Parameterizing Speech Phonemes by Exponential Sinusoidal Model

Jayanth Kumar Talasila and Pradip Sircar *

Department of Electrical Engineering
Indian Institute of Technology Kanpur
Kanpur 208016, India

Abstract. The exponential sinusoidal model (ESM) for parameterizing speech signals is proposed in this paper. The main feature of the ESM is that the amplitude of each sinusoidal component is allowed to vary exponentially with time. A novel variable segmentation strategy is applied first to separate individual transients of a voiced phoneme, which can then be fitted with the ESM. The epoch of a transient is used to separate each speech segment, and the estimation of the ESM parameters is carried out by utilizing the accumulated autocorrelation functions (AACFs) of the speech segment. It is demonstrated that the proposed epoch-ESM representation can be used in analysis-synthesis of voiced phonemes with high accuracy and low computational complexity.

Keywords: Parametric modeling; Speech phonemes; Exponential sinusoidal model; Low complexity speech coder

1 Introduction

Speech coding, signal enhancement, recognition, and more signal processing tasks have been presented in the literature based on the sinusoidal model of speech signals [1–4]. The quasi-periodic nature of voiced speech sounds leads to the harmonic sinusoidal model, and the noise-like feature of unvoiced speech can be generated by a large number of closely-spaced sinusoidal components. Thus, the sinusoidal model turns out to be a powerful tool for model-based speech signal processing because the model can represent both of voiced and unvoiced speech sounds.

However, the main drawback of the sinusoidal model is the requirement of quasi-stationarity for the speech signal, which means that the signal is assumed to be stationary over the time-duration of processing. The speech signal being dominantly non-stationary in nature shows the following points of concern when the signal is fitted with the sinusoidal model:
(i) The modeling is non-parsimonious.
(ii) The parameter estimation has a trade-off between reliability and accuracy, and the choice of the time-duration of processing becomes a crucial issue [5].

* Corresponding author; Email address: sircar@iitk.ac.in

(iii) The model does not fit over transitional speech segments such as speech onsets and voiced/unvoiced transitions [6].
The exponential sinusoidal model (ESM) is found to be suitable for transitional speech segments, typically at boundaries between unvoiced and voiced speech or at speech onsets [6].

The complex amplitude modulated (AM) and frequency modulated (FM) sinusoidal models are found to be well-suited for parametric representation of voiced and unvoiced speech signals, respectively [5,7,8]. The AM and FM models can capture the non-stationarity of speech signals effectively and efficiently, and when the two models are employed together with a time-varying gain parameter [9] for coding of natural speech, we get drastic reduction of bit rates [10]. Thus, the combined method shows potential of high data-compression for speech signals without compromising sound fidelity. However, the method is computation-intensive, and it may not be suitable for some applications where low complexity is primary requirement.

Complexity of speech coding is an important issue in mobile communications because we need to conserve battery power to have reasonable talk time between battery recharging while maintaining handsets of moderate weight [11]. While searching for a speech coder of low complexity, we reconsider the ESM and demonstrate that with a new strategy of signal segmentation and parameter estimation, the model can represent speech phonemes with high accuracy and low computational complexity. We employ the accumulated autocorrelation matrix (AAM) method for estimating the parameters of the ESM [12]. The AAM method is capable of handling the non-stationarity of the modeled signal well, and it can extract model parameters accurately even with short data length.

2 Model Parameter Estimation

A voiced phoneme shows quasi-periodicity, and it comprises of repeated transients. We apply variable segmentation to separate each transient in a voiced phoneme, and the separated transient (speech segment) is modeled by allowing the amplitude of each sinusoidal component to vary exponentially with time in the ESM representation. The speech segment $\{x[n]\}$ in a voiced phoneme is represented by

$$x[n] = \sum_{i=1}^{M} A_i e^{\alpha_i nT} \cos\left(\omega_i nT + \phi_i\right); \quad n = 0, 1, \cdots, N-1 \tag{1}$$

where A_i is the amplitude, ω_i is the angular frequency, and ϕ_i is the random phase of the ith sinusoid; ϕ_i's are assumed to be independent identically distributed (i.i.d.), ϕ_i is uniformly distributed over $[0, 2\pi)$; α_i is the exponential factor, it is usually negative, but it may be positive at speech onset; M is the number of ESM components, N is the segment length in samples, T is the sampling interval, and n is discrete time.

Rewriting Eq. (1) in complex exponential form, we get

$$x[n] = \sum_{i=1}^{2M} B_i e^{(\alpha_i + j\omega_i)nT} e^{j\phi_i} \tag{2}$$

where ω_i and ϕ_i appear in positive-negative pair, and $B_i = A_i/2$. Computing the time-variant autocorrelation function (ACF) r_x of the sequence $x[n]$, we get [12]

$$\begin{aligned} r_x[n,k] &= E\{x^\star[n]x[n+k]\} \\ &= \sum_{i=1}^{2M} B_i^2 e^{2\alpha_i nT} e^{(\alpha_i + j\omega_i)kT} \end{aligned} \tag{3}$$

where E is the expectation operator and * stands for complex conjugation.

By taking the sum of ACFs with same lag k over a fixed time-frame $[n_1, n_2]$, we calculate the accumulated ACF (AACF) c_x which is independent of time,

$$\begin{aligned} c_x[k] &= \sum_{n=n_1}^{n_2} r_x[n,k] \\ &= \sum_{i=1}^{2M} D_i e^{(\alpha_i + j\omega_i)kT}; \quad k = -J, \cdots, 0, \cdots, J \end{aligned} \tag{4}$$

where $D_i = B_i^2 \sum_{n=n_1}^{n_2} e^{2\alpha_i nT}$ with $n_1 = J$ and $n_2 = N - J - 1$ [12]. The maximum lag $J \geq 2M$ and the frame length $(n_2 - n_1 + 1) = (N - 2J)$ should both be large, and there is a trade-off in the choice.

With the calculated values of the AACFs, we form the homogeneous matrix equation [12]

$$\begin{bmatrix} c_x[-J+L] & c_x[-J+L-1] & \cdots & c_x[-J] \\ c_x[-J+L+1] & c_x[-J+L] & \cdots & c_x[-J+1] \\ \vdots & \vdots & & \vdots \\ c_x[J] & c_x[J-1] & \cdots & c_x[J-L] \end{bmatrix} \begin{bmatrix} 1 \\ a_1 \\ \vdots \\ a_L \end{bmatrix} = \mathbf{0} \tag{5}$$

with $J \geq L \geq 2M$, which can be solved for the coefficient vector $\mathbf{a} = [1, a_1, \cdots, a_L]^T$. When speech is noise-corrupted, we form the accumulated autocorrelation matrix (AAM) of Eq. (5) with the extended order L sufficiently larger than $2M$, and the maximum lag J is chosen to be between $N/3$ and $2N/5$ for best result [5]. By utilizing the singular value decomposition (SVD) of the AAM, we determine the vector $\mathbf{a}$ as

$$\mathbf{a} = \frac{\mathbf{e}_1 - \sum_{k=1}^{2M} \mathbf{v}_k^\star(1)\mathbf{v}_k}{1 - \sum_{k=1}^{2M} |\mathbf{v}_k(1)|^2} \tag{6}$$

where $\mathbf{e}_1$ is the vector whose first element is one and the rest zeros, and $\{\mathbf{v}_k; k = 1, \cdots, 2M\}$ are the right singular vectors of the AAM corresponding to the largest $2M$ singular values [12].

Once the coefficient vector is determined, we then form the polynomial equation

$$z^L + a_1 z^{L-1} + \cdots + a_{L-1} z + a_L = 0 \tag{7}$$

and extract the L roots $z_i = e^{(\alpha_i + j\omega_i)T}$. Substituting the values of z_i in Eq. (2) and replacing $2M$ by L, we write the matrix equation with $B_{ci} = B_i e^{j\phi_i}$ as

$$\begin{bmatrix} x[0] \\ x[1] \\ \vdots \\ x[N-1] \end{bmatrix} = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ z_1 & z_2 & \cdots & z_L \\ \vdots & \vdots & & \vdots \\ z_1^{N-1} & z_2^{N-1} & \cdots & z_L^{N-1} \end{bmatrix} \begin{bmatrix} B_{c1} \\ B_{c2} \\ \vdots \\ B_{cL} \end{bmatrix} \tag{8}$$

which can be solved for the vector comprising of B_{ci}. We may use the SVD technique to compute the principal component solution [13]. It is now easy to identify the signal parameters corresponding to the largest $2M$ values of the amplitude B_i.

3 Speech Signal Modeling

Segmentation in the time-domain plays an important role in analysis-synthesis of voiced phonemes consisting of repetitive transients. The voiced sounds are reconstructed accurately by the ESM when each transient of a voiced phoneme begins and ends within the processing time-frame. Hence, we need to employ variable segmentation in the signal waveform such that the signal segments synchronize with the repetitive transients. The lowest (or zero-crossing) point in the signal waveform just preceding the highest peak of a transient is found to be an appropriate location for signal segmentation.

It is to be pointed out that the variable segmentation scheme as introduced in this paper to separate speech segments consisting of individual transients of a voiced phoneme is the key step which makes the ESM suitable for representing speech signals with parsimony of parameters. Note that in this representation, the starting time of each transient or the epoch is one parameter which is to be stored and utilized together with the parameters of the ESM for reconstruction of the speech phoneme.

Once all parameters of the speech phoneme are estimated, we reconstruct the modeled phoneme utilizing the estimated parameters. The goodness of fit of the proposed model is tested by the spectral distance measure proposed by Itakura and Saito (IS), which is defined as [14]

$$d_{IS}(S, \hat{S}) = \left\| \frac{S}{\hat{S}} - \ln\left(\frac{S}{\hat{S}}\right) - 1 \right\|_1 \tag{9}$$

where S and $\hat{S}$ are the spectra (the square-magnitude of Fourier transforms) of the original and reconstructed speech phoneme, respectively, and $\| \; \|_1$represents L^1-norm. The Itakura-Saito distance is a measure of the perceptual difference between the original and reconstructed speech phoneme [14].

4 Simulation Results

The voiced phonemes /aa/, /eh/ and /iy/ of speech selected from the TIMIT database [15] are fitted in the epoch-ESM representation as described in the previous section. A total of 800 samples for each voiced phoneme is available. The sampling rate used is 16kHz. The sampled data are made zero mean and segmented at the epochs of individual transients. The segmentation of the phoneme /aa/ is shown Fig. 1, and the 4 separated transients of the phoneme are shown in Fig. 2. The estimated ESM parameters of the 4 segments of the phoneme are included in Tables 1–4. The original and reconstructed phoneme segments are shown in Figs. 3–6, and all 4 segments connected at the epochs are shown in Fig. 7. The spectra of the original and reconstructed phoneme /aa/ are shown in Figs. 8 & 9 respectively. The Itakura-Saito distance measure for reconstruction of the phoneme /aa/ is computed to be 0.2251.

The parameterization of the phoneme /eh/ is considered next, and the segmentation of the phoneme is shown in Fig. 10. The estimated parameters of the ESM for 4 segments of the phoneme are included in Tables 5–8. the original and reconstructed phoneme /eh/ are shown in Fig. 11, and the Itakura-Saito distance measure for reconstruction is computed to be 0.4011. The segmentation of the phoneme /iy/ is shown in Fig. 12, the estimated ESM parameters are included in Tables 9–12, and the original and reconstructed phoneme are shown in Fig. 13. The Itakura-Saito distance measure for reconstruction is computed to be 0.7042. Note that for all the phonemes considered here, the Itakura-Saito distance measures are found to be well within the perceptible range of values [14].

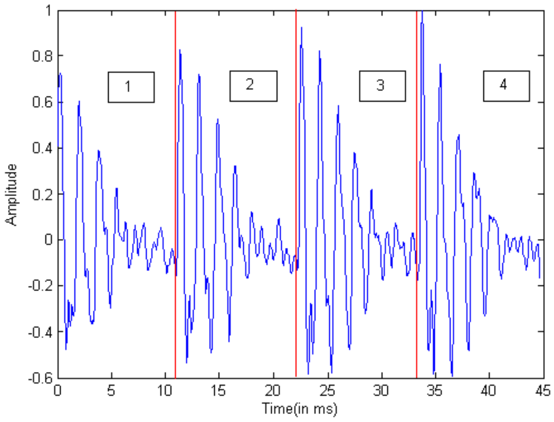

Fig. 1. Voiced phoneme /aa/ showing variable segmentation

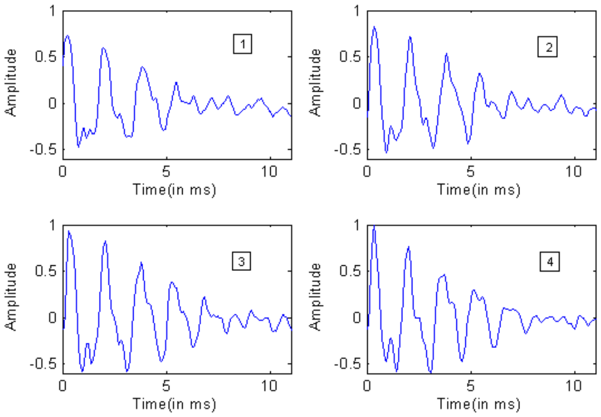

Fig. 2. Voiced phoneme /aa/ separated into 4 segments

Table 1. Estimated parameters of phoneme /aa/: Segment 1

Component	ω	α	A	ϕ
1	±3304.0	-467.59	0.6710	±5.8995
2	±4057.1	-621.40	0.3511	±2.4061
3	±7437.2	-262.81	0.1516	±3.7026

Table 2. Estimated parameters of phoneme /aa/: Segment 2

Component	ω	α	A	ϕ
1	±3458.9	-371.38	0.5742	±5.4601
2	±4178.2	-353.50	0.2270	±2.1237
3	±7747.3	-192.74	0.1598	±2.6489

Table 3. Estimated parameters of phoneme /aa/: Segment 3

Component	ω	α	A	ϕ
1	±3579.4	-334.40	0.5706	±5.4136
2	±4176.8	-244.34	0.2261	±2.3870
3	±7982.7	-192.98	0.1686	±2.4506

Table 4. Estimated parameters of phoneme /aa/: Segment 4

Component	ω	α	A	ϕ
1	±3657.6	-363.97	0.5783	±5.5229
2	±4253.7	-284.52	0.2558	±2.4706
3	±8234.8	-235.70	0.1834	±2.8705

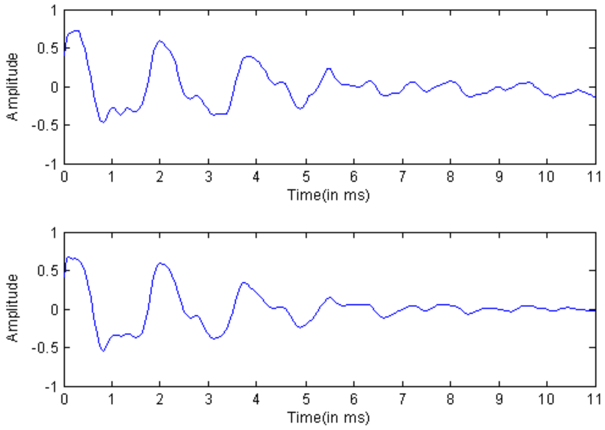

Fig. 3. Original (top) and reconstructed (bottom) phoneme /aa/: Segment 1

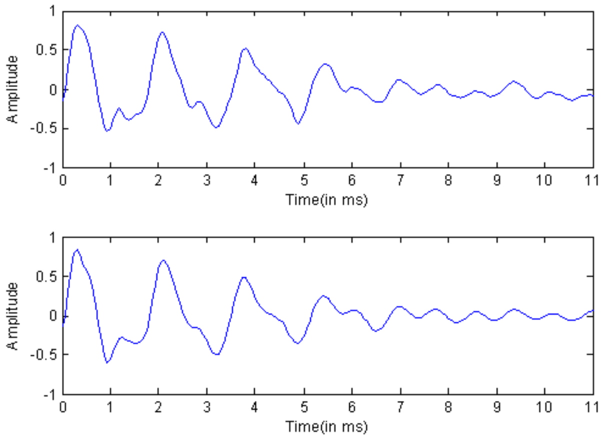

Fig. 4. Original (top) and reconstructed (bottom) phoneme /aa/: Segment 2

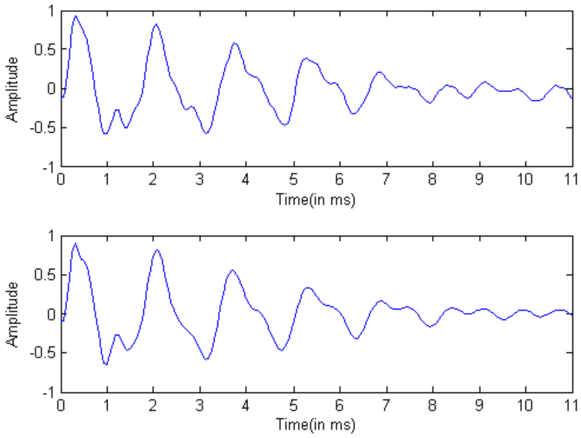

Fig. 5. Original (top) and reconstructed (bottom) phoneme /aa/: Segment 3

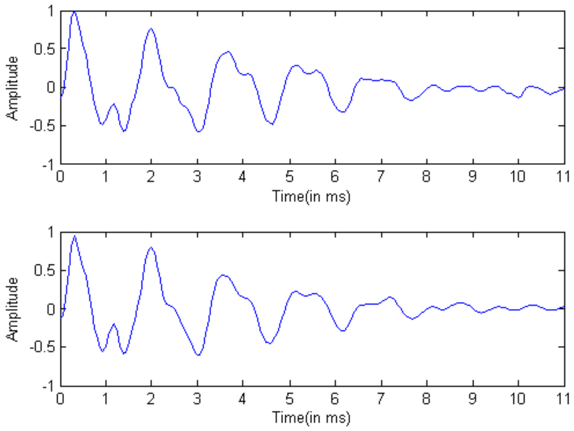

Fig. 6. Original (top) and reconstructed (bottom) phoneme /aa/: Segment 4

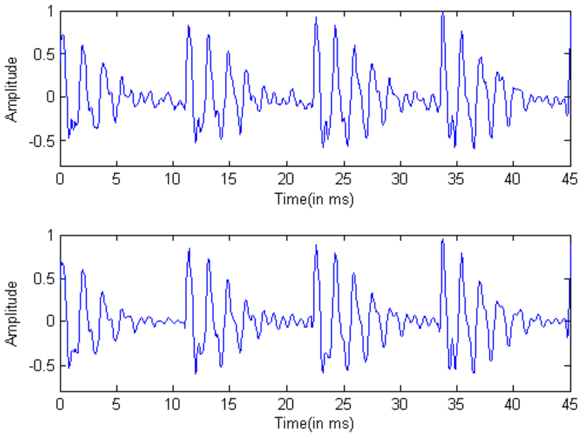

Fig. 7. Original (top) and reconstructed (bottom) voiced phoneme /aa/: all 4 segments

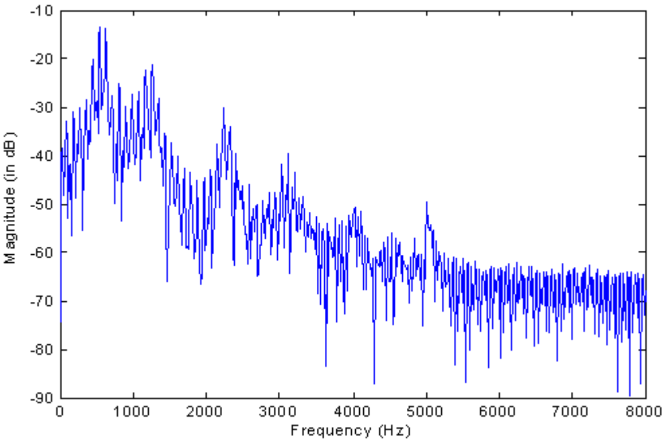

Fig. 8. Spectrum of original voiced phoneme /aa/

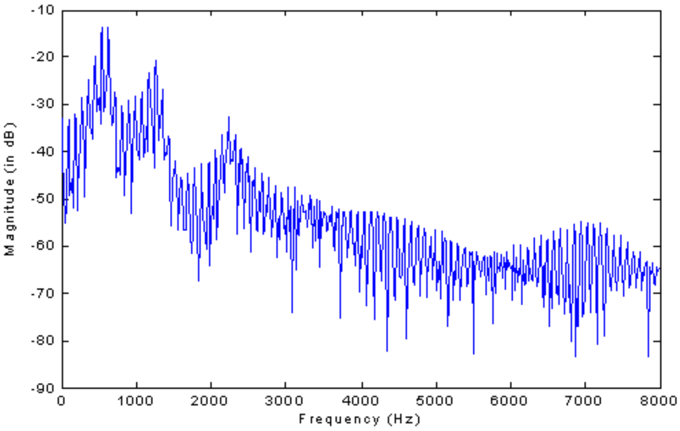

Fig. 9. Spectrum of reconstructed voiced phoneme /aa/

It is to be pointed out that the estimated ESM parameters for 4 segments of the phoneme /aa/ do not change considerably over the segments, which reveals that the signal is a sustained vowel sound. In fact, we can calculate the average of estimated parameters as included in Table 13, and reconstruct the periodic sound signal as shown in Fig. 14. The spectrum of the reconstructed periodic signal is shown in Fig. 15. comparing Figs. 9 & 15, we find that the high frequency components are removed from the reconstructed signal when periodicity is forced in the signal. The Itakura-Saito distance measure for reconstruction of the periodic sound signal is computed to be 0.5758, which is considerably higher than the distance measure computed earlier with the general model. However, it should be noted that the distance measure is still well within the perceptible range with drastic reduction of number of parameters.

Consider now the model fitting of the voiced phonemes /eh/ and /iy/. Note that estimated ESM parameters can not capture the high peaks of segments 1 & 3 of the phoneme /eh/ unless we deviate completely from the quasi-periodicity feature of the voiced phoneme. On the other hand, the number of ESM signals and the estimated parameters vary considerably over the segments for the phoneme /iy/. Therefore, the extracted features of the phonemes /eh/ and /iy/ reveal that the corresponding signals are transitional vowel sounds. The simulation results presented here clearly demonstrate that the epoch-ESM representation provides an efficient means to parameterize sustained and transitional speech phonemes with good fidelity.

5 Concluding Remarks

In this paper, we have proposed a new strategy of epoch-based variable segmentation of speech phonemes, which has made an efficient (accurate with low compu-

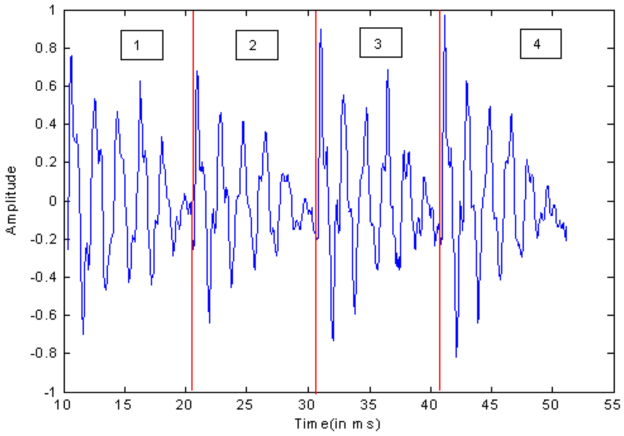

Fig. 10. Voiced phoneme /eh/ showing 4 segments

Table 5. Estimated parameters of phoneme /eh/: Segment 1

Component	ω	α	A	ϕ
1	±3237.8	-175.31	0.3130	±5.3339
2	±3673.2	-139.10	0.1396	±2.8964
3	±10074.9	-224.86	0.1343	±3.2332

Table 6. Estimated parameters of phoneme /eh/: Segment 2

Component	ω	α	A	ϕ
1	±3366.3	-217.32	0.2634	±5.3502
2	±3835.8	-185.87	0.1369	±2.8105
3	±10085.5	-243.17	0.1339	±2.7235

Table 7. Estimated parameters of phoneme /eh/: Segment 3

Component	ω	α	A	ϕ
1	±3472.1	-216.72	0.3446	±5.3654
2	±3932.9	-160.48	0.1509	±2.7256
3	±10301.3	-207.88	0.1372	±2.8091

Table 8. Estimated parameters of phoneme /eh/: Segment 4

Component	ω	α	A	ϕ
1	±3519.1	-251.82	0.3888	±5.1606
2	±3997.5	-175.92	0.1609	±2.5535
3	±10354.1	-254.96	0.1638	±2.4450

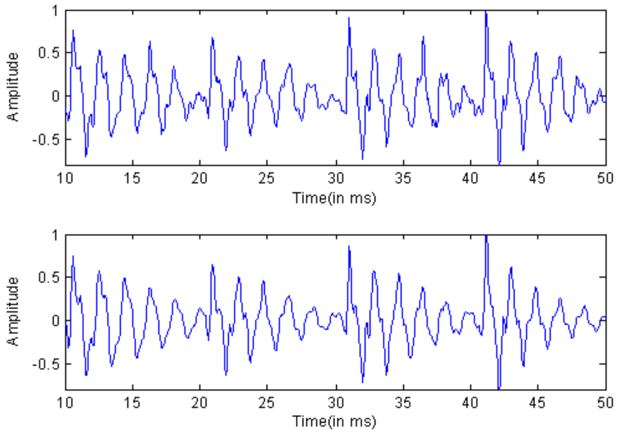

Fig. 11. Original (top) and reconstructed (bottom) voiced phoneme /eh/

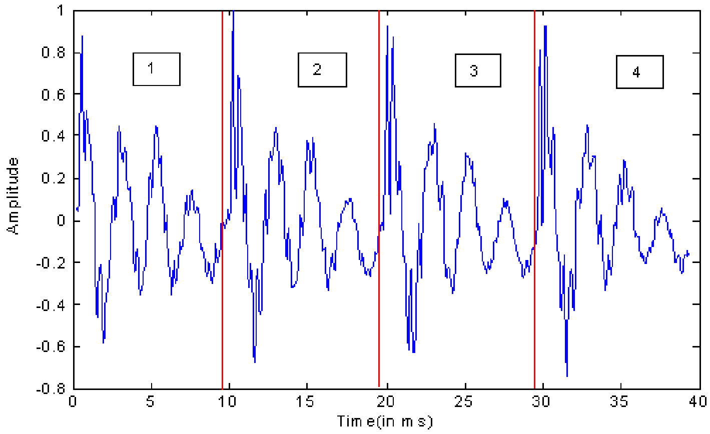

Fig. 12. Voiced phoneme /iy/ showing 4 segments

Table 9. Estimated parameters of phoneme /iy/: Segment 1

Component	ω	α	A	ϕ
1	± 2708.0	-252.08	0.3940	± 5.1351
2	± 3169.0	-154.47	0.1042	± 2.1380
3	± 10927.1	-414.96	0.2112	± 3.0497

Table 10. Estimated parameters of phoneme /iy/: Segment 2

Component	ω	α	A	ϕ
1	±2618.0	−287.43	0.4508	±4.7404
2	±3039.6	−159.41	0.1156	±1.7976
3	±15857.9	−273.04	0.1199	±5.8524

Table 11. Estimated parameters of phoneme /iy/: Segment 3

Component	ω	α	A	ϕ
1	±2528.9	−260.15	0.4054	±4.7421
2	±12404.5	−626.90	0.1237	±2.9867
3	±15793.5	−555.20	0.1498	±1.4519

Table 12. Estimated parameters of phoneme /iy/: Segment 4

Component	ω	α	A	ϕ
1	±2427.0	−350.40	0.4610	±4.9133
2	±2928.8	−209.77	0.1290	±1.9468
3	±12535.4	−381.90	0.1281	±4.2502
4	±16068.3	−442.58	0.1386	±2.6480

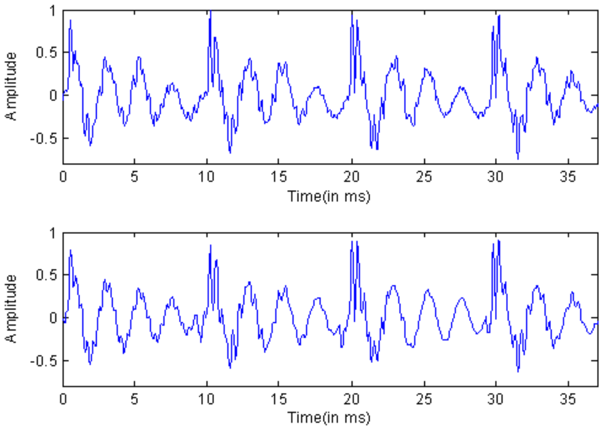

Fig. 13. Original (top) and reconstructed (bottom) voiced phoneme /iy/

tational complexity) parametric modeling possible for voiced speech phonemes. We have used a simple approach to extract epochs of the speech phonemes, which provides satisfactory performance in the simulation study. However, it should be mentioned that there are several methods of epoch extraction available in the literature [16].

It is known that the basic sinusoidal model (BSM) can as well parameterize unvoiced speech phonemes with comparatively large number of sinusoidal signals. It should be emphasized that parameterization of unvoiced speech phonemes by the ESM will be advantageous when the phonemes are transitional at speech onsets or decays. The study of natural speech shows such variations of gain factors appearing frequently on the time axis [9,10]. Therefore, the ESM will be more efficient to parameterize the phonemes of natural speech than the BSM [6].

However, it should be pointed out that for the unvoiced speech phonemes, the epoch and quasi-pariodicity are not meaningful terms. Hence, we have to include a large number of closely spaced aharmonic sinusoids, and we may have to use a variable segmentation strategy based on variation of gain factors or energy levels in the speech phonemes. There is need for further research in this direction.

Table 13. Average of estimated parameters for 4 segments of phoneme /aa/

Component	ω	α	A	ϕ
1	± 3564.6	-356.58	0.5743	± 5.4655
2	± 4202.3	-294.12	0.2363	± 2.3271
3	± 7987.6	-207.00	0.1706	± 2.6560

References

1. McAulay, R.J., Quatieri, T.F.: Low-Rate Speech Coding based on the Sinusoidal Model. In: Furui, S., Sondhi, M.M. (eds.) Advances in Speech Signal Processing, pp. 165–208. Marcel Dekker, New York (1992)
2. Kates,, J.M.: Speech Enhancement based on Sinusoidal Model. J. Speech Hear. Res. 37(2), 449–464 (1994)
3. George, E.B., Smith, M.J.T.: Speech Analysis/Synthesis and Modification using an Analysis-by-Synthesis/Overlap-Add Sinusoidal Model. IEEE Trans. Speech and Audio Processing. 5(5), 389–406 (1997)
4. Virtanen, T., Klapuri, A.: Separation of Harmonic Sound Sources using Sinusoidal Modeling. In: Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, pp. 765–768. Istanbul (2000)
5. Sircar, P., Syali, M.S.: Complex AM Signal Model for Non-stationary Signals. Signal Processing. 53(1), 35–45 (1996)
6. Jensen, J., Jensen, S.H., Hansen, E.: Exponential Sinusoidal Modeling of Transitional Speech Segments. In: Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, pp. 473–476. Phoenix, A.Z. (1999)

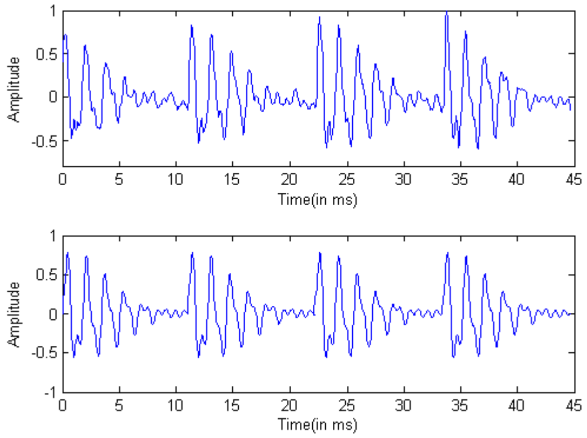

Fig. 14. Original (top) voiced phoneme /aa/ and reconstructed (bottom) periodic phoneme with average parameters

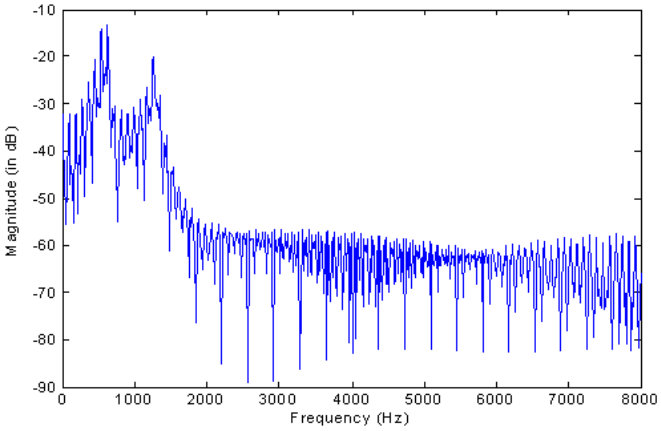

Fig. 15. Spectrum of reconstructed periodic phoneme /aa/ with average parameters

7. Sircar, P., Sharma, S.: Complex FM Signal Model for Non-stationary Signals. Signal Processing. 57(3), 283–304 (1997)
8. Sircar, P., Saini, R.K.: Parametric Modeling of Speech by Complex AM and FM Signals. Digital Signal Processing. 17(6), 1055–1064 (2007)
9. Mukhopadhyay, S., Sircar, P.: Modelling Non-stationary Signals by Time-Dependent AR Process with Time-Varying Gain. IETE J. Research. 43(5), 351–358 (1997)
10. Verma, A.K.: Natural Speech Coding by AM and FM Signal Models. MTech thesis, Elec. Eng. Dept., IIT Kanpur (2003)
11. Budagavi, M., Gibson, J.D.: Speech Coding in Mobile Radio Communications. Proc. IEEE. 86(7), 1402–1412 (1998)
12. Sircar, P., Mukhopadhyay, S.: Accumulated Moment Method for Estimating Parameters of the Complex Exponential Signal Models in Noise. Signal Processing. 45(2), 231–243 (1995)
13. Vaccaro, R.J., Tufts, D.W., Boudreaux-Bartels, G.F.: Advances in Principal Component Signal Processing. In: Deprettere, E.F. (ed.) SVD and Signal Processing: Algorithms, Applications and Architectures, pp. 115–146. Elsevier Science, North Holland (1988)
14. Quackenbush, S.R., Barnwell, T.P., Clements, M.A.: Objective Measures of Speech Quality. Prentice Hall, Englewood Cliffs, N.J. (1988)
15. TIMIT Acoustic-Phonetic Continuous Speech Corpus, `http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC93S1`
16. Murty, K.S.R., Yegnanarayana, B.: Epoch Extraction from Speech Signals. IEEE Trans. Audio, Speech, Language Processing. 16(8), 1602–1613 (2008)