Model-based Monaural Sound Separation by Split-VQ of Sinusoidal Parameters

P. Mowlaee Begzade Mahale and A. Sayadian

Department of Electrical Engineering , Amirkabir University of Technology

Tehran, Iran 15875-4413

P_Mowlaee@ieee.org, eea335@cic.aut.ac.ir

web: http://ele.aut.ac.ir/pejmanmowlaee

Abstract—In many speech separation and enhancement techniques, establishing a statistical model like a Vector Quantization (VQ) is a must to handle the so-called model-based approaches. It is also desirable to establish a trade-off between sparsity and accuracy in the quantizer. To do so, in this paper we present split-VQ for sinusoidal parameters. We observed that sinusoidal parameters including amplitudes and frequencies, are most capable to be used as our features for split-VQ since they can be easily mapped to a tree-like structure. We demonstrate that using such split-VQ along with fixed dimension sparse sinusoidal parameters can significantly result in better source model compared with common STFT feature vectors in terms of objective and subjective measures in model-based approaches like monaural sound separation.

Index Terms—Split-VQ, Sinusoidal parameters, Spectral Distortion, SSNR.

I. INTRODUCTION

Establishing a perfect source model has been introduced as a challenging and difficult topic for decades. For instance, consider the so-called single-channel sound separation (SCSS) problem in which a target speech signal is mixed with other interfering speaker signal recorded by a single channel. In such applications, separating mixture requires using the spectrum amplitude of the *short-time Fourier transform* (STFT) often selected as a primary feature. The objective for separation is to express the spectrum of the mixed signal in terms of the spectra of the underlying speaker signals. This is generally accomplished, by fitting some statistical model such as *Gaussian mixture models* (GMM) [6], or *vector quantization* (VQ) [5] which are commonly used in order to model the underlying speakers features in the training phase. Then in the test phase, two speaker models are combined to model the given mixed signal and the states that best match the mixed signal are decoded based on some criteria e.g., *minimum mean square error* (MMSE). However, each one of the statistical models mentioned above still present a significant amount of distortion, hence model-based approach fails to separate mixtures due to the non-optimality of the source model for each speaker. As a solution, in this paper we demonstrate that using split-VQ structure as our source model for each speaker and employing it along with sinusoidal features including amplitudes and frequencies for each speaker, improves the quality of source model in terms of sparsity as well as more acceptable objective/subjective measures.

The remainder of this paper is organized as follows. In Section 2, we review the Fixed Dimension Modified Sinusoidal Model (FDMSM) recently presented in [8] in order to reach at useful sparse sinusoidal representation for audio signal. Next, in Section 3, the idea of statistical source model with proposed split-VQ structure is introduced. We also derive a new distance measure which is found appropriate to split-VQ structure on sinusoidal parameters obtained from FDMSM. Experimental results are reported in Section 4. Section 5 concludes the paper.

II. SINUSOIDAL PARAMETERS AS FEATURE VECTORS

In this section, we review appropriate features for split-VQ to be introduced in Section 3.

A. Brief review on FDMSM model

Audio signals generally contain either purely periodic (harmonic) parts or non-periodic information which are related to the tonal regions in music signal or the impulsive events or "noise-like" processes occurring in unvoiced regions during in a speech signal, respectively [1]. As a result, a time window segment of the underlying observed audio signal can be accurately modeled as a weighted sum of L sinusoids,

$$\boldsymbol{x}(n) = \sum_{l=1}^{L} a_l \cos(2\pi f_l n + \phi_l) + \epsilon(n) \qquad (1)$$

where $n = 1, \ldots, N_s$ is the sample index, N_s frame length, the triple parameters including $\theta_l = (a_l, f_l, \phi_l)$, denote the amplitude, frequency, and phase of the *l-th* sinusoid, respectively. L is the number of sinusoidal components in the signal, and $\epsilon(n)$ is the observation noise modeled as a zero-mean, additive Gaussian noise sequence. In general, it is of interest to estimate the corresponding sinusoids parameters including frequencies f_l, amplitudes a_l, phases ϕ_l and of course the number of sinusoids, L. The number of sinusoids is of great importance to establish a tradeoff between accuracy versus efficiency. In another phrase, the model should not only hold the sparsity of features but also preserve signal quality as close as possible to the original input signal level. To do so, we recently proposed the *Fixed Dimension Modified Sinusoidal Model* (FDMSM) in [8]. We demonstrated that using only $33 < L < 40$ sinusoids are enough to perfectly reconstruct speech signals in that it results in insignificant difference
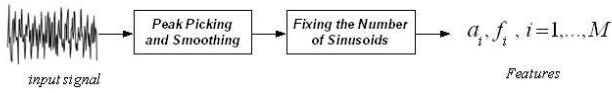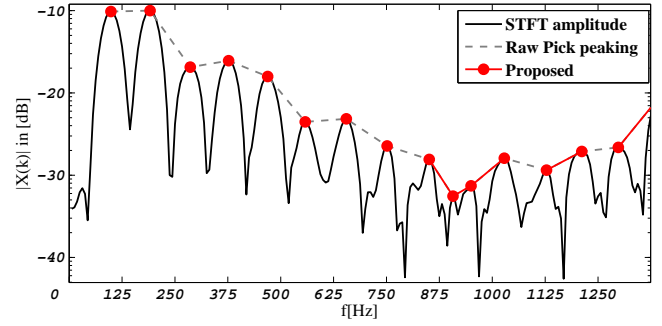
Fig. 1. FDMSM block diagram.

(a)

(b)

Fig. 2. STFT, raw peak picking and the FDMSM approach in Log-scale for (a) voiced and (b) unvoiced speech frame.

compared to the original signal in perception as indicated by listeners. In addition, presenting a model order based approach in [9], it was also shown both theoretically and experimentally that the best choice for L is $L \approx 33$ and $L \approx 20$ in the case of speech and music, respectively.

In this work, we employ sinusoidal parameters including amplitudes and their related frequencies in some Mel-bands obtained by FDMSM [8] as split-VQ features discussed in this paper. Using the fixed dimension features is in agreement with the fact that for a trained VQ using the k-means algorithm all elements in the vector are quantized jointly, and the dimension must be fixed and known a priori as in [12]. Fig.1 illustrates FDMSM model. In the following, we briefly explain how FDMSM results in most efficient peak candidates preserving the signal quality as close as possible to the original level.

Assume that $\mathbf{S}(k)$ be the spectrum of current speech frame. For peak picking process it is common to have, $8 < L < 80$ for $f_s = 8kHz$, however, for practical consideration, the choice of the number of sinusoids, is a tradeoff since larger value of L implies more computations. According to our previous results obtained in [8],[9], in this work we set L=33. On the other hand, window size and frame shift can also play a key role and affect the overall quantization performance. It has been demonstrated that the size of the window is critical since the peaks resolvability depends on the analysis window length, $N_a - 1$ [10]. However, the window must be long enough for individual harmonics to be resolvable. Hence, it is common to assume the max pitch period to be 16ms, and the analysis window, as a result is chosen $40(msec)$ in agreement with 2.5 times the averaged pitch period [10]. The time increment should also be small compared to the window size [11]. As a result, we opt for 8 msec hop size.

A parabolic interpolation is also used to arrive at a more accurate estimate of amplitude and frequency parameters. Furthermore, assuming sampling frequency $f_s = 8kHz$, the frequency components within $f_i < 62.5Hz$ and $f_i > 3840Hz$ are removed from the STFT spectrum amplitude prior to peak picking due to low (50 or 60Hz) and high frequency sensitivity harmful effects. The number of peaks obtained so far may differ from one segment to another. Hence, we also aim to fix the number of sinusoidal parameters while preserving the synthesized signal quality as close as possible to the original signal in terms of perception. This results in significant reduction in complexity. Moreover, handling or saving features with fixed dimension is of much more interest. Our proposed approach for fixing the number of sinusoids is described as follows. The frequency range of $[0, 4kHz]$ is converted into Mel-scale for perceptual purposes. A center frequency in addition to the related bandwidth is calculated
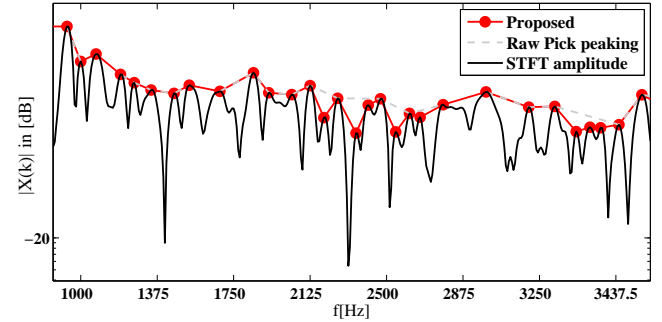
for each frequency range in that the number of sinusoids, M can be selected by the user with the following relationship,

$$M' = \lceil \frac{\Omega(\omega_{finish}) - \Omega(\omega_{start})}{step} \rceil \qquad (2)$$

where M' is the number of Mel-bands, $\Omega(\omega_{finish})$, $\Omega(\omega_{start})$ are Mel-scaled start and finish frequencies and step is the frequency interval, respectively. Note that user can arbitrarily change step in order to reach at different number of bands within a given frequency range. We opt for $\Delta f = 62.5Hz$ with $f_{start} = 62.5Hz$ and $f_{finish} = 3840Hz$ which results in $L = 33$ sinusoids. Next, we search in each band for peak candidates, and decide whether to pick or discard. Fig.2 compares the common STFT amplitude with raw peak picking, and the result obtained by using the FDMSM model discussed here. The x abscissa is set in the range 0 to 1.4kHz for voiced and 0.8 to 3.5kHz for un-voiced for improved visualization. As it is seen, considering the perceptual concepts related to Mel-scale along with the fixing process for the number of sinusoids discussed earlier, can result in choosing the most effective peaks sited in different bands and ignoring the low-level spectral portions that may be due to either noise or side-lobe analysis window effects. The number of sinusoids in each frame as a result is set fixed resulting in more compactness and lower feature dimension.

B. Synthesis at output stage

Once the quantizer finds the indices for the input test signal, the codebook vectors can be used to synthesis the output signal. The FDMSM output will be: $[a_i, f_i, \phi_i], i = 1, \ldots, M$, which can be used to reconstruct the kth synthetic contribution sequence as follows,

$$\hat{\mathbf{x}}(n) = \sum_{j=1}^{M} a_j^k \cos(2\pi f_j^k n + \phi_j^k) \qquad (3)$$

where $\wedge$ denotes estimated sinusoid parameters obtained by FDMSM. Finally $\hat{\mathbf{x}}(n)$, is overlapped and windowed by a hanning window. This accomplishes the synthesis process at split-VQ output used to produce separated signal. Note that by picking sinusoidal peaks within the subbands in (2), it will be enlightened in simulation results that employing such sinusoidal parameters along with the split-VQ structure will successfully improve the overall performance of source model.

III. SPEAKER MODELING BY SPLIT VQ

A. Overall split structure

Having N training vectors of fixed dimension D, our object is to find $M \ll N$ representative vectors. It is assumed that the number of training vectors, N are sufficiently larger than the codebook size, M. Eventually, $L \approx k \times M$ and $k > 10$ and here we consider $k = 30$ to establish a trade off between computational complexity and accuracy. These representative vectors also known as codevectors, are selected by minimizing a cost function such as the Euclidian distance [7].

In many cases, vector quantisers need to operate at higher bitrate and vectors of larger dimension, which both impact on the overall computational complexity and memory requirement in an exponential fashion. therefore, structurally VQs', such as split need to be employed [4]. Split-VQs were first applied to norrowband LPC parameter quantization in [3]. In the following, we present a new distance measure to be employed on the FDMSM amplitude vectors of the underlying speaker signals in the split-VQ structure.

As our aim in this work is to evaluate the performance of the proposed split-VQ on SSCS scenario, the overall segregation process is explained in what follows. For more info see [14]. Let $\mathbf{A}_i(k)$ and $\mathbf{B}_j(k)$ denote the arbitrary amplitude prototypes (codewords) of codebook related to first and second speaker, respectively. As our primary stage, a search is done through the codevectors $\mathbf{A}_i(k)$ and $\mathbf{B}_j(k)$ to find the optimal codevectors $\mathbf{A}_{opt}(k)$ and $\mathbf{B}_{opt}(k)$, that when mixed satisfy a minimum distortion criterion to find which indices i.e. i,j result in closest estimation of mix in terms of minimal spectral distortion. Note that in this work we assume that the estimator is ideal and we only report the quantization performance of the proposed split-VQ in terms of Segmental Signal to Noise Ratio (SSNR) and Segmental Signal to Distortion Ratio (SSDR) measures. Hence, the represented results can be interpreted as separation upper bound performance (Ideal VQ) as indicated in [2].

B. Deriving new Distance Measure

Several distance measures have been proposed for various speech applications including the so called Euclidean distance [7], cepstral and line spectral frequency (LSF) distance in speech coding proposed by Paliwal in [4]. However, up to now no distance measure has been employed to model-based SCSS to achieve at a better quantization of STFT feature vectors. As a result, in this work a new distance measure is presented based on some perceptual cues which is used in our proposed split-VQ. In the following we investigate the procedure to reach at a new distance measure for sinusoidal parameters in split-VQ.

Before any clustering we need to perform some pre-processing as: (1) Normalize the amplitude code-vectors to their maximum value. This will scale the parameter range to lie between [0,1]. In many clustering works, it has been demonstrated that employing features in such a region can improve the classification accuracy. (2) Take the logarithm for the resulting normalized amplitude vectors obtained from the previous stage. This will result in a reduction in code-vectors dynamic range. Note that due to the problems of logarithm scale for parameters with negligible values, a multiplication by a factor called α is used. In addition the multiplicand will be added with unity. This multiplication and addition will result in a scale transformation. For instance, using $\alpha = 1000$ results in a change in scale of the input code-vectors from [0,1] to [0,30] which is more acceptable while calculating the distance between two amplitude vectors. The pre-processes employed here can be summarized as follows,

$$Y(k) = \frac{|Y(k)|}{max(|Y(k)|)} \qquad (4)$$

$$\tilde{Y}(k) = \log\left(1 + \alpha Y_n(k)\right) \qquad (5)$$

where $k = 1, \ldots, N_{FFT}$ is the frequency bin and $100 < \alpha < 1000$ to ensure the log-amplitude range between (0,30). In addition, This pre-process results in neglecting non-significant peaks often occurring in a DFT-amplitude of a certain speech signal. It also ignores the low-level spectral portions that may be a result of noise or side-lobe effects. The normalization to max amplitude given in (5) results in a better performance in VQ process since no more bits are needed to model the vectors gain. To elaborate this, assume that we were to perform a VQ on raw spectrum amplitudes as in [2], then we had,

$$|Y(k)| = max(|Y(k)|)Y_n(k) \qquad (6)$$

$$\log|Y(k)| = \log|Y_n(k)| + \log v \qquad (7)$$

where $Y_n(k)$ is assumed to be the normalized version of the DFT-amplitude vector, and v is its gain. As can be seen an extra term now exists which constrains VQ to use more bits in order to model this dynamic range. The modified spectrum amplitude vector demonstrated in (4),(5) also results in choosing the most effective peaks sited in different frequency bands.

The tree structure for split-VQ is depicted in Fig.3. As our 1st stage in the proposed split-VQ, a VQ is performed on all amplitude parameters obtained by FDMSM explained

earlier. We employed different cluster sizes including $M_a = 1024, 2048, \ldots$. Assume that **TV** consists of amplitude and frequency parameters obtained by FDMSM. As a result, each training vector consists of two different components, namely $TV^a(k)$ and $TV^f(k)$. The distance between the original training vector and approximated vector (codevector) for amplitudes, $dist_w(\mathbf{TV}^a, \hat{\mathbf{CW}}^a)$ is defined by incorporating the Euclidean distance as,

$$d(\mathbf{TV}^a, \hat{\mathbf{CW}}^a) = \sqrt{\sum_{k=1}^{L_{mel}} (TV^a(k) - C\hat{W}^a(k))^2} \quad (8)$$

where $\mathbf{TV}^a$ denotes the amplitude part of the training vector. The number of sinusoids are selected as $L_{mel} = 33$ according to results obtained in [8],[9]. After establishing M_a amplitude reference vectors, we are to obtain the complement part of our split-VQ, i.e. appropriate frequency parameters related to each amplitude codewords. These frequency codewords should be selected such that the resulting joint codeword including amplitude and related frequency parameters represent source model as close as possible. To do so, we deemonstrate a VQ on each amplitude codeword obtained in previous stage. As depicted in Fig.3, a VQ with $M_f = 1, 2\ or\ 4$ is performed. The distance between the original training vector and approximated vector (codevector), $dist_w(\mathbf{TV}^f, \hat{\mathbf{CW}}^f)$ is,

$$dist_w(\mathbf{TV}^f, \hat{\mathbf{CW}}^f) = \sqrt{\sum_{k=1}^{L_{mel}} \mathbf{w}(k)(TV^f(k) - C\hat{W}^f(k))^2}$$

$$(9)$$

where $\mathbf{TV}^f$ denotes the frequency part and $\mathbf{w}$ is,

$$\mathbf{w} = \frac{\mathbf{TV}_i^a}{\sum_{k=1}^{L_{mel}} TV_i^a(k)} \quad (10)$$

dynamic weights. The weight in (10) is selected to finely quantize frequencies located in the vicinity of a spectral peak, resulting in less spectral distortion as stated in [4]. This completes the overall structure of split-VQ. Such distance measure is close to the proposed method which possesses spectral error localization properties. Similar to split-VQ used in [4], employing such structure and using some weighted distance measure, emphasizes specific sinusoidal peaks located near the formant peaks. This directly can lead to nearly transparent performance likewise *Switch Split-VQ*(SSVQ) [4] i.e. lower spectral distortions and percentages of outlier frames. The tree structured nature of split-VQ also provides for lower search complexity required in model-based scenarios.

IV. SIMULATION RESULTS

In this section the performance of our proposed split-VQ source model approach is evaluated. To do so, we conduct experiments to demonstrate the results of employing the split-VQ proposed in section 3 as a source model in order to evaluate the results for SCSS upper-bound Performance. Note that the simulation results presented here can also be considered as the upper bound for single-channel VQ-based separation scenario. The resulting codebooks indices are then used to produce the
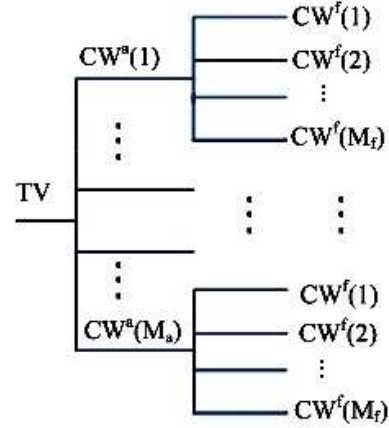

Fig. 3. Split tree structure used in the proposed split-VQ.

separated signals like Ideal-separation (Ideal VQ) reported in [2]. However, in this work, we only present the upper bound performance of SCSS scenario to determine how successfully can the proposed split-VQ model the speakers compared with conventional VQ on STFT.

To achieve reliable results we need a comprehensive database. Recently, Cooke et al. [13] have provided a new database for the performance evaluation of speech separation and enhancement systems. We use this database to conduct the experiments. The database consists of speech files of 34 speakers, each containing 500 utterances. The experiments are only performed for speaker dependent case. The sampling rate is decreased to 8 kHz from the original 25 kHz. Throughout all experiments, a Hamming window is used with a duration of 32 ms. It has been shown that too many outliers in the speech utterance having large SDs' can cause audible distortion even though the average SD is 1dB. Therefore, the more recent studies [3],[4] have tried to reduce the amount of outlier frames, in addition to the average SD. Basically, the outlier frames are divided into two groups including the frames with $2dB < SD < 4dB$ and with $SD > 4dB$ [4]. According to [4] the reasonable accuracy for transparent coding is attainable whenever the average SD is about 1 dB i.e. the coded speech is indistinguishable from original speech through listening tests. The conditions are: (1) The average SD is approximately 1 dB; (2) There is no outlier frame having more than 4 dB of SD; (3) Less than 2% of outlier frames are within 2-4 dB.

Table.1 summarizes the SSDR and SSNR results for different frame shifts. As can be seen from Table.1, the minimum outliers with distortion higher than 4 dB occurs for frame shift of 10 ms. The minimum average distortion among different frame shifts is also related to 10 ms frame shift. It also results in higher performance in terms of SNR measures (SSNR and SSDR). Table.2 illustrates the result of using different α in (5). It is observed that $\alpha = 1000$ results in the best performance at point of both minimum outliers and average distortion. In addition, comparing split-VQ with conventional VQ on STFT in Table. 2, we observe that the former method results in at

TABLE I
SPECTRAL DISTORTION (SD) VS. FRAME SHIFT.

shift (ms)	SSDR	SSNR	Avg	$2-4$dB	> 4	$SD < 2$
16	11.81	8.43	1.02	10.8	2.5	86.7
10	19.81	14.79	0.68	5.4	0.6	94.1
8	14.25	10.03	0.93	9.4	1.8	88.8

TABLE II
SPECTRAL DISTORTION (SD) VS. DIFFERENT FEATURE TYPES.

Feature type	Avg	$2 < SD < 4$dB	$SD > 4$	$SD < 2$
$a = 1000$	0.4	12.16	0.55	87.29
$a = 1$	1	14.81	1.17	84.02
$a = 0$	1.1	15.75	1.52	82.73
$STFT$	1.67	19.91	9.64	70.45

Fig. 4.   Freq trajectories for female speaker using $M_a = 1024$.

Fig. 5.   Histogram vs. SD for split-VQ with $M_a = 1024$ and $M_f = 2$.

least 1.2 dB lower average SD. In addition, the percentage of outliers in STFT-based method is approximately 10 times greater than split-VQ which is unacceptable.

After obtaining amplitude code-vectors, we incorporate split-VQ on training vectors to extract candidate frequencies for some amplitude code-vector. We observed that frequency candidates are highly similar to each other in each mel-scale frequency bands as given in (2). Fig.4 shows the frequency trajectories obtained for candidate frequency code-vectors corresponding to each amplitude code-vector. In this figure, y-axis denotes $L_{mel} = 33$ frequency bands. However, due to assigning higher bandwidths to higher frequencies, the similarity is lower in contrast to lower frequencies. Fig.5 shows the histogram of distortion vs. SD in dB. We observe that the distortion is nearly exponentially distributed (sparse). In addition, as our subjective measures we conducted a *Mean opinion Score* (MOS) informal listening test to measure the perceived quality of the reconstructed signals. Ten people were asked to give score between 0-5 to the reconstructed utterances (5 represents original utterances score). Averaging the listeners' scores resulted in 3.4 out of 5.

V. CONCLUSION

We proposed split-VQ on sinusoidal parameters as a high quality quantizer compared with conventional VQ commonly performed on STFT feature vectors used in model-based applications including speech separation and enhancement. A new distance measure was also derived found suitable for quantization of sinusoidal parameters in the proposed split-VQ. According to simulation results, it was demonstrated that the proposed split-VQ can significantly result in a better source model in terms of both objective and subjective measures.

REFERENCES
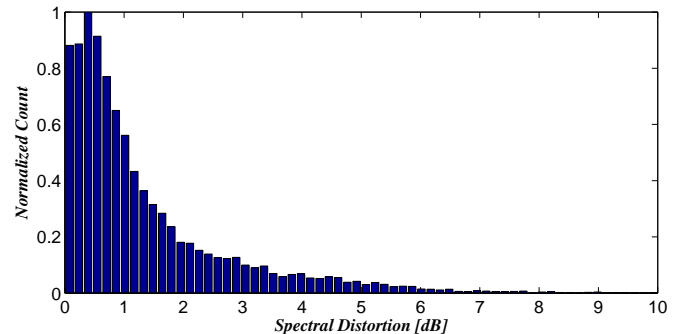
[1] D. O'Shaughnessy, "Speech Communications Human and Machine," IEEE press, New York, 2000.
[2] M. H. Radfar, R. M. Dansereau, and A. Sayadiyan, "A maximum likelihood estimation of vocal-tract-related filter characteristics for single channel speech separation," EURASIP J. Audio, Speech, Music Process., vol. 2007, doi:10.1155/2007/84186, article ID 84186, 15 pages.
[3] K. K. Paliwal and B. S. Atal, "Efficient vector quantization of LPC parameters at 24 bits/frame," IEEE Trans. Speech and Audio Processing, vol. 1, pp. 3-14, Jan. 1993.
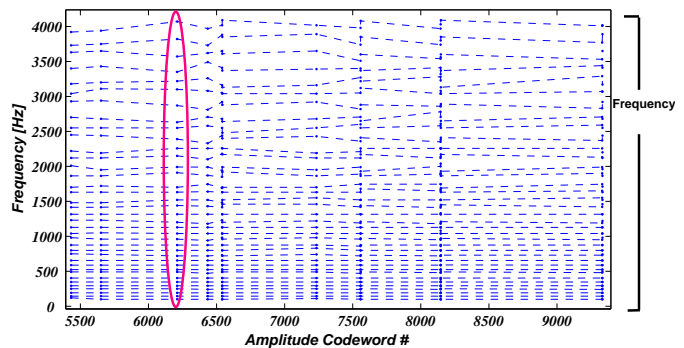[4] S. So and K. K. Paliwal, *A comparative study of LPC parameter representations and quantisation schemes for wideband speech coding*, Elsevier Digital Signal Processing, vol.17, pp.114137, (2007). doi:10.1016/j.dsp.2005.10.002
[5] D. Ellis and R. Weiss, "Model-based monaural source separation using a vector-quantized phase-vocoder representation," ICASSP06, Vol. V, pp. 957-960, May, 2006.
[6] A. M. Reddy and B. Raj, "A minimum mean squared error estimator for single channel speaker separation," in INTERSPEECH-2004, pp. 2445-2448, Oct. 2004.
[7] Y. Linde, A. Buzo, R.M. Gray, "An algorithm for vector quantizer design", IEEE Trans. Commun. COM-28, pp. 84-95, 1980.
[8] P. Mowlaee and A. Sayadian, *A Fixed Dimension Modified Sinusoid Model (FDMSM) for Single Microphone Sound Separation*, International Conference on Signal Processing and Communications (ICSPC), Dubai, U.A.E., pp.1183-1186, Nov., 2007.
[9] P. Mowlaee and A. Sayadiyan, *A Model order based Sinusoid representation for Audio signals*, 6th ACS/IEEE International Conference on Computer Systems and Applications, Qatar, pp. 501-507, May, 2008.
[10] E. B. George and M. J. T. Smith, "Speech analysis/synthesis and modification using an analysis-by-synthesis/overlap-add sinusoidal model,"Speech and Audio Processing, IEEE Trans. Speech and audio processing, Vol.5, Iss.5, pp. 389-406, 1997.
[11] T. Virtanen and A. Klapuri, "Separation of harmonic sound sources using sinusoidal modeling,"ICASSP-2000, Jun., pp.765-768. 2000.
[12] M. H. Larsen, M. G. Christensen, and S. H. Jensen, *Variable dimension trellis-coded quantization of sinusoidal parameters*, IEEE Signal Processing Letters, Vol. 15, Issue, pp. 17-20, 2008.
[13] M. P. Cooke, J. Barker, S. P. Cunningham, X. Shao, *An audiovisual corpus for speech perception and automatic speech recognition,* JASA 120, 24212424, 2006.
[14] P. Mowlaee and A. Sayadiyan *New Distance Measure for Monaural Model-based Sound Separation*, 3rd International Conference on Information and Communication Technologies (ICTTA), April, 2008 (to appear).