators.

The main purpose of this paper is to propose a novel mixture estimator and apply it to modified unconstrained sinusoidal parameters. The separation result of the proposed method is compared with MAX-VQ [7], Wiener filter [8] and model-based VQ system by [4]. The paper is structured as follows: In the next section, we introduce modified unconstrained sinusoidal parameters to be employed as feature parameters. Parameter estimation is presented and followed by the proposed sinusoidal mixture estimator. We also explain the procedure to produce split-VQ speaker models composed of sinusoidal parameters to be used in our proposed method. In Section 3, we present the experimental results with PESQ as an objective measure and MUSHRA test as a subjective measure. Section 4 features the discussions and future work and Section 5 concludes on the work.

2. PROPOSED SEPARATION METHOD

2.1. Sinusoidal model

Each speaker signals is denoted by $s_j(n)$ with $j \in [1, 2]$ and their mixture is shown by $z(n)$ with $n = 0, \ldots, N-1$ as the time sample index where N is the window length in samples. The sinusoidal

model of speech in a fixed signal frame is

$$s(n) = \sum_{i=1}^{M} a_i \cos(2\pi f_i n + \phi_i) + e(n) \quad 0 \leq n \leq N-1, \qquad (1)$$

where $e(n)$ is the sinusoidal modeling error assumed as an additive noise, M is model order and $i \in [1, M]$ is an index used to refer the ith sinusoidal component characterized by f_i, a_i, and ϕ_i as the frequency, amplitude, and phase, respectively. As a parametric feature vector we have $\mathbf{\Theta}$=$[\mathbf{a}, \mathbf{f}, \phi]$ of size $M \times 3$.

2.2. Sinusoidal Modeling and Parameter Estimation

We consider two modifications on unconstrained sinusoidal model developed in [11]. The modifications we made are described as follow; 1) the spectral coefficients are translated to Mel scale to take into account the logarithmic sensitivity of human auditory system, and 2) at each Mel band, the spectral peak with the highest amplitude is selected. By employing these two foundations as our sinusoidal parameter estimation rule, we find one peak per band and end up with three $M \times 1$ vectors of amplitude, frequency and phase for each speaker signal or their mixture. We define $\mathbf{v}_i = [1 \quad e^{j2\pi f_i} \quad \ldots \quad e^{j2\pi f_i (N-1)}]^T$ with $i \in [1, M]$ as the sinusoidal frequency vector of dimension $N \times 1$ and f_i is the selected peak at the ith band. All estimated sinusoidal frequency vectors for each speaker signal are represented in a matrix format as

$$\mathbf{V} = [\mathbf{v}_1 \quad \mathbf{v}_2 \quad \ldots \quad \mathbf{v}_M]^T \quad i \in [1, M] \quad , \qquad (2)$$

where $\mathbf{V}$ is an $M \times N$ Vandermonde matrix whose rows are $\mathbf{v}_i$. Then signal representation in terms of sinusoids is an $N \times 1$ vector given by $\hat{\mathbf{s}} = \mathbf{V}^T \mathbf{a}$ where $\mathbf{a} = [a_1 \quad \ldots \quad a_M]^T$ and $\hat{\mathbf{s}}$ the reconstructed signal by the sinusoidal peaks in (2). Defining the complex amplitude for each sinusoid as $a_i = A_i e^{j\phi_i}$, the objective of the parameter estimation stage is to find peaks characterized by an amplitude frequency pair given by $A_i = |S_i(f_i)|$ and $f_i = \arg\max_{f \in \mathcal{F}_i} \log |S_i(f)|$, respectively where $\mathcal{F}_i$ denotes a set composed of all the frequencies within the frequency spectrum in the ith subband denoted by $S_i(f)$.

2.3. Sinusoidal Mixture Estimator

According to previous Section, we model the mixed signal as $\mathbf{z} = \mathbf{V}_z^T \mathbf{a}_z$ where $\mathbf{V}_z$ is a Vandermonde matrix composed of M frequency vectors of $N \times 1$ as $\mathbf{v}_{z,i} = [1 \quad e^{j\omega_{z,i}} \quad \ldots e^{j\omega_{z,i}(N-1)}]^T$ related to $\{\omega_{z,i}\}$ as the set of sinusoidal frequency peaks retained for the mixture at the ith band. We derive mixture estimator based on unconstrained sinusoidal parameters of the underlying speakers and their mixture. The key idea is to project the mixture to its sinusoidal subspace spanned by the columns of $\mathbf{\Theta}_z$ and attempt to find a cost function to be minimized in mixture estimation stage. Based on the model in (1), for each speaker the power spectrum at the ith band is

$$P(e^{j\omega}) = \sigma_i^2 + A_i^2 \left[\delta(\omega - \omega_i) + \delta(\omega + \omega_i)\right] \quad , \qquad (3)$$

where we can replace ω_i with underlying speakers signals frequency sets given by $\{\omega_{1,i}\}$ and $\{\omega_{2,i}\}$ or the mixture denoted by $\{\omega_{z,i}\}$ to define the related power spectrum. A cost function is defined as the squared error between the power spectra of the given and estimated mixture to be sampled only at sinusoidal peaks defined by set $\{\omega_{z,i}\}$. Sampling at sinusoidal frequencies of the mixed signal $\{\omega_{z,i}\}$ is not necessarily synchronous with $\{\omega_{1,i}\}$ and $\{\omega_{2,i}\}$, bringing the requirement of using 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 given by $E\{\hat{P}(e^{j\omega})\} = P(e^{j\omega}) * W(e^{j\omega})$ where $E\{\cdot\}$ denotes expectation operator. The expected value for the mixture approximation error at the ith band is

$$\begin{aligned} E\{\epsilon_i(e^{j\omega})\} &= E\{\hat{P}_z(e^{j\omega}) - \hat{P}_1(e^{j\omega}) - \hat{P}_2(e^{j\omega})\} \qquad (4) \\ &= \sigma_{\epsilon,i}^2 + A_{z,i}^2 [W(e^{j(\omega - \omega_{z,i})}) + W(e^{j(\omega + \omega_{z,i})})] \\ &\quad - \sum_{k=1}^{2} A_{k,i}^2 [W(e^{j(\omega - \omega_{k,i})}) + W(e^{j(\omega + \omega_{k,i})})], \qquad (5) \end{aligned}$$

where 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 key idea is to sample the expected mixture estimation error in (5) at sinusoidal frequencies of the mixture per ith band defined by set $\{\omega_{z,i}\}$. Replacing ω by $\omega_{z,i}$ in (5) we get

$$\epsilon_i = A_{z,i}^2 - A_{1,i}^2 W(e^{j(\omega_{z,i} - \omega_{1,i})}) - A_{2,i}^2 W(e^{j(\omega_{z,i} - \omega_{2,i})}), \qquad (6)$$

which addresses the mixture approximation error defined between the original and estimated spectra at the ith subband. $A_{1,i}$,$A_{2,i}$ and $A_{z,i}$ indicate the first, second and the mixture sinusoidal amplitude selected at the ith band. According to (1), the mixture approximation error energy converges to zero when the underlying speaker spectra are highly harmonic. Then the mixture estimation error energy termed as d at a given frames is $d = \sum_{i=1}^{M} |\epsilon_i|^2$. Finally, the sinusoidal mixture estimation is accomplished by searching for the optimal states of the composite sources denoted by $\{q^*, t^*\}$ obtained by solving the following minimization problem at each frame

$$\{q^*, t^*\} = \arg\min_{q,t} d_{q,t} \quad , \qquad (7)$$

where q, t can be any possible state in the speaker models and $d_{q,t}$ is a 2D cost function defined based on the mixture approximation error in (6). At each frame, by in-place minimization of $d_{q,t}$ in (7), we achieve two states of the speaker models that when combined best fit the mixture. The selected codebook indices are then sent to a weighted overlap-add (OLA) to reconstruct two separated signals.

2.4. Split-VQ Speaker Codebooks

Recently, we reported improvements by applying perceptually weighted subband on the short-time Fourier transform (STFT) features especially at low SSR [5]. It was observed that the selected feature type along with the statistical model determine the upper bound of separation performance. Therefore, to achieve the upper bound separation quality, the selected feature for SCSS is required to perform a high quantization performance which is in agreement with the results reported in [4], [5], [12]. This is in accordance with the conclusion in [13] stating that the ultimate quality of the model-based speech enhancement system is upper bounded by the performance of the coder used. In this respect it was shown in [12] that by applying the split-VQ codebooks on sinusoidal parameters, it is possible to achieve a better quantization performance compared to the conventionally used STFT features. Due to this, we use split-VQ codebooks on sinusoidal amplitude and frequencies of the underlying signals as our speaker codebooks. Sinusoidal parameters from the training dataset of each speaker in the mixture are extracted and results in matrices whose entries are comprised of two distinctive parts; amplitude and frequency each of dimension $1 \times M$. Similar to [12], we apply two different distance measures to produce codebooks of amplitude and frequency. For the amplitude part we apply $d_a(\mathbf{a}, \hat{\mathbf{a}}) = \frac{1}{\|\mathbf{a}\|} \sum_{i=1}^{M} (a_i - \hat{a}_i)^2$ where $d_a(\cdot)$ denotes the distance measure applied to the amplitude

part, M is the number of sinusoids used, and $\hat{}$ denotes the coded parameters. Let M_a be the codebook size for the amplitude part of our split-VQ codebook. After establishing M_a amplitude reference vectors, we select the most appropriate frequency vectors for each amplitude codeword. Another VQ of a lower size is performed on the frequency candidates for amplitude codeword. A VQ with frequency codebook size of $1, 2$ or 4 bits was found as an appropriate choice [12]. To produce frequency codevectors, we apply a distance measure defined between the frequency part of the trained data matrix (defined by $\mathbf{V}$ in (2)) and their related codevectors denoted by $\hat{\mathbf{V}}$ as $d_w(\mathbf{V}, \hat{\mathbf{V}}) = \sum_{i=1}^{M} w_i(\mathbf{v}_i - \hat{\mathbf{v}}_i)^2$ where $d_w(\cdot)$ is defined as a weighted square error measure with $w_i = \frac{a_i}{\|\mathbf{a}\|}$ as the energy normalized amplitude vector used as a dynamic weighting to weight the Euclidean distance measure proportional to the sinusoidal amplitude at the peak frequencies indicated by $\mathbf{V}_i$. Concatenating the coded amplitude and frequency vectors denoted by $\hat{\mathbf{v}}$ and $\hat{\mathbf{a}}$, respectively, we achieve coded vectors in split-VQ of each speaker model.

3. EXPERIMENTAL RESULTS

3.1. Dataset used and Separation Scenario

To evaluate the proposed separation algorithm, we selected four speakers including two male (speakers 9 and 19) and two female speakers (4 and 23) from the database [14]. Ten minutes of the speech signals of each speaker was used to produce split-VQ [12] and STFT codebooks similarly to [4], [5], [7], all with a codebook size of 2048 (for practical reasons 11 bits are used for amplitude and 3 bits for frequency part in split-VQ codebooks). As our separation scenario, we select two speaker signals, and mixed them together at a certain SSR ranging within $[-18, 18]$. The sampling frequency was decreased to 8 kHz from the original 25 kHz. A Hanning window of duration 32 ms is used with a frame rate of 8 ms. The benchmark methods used in our simulations are the mask based methods both binary mask (log-max) [6], [7] and Wiener filter (soft mask) [8]. Since most separation systems predominantly employ STFT or its logarithm as their signal representation [4], [6–8], we include the results obtained by the model-based VQ in [4], [5], [7].

3.2. Objective and Subjective Results

As a proof of concept, we evaluate the separation performance of the proposed method in a speaker dependent scenario. The core of the separation scenario is composed of two trained codebooks. Simulation results are conducted to assess the separation performance of the proposed method and compared them to those obtained by other separation methods. As our testing phase, fifteen pairs of utterances of each speaker (not used in the training set) were randomly selected to make mixtures. The separation results are quantified using PESQ [15]. The results for the separated signals were averaged at each SSR level over all pairs of test signals. Fig. 1 illustrates the separation results obtained by different methods for each speaker output. We also include the upper bound for the separation performance where it is assumed that the optimal indices are known *a priori*. From Fig. 1 it is observed that the proposed method consistently achieves the highest PESQ score compared to the mask-based approaches of MAX-VQ in [2], [6], [7] and Wiener filter in [8]. The mask-based methods introduce significant mixture estimation error especially at low or high SSR levels. Further, compared to the STFT-based VQ system in [4], [5] denoted by STFT-CB in Fig. 1(a) and (b), the proposed separation approach outperforms the STFT upper bound performance [5]. From curves shown in Fig. 1 it is observed that the proposed mixture estimator asymptotically reaches to the upper bound performance achieved by the split-VQ codebooks [12] while there is on average a large gap between the separation upper bound and those obtained by the mask-based methods [6], [7] and the model-based VQ in [4], [5]. However, all methods exhibit their best performance as SSR increases for the target speaker. The test and the processed signals used in our MUSHRA test are presented on our webpage[1].

Table 1. Labels of the excerpts used in MUSHRA test.

Excerpt	Separation method and SSR scenario
BMssr0	Binary mask at SSR=0 dB
BMssr-18	Binary mask at SSR= -18 dB
SINssr0	Proposed method at SSR= 0 dB
SINssr-18	Proposed method at SSR= -18 dB
FFTssr0	STFT-based VQ at SSR=0 dB
FFTssr-18	STFT-based VQ at SSR= -18 dB

As our second experiment we set up a MUlti-Stimulus test with Hidden Reference and Anchors (MUSHRA) listening test as described in ITU-R BS.1534-1 [16] in order to assess the perceived speech quality of the separated signals. Eight listeners participated in the test (the authors not included) and the items used in our listening test are the separated signals produced by different methods at certain SSRs. Fig. 2 depicts the mean opinion score (MOS) obtained from different speech separation methods averaged over all listeners. The excerpts used are shown in Table 1. All of the played signals were monophonic sampled at 8 kHz of duration 2 sec. For each excerpt the listeners were asked to rank eight different separated signals relative to a known reference on a score from 0 to 100. The excerpts are composed of the hidden reference (denoted by HR), an anchor low-pass filtered at 2 kHz (denoted by Anchor 1). The remaining six excerpts are the separated signals defined in Table 1.

In our listening test, the separated signals produced by binary mask (MAX-VQ) in [6], [7] and the STFT-based VQ system [4], [5] were included. Two extreme cases of SSR level as 0 and -18 dB are included. It is observed that the proposed sinusoidal mixture estimator scores about twenty points higher on average than the mask-based method, and more than 25 points higher than STFT-based method. According to Fig. 2, no overlap exists between the proposed method and the benchmark methods. Hence, it can be concluded that the proposed method can achieve statistically significant improvement compared to other methods and consistently improves the performance of the synthesized speech for both target and interference separated signals. Compared to the mask-based approach, the proposed method shows improvements in the perceived signal quality. As indicated by the listening experiments, the separated output for the MAX-VQ method was found to suffer from severe crosstalk. Tests also revealed that the separation performance of the mask-based methods (especially at SSR=0 dB where their separation performance is often reported) do not necessarily produce the highest perceived quality for the separated signals. This is observed by comparing the MOS in Fig. 2 for the BMssr0 and SINssr0.

4. DISCUSSION AND FUTURE WORK

The results obtained in our simulations are in agreement with [3] stating that the separation quality degrades as the energetic masking takes place at some overlapping time-frequency cells. The sinusoidal features used in this work lead to a high frequency resolution peak picking, reflecting the pitch harmonic structure of single speaker signals and their mixture. In this aspect, the idea is conceptually similar to the motivations behind the use of GF in CASA [2], [3]. By selecting the peak with the highest amplitude we simply exclude peaks

[1] http://kom.aau.dk/∼pmb/IEEE_ICASSP.htm

mainly caused by windowing effect or modulation of low frequency components while still preserving high perceptual quality.

Comparing the upper bound separation performance in Fig. 1 confirms our recent findings in [5] stating that transforming full-band STFT features into perceptually weighted subbands can significantly provide improvements especially at low SSR levels. Correspondingly, the results in this paper show that by using split-VQ codebooks it is possible to achieve a higher separation upper bound compared to the conventionally used STFT features. The results presented here were in agreement with our recent findings in [5], [12], where the upper bound performance in SCSS was evaluated as the performance of the coder when the optimal codebook indices are known *a priori* (ideal separation).

In this paper, we only considered SCSS. Future work should consider the generalization of the proposed estimator for speech enhancement in non-stationary noise (babble or harmonic) where many researchers show growing interest in this field.

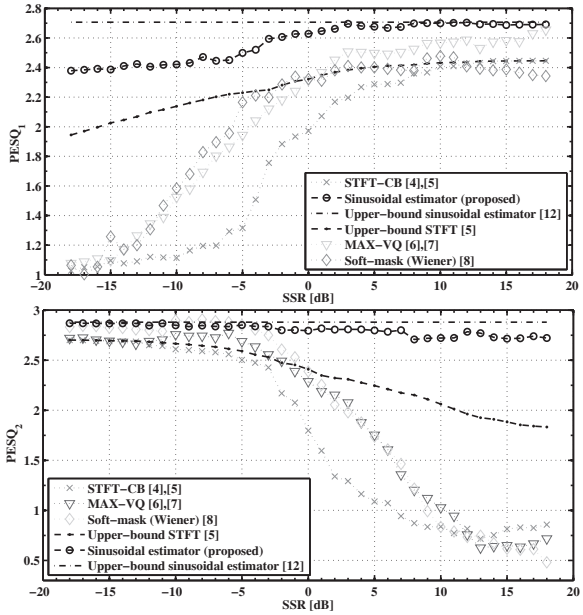

Fig. 1. Evaluation results for different separation methods in terms of PESQ for (a) speaker one (b) speaker two versus SSR.

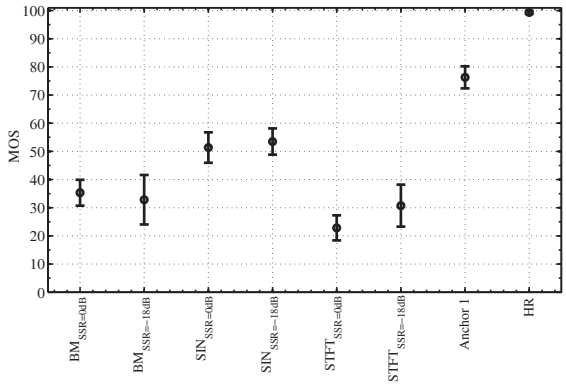

Fig. 2. MOS scores for different separation methods over all excerpts and all listeners. The error bars indicate the 95% confident intervals.

5. CONCLUSION

In this paper, a novel mixture estimator has been proposed and derived based on a modified unconstrained sinusoidal parameters to improve the speech separation performance. The method is independent of pitch estimates and offer a new approach for single-channel speech separation, where pitch estimation is sometimes difficult because of energetic masking occurred at time-frequency cells in a mixture at different SSR. Through several experiments it was observed that the proposed method achieved a higher score compared to mask-based methods of MAX-VQ, Wiener filter and the STFT VQ-based separation system especially at low SSR levels. As SSR increases, the proposed method asymptote its separation upper bound performance where it is assumed that the optimal indices are *a priori* available. According to the MUSHRA listening test, it was observed that the perceived speech quality of the proposed system was the highest. Further, compared to the benchmark methods, the proposed method achieved lower cross-talk and was mostly preferred by the listeners.

6. REFERENCES

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

[2] 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.

[3] G. Hu and D. Wang, "Monaural speech segregation based on pitch tracking and amplitude modulation," *Neural Networks, IEEE Transactions on*, vol. 15, no. 5, pp. 1135–1150, Sept. 2004.

[4] 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.

[5] 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*.

[6] 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*, , no. 84 186, March. 2007.

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

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

[9] M. Radfar, A. Sayadiyan, and R.M. Dansereau, "A new algorithm for two-talker pitch tracking in single channel paradigm," in *Proceedings of International Conference on Signal Processing ICSP*, Nov. 2006.

[10] 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.

[11] R. McAulay and T. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 34, no. 4, pp. 744–754, Aug. 1986.

[12] P. Mowlaee and Sayadiyan, "Model-based monaural sound separation by split-vq of sinusoidal parameters," in *European Signal Processing Conference EUSIPCO*, 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] "Method for the subjective assessment of intermediate quality level of coding systems.," ITU-R BS.1534-1.