Journal of Zhejiang University-SCIENCE C (Computers & Electronics)
ISSN 1869-1951 (Print); ISSN 1869-196X (Online)
www.zju.edu.cn/jzus; www.springerlink.com
E-mail: jzus@zju.edu.cn

Evaluating single-channel speech separation performance in transform-domain[#]

Pejman MOWLAEE, Abolghasem SAYADIYAN, Hamid SHEIKHZADEH

(*Department of Electronic Engineering, Amirkabir University of Technology, Tehran 15875-4413, Iran*)

E-mail: pmowlaee@ieee.org; {eeas335, hsheikh}@aut.ac.ir

Received Feb. 12, 2009; Revision accepted June 25, 2009; Crosschecked Dec. 8, 2009

Abstract: Single-channel separation (SCS) is a challenging scenario where the objective is to segregate speaker signals from their mixture with high accuracy. In this research a novel framework called subband perceptually weighted transformation (SPWT) is developed to offer a perceptually relevant feature to replace the commonly used magnitude of the short-time Fourier transform (STFT). The main objectives of the proposed SPWT are to lower the spectral distortion (SD) and to improve the ideal separation quality. The performance of the SPWT is compared to those obtained using mixmax and Wiener filter methods. A comprehensive statistical analysis is conducted to compare the SPWT quantization performance as well as the ideal separation quality with other features of log-spectrum and magnitude spectrum. Our evaluations show that the SPWT provides lower SD values and a more compact distribution of SD, leading to more acceptable subjective separation quality as evaluated using the mean opinion score.

Key words: Single-channel separation (SCS), Magnitude spectrum, Vector quantization (VQ), Subband perceptually weighted transformation (SPWT), Spectral distortion (SD)

doi:10.1631/jzus.C0910087 **Document code:** A **CLC number:** TN912.3

1 Introduction

High-quality speech enhancement is one of the most competitive ongoing research fields where attempts are conducted to arrive at more robust methods. When the interfering signal is speech, the well-known traditional speech enhancement methods including spectral subtraction, the Wiener filter (Benaroya *et al.*, 2006; Reddy and Raj, 2007) or other alternative methods cannot effectively recover the speech signals from the mixture. Single-channel separation (SCS) (Ellis and Weiss, 2006) was introduced as an effective tool for such applications. However, the challenge in the SCS methods is the ill-conditioning in the solutions that have to be mathematically addressed.

Possible SCS methods are mainly divided into two approaches: source-driven and model-driven. As a major example for the first group, computational auditory scene analysis (CASA) has widely been studied (Hu and Wang, 2004; Barker *et al.*, 2005). Generally speaking, CASA-based methods aim at segregating audio sources based on possible intrinsic perceptual acoustic cues from speech signals (Wang and Brown, 2006). For CASA systems, a reliable multi-pitch tracking component is critical to find pitch trajectories of two interfering speech signals (Wu *et al.*, 2003). In Hu and Wang (2004) and Srinivasan *et al.* (2006), a CASA system was built for the separation of speech mixture based on the estimated pitch. The CASA methods are fast and could be implemented in real time. There are, however, challenges that limit the pitch tracking performance for a mixture (Wang and Brown, 2006) as follows: (1) Most existing pitch estimation methods perform reliably only with clean speech signals that have a single pitch track or harmonically related sinusoids (Christensen and Jakobsson, 2009) with almost no background interference (Tolonen and Karjalainen,

[#] A preliminary version of this paper was presented at the 7th ACS/IEEE International Conference on Computer Systems and Applications, Rabat, Morocco, 2009

2000). (2) It is possible to perform a reliable pitch estimation using a mixture of a dominant (target) and a weaker (masking) signal as long as the pitches of the masking and target speech are different in a short frame (Gu and Stern, 2008). A high similarity between the interference and target pitch trajectories results in performance degradation of CASA methods (Srinivasan and Wang, 2008). (3) Because of energetic masking defined in Srinivasan and Wang (2008), the weaker signal frames are masked by the stronger ones complicating the pitch estimation. Accordingly, at target-dominant time-segments, it is possible to accurately track the pitch contour of only the dominant (target) signal. (4) Pitch tracking performance has not been promising for scenarios where the underlying signals include mixtures of unvoiced and voiced frames and as a result, the separated speech signals include severe cross-talks (Radfar *et al.*, 2007).

As an alternative SCS method, speech fragment decoding (Barker *et al.*, 2005) is known as a combinational method that successfully integrates source- and model-driven approaches. Barker *et al.* (2005) employed the method for a robust speech recognition task. In this method, a top-down search is performed with a bottom-up grouping along with sound fragments. Speech fragment decoding can also be used for the signal reconstruction of the target speech (Barker *et al.*, 2005; 2006). However, separation methods based on speech fragment decoding encounter difficulties when both sources are speech. Fragment decoding as presented in Barker and Shao (2007) requires the testing of all the possible fragment combinations by incorporating full hidden Markov models (HMMs) for speech signals. Among all possible segmentations, the one resulting in the lowest recognition error rate is selected at the decoder. Hence, the word-sequence with the highest likelihood is found. This brings high computational complexity making the method intractable. In addition, according to Barker and Shao (2007) the performance significantly degrades when the signal-to-noise ratio (SNR) is reduced from 6 to −9 dB. Additionally, reliable identification of the speech-dominated regions in the mixture is a challenging task, especially in the case of high energetic portions of the interferer that can easily be mistaken with the target speech (Barker *et al.*, 2005). Hence, the separation performance may significantly degrade when energetic masking occurs at a signal-to-signal ratio (SSR) of lower than −6 dB or so (Radfar *et al.*, 2006a). Additionally, at low SSRs, the segregated signals would suffer from cross-talks.

Due to the difficulties discussed, in this work we do not consider CASA methods including speech fragment decoding for separating speech mixtures at low SSRs.

As an alternative form of SCS, the model-based SCS is commonly employed to deliver an appropriate separation quality at SSRs typically at or above 0 dB. Several model-based separation systems based on vector quantization (VQ) (Ellis and Weiss, 2006), HMMs (Roweis, 2003; Kristijansson *et al.*, 2006), and Gaussian mixture models (GMMs) (Reddy and Raj, 2007) have been introduced. The separation performance of the model-based methods depends on two factors: (1) the statistical model employed for each speaker, and (2) the selected feature type. The feature employed should possess two properties: the additivity property (for example, additivity for the spectrogram magnitude (Bach and Jordan, 2006)) and high quantization performance (Ellis and Weiss, 2006; Mowlaee and Sayadiyan, 2008). To obtain an improved model, it is generally accepted that selecting a proper distortion measure plays a key role in the overall performance in a VQ-based system. According to Hai and Lois (1998), a proper distortion measure should highly correlate with the subjective results and at the same time provide high computational efficiency. In a recent study (Radfar *et al.*, 2006b), three different feature types of modulated lapped transform (MLT) coefficient, log spectrum and a combination of the spectral envelope and pitch frequency parameter were compared in a VQ-based speech separation scenario. It was concluded that the mask-based methods suffer from cross-talk, as reported in Radfar *et al.* (2006b), in segments where energetic masking occurs (low SSRs) (Srinivasan and Wang, 2008). Additionally, according to the results reported in Radfar *et al.* (2006b), it was observed that the integration of pitch and spectral envelope is not an effective SCS solution since extracting pitch frequencies from a mixed signal is challenging (Radfar *et al.*, 2006a). Moreover, large gross error rates of a multi-pitch tracker reported in Radfar *et al.* (2006a) may result in the selection of incorrect indices of VQ codebooks and as a consequence degrade the perceptual quality of the separated signal.

Another example for model-based techniques is the factorial method. In Kristijansson *et al.* (2006) the proposed factorial HMM-based speech separation surpassed human listener performance at SSRs of 0 dB through −6 dB across all speakers. In this case, each of the underlying speech signals in the mixture is modeled with a separate HMM. The mixed signal is modeled with a factorial HMM to model the contribution of the speakers in the mixture. The mixture estimation process is complex, time consuming, intractable, and a huge state space is required to capture every possible signal transition state. For instance, according to Roweis (2003) for a two-speaker separation task, 8000 states are required to accurately model one speaker.

Another example of model-based SCS is the MAX-VQ framework presented in Li *et al.* (2010). The selected codewords in MAX-VQ are mean vectors previously trained with clean speech signals. This inevitably leads to errors in estimating correct masks in the presence of interference (Li *et al.*, 2010). Furthermore, employing the mask signal, the MAX-VQ provides re-synthesis that is corrupted with cross-talks (Radfar *et al.*, 2007).

Analogous to the discrete Fourier transform (DFT) based spectral coefficients, Gammatone filterbank (GF) coefficients have also been used to produce binary masks to segregate speech signals from a mixture (Hu and Wang, 2004). Each element of this mask determines whether the related GF component is reliable or not at each time frame (Hu and Wang, 2004). Pitch tracks are found using the GF responses. The overall separation performance degrades when energetic masking at overlapping frames occurs. Additionally, mask-based methods inevitably introduce cross-talks in the outputs (Hu and Wang, 2004).

This research is aimed at developing a method of model-based SCS that performs satisfactorily at low SSRs. Based on the analysis of available literature, we believe that model-based methods have a higher potential for such a goal. Many separation methods predominantly employ short-time Fourier transform (STFT) vectors as their feature parameters. There has been an increasing research interest in efficient vector quantization approaches to resolve the poor quantization performance of the STFT-based methods. Jensen *et al.* (2003) proposed a perceptual distortion measure to transform the STFT features from the linear domain to a perceptual domain, leading to a norm that highly correlates with subjective quality. This is in contrast to the common L_2-norm distortion measure that is not consistent with the perceptual quality (Jensen *et al.*, 2003). The hybrid transform coder (HTC) (Kondoz and Evans, 1987) together with the perceptual subjective tests (Kondoz and Evans, 1988) revealed that the subband VQ method provides extra robustness in the presence of different speakers and improves the quantization performance. A weighted codebook-mapping (WCBM) method was also presented in Zavarehei *et al.* (2007) to improve the speech enhancement performance by restoring the suppressed information. Based on these three research approaches, it is reasonable to expect that processing in subbands provides improvements over the separation performance of conventional STFT-based methods.

The major objective of this work is to improve the SCS performance for a human listener as measured by subjective evaluations such as the mean opinion score (MOS) or objective evaluations (such as perceptual evaluation of speech quality (PESQ) and weighted spectral slope (WSS)) that correlate well with subjective evaluations. As a starting point, a transformation is designed to be applied to the STFT signal parameters at each frame to obtain a new feature vector. Inspired by subband coding and psychoacoustic foundation of the Mel scale and perception of loudness, one main contribution of this research is to derive an appropriate transformation called subband perceptually weighted transformation (SPWT) to be applied to the magnitude STFT. The new transformation improves the separation performance in the SCS problem. In particular, to investigate the effectiveness of the new feature vector, the separation performance of the proposed method is compared to that obtained using magnitude spectrum and log-spectrum magnitude. The evaluations are made in two steps: (1) in terms of objective measures to assess the upper-bound for the SCS performance, and (2) in terms of the SD statistics by performing a statistical analysis on SD measures (including outlier percentage and average SD) to assess the quantization performance. In addition, the coded reconstructed signals obtained from different methods are evaluated through the MOS, which determines the perceptual quality of the reconstructed signals. Through conducting extensive simulations it is possible to determine which method is more desirable for SCS. It will be clear that by

applying the proposed method, a higher transparent quantization is achievable, leading to a higher separation performance in comparison with the STFT-based methods.

2 Review of model-based approaches

In this section, various major parts of model-based SCS approaches are reviewed. The first part is the VQ-statistical model to generate codebooks. The second part is dedicated to the feature type. The last part is an appropriate distortion measure required to obtain an effective quantizer.

2.1 Vector quantization based speakers statistical models

To apply the model-based approach, first we obtain the DFT spectra from the sentences uttered by speakers. Each utterance is composed of a number of frames. For each frame index j from the overall utterance set, we define $\boldsymbol{S}_j(f)$ as the DFT magnitude spectrum with f=1, 2, …, L and j=1, 2, …, N, where N represents the number of training vectors for all extracted features, and L indicates the DFT dimensionality. In order to apply the model-based approach for SCS, we provide a codebook for each speaker. The main objective of a model-based separation technique is to find M representative vectors where M is the VQ codebook size. In practice, N is selected sufficiently large compared to M, typically $N \approx K \times M$ with k>100. An appropriate selection of M is determined by the required accuracy and the tolerable complexity.

2.2 Feature type selection problem

The speech short-term spectrum is susceptible to frequency response distortions introduced by communication media. Relative spectral processing (RASTA) methodology proposed in Hermansky and Morgan (1994) makes the short-term spectrum-based methods robust to linear spectral distortions (Hermansky, 1990). It is expected that selecting superior features will enable the statistical models to obtain high quality separation results. We propose a transformation (indicated by TD{·}) to map the STFT feature vectors to another domain. The new domain is called the transform-domain and possesses more perceptual information compared to the previous STFT feature set. The transformation, which as shown in next sections is based on solid foundations, results in a better representation.

Most of previous separation methods (Ellis and Weiss, 2006; Kristijansson *et al.*, 2006; Reddy and Raj, 2007) predominantly employed the STFT feature parameters as their feature vectors. However, the STFT feature bears drawbacks as follows: (1) all frequency bins in the STFT are uniformly weighted (Ellis and Weiss, 2006); (2) no perceptual considerations of the human auditory system are taken into account (Ellis and Weiss, 2006; Mowlaee and Sayadiyan, 2008); and (3) the magnitude DFT-spectrum feature vectors introduce weak quantization behaviors as indicated in Mowlaee and Sayadiyan (2008). This is mainly due to the fact that magnitude spectrum possesses many perceptually irrelevant peaks and valleys that consequently bias the centroids in a vector quantizer toward an erroneous direction (Mowlaee and Sayadiyan, 2008). This in turn prevents the quantizer from effectively modeling the most perceptually important STFT bins to minimize SD, as detailed in Mowlaee and Sayadiyan (2008). As a result, applications of the STFT-based mixture estimators including VQ-based, the binary mask (Roweis, 2003; Hu and Wang, 2004; Radfar *et al.*, 2007), and the Wiener filter (Benaroya *et al.*, 2006; Srinivasan *et al.*, 2006; Reddy and Raj, 2007) would cause inevitable errors, especially at ambiguous overlapping regions. These drawbacks of STFT introduce challenges to the model-based techniques severely degrading the separation performance. In this work, we propose a new transformation that is combined with a distortion measure (as described in the following sections), leading to significant improvements of the perceptual quality of the VQ-based SCS.

2.3 Distortion measure

The main objective for a VQ-based system is to minimize the average distortion between the uncoded and coded vectors in a mean-square error (MSE) sense (using L_2-norm) or perceptual measures. An effective distortion measure should be subjectively meaningful and highly correlate with human perceptual quality. Various speech applications necessitate different distortion measures specifically designed and optimized for the application. In this research, we follow a procedure to propose a new distortion measure in a transform domain called the SPWT.

2.4 The proposed feature parameters

A suitable transformation for a separation scenario should possess several characteristics: (1) producing low quantization errors in terms of the SD, and (2) appropriately weighting the perceptually important regions of the magnitude spectrum (similar to subband coding (Kondoz and Evans, 1988)), HTC (Kondoz and Evans, 1987), and WCBM for speech enhancement (Zavarehei *et al.*, 2007).

In the following, we present our newly proposed feature set to be used in the separation scenario. The proposed features (also called transform-domain features) are based on our recent research in Mowlaee and Sayadiyan (2008; 2009). The new feature is attained by taking the following two steps: (1) Normalize each code-vector (STFT magnitude spectrum) to its maximum value. This scales the feature values to lie within (0, 1]. Moreover, in clustering techniques it has been shown that employing normalized features improves the classification accuracy (Bishop, 2006). (2) Utilize the logarithm of the normalized magnitude spectrum in step (1) to reduce the dynamic range. These two steps are implemented as

$$\tilde{\boldsymbol{S}}_j(f)=\lg\left(1+\alpha\left|\boldsymbol{S}_j(f)\right|\big/G_j\right), \tag{1a}$$

$$G_j=\sqrt{\sum_{f=1}^{L}\left|\boldsymbol{S}_j(f)\right|^2}, \tag{1b}$$

where G_j is the energy of the magnitude spectrum for the jth frame with $f\in[1, L]$, and L indicates the total number of DFT-points. In Eq. (1a), α is a balancing parameter to be optimized with $\alpha>1$, $S_j^{\mathrm{n}}(f)=|\boldsymbol{S}_j(f)|/G_j$ is the normalized magnitude spectrum and $\tilde{\boldsymbol{S}}_j$ is our proposed transformed feature vector. In the rest of the paper by each formula or definition we are implicitly allowed to substitute $\boldsymbol{S}$ (magnitude spectrum) with $\tilde{\boldsymbol{S}}$ (the proposed feature in Eq. (1)). As indicated in Kondoz and Evans (1988) each magnitude spectrum $|\boldsymbol{S}_j(f)|$, before searching the codebook, should first be normalized to its gain G_j. At the decoder end, the selected normalized magnitude spectrum $S_j^{\mathrm{n}}(f)$ is multiplied by the original energy (G_j) for rescaling it to obtain $\hat{\boldsymbol{S}}_j(f)=G_jQ(\boldsymbol{S}_j(f))$, where $Q(\boldsymbol{S}_j(f))$ is the best match of $\boldsymbol{S}_j(f)$ from the codebook, and $\hat{\boldsymbol{S}}_j(f)$ is the decoded magnitude spectrum. The new feature parameter given in Eq. (1) provides three useful properties: (1) emphasizing the spectral peaks, (2) reducing the dynamic range of the magnitude spectrum due to the use of the log function, and (3) balancing the magnitude spectrum by incorporating α.

2.5 On spectral distortion measure

In this subsection, we briefly review the effectiveness of an appropriate distortion measure. This is useful in describing how the proposed transform-domain feature parameters improve the quantization performance of the STFT features. An SD measure is employed as our performance index to assess the quantization performance. The objective is to find a weighting function that can be used to minimize the averaged spectral distortion and lower the number of outliers within SD>4 dB (Paliwal and Kleijn, 1995; So and Paliwal, 2007). The SD measure for the jth frame, SD_j (in dB), is defined as (Paliwal and Kleijn, 1995)

$$\mathrm{SD}_j=\sqrt{\frac{1}{F}\sum_{f=0}^{F-1}\left(\boldsymbol{S}_j(f)-\hat{\boldsymbol{S}}_j(f)\right)^2}, \tag{2}$$

where F is around 3 kHz at a sampling rate of 8 kHz, and $\boldsymbol{S}_j(f)$ and $\hat{\boldsymbol{S}}_j(f)$ are the uncoded and coded DFT-spectra for the jth feature vector (of length L), respectively. Employing the Linde-Buzo-Gray (LBG) algorithm in VQ training (Gersho and Gray, 1992), $\boldsymbol{S}_j$ chosen from the training set is assigned to the mth cell if $d(\boldsymbol{C}_{\mathrm{ix}(j)}, \boldsymbol{S}_j)\le d(\boldsymbol{C}_{\mathrm{ix}(j)}, \boldsymbol{S}_i)\ \forall i\neq j$, where $\boldsymbol{C}_{\mathrm{ix}(j)}$ is the vector pointing to the optimal cell number for frame $j\in[1, N]$ and $\mathrm{ix}(j)\in[1, M]$, and M denotes the number of cells. The objective is to minimize the sum of distortions for $d(\boldsymbol{C}_{m,j}, \boldsymbol{S}_j)$ at each cell of $m\in(1, M)$ and for the overall quantization distortion we have

$$D_{\mathrm{total}}=\sum_{j=1}^{N}d(\boldsymbol{C}_{\mathrm{ix}(j)}, \boldsymbol{S}_j). \tag{3}$$

The codebook design is started by initializing the centroids using random selection initialization. Next, the LBG algorithm is applied to update these centroids to produce representative vectors as entries of the final codebook. This is accomplished by minimizing a cost function often in L_2-norm to find the most likely codeword. Dependent on the distance measure, different quantization outcomes are

achievable. This motivates us to find a new method to improve the quantization performance through optimizing the distortion measure.

2.6 Principles of transform-domain vector quantization

In our transform-domain VQ, a mapping is accomplished by dividing the magnitude spectrum into subbands in Mel scale (also called Mel band grouping). The Mel scale conversion is employed to take into account the logarithmic frequency sensitivity of the human auditory system. The transform-domain quantization steps are considered as $\boldsymbol{S}_j \xrightarrow{\mu} \mathrm{TD}\{\cdot\} \rightarrow Q(\cdot) \rightarrow \mathrm{TD}^{-1}\{\cdot\} \xrightarrow{\beta} \hat{\boldsymbol{S}}_j$, where μ and β denote the encoder and the decoder, respectively. The transform-domain distortion is defined as $d^{\mathrm{TD}}(\boldsymbol{S}_j, \hat{\boldsymbol{S}}_j) = \mathrm{SD}\{\boldsymbol{S}_j, \mathrm{TD}^{-1}(Q(\mathrm{TD}(\boldsymbol{S}_j)))\}$, where $\mathrm{TD}\{\cdot\}$ is the transformation, $Q(\cdot)$ denotes the quantizer, and $\mathrm{TD}^{-1}\{\cdot\}$ is the inverse transformation. The performance of a vector quantization system can be quantified by calculating an average distortion $E\{d(\boldsymbol{S}_j, \hat{\boldsymbol{S}}_j)\}$, where $E\{\cdot\}$ denotes the expectation. The smaller the average distortion is, the better the quantization system will be. The expected distortion for an ergodic stationary process can be expressed as (Gray, 1990): $E\{d(\boldsymbol{S}_j, \hat{\boldsymbol{S}}_j)\} = \lim_{N\to\infty} \frac{1}{N} \sum_{j=1}^{N} d(\boldsymbol{S}_j, \hat{\boldsymbol{S}}_j)$, where $d(\boldsymbol{S}_j, \hat{\boldsymbol{S}}_j)$ is the Euclidean distance between the coded and uncoded magnitude spectra $\boldsymbol{S}_j$ and $\hat{\boldsymbol{S}}_j$.

3 The proposed transform-domain vector quantization

In this section, a theoretical analysis is performed to derive the mathematical formulation for the transform-domain SD. The derivation for an appropriate weighted distortion measure for the proposed SPWT is described in detail.

3.1 Mathematical derivations for the proposed method

We derive and propose a distortion measure for the weighted minimization of the SD per Mel scale subbands. Define d^{TD} as the distortion measure in the transform domain and $\mathrm{TD}\{\cdot\}$ as the mapping from STFT into the desired perceptual space. The distortion measure in the transform-domain is given as

$$\sum_{j=1}^{N} d^{\mathrm{TD}}(\boldsymbol{C}_{\mathrm{ix}(j)}, \boldsymbol{S}_j) = \sum_{j=1}^{N} \left\| \mathrm{TD}\{\boldsymbol{C}_{\mathrm{ix}(j)}\} - \mathrm{TD}\{\boldsymbol{S}_j\} \right\|_2$$

with $\mathrm{TD}\{\boldsymbol{S}_j\}=\lg(1+\alpha\boldsymbol{S}_j)$. As described in Mowlaee and Sayadiyan (2008; 2009), the transformation $\mathrm{TD}\{\cdot\}$ is an efficient representation for harmonic spectra of speech frames. Then $d^{\mathrm{TD}}(\boldsymbol{C}_{\mathrm{ix}(j)}, \boldsymbol{S}_j)$ is reduced to

$$d^{\mathrm{TD}}(\boldsymbol{C}_{\mathrm{ix}(j)}, \boldsymbol{S}_j) = \left\| \lg\left(\frac{1+\alpha\boldsymbol{C}_{\mathrm{ix}(j)}}{1+\alpha\boldsymbol{S}_j} \right) \right\|_2. \quad (4)$$

To proceed our mathematical derivation, without loss of generality, let α=1 for simplicity. $d^{\mathrm{TD}}(\cdot)$ can be further approximated by the Taylor series expansion for the natural logarithm as (Spiegel *et al.*, 1998)

$$\ln(1+x) = \frac{1}{n}\sum_{n=1}^{\infty}(-1)^{n+1}x^n = x - \frac{x^2}{2} + O(x^3),$$

with $|x|\le 1$, $x\neq -1$ and $O(x^3)$ denoting the higher order terms. Substituting the first two terms of the Taylor series into Eq. (4), we have

$$\begin{aligned} d^{\mathrm{TD}}(\boldsymbol{C}_{\mathrm{ix}(j)}, \boldsymbol{S}_j) &= \left\| \boldsymbol{S}_j - \frac{\boldsymbol{S}_j^2}{2} - \boldsymbol{C}_{\mathrm{ix}(j)} + \frac{\boldsymbol{C}_{\mathrm{ix}(j)}^2}{2} \right\|_2 \\ &= \sum_f \boldsymbol{W}_j(f) \underbrace{\left\| \boldsymbol{S}_j(f) - \boldsymbol{C}_{\mathrm{ix}(j)}(f) \right\|_2}_{d(\boldsymbol{C}_{\mathrm{ix}(j)}, \boldsymbol{S}_j)}, \end{aligned} \quad (5)$$

where $\boldsymbol{W}_j(f)=1-(\boldsymbol{S}_j(f)+\boldsymbol{C}_{\mathrm{ix}(j)}(f))/2$ is the average magnitude spectrum between the optimal centroid $\boldsymbol{C}_{\mathrm{ix}(j)}$ for the jth training vector $\boldsymbol{S}_j$. Then the spectral distortion, $\boldsymbol{D}$, as an $N\times 1$ column vector, is defined as

$$\boldsymbol{D} = [d(\boldsymbol{C}_{\mathrm{ix}(1)}, \boldsymbol{S}_1),\ d(\boldsymbol{C}_{\mathrm{ix}(2)}, \boldsymbol{S}_2),\ \ldots,\ d(\boldsymbol{C}_{\mathrm{ix}(N)}, \boldsymbol{S}_N)]^{\mathrm{T}}. \quad (6)$$

The total distortion D_{total} in Eq. (3) is obtained by replacing d^{TD} from Eq. (5), which will result in

$$D_{\mathrm{total}} = \mathrm{tr}\left\{ \underbrace{\begin{bmatrix} w_1(\boldsymbol{S}_1) & \cdots & 0 \\ \vdots & & \vdots \\ 0 & \cdots & w_N(\boldsymbol{S}_N) \end{bmatrix}}_{\boldsymbol{W}(\boldsymbol{S})} \boldsymbol{D}(\boldsymbol{S})\boldsymbol{D}^{\mathrm{T}}(\boldsymbol{S}) \right\}, \quad (7)$$

where $\boldsymbol{W}(\boldsymbol{S})$ is an $N\times N$ diagonal matrix with diagonal elements of $w_j(\boldsymbol{S}_j(f))=1-(\boldsymbol{S}_j(f)+\boldsymbol{C}_{\mathrm{ix}(j)}(f))/2$. This completes the derivation for the overall distortion in the transform-domain. Simply speaking, the transformation in Eq. (1) leads to a weighted spectral distortion measure proposed in Eq. (7). In the following section it is shown that the proposed SPWT can be implemented by minimizing the D_{total} in Eq. (7), which can also be interpreted as using a sensitivity matrix $\boldsymbol{W}(\boldsymbol{S})$ similar to Gardner and Rao (1995), where the weighted square Euclidean distance (WSED) was incorporated as an approximation to SD formulated as

$$d_w(\boldsymbol{S}_j,\hat{\boldsymbol{S}}_j)=(\boldsymbol{S}_j-Q(\boldsymbol{S}_j))^{\mathrm{T}}\boldsymbol{W}(\boldsymbol{S}_j)(\boldsymbol{S}_j-Q(\boldsymbol{S}_j)),$$

where $\boldsymbol{S}_j$ and $\hat{\boldsymbol{S}}_j$ indicate the uncoded and coded vectors respectively, and $\boldsymbol{W}(\boldsymbol{S}_j)$ denotes a diagonal weighting matrix dependent on the magnitude spectrum matrix, $\boldsymbol{S}_j$, and can also be interpreted as a transformation similar to our proposed transformation. According to the high-rate quantization theory in Chatterjee and Sreenivas (2008), the spectral sensitivity coefficients are the optimal weights in a minimum MSE (MMSE) sense but in general do not guarantee the minimization of SD in Eq. (2). For small spectral distances, however, the SD in Eq. (7) approaches a simplified quadratically weighted Euclidean error where the weighting matrix is a sensitivity matrix (Gardner and Rao, 1995).

Based on the method in Gardner and Rao (1995), we present a theoretical analysis to obtain the optimal distortion in the transform domain. By considering $\boldsymbol{C}_{\mathrm{ix}(j)}$ as a constant, we write a Taylor series expansion for $d^{\mathrm{TD}}(\boldsymbol{S}_j,\boldsymbol{C}_{\mathrm{ix}(j)})$ around $\boldsymbol{C}_{\mathrm{ix}(j)}$ as

$$d^{\mathrm{TD}}(\boldsymbol{S}_j,\boldsymbol{C}_{\mathrm{ix}(j)})=d^{\mathrm{TD}}(\boldsymbol{C}_{\mathrm{ix}(j)},\boldsymbol{C}_{\mathrm{ix}(j)})+d_j^{\mathrm{TD}}(\boldsymbol{C}_{\mathrm{ix}(j)})(\boldsymbol{S}_j-\boldsymbol{C}_{\mathrm{ix}(j)})$$
$$+\frac{1}{2}(\boldsymbol{S}_j-\boldsymbol{C}_{\mathrm{ix}(j)})^{\mathrm{T}}\boldsymbol{G}(\boldsymbol{C}_{\mathrm{ix}(j)})(\boldsymbol{S}_j-\boldsymbol{C}_{\mathrm{ix}(j)})+O(\|\boldsymbol{S}_j-\boldsymbol{C}_{\mathrm{ix}(j)}\|^3), \tag{8}$$

where $d^{\mathrm{TD}}(\boldsymbol{C}_{\mathrm{ix}(j)})$ is a $1\times L$ row vector and $\boldsymbol{G}(\boldsymbol{C}_{\mathrm{ix}(j)})$ is an $L\times L$ matrix, defined as

$$d_j^{\mathrm{TD}}(\boldsymbol{C}_{\mathrm{ix}(j)})=\left.\frac{\partial d^{\mathrm{TD}}(\boldsymbol{S}_j,\boldsymbol{C}_{\mathrm{ix}(j)})}{\partial \boldsymbol{S}_j}\right|_{\boldsymbol{S}_j=\boldsymbol{C}_{\mathrm{ix}(j)}}, \tag{9}$$

$$\boldsymbol{G}(\boldsymbol{C}_{\mathrm{ix}(j)})=\left.\frac{\partial^2 d^{\mathrm{TD}}(\boldsymbol{S}_j,\boldsymbol{C}_{\mathrm{ix}(j)})}{\partial \boldsymbol{S}_j^2}\right|_{\boldsymbol{S}_j=\boldsymbol{C}_{\mathrm{ix}(j)}}. \tag{10}$$

From the definition of the distortion function in transform-domain in Eq. (8), it is observed that $d^{\mathrm{TD}}(\boldsymbol{C}_{\mathrm{ix}(j)},\boldsymbol{C}_{\mathrm{ix}(j)})$ is zero. Furthermore, to have a local minima we should have $d_j^{\mathrm{TD}}(\boldsymbol{C}_{\mathrm{ix}(j)})=0$. Therefore $d^{\mathrm{TD}}(\boldsymbol{S}_j,\boldsymbol{C}_{\mathrm{ix}(j)})$ is reduced to

$$d^{\mathrm{TD}}(\boldsymbol{S}_j,\boldsymbol{C}_{\mathrm{ix}(j)})=\frac{1}{2}(\boldsymbol{S}_j-\boldsymbol{C}_{\mathrm{ix}(j)})^{\mathrm{T}}\boldsymbol{G}(\boldsymbol{C}_{\mathrm{ix}(j)})(\boldsymbol{S}_j-\boldsymbol{C}_{\mathrm{ix}(j)})+O(\boldsymbol{S}_j^3). \tag{11}$$

Note that for a high-rate quantization scenario, the distance $\|\boldsymbol{S}_j-\boldsymbol{C}_{\mathrm{ix}(j)}\|$ between $\boldsymbol{C}_{\mathrm{ix}(j)}$ and $\boldsymbol{S}_j$ in $\boldsymbol{G}$ gets smaller and we have

$$d^{\mathrm{TD}}(\boldsymbol{S}_j,\boldsymbol{C}_{\mathrm{ix}(j)})=\frac{1}{2}(\boldsymbol{S}_j-\boldsymbol{C}_{\mathrm{ix}(j)})^{\mathrm{T}}\boldsymbol{G}(\boldsymbol{C}_{\mathrm{ix}(j)})(\boldsymbol{S}_j-\boldsymbol{C}_{\mathrm{ix}(j)}), \tag{12}$$

where $\boldsymbol{G}(\boldsymbol{C}_{\mathrm{ix}(j)})$ is denoted here as the sensitivity matrix for the transform domain. Note that $\boldsymbol{W}_j$ in Eq. (5) leads to $\boldsymbol{W}(\boldsymbol{S})$ defined in Eq. (7), which serves as a sensitivity matrix. Furthermore $\boldsymbol{W}_j$ is comparable to $\boldsymbol{G}(\boldsymbol{C}_{\mathrm{ix}(j)})$ in Eq. (12) as the sensitivity matrix in the transform-domain of Eq. (1). By comparing Eq. (12) with Eq. (7) it is observed that the proposed transformation (log function) leads to a weighted spectral distortion, where the weighting matrix $\boldsymbol{W}(\boldsymbol{S})$ serves as a sensitivity matrix. Accordingly, by comparing Eqs. (7) and (12) it is concluded that the proposed log transformation given in Eq. (1) leads to a sensitivity matrix $\boldsymbol{G}(\boldsymbol{C}_{\mathrm{ix}(j)})$ that addresses minimization problems to achieve the best matches in transform-domain VQ.

The high-rate quantization scenario leads to a small volume of Voronoi regions (Gersho and Gray, 1992; Gardner and Rao, 1995) and the minimal distortion is achieved by employing a standard Lagrange multiplier technique. The result is a function of the sensitivity matrix $\boldsymbol{G}(\boldsymbol{C}_{\mathrm{ix}(j)})$ and the probability density function (PDF) of the transform-domain features (Gardner and Rao, 1995).

3.2 Mean-square error distortion measure for SPWT

According to Ellis and Weiss (2006), the STFT magnitude spectrum plays a key role in mask-based methods as in mixmax (binary mask), Wiener (soft-mask) and the model-based SCS. However, due to the poor quantization performance of STFT feature vectors, the quality of the separation method would

typically degrade. To improve the quantization behavior of the model-based methods and to cope with the difficulties of the STFT, the SPWT on all possible $\boldsymbol{S}_j(f)$ is proposed as the following optimization scenario:

$$\boldsymbol{S}_{\text{opt}}(f)=\arg\min_{\boldsymbol{S}_j(f)}\text{SD}_{\text{SPWT}}(\boldsymbol{S}_j,\hat{\boldsymbol{S}}_j), \quad (13)$$

where $\text{SD}_{\text{SPWT}}(\cdot)$ is our proposed spectral distortion based on Eqs. (5)−(7), defined as

$$\text{SD}_{\text{SPWT}}(\boldsymbol{S}_j,\hat{\boldsymbol{S}}_j)=\sum_{k=1}^{K}\boldsymbol{W}_k\left|\boldsymbol{S}_j^k-\hat{\boldsymbol{S}}_j^k\right|^2, \quad (14)$$

where $\boldsymbol{W}_k$ is the perceptual weighting function having different values at the kth subband, and $\boldsymbol{S}_j^k$, $\hat{\boldsymbol{S}}_j^k$ are the uncoded and coded magnitude spectra within the kth subband, respectively. The subband frequencies are determined by employing Mel grouping on a frequency range in Mel scale. The motivation behind using subbands is to emphasize different frequency divisions in an uneven manner. To obtain the $\boldsymbol{W}_k$ proposed in Eq. (14), the subbands are designed by grouping Mel-frequency bands so that all subbands include an equal number of Mel-frequency bands. The objective is to minimize the distortion as

$$D_{\text{SPWT}}(\boldsymbol{S}_j(f),\hat{\boldsymbol{S}}_j(f))=\sum_{k=1}^{K}g_k^p d_{\Omega_k}(\boldsymbol{S}_j^k,\hat{\boldsymbol{S}}_j^k)\Big/\sum_{k=1}^{K}g_k, \quad (15)$$

where D_{SPWT} denotes the overall perceptually weighted spectral distortion in subbands, $d_{\Omega_k}(\cdot)$ indicates the Euclidean distance measured in the kth Mel band, and g_k ($k\in[1, K]$) is the subband energy of $\boldsymbol{S}_j^k$ at the kth subband, defined as

$$g_k=\sqrt{\frac{1}{\Omega_{\text{BW}}^k}\sum_{f=\Omega_k}^{\Omega_{k+1}}\boldsymbol{S}_j^k(f)}, \quad (16)$$

where Ω_{BW}^k denotes the frequency bandwidth. According to Kondoz and Evans (1988), the exponent p in Eq. (15) accounts for the noise shaping factor selected through subjective listening tests for p within $0\le p\le 1$. In Eq. (15) we select p=0.66, which highly correlates with perception (Stevens and Marks, 1965; Hermansky, 1990). As an example, for three subbands, the subband boundaries are given as: Ω_1=100 Hz, Ω_2=890 Hz, Ω_3=1929 Hz, Ω_4=3750 Hz. The weighting g_k employed in Eq. (15) functions within subbands and preserves the spectral shape of subbands.

In a related work, Zavarehei *et al.* (2007) employed a WCBM as an effective tool for speech enhancement. Using the parameters of a harmonic plus noise model (HNM) obtained for the enhanced signal, their method defines a weighted L_2-norm distance between the codewords and the enhanced signal as $D_m(\boldsymbol{B}_{\text{NR},k},C_{k,m})=\sum_{k=1}^{N}[\boldsymbol{W}_k(\boldsymbol{B}_{\text{NR},k}-C_{k,m})]^2$, where $\boldsymbol{W}_k$ denotes the weight for the kth harmonic amplitude in the range [0, 1], $C_{k,m}$ indicates the kth harmonic of the mth codebook entry (codeword), $\boldsymbol{B}_{\text{NR},k}$ is the kth harmonic amplitude for the noise reduced frame (denoted by subscript 'NR') and is normalized to its energy defined as $\boldsymbol{B}_{\text{NR},k}=\boldsymbol{A}_{\text{NR}}\Big/\sqrt{\sum_k\boldsymbol{A}_{\text{NR},k}^2}$, where $\boldsymbol{A}_{\text{NR}}$ indicates the HNM amplitude vector, and N is the number of frames. Then the best codewords (with the lowest distance) are weighted based upon the inverse of the codeword distance $\boldsymbol{C}_m$ from HNM normalized harmonic amplitudes ($\boldsymbol{B}_{\text{NR}}$) and $\boldsymbol{A}_{\text{NR}}=\alpha\boldsymbol{B}_{\text{NR}}$. The optimum α is obtained as (Zavarehei *et al.*, 2007)

$$\alpha_{\text{opt}}=\sum_k\boldsymbol{W}_k\boldsymbol{A}_{\text{NR},k}\boldsymbol{B}_{\text{CB},k}\Big/\sum_k\boldsymbol{W}_k\boldsymbol{B}_{\text{CB},k}^2. \quad (17)$$

The energy normalization in Eq. (17) was used to effectively restore harmonic components within subbands. In this research, we employed a similar energy normalization given by g_k but applied to subbands based on groups of Mel-scale bands. The objective of the proposed SPWT is to minimize the SD between the $\boldsymbol{S}_j$ and $\hat{\boldsymbol{S}}_j$ in subbands. More specifically, g_k^p in Eq. (15) is similar to the weighting function $\boldsymbol{W}_k$ in Eq. (17), while spectral distortion $d_{\Omega_k}(\cdot)$ in the SPWT is comparable to $D_m(\cdot)$ in Zavarehei *et al.* (2007). $\boldsymbol{S}_j^k$ and $\hat{\boldsymbol{S}}_j^k$ are also comparable to $\boldsymbol{A}_{\text{NR},k}$ and $\boldsymbol{B}_{\text{CB},k}$, respectively, each pair denoting the original and codeword spectrum magnitude vectors. Our rationale for the normalization to the subband energies given by g_k is that the codebook entries will be independent of their energy variation across speech frames. The weighted distance SD_{SPWT} in Eq. (14) still preserves the shape of the spectrum and the frequency mapping

to Mel leads to a performance that highly correlates to the perceptual quality. Our simulations show that the transformation in Eq. (1) leads to improvement in separation performance.

4 Simulation results

In this section, we first describe the methods and the data employed for simulations to evaluate the performance of the proposed SPWT in an SCS scenario. The evaluation results are then compared to those obtained using previous mask-based approaches of mixmax (Roweis, 2003; Radfar *et al.*, 2007) and Wiener (Benaroya *et al.*, 2006; Srinivasan *et al.*, 2006; Reddy and Raj, 2007). We set two objective experiments to evaluate the separation performance as well as the lower bound on SD, comparing the proposed SPWT with other mask-based alternatives (mixmax and Wiener). The separation upper bound was measured in terms of the log likelihood-ratio (LLR), WSS, and PESQ (ITU-T P.862, 2001). Through evaluations, the SD statistical properties of the SPWT were compared to those obtained using the STFT feature. Finally, the MOS was employed for subjective evaluations and comparisons.

4.1 Separation scenario

To evaluate the performance of the methods proposed in this research, the following experiments were conducted. The speech corpus utilized in our simulations was composed of 100 sentences of male and female speech data chosen from the Cooke database (Cooke *et al.*, 2006). The signal was downsampled from 25 kHz to 8 kHz. The frame length was 32 ms and a frame shift of 8 ms was used. We considered only the speaker-dependent scenario in this work. For each speaker, 85 sentences were used for training and the remaining 15 sentences were chosen for testing. The training vector size (N) was chosen to be 15 times the codebook size (M) to provide the required accuracy and a tolerable complexity. Additionally, to determine the optimal codebook size and the number of subbands, the evaluations were conducted at M=128, 256, 512, 1024, 2048 and $k\in[3, 6]$.

Fig. 1 depicts the separation scenario used for experimental results in this work. First, two speaker signals, $\boldsymbol{s}_1(n)$ and $\boldsymbol{s}_2(n)$, were mixed together at a certain SSR determined by the attenuation of the acoustic transfer function (ATF) of $\boldsymbol{H}(f)$ in Fig. 1.

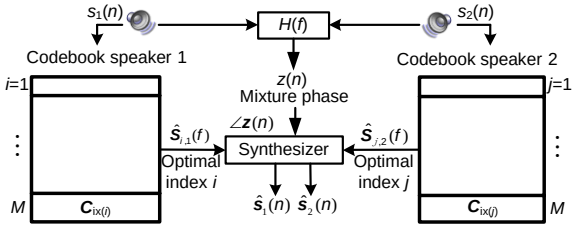

Fig. 1 Block diagram for the simulations conducted in a single-channel separation scenario

The core of the SCS was composed of two trained VQ-codebooks, one for each speaker. The objective was to evaluate the separation upper-bound versus the SSR levels ranging within [−18, 18] dB. Let i and j be the optimal indices for speakers 1 and 2, respectively with $i, j\in[1, M]$ where M denotes the VQ-codebook size. The estimated magnitude spectra, $\hat{\boldsymbol{S}}_{i,1}$ and $\hat{\boldsymbol{S}}_{j,2}$, were obtained by replacing the centroid of the optimal index for each codebook, denoted by $\boldsymbol{C}_{\mathrm{ix}(i)}$ and $\boldsymbol{C}_{\mathrm{ix}(j)}$ respectively (Fig. 1). Next, the mixture approximation searched for the best match for codeword indices (i, j) and the two speaker VQ-models were combined to estimate the mixture. In this work, we assumed that the optimal indices were known a priori and we addressed no estimation error in the mixture estimation. This is also known as the ideal separation that implies the separation upper-bound. For signal reconstruction, the mixture phase $\angle\boldsymbol{z}(n)$ was used along with the best indices decoded as the estimated magnitude spectra ($\hat{\boldsymbol{S}}_{i,1}$, $\hat{\boldsymbol{S}}_{j,2}$) to obtain $\hat{\boldsymbol{s}}_1(n)$ and $\hat{\boldsymbol{s}}_2(n)$.

Note that in this study we do not allow any estimation errors induced at the separation stage of the SCS by using the original signals $\boldsymbol{s}_1(n)$ and $\boldsymbol{s}_2(n)$ in the separation. Hence, we assume that the optimal indices are available to the mixture estimator (this work is aimed to evaluate the upper-bound separation for the proposed SPWT and to compare it with STFT-based methods). The ultimate goal is to compare the quantization performance by evaluating the quality of the separated output obtained using each method. In this aspect, the evaluations were conducted with respect to the separation performance obtained from the proposed SPWT and those attained by log spectrum, and

magnitude spectrum. On the other hand, it is quite useful to assess the upper-bound approximation for the separation task (called the ideal estimation) (Radfar *et al.*, 2007). Hence, this study reports a comparison of the upper-bound separation performance for different methods.

4.2 Benchmark methods

To demonstrate the quality of the proposed SPWT and to compare it with the performance of other mask-based methods, mixmax (Roweis, 2003; Radfar *et al.*, 2007) and the Wiener filter (Benaroya *et al.*, 2006; Reddy and Raj, 2007) were employed as the benchmark methods. In the following, the methods are briefly reviewed. The main objective of these methods is to design two binary masks to be applied to the spectrum of the mixed signal to recover individual signals. These methods utilize the fact that within a critical band the weaker signal is masked by the stronger one (Moore, 1997), i.e., the masking phenomenon. From DFT-spectra $\boldsymbol{S}_1(f)$ and $\boldsymbol{S}_2(f)$ two ideal binary masks were generated as $\text{mask}_1^{\text{ideal}}(f)=1$ if $\boldsymbol{S}_1(f)>\boldsymbol{S}_2(f)$ and 0 otherwise. Similarly, $\text{mask}_2^{\text{ideal}}(f)=1$ if $\boldsymbol{S}_2(f)>\boldsymbol{S}_1(f)$ and 0 otherwise, with $f\in[1, L]$ where L is the DFT size. Multiplying the ideal masks by the mixed DFT-spectrum $\boldsymbol{Z}$ gave the estimated spectra and we have $\hat{\boldsymbol{S}}_1^{\text{ideal}}=\boldsymbol{Z}\times\text{mask}_1^{\text{ideal}}$, $\hat{\boldsymbol{S}}_2^{\text{ideal}}=\boldsymbol{Z}\times\text{mask}_2^{\text{ideal}}$. The estimated spectra along with the phase of the mixture, φ_z, was used to recover the time domain signals as $\hat{\boldsymbol{S}}_t^{\text{ideal}}=\text{FD}_L^{-1}(\boldsymbol{S}_t^{\text{ideal}}\text{e}^{\text{j}\angle\varphi_z})$, $t\in\{1, 2\}$, where FD_L^{-1} is the L-point inverse-DFT. Note that the binary mask aims to retain the stronger signal in a mixture by removing the interference-dominant units instead of recovering both target and masked signals (Roweis, 2003; Radfar *et al.*, 2007; Srinivasan and Wang, 2008).

As another benchmark method, the Wiener filter is employed as a spectral-domain filter (Benaroya *et al.*, 2006; Reddy and Raj, 2007). Wiener-based separation algorithms utilize mostly the additivity of the power spectrum of the independent sources to separate a mixture. The Wiener estimation is formulated as

$$\text{FD}_L(\hat{\boldsymbol{S}}_t)=\frac{\boldsymbol{S}_1}{\boldsymbol{S}_1+\boldsymbol{S}_2}\boldsymbol{S}_z,\quad t\in\{1, 2\},$$

where $\text{FD}_L(\cdot)$ represents the L-point DFT, and $\boldsymbol{S}_z$ denotes the power spectrum of the mixed signal. Note that the Wiener filter basically filters the mixture signal according to the SSR. This implies that a Wiener filter mask for the SCS (Loizou, 2007) is based upon soft-decisioning.

Experiment 1 (Separation performance) Fig. 2 depicts the separation upper-bound performance for different methods, where PESQ is used as our objective measure to assess the quality of the separated outputs for different methods.

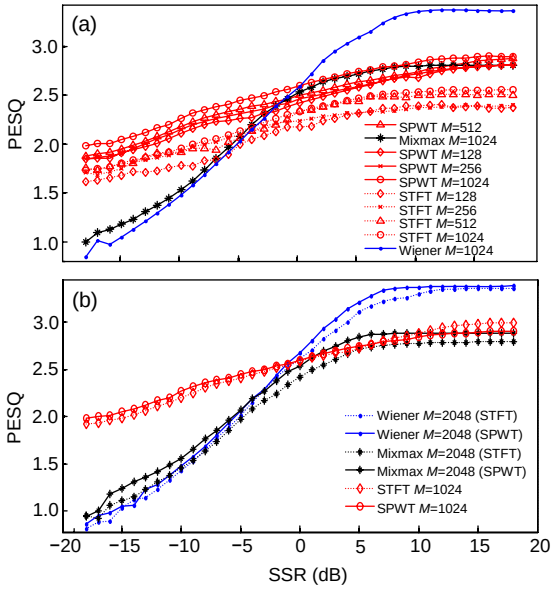

Fig. 2 (a) Separation upper-bound for STFT, SPWT and mask-based methods in terms of PESQ with different codebook sizes; (b) Comparing the methods with their optimal codebook sizes

The separation was conducted at SSR levels ranging from −18 to 18 dB. For a fair comparison between the mask- and VQ-based methods, in all methods it was assumed that the optimal indices were known a priori. The only difference was in the synthesis stage: the SPWT employed a perceptually relevant distortion measure defined in Eq. (15), while the mask-based methods applied a mask to the mixed signal to produce the separated outputs (as discussed in Section 4.1). Remember that both methods employ the phase information of the mixture to reconstruct their separated output signals. In Fig. 2, the results are shown for speakers chosen from the Cooke database.

Several important observations can be made from Fig. 2a: (1) At large SSRs, the proposed

VQ-based method SPWT asymptotically reached the performance achieved using the mask-based (ideal mask) method especially for codebook sizes larger than 512 (comparing the PESQ scores of the SPWT at M=1024 with those obtained using mixmax at M=1024 or 2048, Fig. 2b). (2) The VQ-based methods were less sensitive to SSR variation. This was verified by a PESQ difference of 0.8 as compared to 1.8 for mixmax and 2.85 for Wiener. This provides a significant improvement over undesirable large fluctuation experienced by the mask-based methods. (3) The VQ-based methods provided a higher PESQ compared to the mask-based methods at low SSRs. The best result for the SPWT was obtained by employing M=1024.

To conduct a better comparison between different SCS methods based on their feature, the simulation explained was repeated for the methods using the optimal codebook size for each method. Fig. 2b compares the approaches of the proposed SPWT with M=1024, STFT with M=1024, and mask-based methods with M=1024 and 2048. It was observed that the SPWT with M=1024 outperformed the STFT with M=1024. Additionally, in mask-based methods, using SPWT led to higher PESQ results compared to the STFT (Fig. 2b). Finally, compared to the STFT-based SCS, the SPWT led to a higher performance in PESQ for each selection of the codebook size (M=128, 256, 512, 1024).

The presented results showed that the proposed SPWT contributed to a higher upper-bound performance of speech separation. It can be inferred that the VQ-based approaches improve the separation result especially at low SSRs where the mask-based methods result in poor separation quality. This is evident from the PESQ curves for mixmax or Wiener at SSR<0 dB in Fig. 2. The proposed SPWT was also evaluated for various numbers of subbands from 3 to 6. It was observed that with a codebook size of M=2048, choosing k=3 and k=4 achieved the best PESQ score for a female and a male speaker, respectively.

Experiment 2 (Spectral distortion analysis) In order to assess the quantization performance of the methods discussed in this research, the average SD and the number of outliers were evaluated. According to Paliwal and Kleijn (1995), to remove any audible distortion, the average SD for all frames should be limited to about 1 dB. Under these conditions, the quantizer introduces negligible distortion (for transparent coding). In addition, there should be no outlier frames with SD>4 dB and the percentage of frames with 2 dB<SD<4 dB should be lower than 2% (Paliwal and Kleijn, 1995).

The quantization performance and details of SD statistics (SD<2 dB, SD>4 dB, and 2 dB<SD<4 dB) are shown in Tables 1 and 2 for the features discussed in this research. Tables 1 and 2 summarize the SD statistics for a female and a male speaker, respectively. Three subbands with M=2048 codewords sufficed to obtain the best performance for the proposed SPWT. As is evident from Table 1, the SPWT significantly outperformed the STFT with SD<2 dB (98.91%) and outlier=0.02% for the SPWT as compared to SD< 2 dB (80.49%) and outlier=5.46% for STFT. Similarly, for a male speaker at SD<2 dB we obtained 96.63% and outlier=0.03% for the SPWT as compared to 63.89% and outlier=7.24% for STFT.

Table 3 presents the speech separation upper-bound for the SPWT versus the number of subbands. The separation performance was assessed in terms of objective measures: LLR, WSS, and PESQ. The best performance for the female speaker was obtained using 4 subbands and 2048 codewords. Other objective measures were LLR=0.68, WSS=19.71, and PESQ=2.5. For the male speaker the best result was obtained at k=3 and M=2048 with LLR=0.66, WSS=19.24, and PESQ=2.41.

Table 4 shows the evaluation results for different feature parameters discussed in this work: log-spectrum amplitude, the STFT, and the SPWT. The results were obtained for a female speaker. It was again observed that the SPWT showed the largest percentage (99.46%) for SD<2 dB and the lowest outliers (0.04% for SD<4 dB). The results showed significant improvements over those obtained using the STFT (70.45% for SD<2 dB, outlier=9.64%) and log-spectrum amplitude (25.66% for SD<2 dB, outlier=12.63%). The results for the SPWT were reported versus various codebook sizes $M\in\{256, 512, 1024, 2048\}$. We found that choosing a codebook size of M=2048 balanced the required accuracy and the computational complexity. The number of outliers within 2 dB<SD<4 dB was also critical. According to Table 4, the SPWT introduced only 0.5% of the whole spectral distortion within 2 dB<SD<4 dB, compared

Table 1 SD statistics for a female speaker versus different *M* and subbands (for SD<2 dB and SD>4 dB)

Scenario	SD outliers (%)									
	SD<2 dB					SD>4 dB				
	M=128	256	512	1024	2048	*M*=128	256	512	1024	2048
SPWT, *k*=3	96.73	97.14	93.71	99.25	**98.91**	0.10	0.20	0.27	0.03	**0.02**
SPWT, *k*=4	96.10	96.63	96.33	**99.30**	97.51	0.10	0.15	0.15	**0.03**	0.05
SPWT, *k*=5	97.34	94.63	94.43	98.15	96.18	0.10	0.31	0.34	0.08	0.13
SPWT, *k*=6	88.31	88.23	94.39	98.94	99.46	0.72	0.67	0.10	0.05	0.02
STFT	69.59	69.21	69.59	81.21	**80.49**	10.08	11.68	5.59	4.85	**5.46**

Bold numbers represent the best results

Table 2 SD statistics for a male speaker versus different *M* and subbands (for SD<2 dB and SD>4 dB)

Scenario	SD outliers (%)									
	SD<2 dB					SD>4 dB				
	M=128	256	512	1024	2048	*M*=128	256	512	1024	2048
SPWT, *k*=3	91.03	90.67	88.94	93.84	**96.63**	0.40	0.45	0.25	0.18	**0.03**
SPWT, *k*=4	88.08	87.92	87.43	94.45	93.86	1.11	0.65	0.45	0.12	0.17
SPWT, *k*=5	85.47	93.18	98.11	97.93	95.49	0.91	0.40	0.07	0.03	0.08
SPWT, *k*=6	89.77	93.57	98.43	93.13	92.81	0.41	0.30	0.05	0.21	0.17
STFT	44.86	46.73	49.85	62.43	**63.89**	20.26	19.37	16.13	8.55	**7.24**

Bold numbers represent the best results

Table 3 Speech separation upper-bound versus different *M* and subbands for a female and a male speaker in terms of objective measures LLR, WSS, and PESQ

Scenario	SD outliers (%)											
	Female						Male					
	M=1024			*M*=2048			*M*=1024			*M*=2048		
	LLR	WSS	PESQ	LLR	WSS	PESQ	LLR	WSS	PESQ	LLR	WSS	PESQ
SPWT, *k*=3	0.69	20.06	2.42	0.69	19.95	2.49	0.77	21.07	2.18	**0.66**	**19.24**	**2.41**
SPWT, *k*=4	0.69	20.85	2.40	**0.68**	**19.71**	**2.50**	0.76	20.9	2.22	0.67	19.49	2.41
SPWT, *k*=5	0.68	20.56	2.39	0.71	20.27	2.49	0.81	21.27	2.12	0.68	19.31	2.41
SPWT, *k*=6	0.70	20.33	2.31	0.72	19.89	2.49	0.78	20.94	2.17	0.67	19.30	2.42

Bold numbers represent the best results

Table 4 Comparison of the statistics of the SD of different feature types

Feature	*M*	SD outliers (%)		
		SD<2 dB	2 dB<SD<4 dB	SD>4 dB
SPWT	256	88.94	8.00	3.06
	512	96.88	2.79	0.33
	1024	98.92	0.88	0.19
	2048	99.46	0.50	0.04
Log	2048	25.66	61.71	12.63
STFT	2048	70.45	19.91	9.64

to 19.91% for STFT and 61.71% for log-spectrum amplitude. The results indicated that the SPWT significantly reduced the number of outliers and provided a more appropriate statistical model.

As another important measure, the histogram of distortion of SD was evaluated for each feature. After the training of the VQ-codebooks for each feature, 15% of the sentences, which were not trained by the codebook, were chosen for testing. The SD statistics were obtained by averaging the SD results for these test sentences. Fig. 3 depicts the histogram of SD for three features: (1) STFT magnitude spectrum (Fig. 3a), (2) log-spectrum amplitude (Fig. 3b), and (3) SPWT (Fig.3c). As observed from the figures, the number of outliers in STFT was unacceptably high (5.75%) compared to that obtained by the proposed SPWT (0.06%). The averaged SD (1.3) obtained by STFT was also unacceptably higher than that obtained by the SPWT (0.32). Accordingly, it was found that,

compared to the almost Gaussian PDF obtained by log-spectrum (Fig. 3b), the statistical distribution for STFT and SPWT (Figs. 3b and 3c) were more compact. The proposed SPWT consequently led to a highly compact PDF that was very desirable for quantization purposes in model-based SCS (Fig. 3). In contrast, using the log-magnitude spectrum resulted in a Gaussian distribution for spectral distortion.

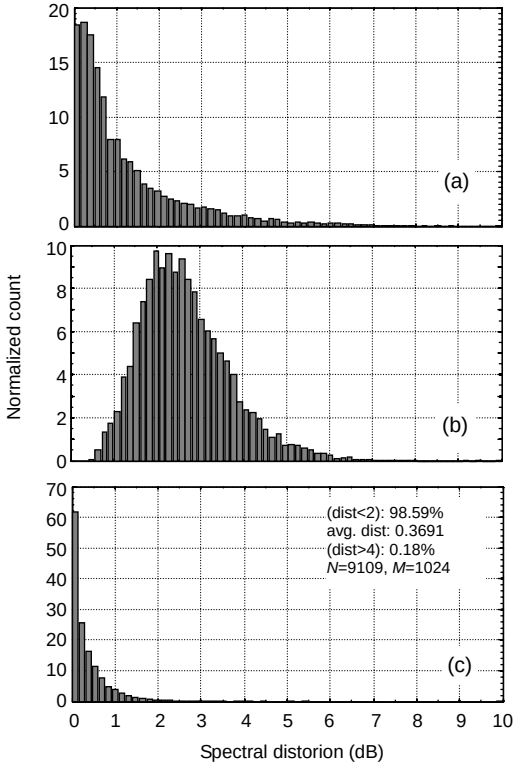

Fig. 3 The histogram of the spectral distortion (SD) for (a) magnitude spectrum, (b) log spectrum, and (c) SPWT (Mowlaee and Sayadiyan, 2009)

According to (Martin, 2005; Hendriks and Rainer, 2007), the normal inversed Gaussian (NIG) distribution provides a more accurate fit to the histogram of speech DFT coefficients than the Gaussian density (Fig. 3b). By considering both coded and uncoded speech spectra as NIG-distributed independent random variables, their difference (denoted by the SD in Eq. (2)) is also NIG-distributed (Chhikara and Folks, 1989). Similarly, from Fig. 3c we conclude that the spectral distortion for SPWT is similar to NIG distribution. Accordingly, the histogram of SD for the SPWT (Fig. 3c) leads to the improved quantization performance, and consequently guarantees low SD compared to other feature types while being employed in the decoding stage of SCS.

Note that the proposed distortion measure (SPWT) given in Eq. (15) is very similar to the signal and segment-dependent band importance function recently proposed as a weighting function in calculating articulation index measure in Ma *et al.* (2009), where such measure highly correlates to the intelligibility of speech in the presence of other talker signal. Similarly, in this paper we demonstrated that by using a subband transformation (called SPWT) it is possible to significantly improve the perceptual quality in the VQ-based single-channel speech separation framework.

4.3 Subjective evaluations

A mean opinion score (MOS) listening test was used as a subjective measure to assess the perceived quality of the separated signals obtained using the benchmark and the proposed methods. A total of 10 persons of various ages participated and were trained for the test. The subjects included 5 women and 5 men with graduate-level educations. The participants listened to the original and synthesized signals by the STFT, SPWT and mask-based methods. Then they were asked to give an opinion score from 1 to 5 (1=bad and 5=excellent quality). The MOS was obtained by averaging the results.

The proposed SPWT provided superior results at SSR<0 dB. For an SSR of −6 dB, it was observed that the SPWT obtained a MOS of 3 out of 5 while the magnitude spectrum and the log spectrum resulted in a MOS of 2.5 and 2, respectively. By further examining the synthesized qualities, we found that in general the mask-based methods led to two deficiencies: (1) some portions of speech in the masked signal, with high perceptual importance, may be completely masked by the dominant speaker (Srinivasan and Wang, 2008), and (2) portions of the other speaker signal is perceptible in parts of the separated output signals, a detrimental phenomenon called cross-talk as already reported in Radfar *et al.* (2007) or intrusion percentage (Hu and Wang, 2004). To assess the quality of the reconstructed signals, please visit http://ele.aut.ac.ir/~pejmanmowlaee/jzus.htm where downloadable audio files are available.

5 Conclusion and future work

In this paper, a subband perceptually weighted transformation called SPWT is proposed to improve the separation performance of the SCS scenario. The objective of the proposed method is to effectively map the conventional STFT features into a more perceptually relevant transform-domain feature to offer an attractive alternative feature in SCS. The proposed method is based on two solid psychoacoustic foundations: (1) Mel-band grouping, and (2) perception of loudness. The mathematical formulations are presented to derive distortion measure in transform domain. The proposed SPWT was compared with other mask-based methods (the mixmax and Wiener) in terms of two measures: (1) separation upper-bound, and (2) SD statistics. Extensive evaluations were conducted to determine the upper-bound of separation performance for the VQ-based SCS. It was observed that the SPWT led to higher synthesis and separation performance in terms of PESQ compared to STFT and mask-based methods. Moreover, the SPWT achieved a more appropriate quantization performance compared to those obtained by log-spectrum amplitude and magnitude spectrum. Also, using the proposed method in the SCS framework significantly improved the synthesized output quality as indicated by PESQ and the MOS.

For future work, we aim to add dynamic information to improve the VQ-based SCS and to reduce the ambiguity in index finding. Two factors, continuity and dynamic of the underlying speaker signals, will be considered by incorporating memory in the mixture approximation stage. In addition, as a remaining issue with the VQ-based SCS, we have recently incorporated the mixture estimation stage (separation part) into the system proposed in this work. The separation stage is implemented by minimizing a subband perceptually weighted distortion measure. Then the most likely indices for each speaker codebook are found. The primary results are promising and superior compared to those obtained using the STFT mask-based methods. By employing the proposed transform-domain separation system we expect to make significant improvements in the separation performance at low SSRs. The proposed ideas of employing dynamics as well as reducing the ambiguity in index finding (at the separation stage) will also serve as further improvements to fine-tune the separation stage results.

Acknowledgements

The authors would like to thank the anonymous reviewers for their extensive review of the paper and quite valuable comments.

References

Bach, F.R., Jordan, M.I., 2006. Learning spectral clustering, with application to speech separation. *J. Mach. Learn. Res.*, **7**(1):1963-2001.

Barker, J., Shao, X., 2007. Audio-Visual Speech Fragment Decoding. Proc. Int. Conf. on Auditory-Visual Speech Processing, p.37-42.

Barker, J., Cooke, M., Ellis, D., 2005. Decoding speech in the presence of other sources. *Speech Commun.*, **45**(1):5-25. [doi:10.1016/j.specom.2004.05.002]

Barker, J., Coy, A., Ma, N., Cooke, M., 2006. Recent Advances in Speech Fragment Decoding Techniques. 9th Int. Conf. on Spoken Language Processing, p.85-88.

Benaroya, L., Bimbot, F., Gribonval, R., 2006. Audio source separation with a single sensor. *IEEE Trans. Audio Speech Lang. Process.*, **14**(1):191-199. [doi:10.1109/TSA.2005.854110]

Bishop, C.M., 2006. Pattern Recognition and Machine Learning. Information Science and Statistics Series. Springer, New York, USA, p.2-3. [doi:10.1007/978-0-387-45528-0]

Chatterjee, S., Sreenivas, T.V., 2008. Predicting VQ performance bound for LSF coding. *IEEE Signal Process. Lett.*, **15**(1):166-169. [doi:10.1109/LSP.2007.914786]

Chhikara, R., Folks, L., 1989. The Inverse Gaussian Distribution: Theory, Methodology and Applications. CRC Press, Marcel Dekker Inc., New York, USA, p.39-52.

Christensen, M.G., Jakobsson, A., 2009. Multi-Pitch Estimation. Synthesis Lectures on Speech and Audio Processing. Morgan and Claypool Publishers, San Rafael, CA, USA, p.1-24. [doi:10.2200/S00178ED1V01Y200903SAP005]

Cooke, M.P., Barker, J., Cunningham, S.P., Shao, X., 2006. An audiovisual corpus for speech perception and automatic speech recognition. *J. Acoust. Soc. Am.*, **120**(5):2421-2424. [doi:10.1121/1.2229005]

Ellis, D.P.W., Weiss, R.J., 2006. Model-Based Monaural Source Separation Using a Vector-Quantized Phase-Vocoder Representation. IEEE Int. Conf. on Acoustics, Speech and Signal Processing, p.957-960. [doi:10.1109/ICASSP.2006.1661436]

Gardner, W., Rao, B., 1995. Theoretical analysis of the high rate vector quantization of LPC parameters. *IEEE Trans. Speech Audio Process.*, **3**(5):367-381. [doi:10.1109/89.466658]

Gersho, A., Gray, R.M., 1992. Vector Quantization and Signal Compression. Kluwer Academic Publishers, Boston, USA, p.345-372.

Gray, R.M., 1990. Source Coding Theory. Kluwer International Series in Engineering and Computer Science. Kluwer Academic Publishers, Boston, USA, p.43.

Gu, L.Y., Stern, R.M., 2008. Single-Channel Speech Separation Based on Modulation Frequency. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing, p.25-28.

Hai, L.V., Lois, L., 1998. A New General Distance Measure for Quantization of LSF and Their Transformed Coefficients. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing, p.45-48.

Hendriks, R.C., Rainer, M., 2007. MAP estimators for speech enhancement under normal and Rayleigh inverse Gaussian distributions. *IEEE Trans. Audio Speech Lang. Process.*, **15**(3):918-927. [doi:10.1109/TASL.2006.889753]

Hermansky, H., 1990. Perceptual linear predictive (PLP) analysis of speech. *J. Acoust. Soc. Am.*, **87**(4):1738-1752. [doi:10.1121/1.399423]

Hermansky, H., Morgan, N., 1994. RASTA processing of speech. *IEEE Trans. Speech Audio Process.*, **2**(4):578-589. [doi:10.1109/89.326616]

Hu, G., Wang, D., 2004. Monaural speech segregation based on pitch tracking and amplitude modulation. *IEEE Trans. Neur. Networks*, **15**(5):1135-1150. [doi:10.1109/TNN.2004.832812]

ITU-T P.862, 2001. Perceptual Evaluation of Speech Quality (PESQ): An Objective Method for End-to-End Speech Quality Assessment of Narrow-Band Telephone Networks and Speech Codecs. International Telecommunication Union, Geneva.

Jensen, J., Heusdens, R., Jensen, S.H., 2003. A Perceptual Subspace Method for Sinusoidal Speech and Audio Modeling. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing, p.401-404.

Kondoz, A.M., Evans, B.G., 1987. Hybrid Transform Coder for Low Bit Rate Speech Coding. Proc. European Conf. on Speech Technology, p.105-108.

Kondoz, A.M., Evans, B.G., 1988. A Robust Vector Quantized Sub-Band Coder for Good Quality Speech Coding at 9.6 Kb/s. IEEE 8th European Conf. on Area Communication, p.44-47.

Kristijansson, T., Hershey, J., Olsen, P., Rennie, S., Gopinath, R., 2006. Super-Human Multi-Talker Speech Recognition: The IBM Speech Separation Challenge System. 9th Int. Conf. on Spoken Language Processing, p.97-100.

Li, P., Guan, Y., Wang, S., Xu, B., Liu, W., 2010. Monaural speech separation based on MAXVQ and CASA for robust speech recognition. *Comput. Speech & Lang.*, **24**(1):30-44. [doi:10.1016/j.csl.2008.05.005]

Loizou, P., 2007. Speech Enhancement Theory and Practice. CRC Press, Boca Raton, FL, USA, p.143.

Ma, J., Hu, Y., Loizou, P., 2009. Objective measures for predicting speech intelligibility in noisy conditions based on new band-importance functions. *J. Acoust. Soc. Am.*, **125**(5):3387-3405. [doi:10.1121/1.3097493]

Martin, R., 2005. Speech enhancement based on minimum square error estimation and super-Gaussian priors. *IEEE Trans. Speech Audio Process.*, **13**(5):845-856. [doi:10.1109/TSA.2005.851927]

Moore, B.C.J., 1997. An Introduction to the Psychology of Hearing (4th Ed.). Academic Press, New York, San Diego, USA, p.89-103.

Mowlaee, P., Sayadiyan, A., 2008. Model-Based Monaural Sound Separation by Split-VQ of Sinusoidal Parameters. 16th European Signal Processing Conf., p.1-5.

Mowlaee, P., Sayadiyan, A., 2009. Performance Evaluation for Transform Domain Model-Based Single-Channel Speech Separation. 7th ACS/IEEE Int. Conf. on Computer Systems and Applications, p.935-942. [doi:10.1109/AICCSA.2009.5069444]

Paliwal, K.K., Kleijn, W.B., 1995. Quantization of LPC Parameters. *In*: Kleijn, W.B., Paliwal, K.K. (Eds.), Speech Coding and Synthesis. Elsevier, Amsterdam, the Netherlands, p.443-466.

Radfar, M.H., Sayadiyan, A., Dansereau, R.M., 2006a. A New Algorithm for Two-Talker Pitch Tracking in Single Channel Paradigm. Int. Conf. on Signal Processing.

Radfar, M.H., Dansereau, R.M., Sayadiyan, A., 2006b. Performance Evaluation of Three Features for Model-Based Single Channel Speech Separation Problem. 8th Int. Conf. on Spoken Language Processing, p.2610-2613.

Radfar, M.H., Dansereau, R.M., Sayadiyan, A., 2007. A maximum likelihood estimation of vocal-tract-related filter characteristics for single channel speech separation. *EURASIP J. Audio Speech Music Process.*, **2007**:Article ID 84186, p.1-15. [doi:10.1155/2007/84186]

Reddy, A.M., Raj, B., 2007. Soft mask methods for single-channel speaker separation. *IEEE Trans. Audio Speech Lang. Process.*, **15**(6):1766-1776. [doi:10.1109/TASL.2007.901310]

Roweis, S., 2003. Factorial Models and Refiltering for Speech Separation and Denoising. 8th European Conf. on Speech Communication and Technology, p.1009-1012.

So, S., Paliwal, K., 2007. A comparative study of LPC parameter representations and quantisation schemes for wideband speech coding. *Dig. Signal Process.*, **17**(1):114-137. [doi:10.1016/j.dsp.2005.10.002]

Spiegel, M.R., Lipschutz, S., Liu, J., 1998. Schaum's Mathematical Handbook of Formulas and Tables. McGraw-Hill, New York, USA, p.111.

Srinivasan, S., Wang, D., 2008. A model for multitalker speech perception. *J. Acoust. Soc. Am.*, **124**(5):3213-3224. [doi:10.1121/1.2982413]

Srinivasan, S., Shao, Y., Jin, Z., Wang, D.L., 2006. A Computational Auditory Scene Analysis System for Robust Speech Recognition. 9th Int. Conf. on Spoken Language Processing, p.73-76.

Stevens, J.C., Marks, L.E., 1965. Cross-modality matching of brightness and loudness. *PNAS*, **54**(2):407-411. [doi:10.1073/pnas.54.2.407]

Tolonen, T., Karjalainen, M., 2000. A computationally efficient multipitch analysis model. *IEEE Trans. Speech Audio Process.*, **8**(6):708-716. [doi:10.1109/89.876309]

Wang, D.L., Brown, G.J., 2006. Computational Auditory Scene Analysis: Principles, Algorithms, and Applications. Wiley/IEEE Press, New Jersey, USA, p.1-72.

Wu, M., Wang, D.L., Brown, G.J., 2003. A multipitch tracking algorithm for noisy speech. *IEEE Trans. Speech Audio Process.*, **11**(3):229-241. [doi:10.1109/TSA.2003.811539]

Zavarehei, E., Vaseghi, S., Qin, Y., 2007. Noisy speech enhancement using harmonic-noise model and codebook-based post-processing. *IEEE Trans. Audio Speech Lang. Process.*, **15**(4):1194-1203. [doi:10.1109/TASL.2007.894516]