A Model Order Based Sinusoid Representation for Audio Signals

Pejman Mowlaee Begzade Mahale and Abolghasem Sayadiyan
Amirkabir University of Technology,
Dept. of Electrical Eng, 15875-4413, Hafez, Tehran, Iran
Emails: pejman.mowlaee@ieee.org and eea335@aut.ac.ir

Abstract

This work investigates the problem of sinusoidal parameter estimation including amplitude, frequency and phase in all kinds of audio signals including speech, music and mixtures using a nonlinear least-squares (NLS) estimator where the number of sinusoids is in general unknown. A solution to the sinusoid order selection problem is presented by employing the NLS method along with some information criteria (IC) such as minimum description length (MDL) or AIC. The simulation results demonstrate that the presented approach is successful in determining the number of sinusoids used in sinusoidal representation by using the information provided by well-known ICs'. It is demonstrated that the original audio signal can successfully be reconstructed in terms of both objective (SSNR) and subjective results (MOS).

Keywords— MDL, NLS, AIC, IC, Model order selection.

1. Introduction

In many applications, the signals of interest are complex exponentials corrupted in noise for which a sinusoidal or harmonic model is known to be the most appropriate one. In addition, signals consisting of complex exponentials are often found in different applications such as formant frequencies in speech processing. Moreover, one of the promising methods in audio signal analysis is sinusoidal modeling since many musical instruments produce harmonic or nearly harmonic signals with relatively slowly varying sinusoidal partials [1]. Sinusoidal modeling 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]. It has also proved to be a favorable choice for analysis of all audible signals as reported in [3],[10]. Since the frequency estimation in sinusoid models is a nonlinear process, some estimation for frequencies are usually assumed a priori, and linear techniques can then be used to estimate the amplitude parameters. However, the accuracy of such estimates highly depends on this pre-assumption on frequencies which in turn could result in significant performance degradation. However, for estimators based on the minimization of the squared error, the difficulty lies in the multimodal shape of the error surface (or loss function which is defined as criteria). It is observed that for the sinusoid model, the loss function will have deep troughs at multiples of the true pitch period (the fundamental frequency which is assumed to be known a priori). The relative levels of the troughs depend on the number of sinusoids in the candidate signal as well as the amplitudes of the sinusoids. As a result, the mis-estimation of a pitch frequency as a candidate frequency results in large estimate variances which influence on estimation of other parameters including amplitudes and phase of the related sinusoids. This dependency of the loss function on the pitch and the number of sinusoids is a motivating factor for the work we present here. Hence the parametric or model-based approaches of signal processing are employed as vehicles for selection of model parameters that are equally important to specify sinusoidal representation in our work. In addition, the sinusoidal parameter estimation approach proposed in this work does not suffer from the gross errors and other difficulties related to pitch estimation.

Some previous works, assume the number of sinusoids is fixed or known a priori [4-8]. We develop an approach in which the amplitudes and phases of the sinusoids are estimated. In addition, several order selection strategies are considered to verify the correct number of sinusoids. Hence, a detailed combined detection-estimation algorithm for determining the sinusoidal parameters for speech signals is presented. Finally, the appropriate number of sinusoids is checked through both objective and subjective tests.

The remainder of this paper is organized as follows. In Section 2, the problem formulation is presented. In Section 3, we derive a new framework to estimate the sinusoidal model parameters. In Section 4, some information criteria are employed to determine the best choice for the number of sinusoids. Experimental results are reported and verified with the theoretical aspects in Section 5 and, finally, Section 6 concludes.

2. Problem Formulation

Consider the linear regression model as follows:

\[ Y = \mu(\theta) + \varepsilon \]  

(1)

where \( y \) is the vector of observed data, \( \theta=[\theta_1, \ldots, \theta_n]^T \) is the unknown coefficient vector and \( \epsilon(n)~N(0, \sigma^2) \) is a length \( N \) vector of zero-mean Gaussian white noise with variance \( E[\epsilon^2] = \Sigma \), and \( \epsilon \in L \) model order. A special case of (1), addressed here, is the sinusoidal signal model. To proceed, the sum-of-sinusoids plus noise model is presented. Since, speech signal contains both periodic and non-periodic parts due to the impulsive nature of events or “noise-like”
processes occurring in unvoiced, the observed signal. As a result, each speech frame can be written as weighted sum of several sinusoids plus noise with different parameters as follows:

\[ y(n) = \sum_{i=1}^{L} a_i \cos(2\pi f_i n + \phi_i) + \varepsilon(n) \] (2)

where \( n = 1, \ldots, N \) is the sample index, \( \{a_i, f_i, \phi_i\} \) denote the amplitude, frequency, and phase of the \( i \)th sinusoidal component; \( N \) is the number of speech frames; \( L \) is the number of sinusoidal components present in the signal, and \( \varepsilon(n) \) is the observation noise modeled as a zero-mean, additive Gaussian noise sequence. In general, it is of interest to estimate the corresponding sinusoidal parameters including frequencies \( f_i \), the amplitudes \( a_i \), phases \( \phi_i \), in addition to number of sinusoids, \( L \). However, the signal is a nonlinear function of the phase \( \phi_i \) which is the squared norm of the difference between the measurement and the signal model. Note that the constant term due to the noise variance is ignored since it does not impact the minimization of \( J(\theta) \). For the chosen model of \( \mathbf{S}(\theta) \), the estimation of the parameter vector \( \theta \) in (4) is a highly nonlinear process. However, the minimization of (7) can be achieved by the method of least-squares (LS) if \( f \) and \( L \) are known apriori. For a candidate \( f \) and a fixed \( L \), the Cartesian amplitude estimates are found by:

\[
\{a(i)\} = \arg\min_{a} \| y - \mathbf{C}(f)a \|^2
\] (10)

the LS solutions to Equation (9) can be obtained by:

\[
\hat{\theta} = (\mathbf{C}^T\mathbf{C})^{-1}\mathbf{C}^T y
\] (11)

where the dependence on the frequency, \( f \), has been suppressed to simplify the notation. Given \( a \), the amplitude estimates \( \{a_i\} \) and phase estimates \( \{\phi_i\} \) are found using Equation (11). Using the LS amplitude estimates in (9-10), the loss function for frequency estimates is given by:

\[
L(f) = \| y - \mathbf{C}a \|^2
\] (12)

\[
= \| y - \mathbf{C}(\mathbf{C}^T\mathbf{C})^{-1}\mathbf{C}^T y \|^2,
\] (13)

\[
= y^T P^\perp y
\] (14)

where \( P^\perp = \mathbf{I} - \mathbf{C}(\mathbf{C}^T\mathbf{C})^{-1}\mathbf{C}^T \) denotes noise subspace and projects the observation into the null space of \( \mathbf{C} \). Finally, the maximum likelihood (ML) estimation for frequencies can be obtained as:

\[
f = \arg\min_{f} \| L(f) \|^2
\] (15)

\[
= \arg\min_{f} y^T P^\perp y
\] (16)

this minimization procedure is also known as the nonlinear least-squares (NLS) method. In determining \( f \), the estimator attempts to minimize (maximize) the energy in the null space (column space) of \( \mathbf{C} \). Although there are 3L free parameters, once \( f \) has been computed, the ML Cartesian amplitude estimates are simply given by (10-11).
4. Model Order Selection

Several standard model order selection techniques are well suited to this task including Akaike’s information criterion (AIC) [6], and Rissanen’s minimum description length (MDL) [7]. They are derived from Information Theoretic Criterion (ITC). In the following we derive ICs’ for sinusoidal parameter estimation. To proceed, substituting $\varepsilon$ with its equivalent term $y - \mu(\theta)$, the pdf for error term in (2) can be calculated as follows:

$$p_\varepsilon(\varepsilon) = p(y, \theta) = \frac{1}{(2\pi)^{N/2} (\sigma^2)^{N/2} e^{-\frac{1}{2\sigma^2} \|\varepsilon - \mu(\theta)\|^2}}$$

(17)

$$\theta = \left[ \begin{array}{c} \alpha \\ \varphi \\ \sigma^2 \end{array} \right]^T$$

(18)

we deduce from (3) that

$$-2\ln p(y, \theta) = N \ln 2\pi + N \ln \sigma^2 + \frac{\|y - \mu(\theta)\|^2}{\sigma^2}$$

(19)

a simple calculation shows that the ML estimates of $\mu$ and $\sigma^2$ are given by

$$\hat{\theta} = \arg \min_\theta \|y - \mu(\theta)\|^2$$

(20)

$$\hat{\sigma} = \frac{1}{N} \|y - \mu(\theta)\|^2$$

(21)

the corresponding value of the likelihood function from (17) and the resulting AIC criterion will be

$$-2\ln p(y, \hat{\theta}) = cte + N_s \ln \hat{\sigma}^2$$

(22)

$$AIC = 2N_s \ln \hat{\sigma}^2 + 2(3L + 1)$$

(23)

for the sinusoidal signal model with $L$ components and $N_s$ speech frames an unbiased, consistent estimate of $\sigma^2$ and can be obtained by taking

$$\hat{\sigma}^2 = \frac{1}{N} \sum_{i=1}^{N} \left| y(t) - \sum_{i=1}^{L} \tilde{a}_i e^{j(2\pi f_i + \phi_i)} \right|^2$$

(24)

where $\{y(t)\}$ is the speech signal. As can be seen, in the present case the maximum likelihood method indeed reduces to the NLS. According to IC employed; each method has a similar form with a data term and a penalty term which accounts for the reduced fit error when the model order is overestimated. As a result, the order selection criteria for the sinusoid model in (2), will be:

$$\hat{L}_{AIC} = \arg \min_\theta \left\{ N \ln J(\hat{\theta}) + 3L \right\}$$

(25)

$$\hat{L}_{MDL} - \arg \min_\theta \left\{ N \ln J(\hat{\theta}) + \frac{3L}{2} \ln N \right\}$$

(26)

where $J(\theta)$ is the negative log-likelihood function evaluated at the ML parameter vector $\theta$ and $L$ is the estimate of the number of sinusoids. The number of free parameters is $3L$.

In the following section, conducting several simulations we demonstrate that appropriate number of sinusoids $L$ obtained by information criteria agrees with the significant eigen-values obtained by Principle Component Analysis (PCA) for audio signal.

5. Simulation Results

In this section, using model order criteria discussed in previous Section, we demonstrate that employing FFT signal representation introduce considerable redundancy which results in an inappropriate or useless signal space. This motivates us to use some other signal representation which is sparser in dimensions. To do so, evaluating different information criteria in Section 4, the actual number of signal eigenvectors are determined for both speech and music signals. This is also verified by eigenvalue decomposition and neglecting those eigenvectors whose eigenvalues are negligible. After determining the appropriate number of sinusoids, we will conduct a Mean Opinion Score (MOS) as subjective test to ensure how this choice affects the perceptual results of the synthetic signal.

5.1. Determining Error Distribution

Calculating the error signal from different frame analysis, the following histogram is obtained. Using a computer fitting tool, we observed that the error distribution is closely related to Normal distribution with parameters Mean, $\mu=0.002$ and Variance $\sigma^2=0.0072$.

![Fig.1. Error distribution and histogram fitting for ensemble averaged speech frames.](image)

The sinusoidal term and noise are considered statistically independent. The error encountered in sinusoidal modeling is demonstrated in Fig.1. As it is seen, the noise can be modeled as a normal distribution. As a result, a vector of noise samples distributed as $\varepsilon \sim N(0; \Sigma)$, the last term in right hand side of (2). Hence, $\varepsilon(n)$ is then given as:
e \sim N(0; \Sigma), \quad (27)
\theta_{\text{noise}} = \Sigma = \sigma^2 \mathbf{I}

If the Normal distribution parameters are not known, then they must be estimated. Assuming $\theta_{\text{noise}}$ is known. Given $\theta_{\text{noise}}$, the sinusoid model is given by
\[ \tilde{y} = C(f_k) \tilde{a} + \tilde{e} \quad (28) \]

where $\tilde{e}(n) \sim N(0, \sigma^2 \mathbf{I})$. After computing the estimates of $\tilde{a}$, the estimates of $C(f)$ are determined by using Equation (3). Consequently, without loss of generality, it is assumed $\Sigma = \sigma^2 \mathbf{I}$ in the following derivations.

5.2. KLT transform and sparsity of Audio Signals

To show the redundancy of the FFT representation for audio signals, KLT transform 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.2.a,b depict the eigen-values and eigen-vectors for several eigen-values, respectively.

As it is seen from Fig.2.a, the Eigen vectors obtained from Singular Value Decomposition (SVD) of frame covariance ensemble averaged matrix $\mathbf{R}$, the variance related to first few vectors are much notable. In contrast, ignoring vectors over i=33 results in an insignificant error (0.1%). In addition, Fig.2.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 addressed in [11], 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.

The simulation is repeated for music signal and the results are shown in Fig.3.a.b. As a result, the eigen-vectors are more sinusoid like signals than the speech case. This inter correlation between the eigen-vectors in time domain motivates us to employ a more compact sinusoidal representation. In addition, the steeply decrease in eigen-spectrum shown in Fig.3.b, verifies that only 15-20 eigen values are considerable and the rest could be ignored without perceivable performance degradation.
5.3. Determining number of sinusoids based on ICs’

As a result of redundant characteristics of FFT analysis, another useful representation seems promising. In this section, it is demonstrated how the number of sinusoids can be determined using ICs’. The order of sinusoidal signal representation is unknown a priori, hence model order selection methods can be employed as a solution. Fig.4,5 demonstrate the model order selection results and the related reconstructed signal for a voiced frame of a male speaker. Fig.4 depicts the log-likelihood vs. number of sinusoids (L).

\[
\text{SSNR} = \frac{1}{M} \sum_{j=0}^{M-1} \log_{10} \left( \frac{\sum_{n=1}^{N_s} s[n+N_s] - \hat{s}[n+N_s]}{\sum_{n=1}^{N_s} s[n+N_s]} \right) \quad (dB)
\]

where \( s(n) \) represents the original segmented signal while \( \hat{s}(n) \) each segment \( m \) is of length \( N_s \) samples. Typical values of \( N_s \) range between 100 to 200 samples (15-25 msec). As a result, here we select \( N_s=200 \). Fig.6. shows that using more than \( L=33 \) sinusoids can completely reconstruct the original signal and no SSNR improvement is detected.

As it is seen, AIC performs well partly because of its tendency to select models with relatively large orders. As a result, we tend to prefer BIC over AIC on the grounds that BIC is an asymptotic approximation of the optimal MAP rule. The model order is estimated \( L=27 \) using the sinusoidal representation for the speech signal. Using the sinusoidal representation for the speech signal, the resulting (a) original, (b) synthesized, spectrograms and (c) time domain signals are illustrated in Fig.7.

The simulation is repeated for music signal and the results are shown in Fig.8.a,b. Again, using more than \( L=30 \) sinusoids can completely reconstruct the original music signal as Fig.8.a,b indicate. In addition, the SSNR plot shown in Fig.9 confirms that this selection of \( L \) is optimum.

We also conducted an experiment to confirm that sparse representation for audio signal is more effective. To proceed, a voiced speech frame was selected. Then sinusoidal analysis was performed for different number of sinusoids. Next, the eigen-decomposition was performed to obtain the eigenvalues and the eigen-vectors. The results are shown in Fig.10.a,b. It is observed that the proposed sinusoidal approach requires \( 25<L<35 \) to have a negligible difference in terms of reconstruction error and the resulting SSNR (see Fig.6 and 9).
for speech and music, respectively). This is also verified by investigating information criteria illustrated in Fig. 4,8 for speech and music, respectively.

![Fig. 7](image1)

Fig. 7. Showing spectrograms for (a) original, (b) synthesized, (c) time signals.

![Fig. 8](image2)

Fig. 8. (a) log-likelihood vs. number of sinusoids (L), (b) music signal reconstruction with L=31.

![Fig. 9](image3)

Fig. 9. SSNR vs. the number of sinusoids, L for Music frame (top panel: different frames, bottom panel: ensemble average for 10 frames).

![Fig. 10](image4)

Fig. 10. (a) Reconstruction a speech frame using different number eigen vectors (b) Eigen spectrum, Choosing L largest singular values for signal reconstruction (the dashed line) and the rest (in dotted line).

### 5.4. Subjective Measures

For subjective results, a Mean Opinion Score (MOS) [9] test is conducted to measure the perceived quality for 8kHz speech for different speakers (20 male and female). The results are presented for different groups of listeners versus several number of sinusoids, L∈[21,40] in Table.1. Next, 20 listeners were asked to score between 0-5 to reconstructed utterances. It is observed that the proposed sinusoidal approach requires 25<L<35 to have a negligible difference of in terms of MOS. We found using that L=33 parameters are enough to establish trade-off between low dimensionality and high perceptual quality. In addition, comparing the subjective and objective results, we conclude...
that MDL proposes $L=27$ while MOS results indicate that a perfect reconstruction of speech signal is possible when $25<L<35$ which is in harmony with SSNR results in Fig.6. As a result we opt for $L=33$ to have an indistinguishable speech signal representation using the sinusoidal modeling. The original and synthesized audio files can be downloaded http://ele.aut.ac.ir/pejmanmowlaei/.

<table>
<thead>
<tr>
<th>L</th>
<th>Category</th>
<th>MOS</th>
<th>L</th>
<th>Category</th>
<th>MOS</th>
</tr>
</thead>
<tbody>
<tr>
<td>25</td>
<td>Adult male</td>
<td>3.8</td>
<td>21</td>
<td>Adult male</td>
<td>3.6</td>
</tr>
<tr>
<td></td>
<td>Child</td>
<td>4</td>
<td></td>
<td>Child</td>
<td>3.85</td>
</tr>
<tr>
<td></td>
<td>Old Man</td>
<td>3.9</td>
<td></td>
<td>Old Man</td>
<td>3.7</td>
</tr>
<tr>
<td></td>
<td>Old woman</td>
<td>4.1</td>
<td></td>
<td>Old woman</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>Adult female</td>
<td>4.1</td>
<td></td>
<td>Adult female</td>
<td>4</td>
</tr>
<tr>
<td>33</td>
<td>Adult male</td>
<td>4.4</td>
<td>29</td>
<td>Adult male</td>
<td>4.1</td>
</tr>
<tr>
<td></td>
<td>Child</td>
<td>4.5</td>
<td></td>
<td>Child</td>
<td>4.3</td>
</tr>
<tr>
<td></td>
<td>Old Man</td>
<td>4.4</td>
<td></td>
<td>Old Man</td>
<td>4.1</td>
</tr>
<tr>
<td></td>
<td>Old woman</td>
<td>4.6</td>
<td></td>
<td>Old woman</td>
<td>4.5</td>
</tr>
<tr>
<td></td>
<td>Adult female</td>
<td>4.6</td>
<td></td>
<td>Adult female</td>
<td>4.5</td>
</tr>
<tr>
<td>40</td>
<td>Adult male</td>
<td>4.6</td>
<td>37</td>
<td>Adult male</td>
<td>4.5</td>
</tr>
<tr>
<td></td>
<td>Child</td>
<td>4.75</td>
<td></td>
<td>Child</td>
<td>4.62</td>
</tr>
<tr>
<td></td>
<td>Old Man</td>
<td>4.65</td>
<td></td>
<td>Old Man</td>
<td>4.5</td>
</tr>
<tr>
<td></td>
<td>Old woman</td>
<td>4.8</td>
<td></td>
<td>Old woman</td>
<td>4.7</td>
</tr>
<tr>
<td></td>
<td>Adult female</td>
<td>4.8</td>
<td></td>
<td>Adult female</td>
<td>4.7</td>
</tr>
</tbody>
</table>

6. Conclusion

In order to determine the appropriate number of sinusoids for signal representation, some information criteria including AIC and BIC were employed. The choice of the number of sinusoids was confirmed using different model order selection approaches. It was demonstrated that using a smaller number of sinusoids could result in perfect audio signal reconstruction in term of both objective (SSNR) and subjective results (MOS).

References