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

Split vector quantization for sinusoidal amplitude and frequency

Pejman MOWLAEE, Abolghasem SAYADIAN, Hamid SHEIKHZADEH

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

E-mail: pmb@es.aau.dk; eeas335@aut.ac.ir; hsheikh@aut.ac.ir

Received Jan. 28, 2010; Revision accepted Aug. 19, 2010; Crosschecked Dec. 30, 2010

Abstract: In this paper, we suggest applying tree structure on the sinusoidal parameters. The suggested sinusoidal coder is targeted to find the coded sinusoidal parameters obtained by minimizing a likelihood function in a least square (LS) sense. From a rate-distortion standpoint, we address the problem of how to allocate available bits among different frequency bands to code sinusoids at each frame. For further analyzing the quantization behavior of the proposed method, we assess the quantization performance with respect to other methods: the short-time Fourier transform (STFT) based coder commonly used for speech enhancement or separation, and the line spectral frequency (LSF) coder used in speech coding. Through extensive simulations, we show that the proposed quantizer leads to less spectral distortion as well as higher perceived quality for the re-synthesized signals based on the coded parameters in a model-based approach with respect to previous STFT-based methods. The proposed method lowers the complexity, and, due to its tree-structure, leads to a rapid search capability. It provides flexibility for use in many speaker-independent applications by finding the most likely frequency vectors selected from a list of frequency candidates. Therefore, the proposed quantizer can be considered an attractive candidate for model-based speech applications in both speaker-dependent and speaker-independent scenarios.

Key words: Short-time Fourier transform, Split vector quantization, Sinusoidal modeling, Spectral distortion
doi:10.1631/jzus.C1000020 **Document code:** A **CLC number:** TN912

1 Introduction

In many audio and speech applications, including speech enhancement and single-channel speech separation (SCSS), using a compact and efficient speaker model is often of high interest. In this aspect, vector quantization (VQ) has been widely used in many model-based applications including speech separation (Ellis and Weiss, 2006) in particular, or speech enhancement (Hu and Loizou, 2007; Zavarehei *et al.*, 2007) in general. According to the rate-distortion theorem, the objective of using a vector quantizer is to achieve a transparent quality of speech at the lowest possible bit-rate for transmission and storage purposes. Hence, it is a highly important issue that the quantization should not affect the subjective quality of the coded speech at the receiver end. In general, a speech coder is designed based on available noise-free speech to the quantizer. However, in many realistic noisy environments, the noise-free assumption is not fully satisfied and the received signals are often corrupted with some types of colored noise as interfering signals. Hence, spectral distortion (SD) is inevitable, and as a consequence, there is an increasing demand in many model-based applications to find efficient and robust speaker models in the form of a quantizer.

Many speech applications apply either short-time Fourier transform (STFT) features or its log version as their selected feature vectors. One of these areas is the model-based SCSS (Roweis 2003; Ellis and Weiss, 2006) and speech enhancement (Ephraim, 1992), where the STFT features are used mostly

for both maintaining simplicity and satisfying the additivity constraint. In this spirit, there is much interest for finding efficient quantizers coping with poor speech quality of the STFT-based speaker models (Ellis and Weiss, 2006; Mowlaee *et al.*, 2010a). Furthermore, quantizing the linear STFT magnitude in a mean-squared error (MSE) sense presents some problems: (1) All frequency bins are uniformly weighted (Ellis and Weiss, 2006) and generally no perceptual consideration of the human auditory system is taken into account. Furthermore, there is an undesirable over-emphasis on high frequency coefficients, which directly arises from the uniform bandwidth provided by the STFT bins that often poorly matches the logarithmic frequency sensitivity of the human auditory system (Ellis and Weiss, 2006). (2) The high dimensionality of the STFT features introduces high computational complexity. (3) The STFT represented in a time-frame consists of spurious peaks that are mostly perceptually irrelevant (Mowlaee and Sayadiyan, 2008; Mowlaee *et al.*, 2010a). These peaks may easily bias the decision made by the distortion measures towards extracting insignificant information in speech signals (Mowlaee and Sayadiyan, 2008; Mowlaee *et al.*, 2010a). Therefore, the STFT bins are likely to introduce weak quantization behaviors. Taking all these factors into account, using the STFT features for a model-based speech separation or speech enhancement is computationally expensive. According to Mowlaee *et al.* (2010a), by applying a transformation to the STFT features, it is possible to efficiently represent the most useful speaker information embedded in speech frames. As a consequence, compared to the common STFT-based model-driven methods, it is expected that one will achieve a higher upper-bound separation quality in the model-based SCSS (Ephraim, 1992; Mowlaee *et al.*, 2010a).

The limitations introduced by the STFT features serve as motivations for seeking efficient statistical models for speech enhancement applications to satisfy at least two requirements: (1) The selected model should address high quality when used for recovering the enhanced signals at the signal reconstruction stage. This means that the selected feature should provide a synthesis quality highly correlated with the subjective measures. (2) The proposed model is required to have low computational complexity and show flexibility in its structure. These requirements can be met by taking into account psychoacoustic foundations, mapping the STFT time-frequency cells into a more perceptually relevant domain, called subband perceptually weighted transformation as proposed in Mowlaee *et al.* (2010a). Another possibility is to employ a weighted distortion measure in the quantization step (Heusdens and van de Par, 2002; Christensen, 2008; van Schijndel *et al.*, 2008). The weighted squared error distortion (WSED) measure has been applied in Gardner and Rao (1995) or as a perceptual distortion measure in Christensen (2008) and van Schijndel *et al.* (2008). By employing these perceptually weighted distortion measures, the quantizer encodes the most perceptually relevant speech features while avoiding wasting bits on irrelevant components of the signal.

In general, there are two groups in speech coding that achieve promising quantization performance, i.e., reaching lower bit rates while preserving high perceptual quality: (1) linear predictive coding (LPC) based speech coding, (2) sinusoidal coders. The most popular feature used in the former category is line spectral frequency (LSF) widely used in low-bit-rate narrow-band speech coding (Paliwal and Atal, 1993; Paliwal and Kleijn, 1995). The LSF parameters are compact and highly representative for speech frames at formants. In contrast, the second group is composed of a VQ system based on harmonic or sinusoidal parameters. Both categories offer attractive candidates to arrive at a high fidelity quantization performance. In this paper, we focus only on the latter group. There has been a high demand and focus on the second quantization group. To name a few, the idea of polar quantization was developed in Moo and Neuhoff (1998), and it was shown that sinusoidal parameters can be coded in polar format. According to Ahmadi and Spanias (2001) and van Schijndel *et al.* (2008), sinusoidal coding has proved rather effective for representing audio signals in low-bit-rate speech coding. In this respect, high-resolution theory was adopted for frequency quantization in Heusdens *et al.* (2007). The spherical VQ was presented in Korten *et al.* (2007). In Mowlaee and Sayadiyan (2008), split-VQ tree structure was applied on sinusoidal amplitude and frequency parameters. Recently, we employed split-VQ as the speaker model for SCSS (Mowlaee *et al.*, 2010b; 2010c) and joint speech separation and speaker identification (Mowlaee *et al.*, 2010d).

In this paper, we develop ideas in favor of presenting a new quantization framework based on sinusoids. We derive the mathematical proof for minimizing the distortion function in split-VQ and the total distortion function in a least square (LS) sense based on the sinusoids. Through extensive simulations, we show that the proposed structured VQ outperforms the quality obtained by the STFT feature, predominantly used in model-based speech applications. The quantization performance reported in experiments is interpretable as the separation upper-bound performance. Through experiments we evaluate the effectiveness of the proposed quantizer in terms of speech coding and speech separation performance. As objective measures, we employ perceptual evaluation of speech quality (PESQ) (ITU-T P.862, 2001), segmental signal-to-noise ratio (SSNR), and weighted-spectral slope (WSS) measures. The results are reported for speaker-dependent and speaker-independent scenarios. The proposed quantizer reduces the computational complexity and searching time often introduced as two difficulties in model-based applications.

2 Sinusoidal parameter estimation

In this section, we present the parameter estimation used to find the sinusoidal components composed of amplitude, frequency, and phase values at each frame of the speech signal. Given a real observed speaker signal at an arbitrary frame as $s(n)$ for n=0, 1, ..., $N-1$, the parameters of the signal of interest in additive noise $w(n)$ can be demonstrated as

$$s(n)=\sum_{i=1}^{M} a_i\cos(2\pi f_i n+\phi_i)+w(n), 0\le n\le N-1, \tag{1}$$

where $i\in[1,M]$ is the index for the ith sinusoidal component, n is the time sample index, N is the analysis time window length in samples, f_i, a_i, and ϕ_i denote the frequency, amplitude, and phase of the ith sinusoidal component, respectively, and M is the sinusoidal model order not known a priori. As our parameter estimation for the sinusoidal components in Eq. (1), we use a sinusoidal modeling similar to McAulay and Quatieri (1986), but with two modifications: (1) the spectral coefficients of the given speech signal are first translated to mel-scale, to take into account the logarithmic sensitivity in human perception; (2) at each frequency band, only the spectral peaks with the largest amplitude are retained. This decision making agrees well with the so-called masking principle, stating that a louder signal masks a weaker one and makes it inaudible at a critical band (Moore, 1997). The goal of the parameter estimation is to obtain triplet of vectors $\boldsymbol{a}$ and $\boldsymbol{f}$, each as an $M\times 1$ column vector. As a parametric matrix, we have $\boldsymbol{\Lambda}=[\boldsymbol{a}\ \ \boldsymbol{f}]$ of dimension $M\times 3$. Each sinusoidal frequency vector involving f_i refers to the selected peak in the ith band represented by

$$\boldsymbol{v}_i=[1\ \ \mathrm{e}^{\mathrm{j}2\pi f_i}\ \ \ldots\ \ \mathrm{e}^{\mathrm{j}2\pi f_i(N-1)}]^{\mathrm{T}},\quad i\in[1,M], \tag{2}$$

where $\boldsymbol{v}_i$ is the ith frequency vector of the discrete-time Fourier transform (DFT) of dimension $N\times 1$, and f_i is the frequency of the selected peak at the ith band. Each of the sinusoidal frequency vectors in Eq. (2) is reformulated in a compact matrix format:

$$\boldsymbol{V}=[\boldsymbol{v}_1\ \ \boldsymbol{v}_2\ \ \cdots\ \ \boldsymbol{v}_M]^{\mathrm{T}},\quad i\in[1,M], \tag{3}$$

where $\boldsymbol{V}$ is an $M\times N$ matrix whose columns are $\boldsymbol{v}_i$. The ith sinusoidal frequency vector, $\boldsymbol{v}_i$, is comprised of the selected peak at the ith band. We define the complex amplitude for each sinusoid as $a_i=A_i\mathrm{e}^{\phi_i}$, $i\in[1,M]$. Then the reconstructed signal based on sinusoids is $\{s(n)\}_{n=1}^{N}=\mathrm{Re}\{\boldsymbol{V}^{\mathrm{T}}\boldsymbol{a}\}$, where $\mathrm{Re}\{\cdot\}$ denotes the real part, and we define $\boldsymbol{a}=\{A_i\}_{i=1}^{L}$ as an $M\times 1$ vector composed of sinusoidal amplitudes selected in L frequency bands and $(\cdot)^{\mathrm{T}}$ is the transpose operator. We define $S(f)$ as the complex spectrum. The objective of the sinusoidal parameter estimation considered here is to find peaks by solving the following optimization problem:

$$f_i=\arg\max_{f\in\mathcal{F}_i}\log|S(f)|,\quad A_i=\max|S(f_i)|, \tag{4}$$

with $i\in[1,M]$ and $\mathcal{F}_i\in(\Omega_{i-1},\Omega_i)$ denotes a set composed of all the frequencies within the ith band. This parameter estimation has successfully been used for compact signal representation (Mowlaee and Sayadiyan, 2008; Mowlaee *et al.*, 2009; 2010b).

3 Sinusoidal coder

3.1 Split-VQ for sinusoidal parameters

Conventional VQ algorithms cannot be directly applied to the sinusoidal components since each sinusoid is composed of two different feature types,

namely amplitude and frequency. In the following, we propose a split version VQ to effectively quantize sinusoidal parameters. The procedure we present here is a development of our recent idea on applying split-VQ on the sinusoids already presented in Mowlaee and Sayadiyan (2008).

Fig. 1 depicts the schematic block diagram for the split-VQ tree-structure based on sinusoidal parameters. We use two sub-vectors each of them having a length the same as the sinusoidal model order, M. The split-VQ codebooks are produced by the following two steps (Fig. 1):

1. Establish a codebook composed of amplitude codevectors.

2. Establish a smaller codebook for each amplitude codevector obtained in the first step.

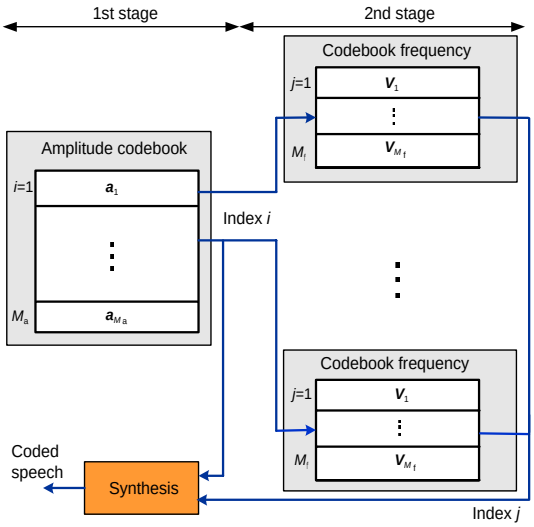

Fig. 1 The block diagram for the proposed sinusoidal coder, split-VQ composed of M_a amplitude codewords $\{\boldsymbol{a_1}, \boldsymbol{a_2}, \cdots, \boldsymbol{a_{M_\mathrm{a}}}\}$ followed by M_f candidates for frequencies $\{\boldsymbol{V_1}, \boldsymbol{V_2}, \cdots, \boldsymbol{V_{M_\mathrm{f}}}\}$

At the first step, we aim to form a codebook of sinusoidal amplitude codewords. The codebook generation process starts with extracting a sinusoidal parametric vector of the form $\boldsymbol{\Theta} = [\boldsymbol{a}|\,\boldsymbol{V}]$ where $\boldsymbol{a}$ and $\boldsymbol{V}$ are two matrices composed of the sinusoidal amplitudes and frequencies, respectively, obtained from all frames in the dataset in the training stage. The codebook size is denoted by M_a. The following distortion measure is used for quantizing the sinusoidal amplitude denoted by $\boldsymbol{a}$ and we obtain

$$d_\mathrm{a} = \sum_{i=1}^{L} \left(\frac{A_i}{\|\boldsymbol{a}\|_2^2} - \frac{\hat{A}_i}{\|\hat{\boldsymbol{a}}\|_2^2} \right)^2, \tag{5}$$

where we define $\|\boldsymbol{a}\|_2^2$ as the root mean square (RMS) value for the amplitude vector with $\|\cdot\|_2^2$ being the l_2-norm, $\hat{\boldsymbol{a}} = \{\hat{A}_i\}_{i=1}^{L}$ is the coded amplitude codevector with $\hat{A}_i$ being the coded amplitude for the sinusoidal peak selected at the ith band, and $d_\mathrm{a}(\cdot)$ denotes the distance measure applied to the amplitude part. In Eq. (5), index i refers to the ith frequency band. Note that both $\boldsymbol{a}$ and $\hat{\boldsymbol{a}}$ are first normalized to their respective RMS value. By applying a quantizer with the distortion measure in Eq. (5), we obtain a set of amplitude codewords denoted by $\hat{\boldsymbol{a}}_i$, where $i \in [1, M_\mathrm{a}]$ (Fig. 1). To select a proper codebook size M_a, we are required to achieve a trade-off between the accuracy and computational complexity. The possible range for the amplitude codebook is $M_\mathrm{a} \in \{256, 512, 1024, 2048, 4096\}$.

After forming the codebook for the amplitude part, the frequency codebook is produced by undertaking the following procedure. For each of the amplitude entries found in the previous stage, a frequency candidate is provided. To this end, a search is performed within all vectors in the training set, to find the closest vectors in terms of amplitude distance in Eq. (5). Then we attempt to find M_f nearest neighbors for each amplitude centroid indicated by $\hat{\boldsymbol{a}}$ with $i \in [1, M_\mathrm{a}]$. Finally, for each entry in the amplitude codebook entry, we provide a smaller codebook comprised of candidate frequency vectors denoted by $\{\hat{\boldsymbol{V}}_1, \hat{\boldsymbol{V}}_2, \ldots, \hat{\boldsymbol{V}}_{M_\mathrm{f}}\}$ where M_f is the frequency codebook size, $M_\mathrm{f} \in \{1, 2, 4, 8\}$. For quantizing the frequency part, we apply the following weighted distortion measure:

$$d_\mathrm{w}(\boldsymbol{V}, \hat{\boldsymbol{V}}) = \sqrt{\sum_{i=1}^{M} w_i (\boldsymbol{v}_i - \hat{\boldsymbol{v}}_i)^2}, \tag{6}$$

which applies a dynamic weighting, where $\boldsymbol{V}$ is the frequency part and $w_i = A_i / \|\boldsymbol{a}\|_2^2$ refers to the energy normalized amplitude used for dynamic weighting of the Euclidean distance measure to make it proportional to the sinusoidal amplitude at the peak frequency in the ith frequency band (Mowlaee and Sayadiyan, 2008; Mowlaee *et al.*, 2009). Then we obtain

$$d_\mathrm{w}(\boldsymbol{V}, \hat{\boldsymbol{V}}) = \sum_{i=1}^{L} w_i \|\boldsymbol{v}_i - \hat{\boldsymbol{v}}_i\|_2^2, \tag{7}$$

where $d_\mathrm{w}(\cdot)$ is a weighted square error measure. The role of $\boldsymbol{w}$ is to weight the Euclidean distance mea-

sure by emphasizing the spectral peaks of the power spectral density (PSD).

3.2 Theoretical derivations

We derive the mathematical formulations for maximizing a likelihood measure to be defined in the sinusoidal domain (equivalently minimizing a weighted distortion function in bands) through centroid update iterations. The objective of a vector quantizer is to maximize a maximum a posteriori (MAP) criterion as a function of the probability distribution function (pdf) of the speech data used in the training stage. We define the maximization step as

$$\boldsymbol{\Theta}_{\mathrm{MAP}} = \arg\max_{\boldsymbol{\Theta}} p(\boldsymbol{\Theta}|\boldsymbol{X}) = \arg\max_{\boldsymbol{\Theta}} p(\boldsymbol{X}]|\boldsymbol{\Theta}) g(\boldsymbol{\Theta}), \tag{8}$$

where $g(\boldsymbol{\Theta})$ is the a priori pdf of the parameters while $p(\boldsymbol{X}|\boldsymbol{\Theta})$ is the likelihood function for the training vectors $\boldsymbol{X}$ defined as $\boldsymbol{X} = \{\boldsymbol{x}_i\}_{i=1}^{R}$, R is the number of utterances used in the training stage, $\boldsymbol{x}_i$ is the ith training vector, and $\boldsymbol{\Theta}$ is the sinusoidal parametric vector. The density function for the kth centroid, $\boldsymbol{c}_k$, is given by

$$p(\boldsymbol{X}|\boldsymbol{c}_k, \boldsymbol{\Sigma}_k) = C(\epsilon)\exp\left(-\frac{d(\boldsymbol{X}, \boldsymbol{c})}{2}\right), \tag{9}$$

where we define $C(\epsilon) = 1/\sqrt{2\pi\varepsilon|\boldsymbol{\Sigma}_k|}$ as a constant and $d(\boldsymbol{X}, \boldsymbol{c}) = -(\boldsymbol{X} - \boldsymbol{c})\boldsymbol{\Sigma}_k^{-1}(\boldsymbol{X} - \boldsymbol{c})^{\mathrm{T}}/2$ with $\boldsymbol{\Sigma}_k$ being the covariance matrix for the kth centroid. We also consider that for each centroid in the quantizer $\boldsymbol{\Sigma}_k = \epsilon\boldsymbol{I}$, ϵ is a small constant and $d(\cdot)$ reduces to a Euclidean distance measure. The optimal $\boldsymbol{\Theta}$ is obtained by maximizing the log-likelihood function given by

$$Q(\boldsymbol{\Theta}_i, \hat{\boldsymbol{\Theta}}_k) = -\sum_{i=1}^{N}\sum_{k=1}^{K} m_{i,k} d(\boldsymbol{x}_i, \boldsymbol{c}_k), \tag{10}$$

where $\hat{\boldsymbol{\Theta}}_k$ is the parameters for the kth centroid estimated in the preceding iteration, $\boldsymbol{\Theta}_i$ contains the parametric vector for the ith sinusoid, and $m_{i,k}$ indicates a binary membership function defined as

$$m_{i,k} = \begin{cases} 1, & \text{if } \boldsymbol{x}_i \in \boldsymbol{c}_k, \\ 0, & \text{otherwise}, \end{cases} \tag{11}$$

indicating whether the ith training vector $\boldsymbol{x}_i$ belongs to the kth centroid. By letting $m_{i,k} = 1$ for those neighbors that satisfy $d(\boldsymbol{x}_i, \boldsymbol{c}_k) < \epsilon$, we re-define the log-likelihood function in Eq. (10) as a weighted sum of the distortions experienced by each centroid $\boldsymbol{c}_k$ as

$$Q(\boldsymbol{\Theta}_i, \hat{\boldsymbol{\Theta}}_k) = \sum_{i=1}^{N}\sum_{j=1}^{K} m_{i,j} d(\boldsymbol{x}_i, \boldsymbol{c}_j), \tag{12}$$

where K is the number of centroids used in the quantizer. From Eq. (5), we replace $\boldsymbol{x}_i$ and $\boldsymbol{c}_j$ with $i \in [1, R]$ and $j \in \{1, K\}$ with their sinusoidal modeled equivalents. Then the likelihood function can be reformulated as

$$Q(\boldsymbol{\Theta}_i, \hat{\boldsymbol{\Theta}}_k) = \sum_{k=1}^{K} m_{i,k} \| \boldsymbol{V}_i^{\mathrm{T}} \boldsymbol{a}_i - \hat{\boldsymbol{V}}_k^{\mathrm{T}} \hat{\boldsymbol{a}}_k \|_2^2, \tag{13}$$

which addresses a sum of least square minimizations defined in frequency bands that can be solved by setting the derivative of Q equal to $\mathbf{0}$, i.e., $\dfrac{\partial Q(\boldsymbol{\Theta}_i, \hat{\boldsymbol{\Theta}}_k)}{\partial \hat{\boldsymbol{V}}} = \mathbf{0}$. We obtain

$$\hat{\boldsymbol{a}}_k = (\hat{\boldsymbol{V}}_k \hat{\boldsymbol{V}}_k^{\mathrm{T}})^{-1} \hat{\boldsymbol{V}}_k \boldsymbol{V}_i^{\mathrm{T}} \boldsymbol{a}_i. \tag{14}$$

The likelihood function is calculated by substituting Eq. (14) into $Q(\boldsymbol{\Theta}_i, \hat{\boldsymbol{\Theta}}_k)$ in Eq. (13) and we obtain

$$\begin{aligned} Q(\boldsymbol{\Theta}_i, \hat{\boldsymbol{\Theta}}_k) =& \| \boldsymbol{V}_i^{\mathrm{T}} \boldsymbol{a}_i - \hat{\boldsymbol{V}}_k^{\mathrm{T}} (\hat{\boldsymbol{V}}_k \hat{\boldsymbol{V}}_k^{\mathrm{T}})^{-1} \hat{\boldsymbol{V}}_k \boldsymbol{V}_i^{\mathrm{T}} \boldsymbol{a}_i \|_2^2 \\ =& \|(\boldsymbol{I} - \boldsymbol{\Pi}_k) \boldsymbol{V}_i^{\mathrm{T}} \boldsymbol{a}_i\|_2^2 \\ =& \|\boldsymbol{\Pi}_k^{\perp} \boldsymbol{V}_i^{\mathrm{T}} \boldsymbol{a}_i\|_2^2, \end{aligned} \tag{15}$$

where we define $\boldsymbol{\Pi}_k$ as a projection matrix, which projects any given signal denoted as $\boldsymbol{x}_i = \boldsymbol{V}_i^{\mathrm{T}} \boldsymbol{a}_i$ from the training set into the sinusoidal subspace spanned by the centroid $\boldsymbol{V}_k$. The matrix $\boldsymbol{\Pi}_k^{\perp}$ projects a given feature vector onto the space that is orthogonal to the space spanned by the columns of $\boldsymbol{V}_k$.

3.3 Relation to previous sinusoidal coders

The proposed split-VQ is close to the previous sinusoidal quantizers presented in Vafin and Kleijn (2005) and Korten *et al.* (2007). However, split-VQ considers a dynamic weight while minimizing the amplitude and frequency distortion functions in Eqs. (5) and (6), respectively. This is similar to the joint distortion function in spherical quantizers presented by Korten *et al.* (2007) as

$$d_{\pi} = K\left(\frac{\|\boldsymbol{w}\|_2^2}{12}(\Delta \boldsymbol{a}^2 + \hat{\boldsymbol{a}}(\sigma^2 \Delta \boldsymbol{\nu}^2))\right), \tag{16}$$

where $\boldsymbol{\nu}$ is an $M \times 1$ vector composed of the angular frequencies measured in rad/s and $\nu = 2\pi \boldsymbol{f}$, and d_π is the distortion function for the entropy-constrained strictly spherical quantization (ECSSQ) in Korten *et al.* (2007) with the distortions for amplitude and frequency parts defined as

$$\Delta \boldsymbol{a}^2 = \|\boldsymbol{a}_i - \hat{\boldsymbol{a}}_k\|_2^2, \quad \Delta \boldsymbol{\nu}^2 = \|\boldsymbol{\nu}_i - \hat{\boldsymbol{\nu}}_k\|_2^2. \tag{17}$$

Putting the equivalent terms of $d_{\boldsymbol{a}}(\boldsymbol{a}_i, \hat{\boldsymbol{a}}_k)$ and $d_{\boldsymbol{w}}(\boldsymbol{V}_i, \hat{\boldsymbol{V}}_k) = \|\boldsymbol{V}_i - \hat{\boldsymbol{V}}_k\|_2^2$ in Eq. (17), we obtain

$$d_{\text{split}-\text{VQ}} = K\left(\Delta \boldsymbol{a}^2 + \boldsymbol{a}_i^2 \Delta \boldsymbol{V}^2\right), \tag{18}$$

where we define $d_{\text{split}-\text{VQ}}$ as the total distortion to be minimized in the two steps as explained in the previous section. This is also comparable to the distortion function already proposed by Vafin and Kleijn (2005), defined as

$$d(\boldsymbol{a}, \Delta \boldsymbol{a}, \Delta \psi) \approx \frac{\Delta \boldsymbol{a}^2 + \hat{\boldsymbol{a}}^2 \Delta \psi^2}{12}. \tag{19}$$

We define $\psi = 2n\pi \boldsymbol{f} + \phi$. By neglecting the phase term $\boldsymbol{\phi}$ in Eq. (19) it is observed that the proposed distortion function in Eq. (19) reduces to Eq. (18).

The required steps in the proposed split-VQ are as follows:

1. Estimate the sinusoidal parameters for the whole training dataset. Find $[A_1, A_2, \ldots, A_M]$, $[v_1, v_2, \ldots, v_M]$.

2. Update the centroid to find the amplitude part.

3. Establish the tree-structure's $M_\text{a} \times M_\text{f}$ connections. Find M_f frequency candidates for M_a codewords.

4. Update the centroid to find the coded sinusoidal frequency.

4 Rate-distortion and spectral distortion

We conduct an experiment to study the distribution of the sinusoidal amplitudes obtained in the parameter estimation step given by Eq. (5). We used the corpus in Cooke *et al.* (2006) provided for separation purposes (see Section 5). We extracted 100 000 vectors following the training utterances. We independently calculated the histogram for each frequency band. Fig. 2 illustrates the histogram for bands 1, 5, 10, 20, and 25. It is observed that the sinusoidal amplitude selected per band follows a Rayleigh distribution with different variances σ_i where $i \in [1, M]$. Therefore, in the following we assume that each component of the amplitude part, A_i, has a Rayleigh distribution, indicated by $A_i \sim R(0, \sigma_i)$.

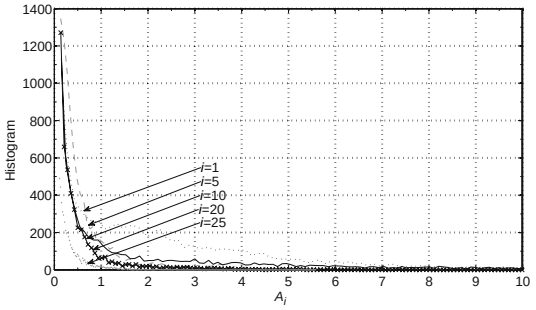

Fig. 2 Histogram for sinusoidal amplitudes selected per each frequency band. It is observed that at each band, the sinusoidal amplitude follows a Rayleigh distribution characterized by a Rayleigh distribution, i.e., $A_i \sim R(0, \sigma_i)$ with $i \in [1, M]$

4.1 Rate-distortion for sinusoidal amplitudes

Let D_A be the amplitude distortion. Then from the rate-distortion theorem in Cover and Thomas (2006) we have

$$R_A(D_A) = \min_{E(A-\hat{A})^2 \le D_A} I(A; \hat{A}), \tag{20}$$

where $I(A; \hat{A})$ indicates the mutual information between the uncoded and coded sinusoidal amplitude parameters, $R_A(D_A)$ is the rate function for the amplitude part calculated for amplitude distortion (D_A), and $\hat{A}$ denotes the coded amplitude. In the following, we aim to find the rate-distortion function for the amplitude part. Assuming that the amplitudes of sinusoids follow a Rayleigh distribution, we obtain

$$f_A(A) = \frac{A}{\sigma_A^2} \exp\left(-\frac{A^2}{2\sigma_A^2}\right), \quad A > 0,\ \sigma_A > 0, \tag{21}$$

where σ_A is the variance of the distribution, $f_A(\cdot)$ is the pdf for the sinusoidal amplitude, and the expected value for the amplitude vector is $E[A] = \sigma_A \sqrt{\pi/2}$ with $E[\cdot]$ being the expectation operator. The differential entropy will be (Cover and Thomas, 2006)

$$h_A(A) = 1 + \frac{\gamma}{2} + \ln \frac{\sigma_A^2}{\sqrt{2}}, \tag{22}$$

where γ is the Euler constant. From the mutual information theorem we obtain

$$\begin{aligned} I(A;\hat{A}) =& h_A(A) - h(A|\hat{A}) \\ =& 1 + \frac{\gamma}{2} + \ln\frac{\sigma_A}{\sqrt{2}} - h_A(A-\hat{A}|\hat{A}). \end{aligned} \quad (23)$$

Conditioning reduces the entropy; hence, $h(A-\hat{A}|\hat{A}) \leq h(A-\hat{A})$ and from Eq. (23) we have

$$\begin{aligned} I(A;\hat{A}) \geq& K + \ln\frac{\sigma}{\sqrt{2}} - h(A-\hat{A}) \\ \geq& K + \ln\frac{\sigma}{\sqrt{2}} - h(f_A(E[(A-\hat{A})^2])) \\ =& \ln\frac{\sigma_A}{\sqrt{2}} - \ln\frac{E(A-\hat{A})^2}{\sqrt{2}} \\ =& \ln\frac{\sigma_A}{D_A}, \end{aligned} \quad (24)$$

where $K = 1+\gamma/2$ and $D_A = E[(A-\hat{A})^2]$. The right hand side in Eq. (24) is the lower bound of the mutual information; by plugging it to the rate-distortion function for the amplitude parameter distribution we obtain

$$R_A(D_A) = \ln\left(\frac{\sigma_A}{D_A}\right). \quad (25)$$

As an important special case, assume that the amplitude parameters per band found by Eq. (5) are independent (but not identically distributed) random variables indicated by $\boldsymbol{a} = \{A_i\}_{i=1}^{M}$. For each sinusoidal component, we assume that $A_i \sim R(0, \sigma_{A_i})$ with $i \in [1, M]$. Assume that we are given R bits to present a random vector $\boldsymbol{a}$. The question is how to allocate bits among different bands to code the amplitude components denoted by A_i with the constraint of minimizing the total quantization error defined by $D_a = E[(\boldsymbol{a}-\hat{\boldsymbol{a}})^2]$, where $\hat{\boldsymbol{a}} = \{\hat{A}_i\}_{i=1}^{M}$ is the coded sinusoidal amplitude vector. This fundamental question is addressed in split-VQ as explained in the following. According to the generalized definition for rate-distortion in a vector format (here a $1 \times M$ vector), we have

$$I(\boldsymbol{a};\hat{\boldsymbol{a}}) = \sum_{i=1}^{M} I(A_i;\hat{A}_i) \geq \sum_{i=1}^{M} R_A(D_{A_i}) \geq \sum_{i=1}^{M} \ln\frac{\sigma_{A_i}}{D_{A_i}}, \quad (26)$$

where we define D_{A_i} as the distortion for the ith sinusoidal component. From Eq. (26), it is observed that given the independence assumption for sinusoidal amplitudes, A_i (each having a Rayleigh distribution σ_{A_i}), the total rate-distortion is expressed in the form of the summation of the rates experienced at each band.

4.2 Lower-bound on entropy of amplitudes

Let $S(\mathrm{e}^{\mathrm{j}\omega})$ be the power spectrum with an autocorrelation matrix of size $N \times N$ and let $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_N \geq 0$ be the corresponding eigenvalues. From the Szegö theorem (Grenander and Szegö, 1984) we have

$$H(A) = \int_{-\pi}^{\pi} g(S(\mathrm{e}^{\mathrm{j}\omega}))\frac{\mathrm{d}\omega}{2\pi} > \lim_{N\to\infty}\frac{\sum_{k=1}^{N} g(\lambda_k)}{N}, \quad (27)$$

where $g(\cdot)$ can be any arbitrary continuous real-valued function. Here we replace $g(\cdot)$ by the natural logarithm function and have

$$H(A) > \lim_{N\to\infty}\frac{\sum_{k=1}^{N}\ln\lambda_k}{N} > \lim_{N\to\infty}\frac{\sum_{k=1}^{M}\ln\lambda_k}{N}, \quad (28)$$

where M is the model order of sinusoids and N is the time window length in samples. Based on Eq. (28), the entropy is lower-bounded by summation of $\{\ln\lambda_k\}_{k=1}^{M}$ defined in the ith band. Hence, to maximize the lower-bound, one is required to maximize the entropy $H(A)$ on the left side of Eq. (28). In order to maximize the lower-bound, it is required to select the largest eigenvalues per band. Given the independence assumption for the frequency bands, each selected amplitude per band is proportional to the eigenvalues λ_i with $i \in [1, M]$. Following that the sinusoidal parameter estimation in Eq. (5) takes the largest peak from each of the M frequency bands, as a conclusion, we end up with maximizing $\sum_{i=1}^{M}\ln\lambda_i$ in the right hand side of Eq. (28), which maximizes the entropy, $H(A)$. We showed that the entropy of the split-VQ coder is maximized by selecting the highest amplitude per band. Accordingly, by following the decision of taking the highest peak per band it is possible to spend most bits in those bands that contribute most to the perceived speech quality.

4.3 Spectral distortion for split-VQ

From the definition, the spectral distortion between the coded and uncoded sinusoidal amplitudes in the proposed split-VQ is defined as

$$\begin{aligned} \mathrm{SD}_a(\boldsymbol{a},\hat{\boldsymbol{a}}) =& \sum_{i=1}^{M}(A_i - \hat{A}_i)^2 \\ =& \sum_{i=1}^{M}\mathrm{SD}_{\mathrm{F}}^2(f)|_{f=\hat{f}_i} < \mathrm{SD}_{\mathrm{F}}^2, \end{aligned} \quad (29)$$

where SD_F is the full-band spectral distortion. Two observations are made from Eq. (29): (1) the SD defined on the sinusoidal amplitudes in split-VQ is upper-bounded by the full-band SD already used in speech coding; and, (2) the SD experienced by the amplitude part is a non-uniform version of the log-spectrum distortion measure, SD_F, sampled at frequencies $\{\hat{f}_i\}_{i=1}^{M}$.

It is already known that too many outliers with large SDs cause audible distortion even though the average SD is 1 dB (Paliwal and Atal, 1993). Therefore, more recent studies have tried to reduce the amount of outlier frames and keep the average SD low. Basically, the outlier frames are divided into two groups including the frames with 2 dB<SD<4 dB and with SD>4 dB (Paliwal and Atal, 1993). According to Paliwal and Atal (1993), reasonable accuracy for transparent coding is attainable whenever the average SD is about 1 dB; i.e., the coded speech is indistinguishable from the original speech through listening tests. Furthermore, the number of outlier frames with more than 4 dB of SD must be kept sufficiently low.

5 Validation through experimental results

The corpus is composed of 34 speakers (18 male, 16 female), with a total number of 34 000 utterances, each following a command-like structure, and all having a unique grammatical structure. Each sentence is formed by different syntaxes of command, color, letter, number, and code, for instance, "bin white by A 3 please".

The performance of the proposed coder is determined by several factors affecting the distortion measures. These parameters are: window size (N), frame-shift, number of sinusoids (M), and split-VQ codebook sizes, determined by amplitude and frequency model orders denoted by M_a and M_f, respectively. We studied the effectiveness of these parameters in detail to determine the best performance by the proposed split-VQ. For evaluating the performance of the proposed split-VQ, we used the comprehensive database provided in Cooke *et al.* (2006) for the SCSS task. Speakers 4, 23, 33, and 34 were selected for female speakers while speakers 9, 19, 30, and 32 were selected as male speakers. The sampling frequency was decreased to 8 kHz from the original 25 kHz. Throughout the experiments, a Hann window of duration 32 ms was used with a frame-shift of 10 ms (except in Experiments 1 and 2 where we varied these parameters to determine their optimal values). The sinusoidal parameters were extracted on the entire training dataset of each speaker. The codebook for each speaker was trained based on the training speech dataset composed of 15 min, extracted from the utterances of each speaker (Cooke *et al.*, 2006). Possible ranges for amplitude and frequency parts were $M_a = \{128, 256, 512, 1024, 2048, 4096\}$ and $M_f = \{2, 4, 8\}$, respectively.

5.1 Experiments

Experiment 1 (Frame-shift selection) Fixing the window size to 32 ms, we obtained the results in terms of SSNR and SD measures versus different choices for the frame-shifts set equal to 8, 10, and 16 ms (Table 1). It is observed that choosing a frame-shift of 10 ms results in the minimum number of outliers with an SD higher than 4 dB. This choice also results in the minimum average distortion and the maximum SSNR.

Table 1 Spectral distortion (SD, in dB) versus frame-shift*

Shift	SD occurrence percentage (%)			SD_{avg}	SSNR
(ms)	SD<2	2<SD<4	SD>4	(dB)	(dB)
8	88.8	9.4	1.8	0.9	7.9
10	**94.0**	**5.4**	**0.6**	**0.7**	**8.5**
16	86.7	10.8	2.5	1.0	7.1

* The window size is fixed at 32 ms. Bold numbers represent the best results

Experiment 2 (Window size selection) We aim to determine the best choice of the window size parameter for the proposed sinusoidal coder. The frame-shift was set equal to 10 ms according to the results given in Table 1. The split-VQ was implemented for different window sizes of $\{25, 32, 40, 50\}$ ms. The results are shown in Table 2. Using a window size of 32 ms with a frame-shift of 10 ms, results in the best SD statistics in terms of achieving a minimum outlier and a high percentage of SD<2 dB.

Experiment 3 (Model order of sinusoids) Table 3 shows the effect of employing different numbers of sinusoids on the quantization performance of the proposed coder. Although the pdf of the spectral distortion is highly compact (>92.0% for

Table 2 Window-size effect of split-VQ*

N	SD occurrence percentage (%)			SSNR
(ms)	SD<2	2<SD<4	SD>4	(dB)
25	80.7	16.4	2.9	6.1
32	**82.2**	**16.0**	**1.8**	**8.5**
40	79.1	18.1	2.8	9.7
50	80.6	15.2	4.2	7.9

* The frame-shift is fixed at 10 ms. SD: in dB. N: window size. Bold numbers represent the best results

SD<2 dB) for M<34, the resulting SD_{avg} score is unacceptably high. In addition, we observe that with 50<M<80 we achieve a trade-off between the model order for the sinusoidal model and the compactness of histogram for SD. The results obtained here are in agreement with those in Mowlaee and Sayadiyan (2008), where it was demonstrated that using 40–50 sinusoids is enough to deal with the trade-off between low dimensionality and good synthesized quality (however, some audible artifacts exist). Some test samples and the processed signals used in experiments are downloadable from http://kom.aau.dk/∼pmb/jzus2.htm.

Table 3 Spectral distortion (SD, in dB) results for split-VQ versus the number of sinusoids (M)

M	SD occurrence percentage (%)			SD_{avg}
	SD<2	2<SD<4	SD>4	(dB)
15	92.6	6.5	0.90	1.07
25	92.8	6.9	0.28	0.08
33	94.8	5.0	0.20	0.036
40	**92.6**	**7.2**	**0.19**	**0.043**
50	91.4	8.1	0.46	0.048
60	90.4	9.1	0.50	0.055
70	90.1	9.4	0.49	0.055
80	89.7	9.5	0.80	0.059

Bold numbers represent the best results

Experiment 4 (Codebook size selection) We studied the performance of split-VQ by evaluating different values for the amplitude codebook size, $M_a \in \{128, 256, 512, 1024, 2048, 4096\}$. Fig. 3 illustrates the SSNR results versus different values of M_a. As can be seen from this figure, choosing M_a=2048 results in an acceptable SSNR, using higher M_a does not necessarily lead to very significant improvement in terms of SSNR, and increasing M_a consistently lowers the value of SD_{avg}, leading to a higher SSNR and a more compact distortion pdf (see the dense concentration for SD<2 dB region in Fig. 4) and finally a lower number of outliers (SD>4 dB region in Fig. 4). As a result, the number of amplitude codevectors was set to M_a=2048.

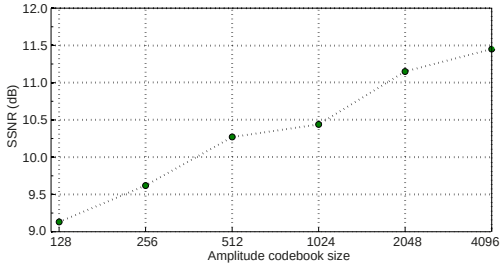

Fig. 3 The segmental signal-to-noise ratio (SSNR) results versus the amplitude codebook size

In order to study the effect of selecting different frequency codebook sizes M_f, we conducted another experiment. According to our previous experiments described above, we fixed the amplitude codebook size M_a=2048, the window size of 32 ms, and the frame-shift of 10 ms. For the frequency codebook size, we used $M_f = \{2, 4, 8\}$. Table 4 shows the quantization results in terms of WSS, PESQ, SNR, and SD. It is observed that increasing the frequency codebook size slightly improves all these measures. This choice of the codebook size will also lower the SD_f and WSS, but increase SNR and PESQ scores.

Table 4 Spectral distortion (SD) results for split-VQ versus different numbers of frequency candidates, M_f, in the frequency codebook*

Gender	M_f	SD_f (dB)	SNR (dB)	WSS	PESQ
Female	2	0.19	8.14	56.15	2.21
	4	0.17	8.62	55.04	2.25
	8	**0.15**	**10.10**	**54.12**	**2.27**
Male	2	0.24	5.14	35.59	2.08
	4	0.21	5.24	35.30	2.09
	8	**0.18**	**5.25**	**34.91**	**2.10**

* Amplitude codebook size M_a=2048, windows size N=32 ms, and frame-shift M=10 ms. SNR: signal-to-noise ratio; WSS: weighted-spectral slope; PESQ: perceptual evaluation of speech quality. Bold numbers represent the best results

5.2 Choice of α

Similar to Mowlaee and Sayadiyan (2008) and Mowlaee *et al.* (2010a), here we employed a transformation step by normalizing the amplitude vector to its maximum value, and then took the logarithm of the resulting normalized amplitude vectors. The

transformation can be summarized as

$$A_{n,i}(k)=\frac{A_i(k)}{\max|\hat{\boldsymbol{a}}|}, \tag{30}$$

$$\tilde{A}_i(k)=\log\left(1+\alpha A_{n,i}(k)\right), \tag{31}$$

where $A_{n,i}(k)$ is the kth frequency bin of the normalized amplitude vector to its maximum, and $\tilde{A}_i(k)$ is the transformed amplitude. The transformation will reduce the amplitude dynamic range and guarantee to keep it positive. Table 5 shows the results of employing different values of α in Eq. (30). It is observed that employing α=1000 results in the best performance in the sense of achieving the minimum outliers and the minimum average distortion.

Table 5 Spectral distortion (SD, in dB) versus different feature types

Feature	SD_{avg}	SD occurrence percentage (%)		
	(dB)	2<SD<4	SD>4	SD<2
Split-VQ				
$\alpha=10^3$	**0.4**	**12.2**	**0.5**	**87.3**
$\alpha=1$	1.0	14.8	1.2	84.1
$\alpha=0$	1.1	15.8	1.5	82.7
STFT	1.67	19.9	9.6	70.5

Bold numbers represent the best results

5.3 Comparison with other speech coders

To study the effectiveness of the proposed quantizer we compared the performance of the split-VQ presented in this work with other two quantization methods, namely LSF (Paliwal and Atal, 1993) and STFT (Ellis and Weiss, 2006). As objective and subjective assessments, we used four measures, namely SD between the uncoded and coded speech spectra, PESQ, SSNR, and the log-likelihood ratio (LLR). According to Hu and Loizou (2007) and Loizou (2007), our main focus is dedicated to PESQ due to its high correlation with subjective measures.

Table 6 summarizes the SD statistics for different quantization methods. In our simulation, the window size was set to 32 ms and a frame-shift of 10 ms was used. The number of sinusoids was set to $M=50$. The percentage of outliers in the STFT case is approximately 10 times larger than the percentage achieved by the proposed split-VQ. These results confirm the inferior performance of the STFT features for being used in SCSS (Kristijansson *et al.*, 2004; Ellis and Weiss, 2006; Mowlaee and Sayadiyan, 2008; Mowlaee *et al.*, 2010a).

Table 6 Spectral distortion (SD, in dB) statistics for different coders

Method	SD occurrence percentage (%)		SD_{avg}
	SD<2	SD>4	(dB)
STFT			
Male	70.50	9.60	1.70
Female	84.30	4.50	1.10
LSF*			
Unweighted	89.34	0.11%	1.37
Weighted	95.57	0.05	1.18
Split-VQ			
α =1	88.60	1.05	0.10
$\alpha=10^3$	**92.60**	**0.19**	**0.40**

* With 24 bits/frame (Paliwal and Atal, 1993). Bold numbers represent the best results

Table 6 shows the SD statistics for the proposed sinusoidal coder compared to other benchmark methods, namely STFT (Ellis and Weiss, 2006) and LSF coders (Paliwal and Atal, 1993; So and Paliwal, 2007). The proposed quantizer achieves a lower SD. The highest occurrence of SD<2 dB (95.57%) is achieved by LSF. However, the average SD in the unweighted LSF domain is around 1.37 dB. It also limits the outliers to be as small as 0.11%. However, using a weighted LSF distance results in outliers of 0.05%. Comparison of these results to those obtained by the split-VQ with α=1000 shows that the proposed sinusoidal coder achieves a performance close to the 24 bits/frame LSF quantizer in Paliwal and Atal (1993). Conducting different simulations, we observe that the pdf for SD becomes more compact compared to the case when no weighting is used (when α=0). Compared to the STFT case, it lowers the average distortion to one tenth showing a significant improvement. Figs. 4a and 4b show the pdf shapes for the SD results of the STFT and split-VQ, respectively. The proposed method achieves a more compact PSD (close to exponential distribution) offering a favorable choice for quantization purposes.

5.4 Evaluating the upper-bound performance

We considered single-channel speech separation as an example for model-based speech applications. We studied the effectiveness of the proposed sinusoidal coder when being used as a speaker model in SCSS. This study can be well generalized to other single-channel model-based speech applications. In general, a model-based SCSS method is at least composed of three blocks: speaker model, likelihood estimator, and a reconstruction stage. The speaker

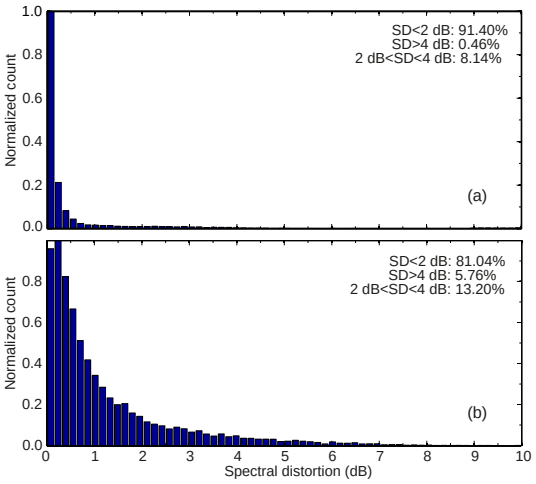

Fig. 4 The probability density function for the spectral distortion (SD) for female speakers using split-VQ with M_a=2048 and M_f=8 (a) and STFT with a codebook size of 2048 (b)

model is to capture speaker specific characteristics, and the likelihood estimator is to find the optimal codevectors (each selected from one speaker codebook) that best fit the observed mixed signal. The reconstruction stage is to re-synthesize the separated output signals (for more details see Mowlaee *et al.* (2010a)).

In this study, we focus only on the speaker modeling stage followed by a signal reconstruction stage, and exclude the errors caused by the likelihood estimation stage. For the reconstruction stage, we used the overlap-and-add procedure (Quatieri, 2002). More specifically, the quantization results given in this study can be interpreted directly as the upper-bound performance obtained in SCSS (Ephraim, 1992; Mowlaee *et al.*, 2010a). Hence, the results reported in this work basically address the separation results when no estimation error is included, called ideal separation (Mowlaee *et al.*, 2010a), where the optimal indices are known a priori. This study enables us to determine how successfully the proposed split-VQ can model the underlying speakers. We report the quality of the synthesized separated signals for two scenarios, namely speaker-dependent and speaker-independent.

According to the above-mentioned validation results, we applied the following setup. The window length was set to 32 ms and a frame-shift of 10 ms was used. The number of sinusoids, M, was set to 50. Figs. 5a–5d depict the SD statistics and average amplitude distortion versus the model order of the split-VQ codebook. By selecting a codebook size of M_a=256, the speech quality for male speakers becomes higher than for female speakers (Fig. 5a). By increasing the amplitude codebook size, the PESQ scores achieve fixed values of 3 for male and 2.8 for female. The percentage of SD<2 dB is higher, while the outlier percentage becomes lower as M_a increases (Fig. 5b). Fig. 5c shows that, for both the male and female cases, the fraction of outliers (SD>4 dB) asymptotically reaches a small value of about 0.5%. Fig. 5d shows that by increasing the amplitude codebook size to M_a=2048, the average SD will monotonically decrease toward a fixed value.

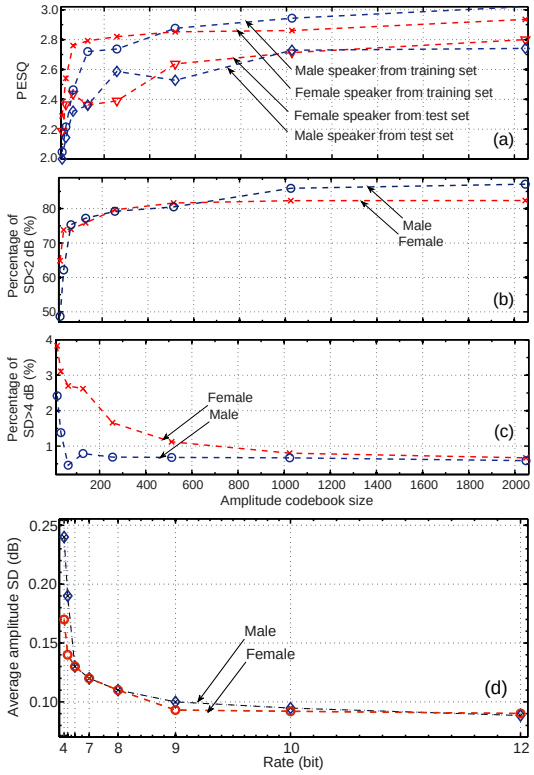

Fig. 5 Evaluating SD statistics for the proposed split-VQ for male and female speakers, showing the PESQ score for the synthesized speech output signals (a), percentage of outliers with SD>4 dB (b), percentage of outliers with SD<2 dB (c), and average amplitude distortion (d)

Table 7 summarizes the SD statistics for different quantization methods. These results are interpretable as separation upper-bound performance. The percentage of outliers using the STFT-based method is approximately 10 times larger than using LSF and split-VQ methods, and thus is unacceptable. This agrees well with the results recently

shown in Mowlaee *et al.* (2010a), confirming the inefficiency of the STFT features for being employed in SCSS as already used in Ellis and Weiss (2006). Furthermore, the average distortion obtained using the STFT-based method is high, which contradicts the transparent coding requirements reported in Paliwal and Atal (1993). In Table 7, the quantization results for split-VQ are also presented. The performance obtained by split-VQ consistently outperforms the STFT-based quantizer in terms of achieving a lower percentage of outliers compared to the STFT-based coder (Mowlaee and Sayadiyan, 2008).

Table 7 Results as separation upper-bounds for STFT and split-VQ for speaker-dependent scenarios

Method	Gender	Train/Test	LLR	PESQ
STFT	Male	Train	0.56	2.62
		Test	0.54	2.61
	Female	Train	0.49	2.86
		Test	0.51	2.72
Split-VQ	Male	Train	0.34	3.24
		Test	**0.44**	**3.19**
	Female	Train	0.47	3.21
		Test	**0.46**	**2.87**

For the speaker-dependent scenario, we chose four male and four female speakers: speakers 4, 23, 33, and 34 for the female case while speakers 9, 19, 26, and 32 for the male case. The quantization results were averaged over 15 test utterances for each gender. The selected test utterances were not used in the training stage. The results are shown in Table 8. Here we chose a codebook size of M_a=2048 and M_f=8.

For the speaker-independent scenario we provided two codebooks, one for male and one for female. We used several speakers to form the codebooks. The speakers labels are {4, 7, 8, 11, 15, 16, 21, 22, 23, 24} for the female scenario and {3, 5, 6, 9, 10, 12, 13, 14, 17, 19} for the male scenario. The quantization results are reported for male speaker 26 and female speaker 31 (Table 9), showing that the proposed split-VQ approach consistently achieves a higher PESQ score compared to the STFT case.

6 Discussion and future work

By integrating the advantages obtained by the proposed quantizer in this work, we believe that a broad range of model-based speech applications comes within reach. In the following, we summarize some of the pros and cons of the sinusoidal coder proposed in this work.

Table 8 Separation upper-bound performance for the speaker-dependent scenario for speakers 9 and 23*

Feature	Gender	Number of bits	SDOP (%) SD<2	SDOP (%) SD>4	PESQ
STFT	Female	7	69.4	9.8	2.9
	Male	7	44.9	20.3	2.2
	Female	8	71.1	9.8	2.8
	Male	8	46.7	19.4	2.3
	Female	10	79.5	6.2	2.9
	Male	10	62.4	8.6	2.4
	Female	11	81.1	5.8	2.9
	Male	11	63.9	7.4	2.6
Split-VQ	Female	10	82.3	0.8	2.9
	Male	10	86.9	0.4	2.7
	Female	11	82.4	0.7	2.9
	Male	11	88.9	0.3	3.1

* With the codebook size M_a=2048 and the number of frequency candidates M_f=8. SDOP: spectral distortion (in dB) occurrence percentage

Table 9 Separation upper-bound performance for speaker-independent scenarios for speakers 26 (male) and 31 (female)

Speaker No.	Feature	Number of bits	PESQ
26	Split-VQ	9	2.2
		10	2.3
		11	**2.6**
	STFT	9	2.1
		10	2.2
		11	2.2
31	Split-VQ	9	2.4
		10	2.4
		11	**2.5**
	STFT	9	2.3
		10	2.4
		11	2.4

The tree structure in split-VQ helps to lower the searching time required to find the optimal codevectors selected from speaker models in a model-based speech application. Additionally, the proposed sinusoidal coder cuts down the memory usage and the computational complexity compared to the STFT-based coder. In the Appendix, we show the efficiency of the split-VQ sinusoidal coder over the STFT-based one by calculating the reduction in memory usage and computational complexity. Therefore, the proposed coder suggests a solution to reducing both the computational complexity and searching time often

introduced as two difficulties in model-based applications. This is of high interest for the search engines to provide the best indices in many model-based speech applications including SCSS or speech enhancements. In general, this achievement enables the proposed method to reach a fast and structured VQ.

Through extensive simulations, it was observed that by applying the proposed sinusoidal coder, the resulting pdf of the quantization distortion becomes compact and more exponentially distributed. We noticed that the proposed coder works best for speech with low pitch period (female case), since lower distortion is introduced when the number of harmonics is small. Similar to the results in Chu (2004), for male speakers, the distortion of the proposed sinusoidal coder was proportionately larger due to the higher number of harmonic components.

To extract the sinusoidal parameters in the feature selection step of the proposed sinusoidal coder, we selected the highest peak per band. This decision making agrees with the concept of maximum likelihood estimation (MLE) for selecting frequency of a single sinusoid in noise per band at each frame, given the Gaussian noise statistics per band.

The distortion measures used in split-VQ agree well with recent findings in psychoacoustics, showing that the human auditory system can integrate distortions ranging within the auditory filters (Heusdens and van de Par, 2002). Similar normalization to either the maximum or RMS of amplitude has often been applied in sinusoidal coding aiming at improving the quantization performance (Heusdens *et al.*, 2007; Korten *et al.*, 2007). Korten *et al.* (2007) showed that the quantization distortion for frequency and phase are both proportional to squared amplitude.

It is important to remember that the SD measure in the split-VQ is different from the commonly used SD measure in the STFT domain. This agrees well with the result reported in Erkelens and Broersen (1996) stating that SD sometimes overlooks models in a codebook that are subjectively good but still have a high SD value. Note that the same codebook sizes were used here while comparing the performance between the split-VQ and the STFT one. However, the split-VQ on sinusoidal parameters employs M_{f} extra bits to encode frequencies. Therefore, the proposed quantizer is higher in bit-rate, and this aspect might be unfavorable in some speech coding tasks. However, the main idea of proposing the new split-VQ structure on the sinusoidal parameters was to incorporate the flexibility to improve the model-based speech separation upper-bound performance over the STFT counterpart. In these applications, the codebook size is not an issue since the inference estimation is offline (see Mowlaee *et al.* (2010a) and the references therein).

The distance measure defined in the two steps of split-VQ presents the spectral error localization properties. More specifically, the distortion measure in split-VQ is a WSED measure and emphasizes specific sinusoidal peaks located near the formant peaks. This follows the observations in psychoacoustics, since the peaks in the PSD of speech play a key role in perceiving audio signals. This weighting emphasizes sinusoidal peaks in strong regions of the PSD, resulting in consideration of the differences in sensitivity of the human ear. The weighting in split-VQ is chosen to provide finely quantized frequencies located at the vicinity of the spectral peaks of the sinusoidal amplitudes. In this regard, the dynamic weighting in Eq. (7) is close to the LSF properties used in speech coding. Through extensive simulations, we show that the proposed technique enables us to keep both the spectral distortion and the percentage of outliers low, both required to have a transparent coding (So and Paliwal, 2007).

As the proposed quantizer works independent of pitch estimates, it can offer an attractive candidate for those scenarios in which estimating the speaker pitch frequency is rather difficult. This is often the case especially when the desired speaker signal is corrupted with some other interfering sources, like another speaker or a noise source. As an example for solving this speech enhancement problem, Zavarehei *et al.* (2007) suggested using a weighted codebook mapping (WCBM) on the amplitude parameters of a harmonic plus noise model. They suggested using the WCBM as an effective tool for speech enhancement. The energy normalization in our distance measures in Eqs. (5)–(7) is similar to the idea suggested in WCBM. However, opposed to Zavarehei *et al.* (2007), the codebook we use is pitch-independent and the segmentation is not pitch-synchronous. Therefore, the overall upper-bound performance is not affected by the multi-pitch estimation errors.

7 Conclusions

We propose split-VQ based on sinusoids by applying a tree-structure on sinusoids composed of amplitude and frequency. The proposed coder brings several advantages: (1) perceptual consideration is considered by including the masking effect and mel-scale concepts; (2) spurious peaks existing in the STFT spectrum are avoided by selecting one peak per frequency band; (3) compared to the STFT scenario, the computational complexity is significantly reduced. These benefits are of high interest for the speaker models used in model-based speech applications, including SCSS and speech enhancement.

From extensive simulation results, we observed that the proposed sinusoidal coder improves the quantization performance and significantly improves the perceived speech quality measured by PESQ, SSNR scores, and SD statistics. As our primary goal, we assessed the effectiveness of the proposed sinusoidal coder as a speech coder. As a secondary goal, yet as an important application, we studied the performance of the proposed method when used in an ideal single-channel speech separation scenario. By ideal separation, we assume that the correct indices were known a priori and we address no estimation error. Therefore, the presented results are directly interpreted as the separation upper-bound (ideal separation). This method shows a significant improvement in the re-synthesized speech quality over the STFT features commonly used in SCSS.

References

Ahmadi, S., Spanias, A.S., 2001. Low bit-rate speech coding based on an improved sinusoidal model. *Speech Commun.*, **34**(4):369-390. [doi:10.1016/S0167-6393(00)00057-1]

Christensen, M.G., 2008. On perceptual distortion measures and parametric modeling. *J. Acoust. Soc. Am.*, **123**(5):3804. [doi:10.1121/1.2935505]

Chu, W.C., 2004. Vector quantization of harmonic magnitudes in speech coding applications—a survey and new technique. *EURASIP J. Adv. Signal Process.*, (17):2601-2613. [doi:10.1155/S1110865704407161]

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]

Cover, T.M., Thomas, J.A., 2006. Elements of Information Theory. John Wiley and Sons, New York.

Ellis, D.P.W., Weiss, R.J., 2006. Model-Based Monaural Source Separation Using a Vector-Quantized Phase-Vocoder Representation. ICASSP, p.957-960. [doi:10.1109/ICASSP.2006.1661436]

Ephraim, Y., 1992. Statistical model based speech enhancement systems. *Proc. IEEE*, **80**(10):1526-1555. [doi:10.1109/5.168664]

Erkelens, J.S., Broersen, P.M.T., 1996. Reconstruction error distortion measure for quantization of LPC models. *Electron. Lett.*, **32**(15):1347-1349. [doi:10.1049/el:19960930]

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]

Grenander, U., Szegö, G., 1984. Topelitz Forms and Their Applications (2nd Ed.). Chelsea Publishing Company, New York.

Heusdens, R., van de Par, S., 2002. Rate-Distortion Optimal Sinusoidal Modeling of Audio and Speech Using Psychoacoustical Matching Pursuits. ICASSP, **2**:1809-1812.

Heusdens, R., Kleijn, W.B., Ozerov, A., 2007. Entropy Constrained High Resolution Lattice Vector Quantization Using a Perceptually Relevant Distortion Measure. Proc. Asilomar.

Hu, Y., Loizou, P., 2007. Subjective comparison and evaluation of speech enhancement algorithms. *Speech Commun.*, **49**(7-8):588-601. [doi:10.1016/j.specom.2006.12.006]

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.

Korten, P., Jensen, J., Heusdens, R., 2007. High-resolution spherical quantization of sinusoidal parameters. *IEEE Trans. Audio, Speech Lang. Process.*, **15**(3):966-981. [doi:10.1109/TASL.2006.885929]

Kristijansson, T., Attias, H., Hershey, J., 2004. Single Microphone Source Separation Using High Resolution Signal Reconstruction. ICASSP, p.817-820. [doi:10.1109/ICASSP.2004.1326383]

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

McAulay, R.J., Quatieri, T.F., 1986. Speech analysis/synthesis based on a sinusoidal representation. *IEEE Trans. Acoust. Speech Signal Process.*, **34**(4):744-754. [doi:10.1109/TASSP.1986.1164910]

Moo, P., Neuhoff, D., 1998. Uniform Polar Quantization Revisited. Proc. IEEE Int. Symp. on Information Theory, p.100. [doi:10.1109/ISIT.1998.708687]

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

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

Mowlaee, P., Sayadiyan, A., Sheikhzadeh, H., 2009. FDMSM robust signal representation for speech mixtures and noise corrupted audio signals. *IEICE Electron. Expr.*, **6**(15):1077-1083. [doi:10.1587/elex.6.1077]

Mowlaee, P., Sayadiyan, A., Sheikhzadeh, H., 2010a. Evaluating single-channel speech separation performance in transform domain. *J. Zhejiang Univ.-Sci. C (Comput. & Elctron.)*, **11**(3):160-174. [doi:10.1631/jzus.C0910087]

Mowlaee, P., Christensen, M.G., Jensen, S.H., 2010b. Improved Single-Channel Speech Separation Using Sinusoidal Modeling. ICASSP, p.21-24. [doi:10.1109/ICASSP.2010.5496263]

Mowlaee, P., Christensen, M.G., Jensen, S.H., 2010c. Sinusoidal Masks for Single Channel Speech Separation. ICASSP, p.4262-4266. [doi:10.1109/ICASSP.2010.5495679]

Mowlaee, P., Saeidi, R., Tan, Z.H., Christensen, M.G., Fränti, P., Jensen, S.H., 2010d. Joint Single-Channel Speech Separation and Speaker Identification. ICASSP, p.4430-4433. [doi:10.1109/ICASSP.2010.5495619]

Paliwal, K.K., Atal, B.S., 1993. Efficient vector quantization of LPC parameters at 24 bits/frame. *IEEE Trans. Speech Audio Process.*, **1**(1):3-14. [doi:10.1109/89.221363]

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.

Quatieri, T.F., 2002. Discrete-Time Speech Signal Processing: Principles and Practice. Prentice-Hall, Upper Saddle River, NJ.

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.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]

Vafin, R., Kleijn, W., 2005. Entropy-constrained polar quantization and its application to audio coding. *IEEE Trans. Speech Audio Process.*, **13**(2):220-232. [doi:10.1109/TSA.2004.840942]

van Schijndel, N.H., Bensa, J., Christensen, M., Colomes, C., Edler, B., Heusdens, R., Jensen, J., Jensen, S.H., Kleijn, W.B., Kot, V., *et al.*, 2008. Adaptive RD optimized hybrid sound coding *J. Audio Eng. Soc.*, **56**(10):787-809.

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

Appendix

In this appendix, we aim to analytically calculate the reduction in the memory usage as well as the computational complexity of using the new sinusoidal coder compared to the STFT-based method. To this end we calculate the number of multiplications and additions per cycle. For the proposed coder, the mathematical operations are divided into two steps: amplitude part and frequency part. For the amplitude part, the number of operations to realize the amplitude distance measure in Eq. (5) is

$$O_{\mathrm{a}} = M_{\mathrm{a}} \cdot [3D_{\mathrm{a}}(\mathrm{mul}) + 2(D_{\mathrm{a}} - 1)(\mathrm{add})], \quad \text{(A1)}$$

where D_{a} indicates the dimensionality of the amplitude part, '(add)' is the number of additions and subtractions, and '(mul)' is the number of multiplications and divisions. The number of operations for the frequency part is

$$\begin{aligned} O_{\mathrm{f}} =& (D_{\mathrm{a}} - 1)(\mathrm{add}) + 2D_{\mathrm{f}}(\mathrm{mul}) \\ &+ M_{\mathrm{f}} \cdot [2D_{\mathrm{f}}(\mathrm{mul}) + (D_{\mathrm{f}} - 1)(\mathrm{add})]. \end{aligned} \quad \text{(A2)}$$

Hence, the overall operations for the proposed split-VQ will be $O_{\mathrm{split}} = O_{\mathrm{a}} + O_{\mathrm{f}}$. Choosing D_{DFT}=1024 DFT-point, for an STFT VQ we have

$$O_{\mathrm{STFT}} = M \cdot [2D_{\mathrm{DFT}}(\mathrm{mul}) + (D_{\mathrm{DFT}} - 1)](\mathrm{add}). \quad \text{(A3)}$$

By considering the symmetric property, the dimensionality of the DFT vector is 512. Now we define $\eta_{\mathrm{opr}} = (1 - O_{\mathrm{split}}/O_{\mathrm{STFT}}) \times 100\%$ as the complexity reduction efficiency in the operations. By replacing the parameters we obtain $\eta_{\mathrm{opr}} = 85.31\%$. In terms of memory usage, the proposed coder requires $M_{\mathrm{a}}(M_{\mathrm{f}}+1)L$ while the STFT quantizer needs $M \cdot D_{\mathrm{DFT}}$; hence, the reduction in memory cost, η_{memo}, will be

$$\begin{aligned} \eta_{\mathrm{memo}} &= \left(1 - \frac{2048 \times (1+8) \times 50}{2048 \times 512}\right) \times 100\% \\ &= 87.89\%. \end{aligned} \quad \text{(A4)}$$

It is observed that, compared to the STFT case, the proposed approach addresses a significant time-saving in the decoding time. This is important for model-based speech applications, where the emission probabilities (required for inference estimation) are required to calculate the likelihood function during the separation stage (to find the optimal indices each selected from one speaker codebook). To study the effectiveness of the proposed coder in speeding up and lowering the computational cost, we conducted ideal SCSS and quantified the computational complexity of the proposed coder for ten 2-s signal. We observed that the STFT-VQ took on average 26.71 s while the proposed one required 5.55 s. Comparing the decoding time averaged over 50 utterances, we observed that it took 43.7 ms for the sinusoidal coder to decode while 8.53 s for the STFT-based coder. It is concluded that the proposed coder significantly lowers the decoding time, which plays a key role in real time speech coding applications. For training, it took 81.59 min for STFT versus 27.66 min for the sinusoidal coder.