

Sparse Sinusoidal Signal Representation for Speech and Music Signals

Pejman Mowlaei Begzade Mahale, Amirhossein Froghani,
and Abolghasem Sayadiyan

Electrical Engineering Department
Amirkabir University of Technology, Tehran, Iran
pejman_mowlaei@ieee.org, pouyafroghani@yahoo.com,
eea335@cic.aut.ac.ir

Abstract. We present a sparse representation called Fixed Dimension Modified Sinusoid Model (FD-MSM) for parametric analysis of audible signals including speech, music and mixtures. Compared with other analysis models, the proposed scheme is both pitch independent and appropriate for sparse signal representation commonly found as a favorable choice for speech enhancement and sound separation. Using the state-of-the-art Principle Component Analysis (PCA) it is demonstrated that FD-MSM signal representation is equivalent to a non-linear mapping into sinusoidal subspace which preserves those components with largest eigenvalues by projecting the signal components into the corresponding eigen-vectors. Conducting subjective experiments, we observed that the resulting signal is perceptually indistinguishable from the original ones.

Keywords: Sinusoidal subspace, STFT, Principle Component Analysis, Sparse Representation, SSNR.

1 Introduction

One of the promising methods in speech processing is proven to be sinusoid model since many musical instruments produce harmonic or nearly harmonic signals with relatively slowly varying sinusoidal partials. On the other hand, Frequency analysis is, roughly speaking, the process of decomposing a signal into frequency components, that is, complex exponential signals or sinusoidal signals. Hence, in many applications finding a sparse representation for the audio signals is a must since it can easily result in lower computational complexity or selecting better features. As a result, sinusoidal modelling offers a parametric representation of audible signal components such that the original signal can be recovered by synthesis and addition of the components [1-2]. The most prominent classic sinusoidal model includes: (1) McAulay and Quatieri [3] and (2) Smith and George [4].

In model presented by McAulay and Quatieri called *Sinusoidal Transformation System* (STS) [3], all peaks from *Short-time Fourier Transform* (STFT) in the spectrum are marked, then the peaks whose occurrences are close to the pitch values and its harmonics are held while the remaining peaks are discarded. However, for mixed

audio signals, it fails to act properly due to its pitch dependency. However, the other sinusoidal model introduced by Smith and George, namely *Analysis-by-Synthesis/Overlap-Add* (ABS/OLA), has proven to be successful [4]. In contrast to STS proposed in [3], it uses a successive approximation-based analysis by synthesis procedure to determine model parameters. However, the computational load for estimating sinusoid parameters remains an obstacle due to exhaustive frequency search. It also requires initial pitch frequency estimation which is susceptible to gross errors for multi-speaker and speech + noise signals [4].

None of the above mentioned sinusoidal approaches are capable to extract fixed number of features for the underlying model. This drawback in turn results in a significant degradation in clustering performance since all model-based speech processing techniques employ statistical modelling techniques. Hence, we have recently proposed a modified version of sinusoidal model called *Fixed Dimension Modified Sinusoid Model* (FD-MSM) in [7], based on model proposed by McAulay and Quatieri [3], including inputs other than speech signals, i.e. music or mixtures. In this paper we study the sparse representation nature of FD-MSM which arrives at a lower dimension while preserving natural signal quality as close as possible.

The paper is organized as follows: The following section summarizes the state-of-the-art sinusoidal models. Section 3 is dedicated to the important concept of sparse representation of audio signals in sinusoidal space. In Section 4, objective and subjective results are reported and Section 5 concludes

2 Sinusoidal Model for Speech

Independent of which approach is used for analysis of speech signals, the spectrum envelope is known as a key feature, generally obtained by method proposed by McAulay [3]. Sinusoidal model represents a sound signal as a set of sinusoids parameterized by amplitude, frequency and phase trajectories carried out over short frames assuming short-time stationarity; under this assumption, frames of speech are modeled as a sum of constant-amplitude and frequency sinusoids [3]. Assuming the analysis frame $\approx 5\text{-}40\text{ms}$, the speech segment of frame k , $\mathbf{s}^k(n)$ will be:

$$s^k(n) = \sum_{j=0}^M A_j^k \cos(2\pi f_j^k n / f_s + \varphi_j^k), \quad n = 0, \dots, 2 \times \text{fl} \quad (1)$$

where f_j is the frequency, M is the number of sinusoids, A_j is the spectrum envelope, φ_j the phase sampled at k^{th} frame, j the index number, and f_s the sampling frequency. The drawback for such representation is its inherent high computational complexity due to storage of $3M$ values of spectrum parameters sampled at each frame.

3 Sparse Representation for Signals in Sinusoidal Space

If we consider the spectrum of a harmonic process, we note that it consists of a set of impulses with a constant background level at the power of the white noise. As a result, the power spectrum of complex exponentials is commonly referred to as a *line spectrum* and such signal can be expressed as:

$$x(n) = \sum_{l=1}^L e^{j2\pi n f_l} + w(n) = s(n) + w(n) \quad (2)$$

where $\mathbf{w}(n)=[w(n) w(n+1) \cdots w(n+L-1)]^T$ is the windowed vector of white noise and

$$\mathbf{V}(f_i) = [1, e^{j2\pi f_i}, \dots, e^{j2\pi f_i(N-1)}]^T, \quad 1 \leq i \leq L \quad (3)$$

is the time-window frequency vector. Note that $\mathbf{v}(f_i)$ is simply a length- N DFT vector at frequency f . We differentiate here between $\mathbf{s}(n)$, consisting the sum of complex exponentials, and the noise component $\mathbf{w}(n)$, respectively. The autocorrelation matrix of the model can be written as

$$\mathbf{R}_x = E\{\mathbf{x}(n)\mathbf{x}^H(n)\} = \mathbf{R}_s + \mathbf{R}_w \quad (4)$$

$$\mathbf{R}_x = \sum_{l=1}^L |\alpha_l|^2 \mathbf{v}(f_l)\mathbf{v}^H(f_l) + \sigma_w^2 \mathbf{I} = \mathbf{V}\mathbf{S}\mathbf{V}^H + \sigma_w^2 \mathbf{I} \quad (5)$$

where:

$$\mathbf{V}(f) = [1, \mathbf{v}(f_1), \dots, \mathbf{v}(f_{L-1})]^T \quad (6)$$

is an $N \times L$ matrix whose columns are the time-window frequency vectors from (3) at frequencies f_i with $i=0, \dots, L$ of the complex exponentials and

$$\mathbf{S} = \begin{bmatrix} |\alpha_1|^2 & 0 & \cdots & 0 \\ 0 & |\alpha_2|^2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & |\alpha_L|^2 \end{bmatrix} \quad (7)$$

is a diagonal matrix of powers for each respective exponentials and autocorrelation matrix of the white noise will be:

$$\mathbf{R}_w = \sigma_w^2 \mathbf{I} \quad (8)$$

which is full rank, as opposed to \mathbf{R}_s which was rank-deficient for $L < N$. In general, we will always choose the time window length, N to be greater than the number of complex exponentials L . The autocorrelation matrix in terms Eigen decomposition is

$$\mathbf{R}_x = \sum_{m=1}^M \lambda_m \mathbf{q}_m \mathbf{q}_m^H = \mathbf{Q}\mathbf{D}\mathbf{Q}^H \quad (9)$$

where λ_m are the eigenvalues in descending order, that is, $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_M$ and \mathbf{q}_m are their corresponding eigenvectors. Here \mathbf{D} is a diagonal matrix made up of the eigenvalues, and columns of \mathbf{Q} are the related eigenvectors. The signal related eigenvalues can be written as the sum of the signal power and noise as follows:

$$\lambda_l = L|\alpha_l|^2 + \sigma_w^2 \quad \text{for } l \leq L \tag{10}$$

and the remaining eigenvalues are due to the noise only components, are

$$\lambda_l = \sigma_w^2 \quad \text{for } l > L \tag{11}$$

therefore, the L largest eigenvalues correspond to signal made up of exponentials and the remaining eigenvalues have equal value and correspond to the noise. Thus, we can partition the correlation matrix into portions due to the signal and noise eigenvectors

$$\begin{aligned} \mathbf{R}_x &= \sum_{l=1}^L (M|\alpha_l|^2 + \sigma_w^2) \mathbf{q}_l \mathbf{q}_l^H + \sum_{l=L+1}^N \sigma_w^2 \mathbf{q}_l \mathbf{q}_l^H \\ &= \mathbf{V} \mathbf{S} \mathbf{V}^H + \sigma_w^2 \mathbf{I} = \mathbf{Q}_s \mathbf{D} \mathbf{Q}_s^H + \sigma_w^2 \mathbf{Q}_w \mathbf{Q}_w^H \end{aligned} \tag{12}$$

where:

$$\mathbf{Q}_s = [\mathbf{q}_1, \mathbf{q}_2, \dots, \mathbf{q}_L], \mathbf{Q}_w = [\mathbf{q}_{L+1}, \mathbf{q}_2, \dots, \mathbf{q}_N] \tag{13}$$

are matrices whose columns consist of the signal and noise eigenvectors, respectively. The matrix \mathbf{D} is $L \times L$ diagonal matrix containing the signal eigenvalues from (10). Thus, the N -dimensional subspace that contains the observations of the time-window signal vector from (6) can be split into subspaces spanned by the signal and noise eigenvectors, respectively. These two subspaces, known as the *signal subspace* and the *noise subspace*, are orthogonal to each other. Recall that the projection matrix from an N -dimensional space onto an L -dimensional subspace ($L < N$) spanned by vectors $\mathbf{Z} = [\mathbf{z}_1 \mathbf{z}_2 \dots \mathbf{z}_L]$ is

$$\mathbf{P} = \mathbf{Z}(\mathbf{Z}^H \mathbf{Z})^{-1} \mathbf{Z} \tag{14}$$

hence, the matrices that project a vector onto the signal and noise subspaces are as:

$$\mathbf{P}_s = \mathbf{Q}_s \mathbf{Q}_s^H, \mathbf{P}_w = \mathbf{Q}_w \mathbf{Q}_w^H \tag{15}$$

since the eigenvectors of the correlation matrix are orthonormal then we have:

$$\mathbf{Q}_s \mathbf{Q}_s^H = \mathbf{I}, \mathbf{Q}_w \mathbf{Q}_w^H = \mathbf{I} \tag{16}$$

since the two subspaces are orthogonal, then all the time-window frequency vectors from (3) must lie completely in the signal subspace, that is,

$$\mathbf{P}_s \mathbf{v}(f_i) = \mathbf{v}(f_i), \mathbf{P}_w \mathbf{v}(f_i) = \mathbf{0} \quad \text{for } 1 \leq i \leq L \tag{17}$$

However, in practice, the correlation matrix is not known and must be estimated from the measured data samples. If we have a time-window signal vector from (6), then we can form the data matrix by stacking the rows with measurements of the time-window data vector at a time n as follows:

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}^T(0) \\ \mathbf{x}^T(1) \\ \vdots \\ \mathbf{x}^T(n) \\ \vdots \\ \mathbf{x}^T(N-2) \\ \mathbf{x}^T(N-1) \end{bmatrix} = \begin{bmatrix} \mathbf{x}(0) & \mathbf{x}(1) & \cdots & \mathbf{x}(L-1) \\ \mathbf{x}(1) & \mathbf{x}(2) & \cdots & \mathbf{x}(L) \\ \vdots & \vdots & \vdots & \vdots \\ \mathbf{x}(n) & \mathbf{x}(n+1) & \cdots & \mathbf{x}(n+L-1) \\ \vdots & \vdots & \vdots & \vdots \\ \mathbf{x}(N-2) & \mathbf{x}(N-1) & \cdots & \mathbf{x}(N+L-3) \\ \mathbf{x}(N-1) & \mathbf{x}(N) & \cdots & \mathbf{x}(N+L-2) \end{bmatrix} \quad (18)$$

which has dimensions of $N \times L$, where N is the number of data records or frames and L is the time-window length. From this matrix, we form an estimate of the correlation matrix, referred to as the sample correlation matrix

$$\hat{\mathbf{R}}_x = \frac{1}{N} \mathbf{X}^H \mathbf{X} \quad (19)$$

and the spectrum of ordered eigenvalues in (19), the “signal eigenvalues” are still identified as the largest ones. Conducting several computer simulations in the following section, we demonstrate that the proposed sparse sinusoidal model (FD-MSM) is a useful analysis model to reduce the feature dimensions while preserving the audio signal quality as close as possible to the original input.

4 Simulation Results

Signal representation can be considered as a mapping from the N -dimensional space to a lower-dimensional feature space say L -dimensional space. Assuming a 1024-point FFT, and considering the symmetric property of FFT and removing the repeated part, we demonstrate that the target space i.e. sinusoidal space will have only a dimension varying between $20 < L < 40$. The only constraint imposed to such sparse representation, is that it should preserve the input signal quality as close as possible.

4.1 PCA and Redundancy of Audio Signals

To show the redundancy of the FFT representation for audio signals, *Principle Component Analysis* (PCA) is performed on both speech and music. The procedure is treated as follows. The window length is set to 32 msec and 100 speech frames were ensemble averaged with a frame shift of 1 msec for stationarity assumption. Fig.1.a,b depict the eigen-values and eigen-vectors for several eigen-values, respectively. The number indicated in right-up section of each plot is related to corresponding eigen-value denoted by c_i where i is the eigen index.

As it is seen from Fig.1.a, the Eigen vectors obtained from *Singular Value Decomposition* (SVD) of frame covariance ensamples averaged matrix \mathbf{R} determined in (19), the variance related to first few vectors are much notable. In contrast, ignoring vectors over $i=33$ results in an insignificant error (about 0.1%). In addition, Fig.4.b demonstrates the Eigen spectrum obtained from SVD decomposition in time-domain and STFT domain, respectively. As it is seen, the Eigen spectrum decreases monotonically in both cases very rapidly in that after about 30 index, the Eigen power attenuates to about -50 dB (=0.001%). In a similar manner for detecting correct model order

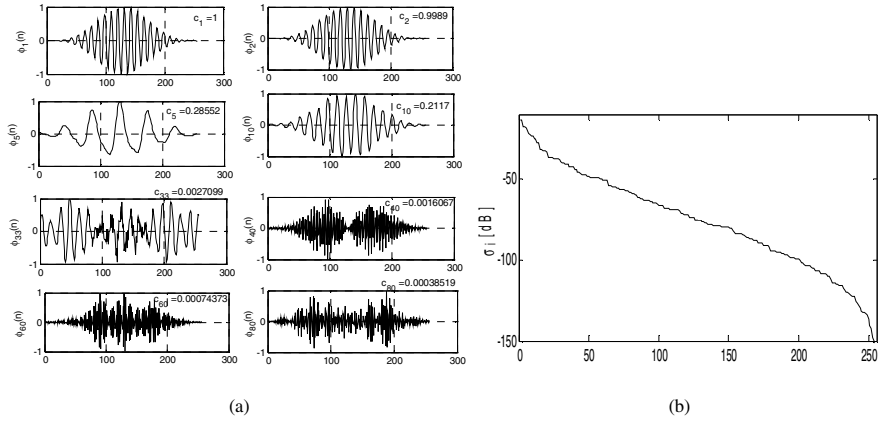


Fig. 1. (a) Eigenvctrs as time domain basis and (b) eigen-values for a male speaker speech

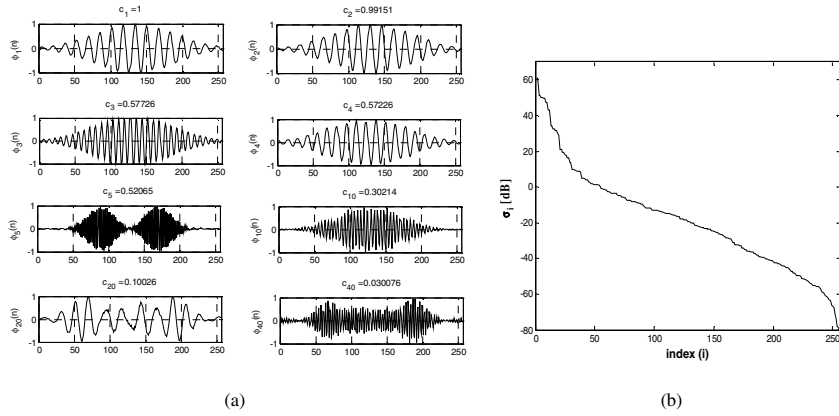


Fig. 2. (a) Eigenvectors as time domain basis functions and (b) eigen-values for a music signal

addressed, here by means of looking directly for a gap between the noise and the signal eigenvalues we observe the appropriate number of sinusoids. As a result, the FFT redundancy is obvious which in turn results in an imperfect signal representation especially for audio signals. Hence, translating speech signal to another domain with less dimension with respect to common STFT, we can expect a more compact and efficient representation which could be accomplished by sinusoidal signal representation discussed in this paper.

Simulation result for music are shown in Fig.2.a,b. Eigen-vectors are more sinusoid like than speech case. This inter correlation among eigen-vectors in time domain motivates us to employ a more compact representation. In addition, the steeply decrease in eigen-spectrum in Fig.2.b, verifies that 15-20 eigenvalues are considerable and the rest could be ignored without perceivable performance degradation.

4.2 Determining Number of Sinusoids Based on Eigen Decomposition

An experiment is conducted to confirm that the FD-MSM approach is in fact a sparse representation for audio signals. To proceed, a voiced frame is selected. Then sinusoidal analysis is performed for different number of sinusoids. Next, PCA is employed to obtain the fundamental eigenvalues. As shown in Fig.3.a and b, the number of prominent eigen-values related to signal subspace is $31 < M' < 35$ in harmony with frame reconstruction in Fig.3.a.

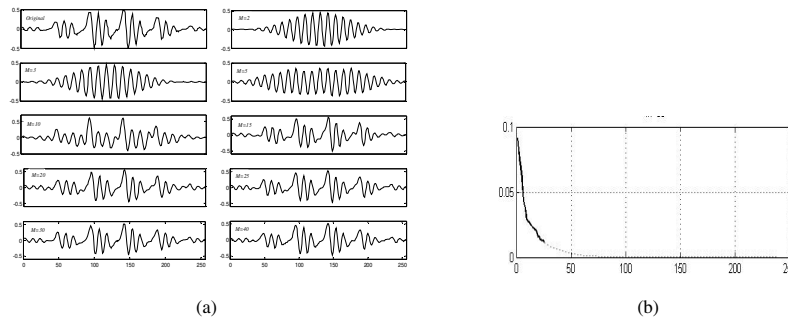


Fig. 3. (a) Reconstruction a speech frame using different number of sinusoids (b) Eigen spectrum, choosing largest singular values for reconstruction (the dashed line) and the rest (in dotted line)

4.3 Subjective and Objective Results

Spectrograms and time signals are illustrated in Fig.4 both for the original input and the synthesized output signal. For subjective results, Mean Opinion Score (MOS) test [5] is conducted to measure the perceived quality for 8^{kHz} speech for 20 male and female speakers and their mixtures. 20 listeners were asked to score between 0-5 to reconstructed utterances. The MOS results are presented for different groups of listeners vs. the number of sinusoids, $M' \in [21,40]$ in Table.1. Waves can be downloaded at: <http://ele.aut.ac.ir/pejmanmowlae>. It is observed that the proposed FD-MSM requires $25 < M' < 35$ to have negligible difference in MOS. Using $M'=33$ parameters are enough to establish trade-off between low dimensionality and high perceptual quality.

Comparing the subjective and objective results, we conclude that eigen-analysis results as shown in Fig.3.a,b propose using $31 < M' < 35$ sinusoids while MOS results indicate that a perfect reconstruction of speech signal is possible when $M' \approx 33$ which is in harmony with results in Fig.3.a. As a consequence, we opt for $M'=33$ to have an indistinguishable speech signal representation using the sinusoidal modelling. In addition evaluating FD-MSM for music signals, no distinguishable difference was inferred by listeners even for $M' \approx 14$. Evaluating these results with the conventional sinusoidal model proposed in [3],[4] they all need 5 to 10 more sinusoidal parameters than FD-MSM. In addition they all suffer from gross error related to pitch estimation while for analyzing audio mixtures as reported in [6]. In this case these methods are unable to extract the pitch value of the weaker speaker if the pitch value is equal to or a multiple of the pitch value of the stronger speaker and hence performance degrades.

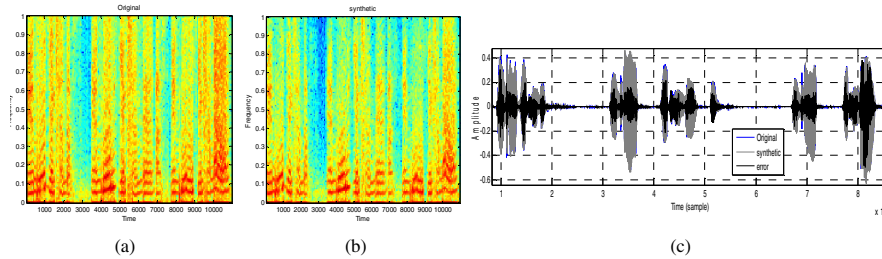


Fig. 4. Showing spectrograms for (a) original, (b) synthesized, (c) related time domain signals

Table 1. MOS results for synthesized speech signals

M'	Category	MOS	M'	Category	MOS	M'	Category	MOS	M'	Category	MOS	
25	Adult male	3.8	21	Adult male	3.6	33	Adult male	4.4	29	Adult male	4.1	
	Child	4		Child	3.85		Child	4.5		Child	4.3	
	Old Man	3.9		Old Man	3.7		Old Man	4.4		Old Man	4.1	
	Old woman	4.1		Old woman	4		Old woman	4.6		Old woman	4.5	
	Adult female	4.1		Adult female	4		Adult female	4.6		Adult female	4.5	
40	Adult male	4.6	37	Adult male	4.5							
	Child	4.75		Child	4.62							
	Old Man	4.65		Old Man	4.5							
	Old woman	4.8		Old woman	4.7							
	Adult female	4.8		Adult female	4.7							

5 Conclusion

In this paper, proposed FDMSM model was studied for sparse representation of audio signals. We demonstrated that the space dimension was significantly reduced and the model can be assumed as a mapping from STFT spectrum space to a more compact one say $25 < L < 40$ called sinusoidal subspace. This reduction is notable since no performance degradation in terms of perception and signal characteristic was. The choice of the number of sinusoids was confirmed using the state-of-the-art PCA.

References

- O’Shaughnessy, D.: Speech Communications Human and Machine. IEEE press, NY (2000)
- Macon, M.W., Clements, M.A.: Sinusoidal modeling and modification of unvoiced speech. IEEE Trans. Speech Audio Process 5(6), 557–560 (1997)
- McAulay, R.J., Quatieri, T.F.: Speech analysis/synthesis based on a sinusoidal representation. IEEE Trans. ASSP 34, 744–754 (1986)
- George, E.B., Smith, M.J.T.: Speech analysis/synthesis and modification using an analysis-by-synthesis/overlap-add sinusoidal model. Speech and Audio Processing, IEEE Trans. 5(5), 389–406 (1997)
- Furui, S., Sondhi, M.M.: Advances in Speech Signal Processing. Marcel Dekker Inc., New York (1992)
- Radfar, M.H., Dansereau, R.M., Sayadiyan, A.: Monaural speech segregation based on fusion of source-driven with model-driven techniques. Speech Comm. 49, 464–476 (2007)
- Mowlae, P., Sayadian, A.: A Fixed Dimension Modified Sinusoid Model (FD-MSM) for Single Microphone Sound Separation. In: International Conference on Signal Processing and Communications, ICSPC 2007, Dubai, United Arab Emirates, pp. 1183–1186 (November 2007)