Potpourri of Research Articles in Voice and Speech - PRAVS 1

Papers on Linear Prediction

Compiled by

T V Ananthapadmanabha

Published by

VOICE AND SPEECH SYSTEMS, BANGALORE

August 2022

Release: May 2023

Free download from URL: www.vagmionline.com
(Use Tab: About Us -> About T V Ananthapadmanabha)

Preface to the Series

Each volume of the series is a collection of research articles written by the author either alone or with some research collaborators that have not been formally published in peer reviewed journals. Some of the articles are the full articles of the Abstracts submitted to conferences. Some articles have appeared earlier in the Proceedings of a Workshop or in some Journals that are not easily accessible etc. Some articles are prepared based on the talks given by the author.

The articles are written in an informal style as Technical Notes with emphasis on the main idea and results. Thus, no attempt has been made to present an exhaustive review of the previous works or to cite all the relevant references. The purpose is to share the ideas or thoughts with co-researchers in the field with the hope that these thoughts or ideas may trigger interest in the researchers ultimately resulting in publications in peer reviewed Journals or in some technological applications.

A series of Volumes is proposed to be brought out and made available on the internet. Each Volume covers a sub-area within the broad area of Voice and Speech. Tentatively, topics in the following sub-areas are planned to be covered:

Linear Prediction
Formants
Voice Source
Phonemics, Phonetics and Phonology
Vocal Tract Acoustics
Acoustic-Phonetics
Miscellaneous

Readers may contact the undersigned for any clarification.

31 August 2022 T. V. Ananthapadmanabha
Release Date: 7th May 2023 Email Address: tva.blr@gmail.com

Visit **www.vagmionline.com** -> About us -> About Ananthapadmanabha: Allied Interests for Free Download of general Articles and Books and Sound-to-Form Transformation (Vagmi Tonoscope) Software

Preface to Volume-1

This volume has four research articles related to LP technique. The first article is a modified method of LP called square-root LP. In the standard autocorrelation formulation of LP, the autocorrelation function is used for solving LPCs. Autocorrelation is the Fourier inverse of magnitude squared spectrum. We refer to the Fourier inverse of magnitude spectrum as square-root autocorrelation, which is then used to solve for square-root LPCs. The standard autocorrelation formulation of LP doesn't give an optimum spectral flattening when the spectral dynamic range of an input signal is large. Since the spectral dynamic range of a magnitude spectrum is one-half of the magnitude squared spectrum, we expect the square-root LP formulation to give a better spectral flattening of the input spectrum.

The second article is on the quantization of parcor or reflection coefficients within the Durbin's algorithm in the context of LPC-10 vocoder. It is shown that such an approach gives a lower error between the spectral envelopes of reconstructed speech and original speech signal compared to the so called optimum quantizer using log-area ratios.

The third article is on the multi-pulse coding of the LP residual utilizing the phase response of LP residual. When the phase component is utilized the mean segmental SNR improves by about 3-dB.

The fourth long article is on the estimation of formant data using the LP technique using synthesized vowels. This article has two main results. The first results is to show that the fundamental frequency or F0 has a very strong influence on the estimated formant data and that the error in the estimation of formant data shows a systematic trend with respect to the harmonic number or harmonicity (F(1)/F0). The second result is that error in F(1) is independent of errors in other formants. Based on these two main results, an analysis-synthesis approach has been proposed to improve the accuracy of initial estimates. A limitation of this study is that a single impulse per pitch period has been used as an excitation function. A similar study using the more appropriate excitation function, viz., the derivative of glottal pulse or voice source pulse is planned to be reported in a forthcoming Volume.

CONTENTS
# of Volume-1

Linear Prediction Analysis based on Square Root Autocorrelation

T. V. Ananthapadmanabha and H. S. Chakravarthy

Article No.1.1.
This work was carried out during 1980.

Abstract: *A new technique to improve the basic accuracy of smooth spectral representation of a speech signal is presented. Improved accuracy is achieved by modelling the short-time 'magnitude spectrum' instead of modelling the 'magnitude squared spectrum' as being done conventionally. Accurate spectral representation is obtained even for cases of spectra with large dynamic range and spectral valleys. Because of an effective whitening of the short-time spectrum of a voiced speech signal, LP residual approximates a sequence of impulses to a better accuracy.*

I. Introduction

Linear prediction (LP) technique [1] gives a computationally efficient scheme for implementing low bit rate (2400 bps) vocoders. Although LP vocoders possess high intelligibility (85%), it has been noted that naturalness is lacking in the quality of reconstructed speech [2]. It is usually assumed that the vocal tract transfer function is adequately represented by an all-pole model of LP technique but the voice source is not properly modelled. Excitation signal other than an impulse with some additive noise components have been proposed for improving the quality of reconstructed vocoder speech [3 - 5]. Improvement in the quality, close to natural sounding speech can be obtained by coding and transmitting LP residual at the cost of increased bit rate (6600-9600 bps) [6]. By coding and transmitting the amplitudes of harmonics of short-time spectrum of LP residual, appropriate excitation signal can be reconstructed [2]. This scheme requires computing the spectrum of LP residual with high resolution and also increases the bit rate.

LP analysis of voiced segments poses greater difficulties compared to the analysis of unvoiced segments. This is because of a large spectral dynamic range and periodicity of the voiced segments. Further, it is known that LP technique gives a better spectral match near the peaks than at the valleys. Hence, we restrict the discussion to analysis of voiced segments. In this paper we propose an effective spectral whitening technique so that the LP residual can be modelled by an impulse sequence to a better accuracy.

II. Proposed Method

II.A. Causes for Inaccuracy in LP Analysis

Voiced speech is a superposition of the responses of vocal tract filter to successive glottal excitations. The problem in the analysis is to estimate the basic component given the composite periodic signal. Consider a composite signal {y(n)} of a basic signal and its echoic component given by

$$\{y(n)\} = \{x(n)\} + G\,\{x(n-n0)\} \qquad (1)$$

where {x(n)} is the basic signal, n0 is the delay and G is the gain of the echoic signal. The autocorrelation of {y(n)} is given by

$$R_y(k) = (1+G^2)\,R_x(k) + G\,R_x(k-n0) + G\,R_x(k+n0)$$

Note that the autocorrelation $R_x(k)$ required to solve for LPCs is distorted by the presence of echoic components at ±n0. This distortion gets severe as n0 decreases (pitch period decreases) and when Rx(k) is slowly decaying; i.e., when {x(n)} has a large dynamic range spectrum. See Fig.1. In fact, it has been noted that for a periodic signal with low n0 (high pitch), LP spectrum may not represent the spectrum of the basic signal accurately [1]. Also see Articles 1.4 and Article 1.5 in this volume.

II. B. Concept of Square Root Autocorrelation

As noted above, the distortion in the autocorrelation of the basic component decreases if the spectral dynamic range is low. We propose modelling the magnitude spectrum of a signal by an all-pole model to achieve the reduction in the spectral dynamic range. That is, if $S(\omega)$ is the Fourier transform of a signal {s(n)}, $|\,S^2(\omega)\,|$ is the power spectrum whose Fourier inverse gives the autocorrelation function $R_s(k)$. Since $|S(\omega)|$ is a positive real function, the Fourier inverse of $|S(\omega)|$ behaves like an autocorrelation function which we refer to as 'square root autocorrelation' denoted by $Q_s(k)$. Since LP formulation can also be viewed as a spectral matching technique, it can be applied on $|S(\omega)|$. An all-pole model 1/B(z) can be assumed to represent the magnitude spectrum $|S(\omega)|$. The coefficients of B(z) can be obtained by solving the normal equations involving $Q_s(k)$ instead of $R_s(k)$.

The dynamic range of $|S(\omega)|$ is one-half of the dynamic range of $|S^2(\omega)|$. Hence, we expect the distortion in $Q_S(k)$ for a periodic signal to be less severe compared to the distortion in $R_S(k)$. The autocorrelation functions and the corresponding log spectra which are modelled by the direct method and square root method are shown in Fig. 1. Note the reduction in dynamic range. Also, two spectral samples at relative levels of 1:100 for $R_S(k)$ will be at level 1:10 for $Q_S(k)$. This shows the relative importance of spectral peaks and valleys in the two autocorrelation functions. To restore the proper dynamic range in the synthesized speech, the filter 1/B (z) should be used in cascade with itself. The use of multiple poles of order two has not produced any noticeable perceptual distortion as tested in informal experiments. The gain term for transmission can be the energy of speech samples per frame. The synthesized samples can then be scaled to match the energy of synthesized speech with the energy of input speech.

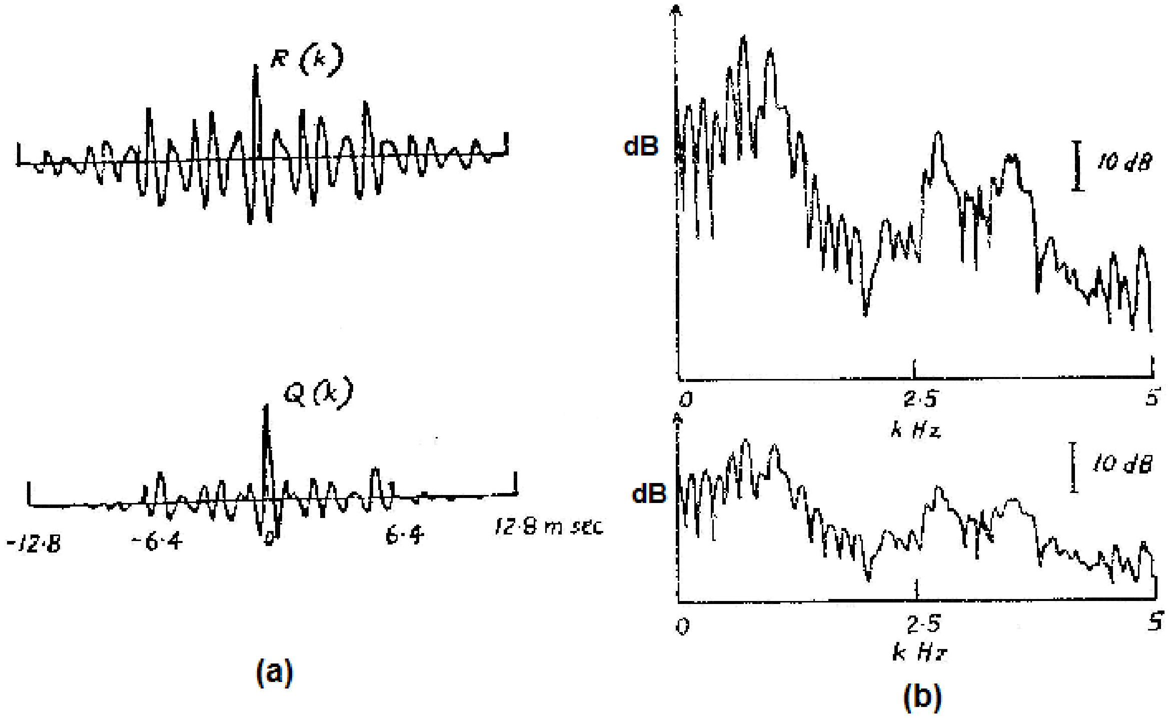


Fig.1. (a) Autocorrelation R(k) and Square root autocorrelation Q(k) of a voiced segment (b) The corresponding log spectra

II.C. Examples

We illustrate the proposed technique with some examples in this section. A segment of vowel waveform and its LP residual obtained by the two methods are shown in Fig.2. LP residual of direct method shows samples with bipolar swings of considerable amplitude near the excitation instants. Also, samples with large amplitudes occur at other instants giving a wrong indication of possible multiple excitations [7]. LP residual of square root method shows sharp impulses at the excitation instants and low values elsewhere. In other words, square root method gives a more accurate inverse filtering. This will also be useful for accurate identification of closed glottis interval [7].

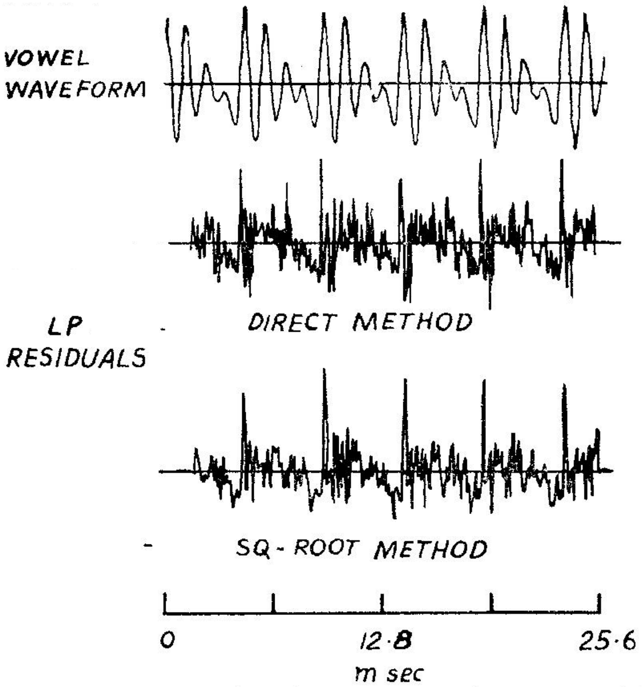


Fig.2. Inverse filtering of a vowel segment.

The estimated vocal tract frequency response and log spectra of LP residuals obtained by the two methods are illustrated in Figs. 3 to 5. The spectral level at pitch harmonics must nearly be the same if inverse filtering is accurate. With this criterion it may be noted that square root method gives better accuracy as seen in the log spectra of LP residuals. The estimated spectral level may be compared near spectral valleys, near the folding frequency and for frequencies beyond 5 kHz (Fig.5). Since the spectral representation over valleys is more accurate in the square root method we expect the technique to give a better spectral representation for nasals, voiced fricatives etc.

The smooth log spectrum shows graceful and consistent changes as M is increased and the filter 1/B(z) remains stable even for large values of M. In the direct method the magnitude of first two reflection coefficients are near unity and pose problems in quantization. Also, the normalized error [1] shows sharp decrease implying that the first few reflection coefficients are relatively more significant. In the square root method, quantization of reflection coefficients don't pose difficulties and all the reflection coefficients are equally important.

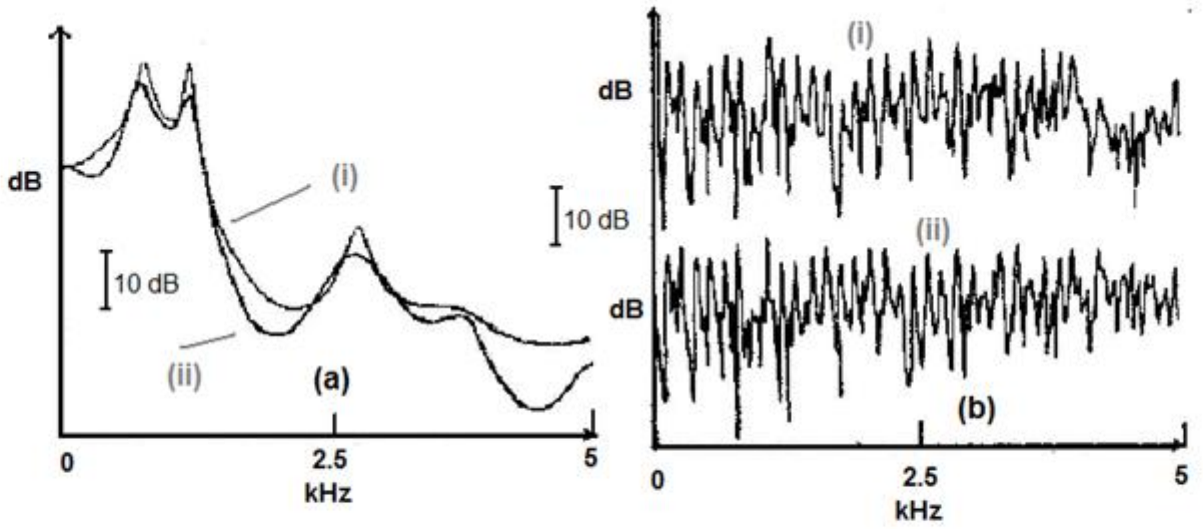


Fig.3. An example of vowel segment of male voice (a) Log spectral envelopes (b) Log spectra of LP residuals. (i) Direct method (ii) Square root method

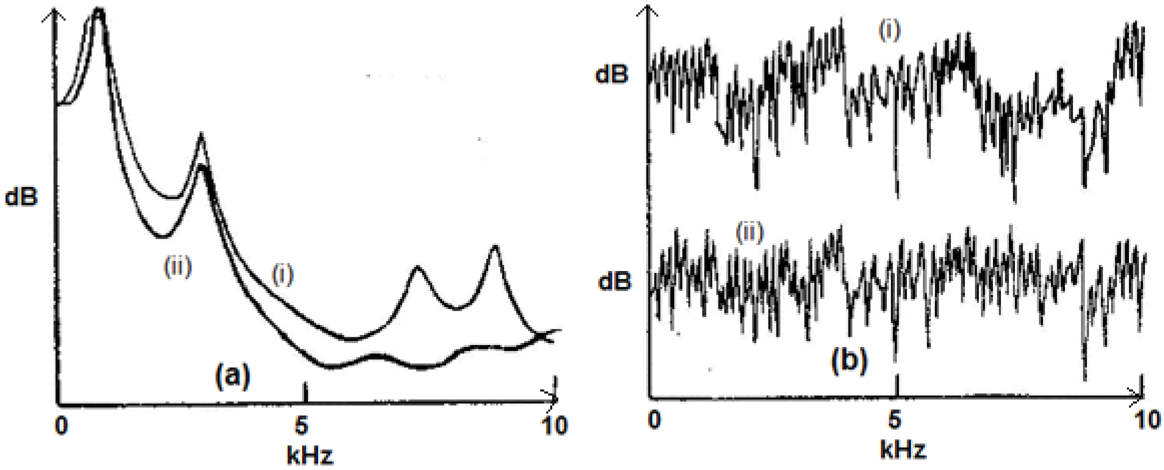


Fig.4. An example of vowel segment of male voice sampled at 20 kHz (a) Log spectral envelopes (b) Log spectra of LP residuals. (i) Direct method (ii) Square root method

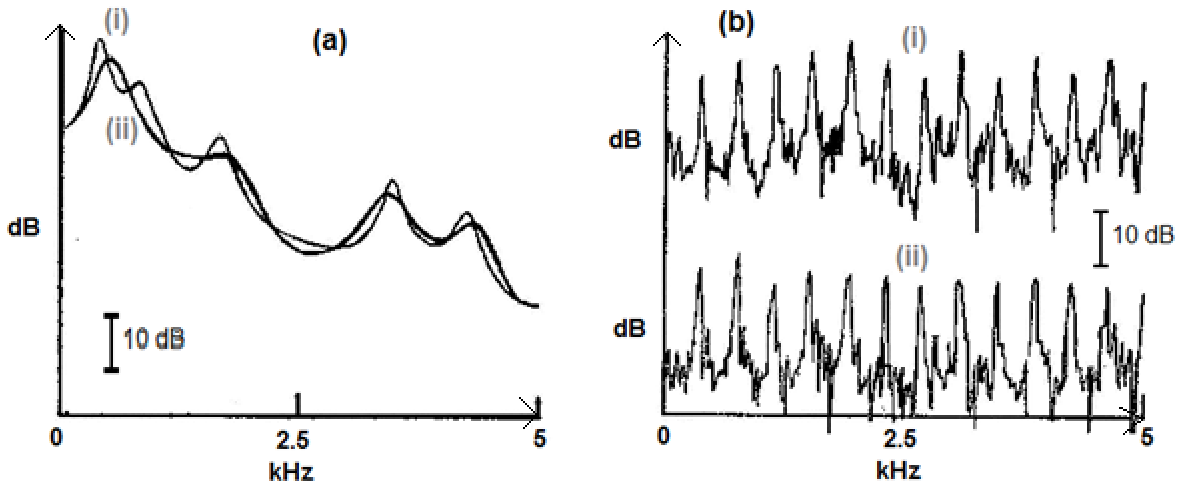


Fig.5. An example of vowel segment of female voice (a) Log spectral envelopes (b) Log spectra of LP residuals. (i) Direct method (ii) Square root method

III. Conclusion

While retaining the computational elegance of linear prediction technique, and without increasing the bit rate, an improved method for spectral modelling in voiced speech analysis has been presented. Computation of square root autocorrelation and synthesis involve additional effort which must be weighed against the improvement in accuracy. Extensive subjective experiments are required to evaluate the technique. However, preliminary experiments have shown better spectral representation especially for low orders of predictor and high sampling frequency (20 kHz).

References

1. Makhoul J., "Linear prediction: A tutorial review", Proc. IEEE, vol.63, Apr 1975, pp.561-580.

2. Atal B. S., and Nancy David, "On synthesizing natural-sounding speech by linear prediction", ICASSP-79, pp.44-47.

3. Sambur M. R. et. al., "On reducing the buzz in LPC Synthesis," J. Acoust. Soc. Am., vol. 63, No.3., 1978, pp. 918-924.

4. Makhoul J., "A mixed-source model for speech compression and synthesis," ICASSP 78, pp.163-166.

5. Akira Kurematsu et. al., "A linear predictive vocoder with new pitch extraction and exciting source," ICASSP-79, pp.69-72.

6. Un G. K., and Magill D. T., "The residual-excited linear prediction vocoder with transmission rate below 9.6 kbps," IEEE Trans. Comm., vol. COM-23, December 1975, pp. 1466-1474.

7. Ananthapamanabha T. V. and Yegnanarayana B., "Epoch extraction from linear prediction residual for identification of closed glottis interval," IEEE Trans. Acoust. Speech and Sig. Proc., vol. ASSP-27, No.4, Aug. 1979, pp. 309-319

Anecdotal Note

This work was presented at a Workshop held at TIFR, Mumbai during 22-25 July 1980. The author met Prof Gunnar Fant at this workshop for the very first time and this meeting led to an opportunity of collaborating with him at KTH, Stockholm.

This paper appeared in the workshop proceedings and it created a minor jolt to Dr Hermansky, who around 1980s was working with Prof. Fujisaki for his doctoral thesis on a very similar concept. In our work the magnitude squared spectrum has been raised to power one-half (1/2). Dr Hermansky instead proposed a general exponent of the form (p/q). He later published his work with the title 'Perceptual Linear Prediction'.

Optimum Quantization Scheme of Vocal Tract Parameters for an LPC-10 Vocoder

T. V. Ananthapadmanabha

Article No.1.2.

This work was carried out during 1984-85.

Abstract: *LPCs and reflection coefficients are obtained from the Autocorrelation coefficients using Durbin's recursive algorithm. In the proposed method, the reflection coefficient obtained at every recursive step is quantized by a linear quantizer and subsequent steps in the algorithm uses the quantized reflection coefficient. This is known as 'quantization within the loop'. The error introduced in the log-spectrum by the proposed method is much lower compared to the error of a LPC-10 coder that proposes an optimum quantizer based on log-area-ratios (LARs).*

I. Introduction

Linear prediction (LP) coding is based on a Source-Filter model of speech production. The 'Filter' part corresponds to the transfer function of vocal tract (VT). Since the articulators are continually moving during speech production, the transfer function of VT is also changing continually. Similarly, the 'source' part is also changing with time. Speech signal is divided into frames of 40 milli-sec in duration, assuming quasi-stationarity.

In LP coding, the transfer function of vocal tract is represented by an all-pole model. The all-pole model is determined by a set of parameters referred to as LP Coefficients (LPCs). LPCs are computed from the autocorrelation coefficients of a speech signal for every frame using Durbin's recursive algorithm. There exists a set of equivalent parameters related to LPCs. One such set corresponds to the 'set of reflection coefficients'. The 'reflection coefficient' is actually an intermediate result that is obtained while solving for the LPCs. However, the term 'reflection coefficient' also has a physical significance. During speech production, at any given instant, VT has a 3-D geometrical shape. When the acoustic waves within VT are assumed to be 1-dimensional, the complex 3-D shape can be approximated by a cascade of cylindrical shapes of uniform width and varying area. When an acoustic wave arrives at a junction of two cylindrical sections of differing area, part of the energy is reflected and remaining part is propagated. The reflected part of energy relative to the total energy at that section is determined by the ratio of the areas of the two cylindrical sections at that junction. This ratio is referred to as the reflection coefficient. Incidentally, the magnitude of reflection coefficient has to be less than or equal to 1. It so happens that the reflection coefficient, the by-product obtained using the Durbin's algorithm also satisfies this property in that its magnitude is bound by unity. If reflection coefficients are known, the areas of cylindrical sections can also be computed assuming an arbitrary reference area at the glottis. Thus a 'set of areas' is also related to 'set of LPCs'. Further, there exist mathematical transformations to compute reflection coefficients given the LPCs and area ratios given the reflection coefficients and vice-versa.

Consider linear PCM representation of a speech signal sampled at 8000 Hz with 8 bits per sample resulting in 64000 bits per second. In LP vocoder, parameters of source and filter parts are extracted for every frame. Typically LP of order 10 is used. The real valued parameters of each fame are quantized for coding purposes. Each real valued parameter is appropriately scaled and converted into an integer. Certain number of bits is allocated based on the range of integer values of each parameter. Thus, a compression is achieved. In the standard LPC-10 coder, the number of bits allocated per frame for both source and filter parts is 48 and the frame rate is 50 frames per second resulting in a bit rate 2400 bits per second for the coded speech compared to 64000 bits per second of the original digital speech signal. Since LPCs are mathematically related to reflection coefficients and area coefficients, any set of parameters may be quantized for compression.

Such a compression in coding is achieved at the cost of some distortion in the quality of decoded speech signal. An objective method to study the distortion is to compare the spectral envelopes of a given frame of the original speech signal and decoded speech signal. Spectral envelope may be computed using LPCs. The mean of the sum of differences in the log spectral envelopes computed before and after quantization is an objective measure of the distortion. The challenge is to come up with a quantization scheme (choice of parameters, number of bits per parameter etc) such that the distortion is as low as possible. Perceptually, the criterion is that the difference spectrum must be as flat as possible and the mean of the sum of differences must be typically below 0.5 dB/Hz. In the literature, it has been reported that the optimum set of parameters that gives the lowest distortion is the set of log-area ratios (LARs).

II. Proposed Method and Results

The steps of the proposed algorithm are shown in Fig.1. These steps are very similar to the standard Durbin's recursive algorithm. In the proposed method, the reflection coefficient k(i) obtained at the i-th iteration is quantized by a linear quantizer. Subsequent steps make use of the quantized q(j), j=1 to i. This approach is well known as the 'quantization within the loop'. The number of bits allocated for each reflection coefficient is the same as in the standard LPC-10.

Results for two different vowel segments are shown in Fig.2. Results for the LAR quantization and proposed method are compared. The difference between the log-spectra before and after quantization may be noted. The performance of the proposed method is better than that obtained with the so called optimum quantization based on LARs.

Proposed Modification to Durbin's Algorithm

```
Step 0:                  ' Initialization
i=0
E(i) = R(i)
a(i) = 1
'----------------------------------------
DO
      i=i+1

     Step 1a: compute partial sum
          sum = 0 if i=1 else sum = $\sum_{j=1 \text{ to } (i-1)} a(i-1)R(i-j)$

     Step 1b: k(i) = -[ R(i) +  sum] / E(i-1)


     Step-2: q(i) = QZN[k(i)]       'q(i) is the quantized reflection coefficient
                                    'QZN is the linear quantizer subroutine

     Step 3a: Make a backup copy of LPCs of the previous iteration
                b(j)=a(j), j= 0 to (i-1)
```

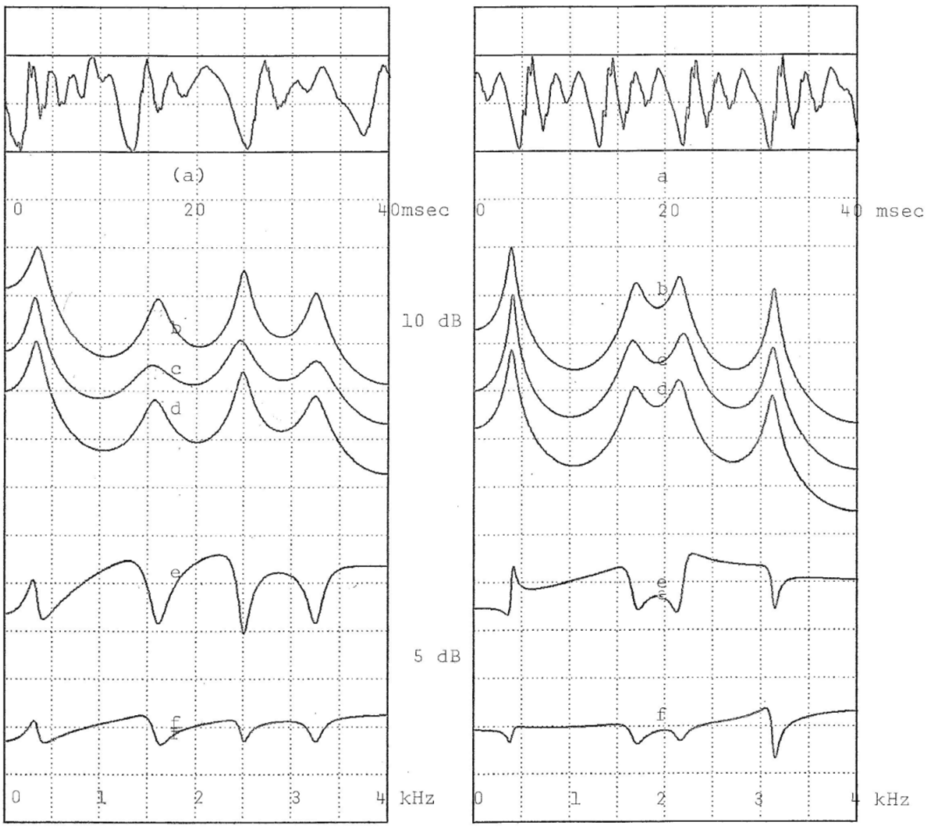
```
        Step 3b: a(i) = q(i)                    'i-th LPC = i-th Reflection coefficient
        Step-3c: a(j) =  b(j) + q(i) b(i-j);  'Update a(j) for j=1 to (i-1)

        Step-4: E(i) = [  1  - $q^2(i)$ ] E(i-1)
Loop till i<p

QZN Subroutine:
save_sgn=sgn(k(i)); 'save the sign
k_Int(i) = FIX(abs(k(i))*2^NBITS(i)); '  NBITS(i): No of Bits allocated for i-th coefficient
q(i)=save_sgn*k_Int(i)/2^NBITS(i);   ' real value after quantization
k_Int(i)=save_sgn*k_Int(i);            'Quantized integer value for transmission
return

------------------------------------------------------------------------------------------------
```

Fig.1. Proposed modification to Durbin's algorithm in a schematic format



a: Waveform of a speech segment

Log Spectra (b) before quantization (c)with LAR qzn (d) Proposed scheme

Error Spectra (e) LAR qzn (f) proposed scheme

10th Order LP

Fig.2. Illustration of the proposed method

III. Conclusion

Quantization within loop of reflection coefficients gives a better performance than the optimum LAR quantization. A study on scalar quantization in the context of LPC-10 may not be very relevant in the contemporary context. However, the author hopes that young researchers may derive some inspiration from this article to the effect that it may be possible to improvise upon an established claim on the most optimum method of quantization.

Key References

Makhoul J., "Linear Prediction: A Tutorial Review", *Proc. IEEE*, vol.63, pp.561-580, Apr. 1975.

Viswanathan P and Makhoul J, "Quantization Properties of Transmission Parameters in Linear Predictive Systems", *IEEE Trans. Acoust. Speech and Sig. Proc.*, vol. ASSP-23, pp.309-321, June 1975.

Gray A. H. and Markel J. D., "Quantization and Bit Allocation in Speech Processing:, *IEEE Trans. Acoust. Speech and Sig. Proc.*, vol. ASSP-24, pp.459-473, Dec. 1976.

Anecdotal Note

The author worked on a consultancy contract awarded by ERICSSON, Sweden during Nov 1984 to Mar 1985. The non-disclosure binding of this contract has long since expired (Mar 1987). The contract award amount served as a seed money to launch the enterprise 'Voice and Speech System' in 1985, which is still continuing. The author is grateful to Dr Tjernlund of ERICSSON for the consultancy contract and to Prof Fant for giving the permission to work on this contract.

The challenge was to improve the quality of LPC-10 coder-decoder without making any major changes in the structure of the vocoder. The work reported in the above article was a part of this contract. Additionally, two modifications were suggested for the synthesizer used in the decoder: (A) A lattice synthesizer that uses reflection coefficients was proposed in place of the direct form that uses LPCs. Also, it was shown that it is optimal to interpolate the reflection coefficients and not the LPCs. (B) For the source part in the synthesis of voiced segments, a voice source model (derivative of glottal pulse) developed by the author was used instead of post-emphasis of impulse (or its Hilbert Transform). The maximum negative amplitude of the model voice source pulse was related to the gain parameter (RMS).

Modeling of Excitation for Linear Predictive Coders of Speech

T. V. Ananthapadmanabha and D. Rajesh

Article No.1.3.

This work was carried out during 2004 when the authors were at MSR School of Advanced Studies, Bangalore. This work was the MSc Dissertation of Mr Rajesh supervised by Dr T V Ananthapadmanabha.

Abstract: *A method for coding of LP residual is proposed. A Wavelet Transform based algorithm is proposed to identify three significant instants of excitation in LPR for every 10 ms. At each instant of significant excitation, an impulse of appropriate amplitude is passed through an all-pass filter with constant phase rotation to obtain the modeled excitation signal (LPR). The optimum phase is determined by analysis-synthesis approach. Segmental SNR is computed in the time domain by comparing the original with the reconstructed speech. An improvement of about 3 dB is obtained compared to the reconstructed speech with multi-pulse excitation without phase rotation.*

I. Introduction

Physiologically, speech sounds are produced by positioning the various articulators at suitable locations thereby producing a complex 3-dimensional vocal tract (VT) shape [1, 2]. The nasal-tract is partially or totally coupled or decoupled to the vocal tract. Vocal tract being passive doesn't produce any audible sounds by itself. By means of aerodynamic processes, excess air from the lungs is modified in three major ways to produce different types of sounds [1, 2]. The three major ways are (a) voiced source as in the case of vowels (b) frication source (noise-like) as in the case of fricatives such as /s/, /sh/ and (c) transient source (impulse-like) as in the case of bursts of stop sounds such as /p/, /t/, /k/. A combination of these types of sources also occurs. For example, sound /z/ is produced by a combination of voice and frication sources; /ch/ is produced by a combination of frication and transient sources.

The radiated acoustic pressure of a speech signal in the atmosphere is modelled as the output of a linear filter excited by a suitable source [1, 2]. In the source-filter model, there are three major components: The source, the filter and the radiation. The major effect of VT is to modify the frequency response of the source and hence acoustically VT is represented in the frequency domain by a 'filter' or a 'transfer function'. In general, VT filter consists of both resonances (or formants or poles) and anti-resonances (or anti-formants or zeros). There two major effects of radiation: (i) The frequency shaping effect of radiation is to increase the bandwidths and shift the resonances of higher formants. This effect is merged with the VT filter. (ii) The second effect is to convert the air-flow at the lips to pressure waves in the atmosphere. This second effect has three parts: (a) Scaling effect that is determined by the inverse square law, i.e., the amplitude of a sound wave decreases inversely as the square of the distance (b) Directionality of the radiated acoustic waves, i.e., amplitude of sound waves is relatively larger along the axis of the lips whereas it is relatively very low towards the back of the head (c) The third most significant part is that the acoustic pressure is a differentiated version of the flow. Note that the acoustic wave reaching a listener or a microphone is the acoustic pressure in the free field (atmosphere). A notable exception is when a speaker talks into a mask with a closely spaced condenser microphone in which case the recorded signal is the air-flow at the lips.

For voiced sounds, when the source is a quasi-periodic sequence of air flow pulses, called glottal pulses, the output is the air-flow at the lips (Fig.1). When the differentiation effect of radiation is included into the source then the source may be considered as the differentiated version of glottal pulses, referred to as voice source pulses, in which case the output is the acoustic pressure wave in the atmosphere [3]. Typical glottal pulses and their derivative (differentiated version) are shown in Fig.2. The source is a broad bandpass type noise source for fricatives and an impulse-like source for stop bursts.

As the articulators move continually, the VT filter frequency response changes continually. The amplitude and the periodicity of voice source pulses change over the voiced segments of a speech signal. The amplitude and the frequency response of the noise source changes depending on the speech sound. In case of stop sounds there are rapidly changing sub-phonetic intervals such as closure (with or without voicing), burst release, voice onset time (with or without aspiration). However, for the purpose of analysis and synthesis (reconstruction), a speech signal is divided into frames of short intervals 20 to 40 ms and over these intervals VT transfer function and source characteristics are assumed to be stationary. In other words, a speech signal is assumed to be made up of a sequence of quasi-stationary frames.

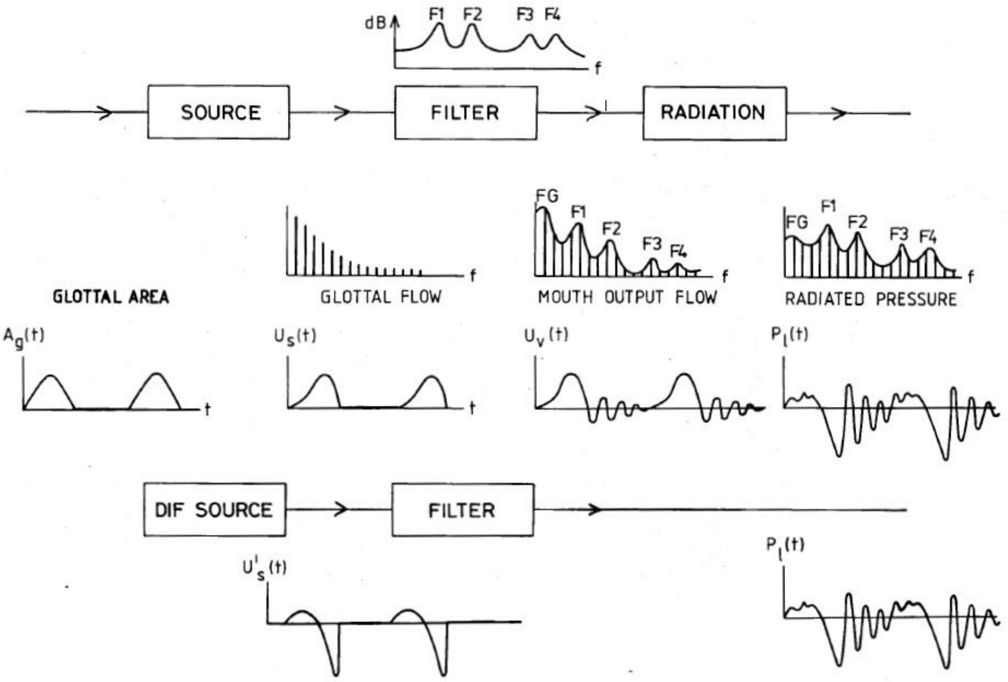


Fig.1. Source-Filter model for voiced sounds. Note that the differentiation effect of radiation can be shifted to be included in the source.

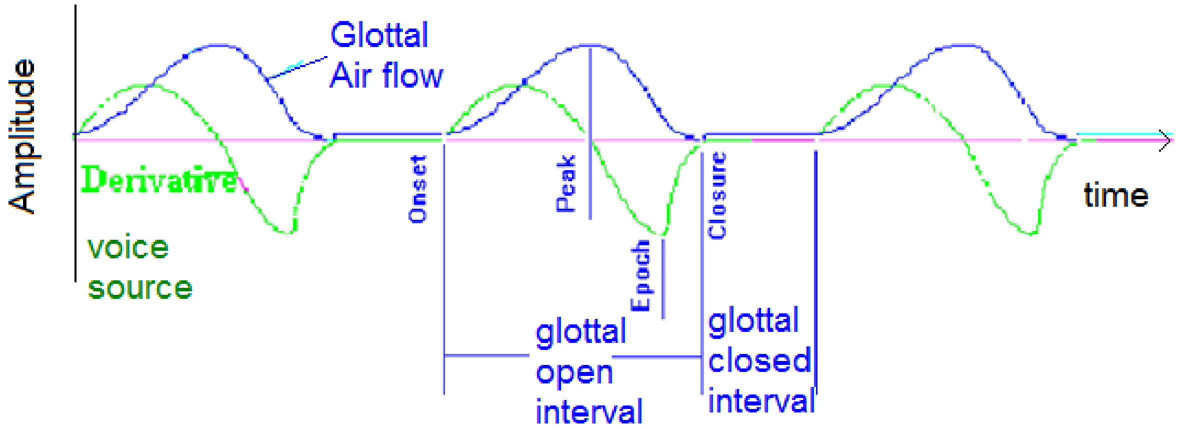


Fig.2. A sequence of typical glottal air flow pulses (in Blue) and its derivative (in Green). The differentiated version of glottal air flow pulses is referred to as voice source.

The earliest form of digital representation of a speech signal was PCM coding sampled at 8000 Hz with 8 bits per sample resulting in 64000 bits per sec (bps). Telephone communication in the past was with overhead lines that could support a digital signal only at 2400 bps. For secure communication of a digital signal over telephone lines it became necessary to compress PCM at 64000 bps to as low as

2400 bps. For the purpose of coding or compression of a speech signal, the source-filter model of speech sounds is grossly approximated in Linear Prediction (LP) vocoders [4, 5]. Here, for all sounds, the VT filter is approximated by a digital all-pole filter. For a speech signal sampled at 8000 Hz, LP (digital IIR) filter of order 10 is used. Source signal is also represented in a gross way. LP vocoders model the pre-emphasized (differentiated) radiated acoustic pressure of a speech signal (Fig.3). During reconstruction of speech signal at the receiver, a post-emphasis is applied to restore spectral balance.

A single pre-emphasized voice source pulse approximates an impulse typically located at the glottal closure instant or epoch (Fig.4). Any possible spectral shaping due to voice source is supposedly represented by the all-pole filter. Fricatives and stops are together referred as unvoiced sounds and the source for the unvoiced sounds is modelled by White noise. Any spectral shaping associated with the source of frication or burst and radiation is assumed to be represented by the all-pole filter.

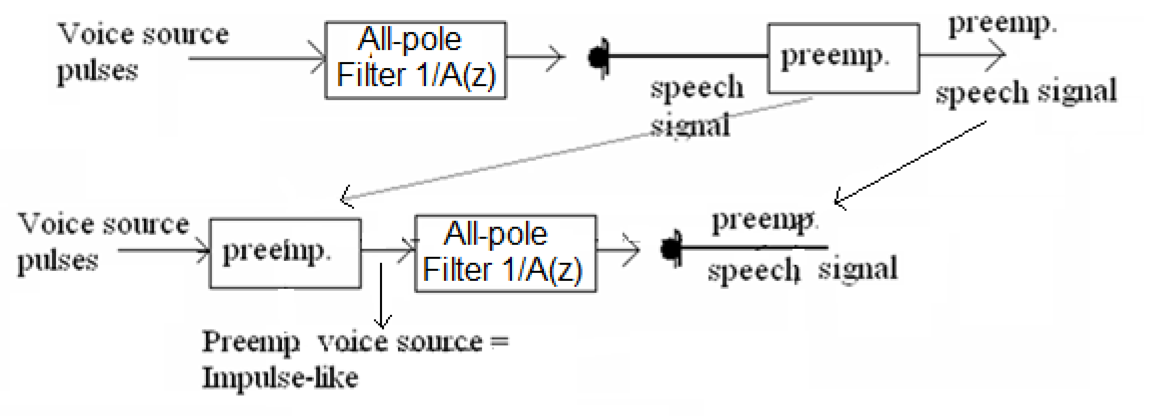


Fig.3. LP model for pre-emphasized voiced speech.

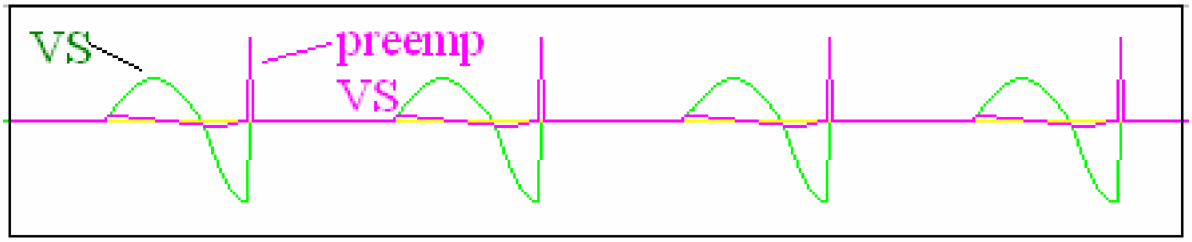


Fig.4. Typical synthetic voice source pulses with abrupt discontinuity at epoch (in Green). Pre-emphasized voice source is dominated by an impulse at epoch (in Pink).

Since the relative occurrence of voiced sounds in a speech signal is much more frequent compared to other sounds and since voiced sounds have relatively higher energy, greater importance is given to the reconstruction of voiced speech segments. In the case of voiced sounds a single impulse per pitch period is used as source or excitation. There is a lack of naturalness in the reconstructed speech signal in LP vocoders operating at 2400 bps. Reconstructed speech sounded 'buzzy', especially for voiced sounds. Also, using an impulse for every cycle (with a positive bias in the source or excitation signal) and subsequent post-emphasis (integration) introduces undesirable low frequency build-up. Various methods have been proposed to improve the naturalness of reconstructed speech [6-9]. For example, the use of zero-mean Hilbert transform of an impulse, sometimes with additive low energy white noise components as excitation signal.

Improved quality in the reconstructed speech signal can be achieved by using a better representation of source signal at the cost of increased bit rate. When a pre-emphasized speech signal is passed through the inverse of an all-pole filter, called the inverse filter, the output signal is referred to as LP residual. When LP residual is used as the input to the all-pole filter, pre-emphasized speech signal is perfectly reconstructed (See Fig.5). A segment of a vowel sound along with its LP residual is shown in Fig.6. Unlike a single impulse per cycle seen for a synthetic voice source (Fig.4), there are a large number of components in the LP residual for a natural voiced speech signal (Fig.6).

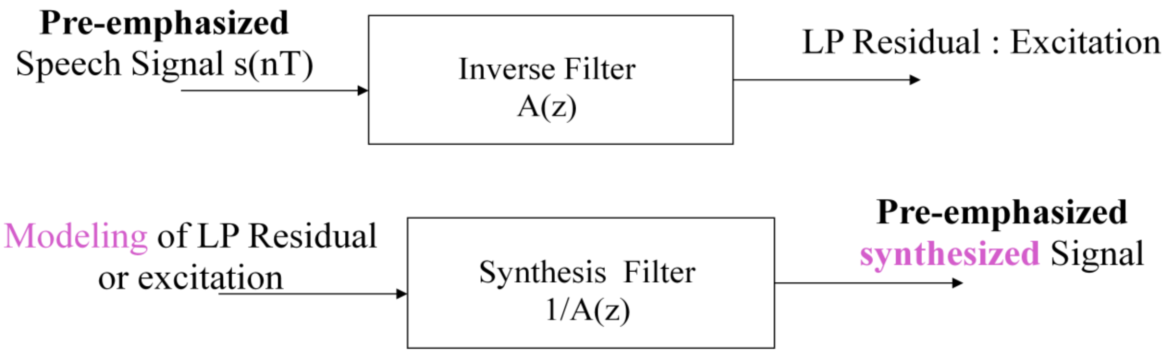


Fig.5. Inverse digital filer and the basis for modelling of LP residual

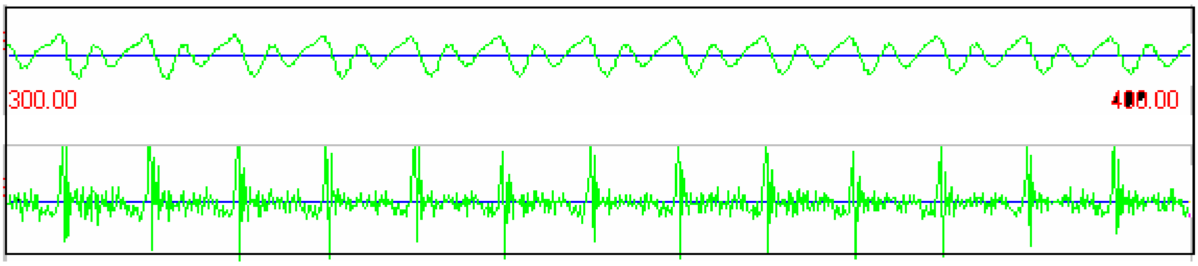


Fig.5. A segment of a vowel sound (upper trace) along with its LP residual (lower trace). Notice the bipolar swings in the LPR around epochal instants. Three or more large amplitude samples seem to be present around an epochal instant.

In order to achieve compression and at the same time to be able to reconstruct natural sounding speech signal, LP residual has been modelled by various researchers. Broadly there are two approaches: (a) open-loop and (b) closed-loop. In open loop coding, LP residual is low pass filtered to 800 Hz and transmitted as a digital signal. The bit rate with this type of vocoders is 9600 bps [10]. At the decoder, the coded residual is converted to a broad band signal before reconstruction. In the closed-loop multi-pulse coding, more than one impulse per cycle is used [11-14]. The optimum location and amplitude of impulse to be used for every 5 ms are determined by analysis-synthesis approach. That is, the parameters are so chosen that the mean square error between the original (pre-emphasized) signal and the reconstructed signal is minimum. The optimum multi-pulse is either determined dynamically or an optimum entry is determined from a pre-determined code-book with a large number of multi-pulse patterns.

Due to an approximation of VT transfer function by an all-pole filter and due to an approximation of the derivative of voice source by an impulse, there is an unmatched phase component in the LP residual. As a consequence, a single impulse of excitation spreads itself and appears with significant bipolar components in the LP residual [15, 16]. In the multi-pulse coding, the basic building block is an impulse. Supposing only one significant impulse is used for every 5 ms, then a large number of components in the LP residual around the excitation instant are left out. For example, a segment of a typical LPR for natural speech is shown in Fig.7. It may be noted that the modelled multi-pulse leaves out significant negative amplitude samples around the instants of significant excitation instants.

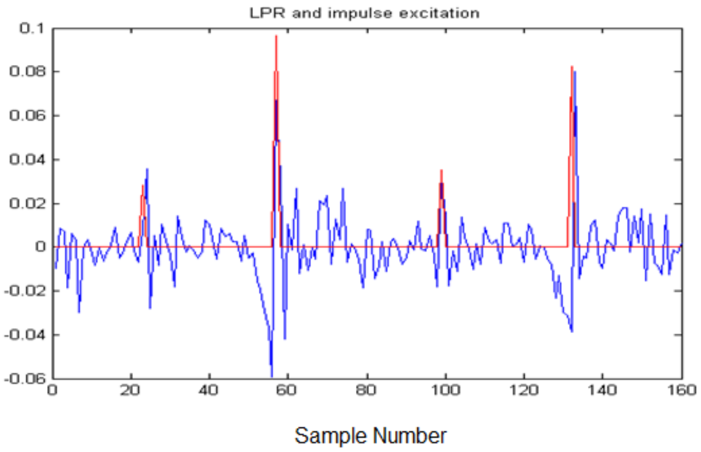


Fig.7. LPR (in Blue) and multi-pulse coding (in Red). Several samples of significant amplitude in the residual are left-out in multi-pulse coding.

II Proposed Method

In this work, an impulse is passed through an all-pass filter with a constant phase response. Such a phase-rotated impulse is used as the building block. By using such a building block, many of the significant components around the single excitation instant are captured. Analysis-synthesis approach has been used to determine the optimum phase angle.

Apply wavelet based method on segments of LPR to identify 3 significant instants of excitation within 10 msec. In the Wavelet based method, 3-level decomposition on LPR as well as on an impulse is used. Cross-correlation between detailed coefficients of level-2 of LPR and that of the impulse is computed. Three most significant peaks in the cross-correlation are selected.

Let ‘x(n)’ be an impulse and ‘x_h(n)’ be its Hilbert transform. Compute $x(n)\cos(\theta)+x_h(n)\sin(\theta)$. Find optimum θ based on signal matching in the residual domain around excitation instants. Take 10 samples around excitation instants from LPR for analysis-synthesis matching. Find the optimum θ.

An Example: LPR of a vowel segment and the modeled sequence of impulses with phase rotation are shown in Fig.8. It may be noted that bipolar components around the significant instants of excitation are captured with the phase rotation.

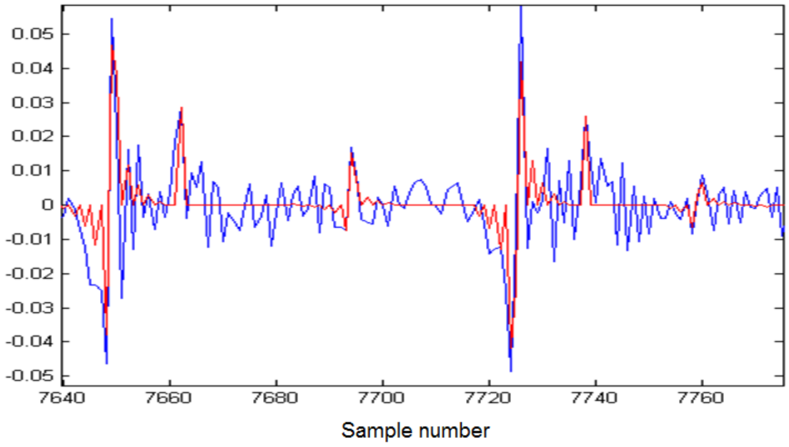


Fig.8. LPR (in Blue) and optimum phase-rotated sequence of impulses (in Red).

Performance: Pre-emphasized voiced speech signal is reconstructed with the optimum number of impulses with and without phase rotation as excitation signal. The difference signal between the original and reconstructed signal is considered as 'noise' and segmental SNR in dB is computed. It has been found that the use of phase rotation of the proposed method gives an improvement of the order of 3 dB for a large number of utterances.

III. Conclusion

Phase rotated impulse is proposed as a building block in multi-pulse coding. Since, the optimum phase angle is also to be transmitted, this would require additional bits. However, quantization of phase angle into eight levels doesn't degrade the performance very significantly.

References

1. Fant, G., *Acoustic Theory of Speech Production*, Mount Hague, 1960

2. Flanagan F. L., *Speech Analysis, Synthesis and Perception*, Springer-Verlag, Sec Edn. 1972

3. Ananthapadmanabha T. V., and Fant, G., Truncation and Superposition, STL-QPSR 2-3, 1982

4. Atal B. S. and Hanauer S. L., “Speech analysis and synthesis by linear prediction of speech wave”, *J. Acoust. Soc. Amer.,* vol. 50, no. 2, pp. 637-655, 1971.

5. Markel J. D., and Gray A. H., "A linear prediction vocoder based upon the autocorrelation method," IEEE Trans. Acoust. Speech and Sig. Proc., vol. ASSP-22, Apr. 1974, pp. 124-134

6. Sambur M. R. et. al., "On reducing the buzz in LPC Synthesis," J. Acoust. Soc. Am., vol. 63, No.3., 1978, pp. 918-924.

7. Makhoul J., "A mixed-source model for speech compression and synthesis," ICASSP 78, pp.163-166.

8. Atal B. S., and Nancy David, "On synthesizing natural-sounding speech by linear prediction", ICASSP-79, pp.44-47.

9. Akira Kurematsu et. al., "A linear predictive vocoder with new pitch extraction and exciting source," ICASSP-79, pp.69-72.

10. Un G. K., and Magill D. T., "The residual-excited linear prediction vocoder with transmission rate below 9.6 kbps," IEEE Trans. Comm., vol. COM-23, December 1975, pp. 1466-1474.

11. Atal B. S. and Remde, J., "A new model for LPC excitation for producing natural sounding speech at low bit rates," Proc. ICASSP-82, pp. 614-617, April 1982.

12. Singhal S. and Atal B. S., "Improving the Performance of Multi-pulse Coders at Low Bit Rates," Proc. ICASSP-84, p. 1.3.1, 1984.

13. Schroeder M. R. and Atal B. S., "Code-Excited Linear Prediction (CELP): High Quality Speech at Very Low Bit Rates," Proc. ICASSP-85, p. 937, Tampa, Apr. 1985.

14. Kroon P, Deprettere E., and Sluyeter R.J., "Regular-Pulse Excitation-A Novel Approach to Effective and Efficient Multi-pulse Coding of Speech," IEEE Trans. ASSP-34(5), Oct. 1986.

15. Ananthapamanabha T. V. and Yegnanarayana B., "Epoch extraction from linear prediction residual for identification of closed glottis interval," IEEE Trans. Acoust. Speech and Sig. Proc., vol. ASSP-27, No.4, Aug. 1979, pp. 309-319

16. Ananthapamanabha T. V. and Yegnanarayana B., "An interpretation of LP residual,", ICASSP- .

Improving the Accuracy of Estimation of Formant Data of Vowel Sounds Obtained Using LP Technique

T V Ananthapadmanabha and D R Raghavendra

Article No.1.4.

This work was carried out at Voice and Speech Systems, Bangalore during the period 2017-18.

Abstract: *The accuracy of estimation of formant data using the autocorrelation formulation of linear prediction (LP) technique is investigated empirically using synthetic vowels. In this article, the source of excitation for synthesis of a vowel is a periodic sequence of impulses. The error in the estimation of formant data decreases sharply with increasing duration of the analysis interval and the error saturates when the duration of analysis interval is greater than 30 ms. The pitch or the fundamental frequency F0 has a very strong influence on the error in the estimation of formant frequency as well as bandwidth. For a range of F0 between 70 to 210 Hz, the maximum error is significantly high (of the order of 10%) for formant frequency below 300 Hz and the percentage error sharply decreases with increasing value of formant frequency. Thus, the greatest inaccuracy is in the estimation of first formant frequency. The error in the estimation of formant frequency as well as the error in the estimation of bandwidth show a systematic oscillatory behaviour with respect to F0.*

The accuracy of estimation of formant data using LP may be improved upon by a simple method. This method uses a major finding that the error in the estimation of first formant is not influenced by the presence of other formants implying that the errors are the same for a four formant vowel as that for a single formant signal. Thus, a single formant signal can be used as a calibration signal. A vowel signal using known formant data (calibration data) and F0 is synthesized. Formant data of the synthesized signal is estimated by LP method. Thus, a relationship is established between the estimated data and the original calibration data. Given the estimated formant data and F0 of a natural vowel, calibration formant data are used as the revised estimates. The revised estimates are more accurate than the first-pass estimates. Both the formant frequency and bandwidths may be re-estimated in this method.

I. Introduction

Scope of the present study: The objective of the present study is to test the accuracy in the estimation of formant frequencies and bandwidths using the autocorrelation formulation of Linear Prediction (LP) technique [1-3]. The influence of F0 on the estimation of formant frequency has been reported previously [4, 5]. Compared to previous studies on this very same topic, the present study is much more detailed and exhaustive.

In this experimental study, vowel sounds with known formant data are synthesized. LP technique is applied on a chosen analysis interval of a synthesized vowel. Formant data are estimated from the LP all-zero polynomial using root-solving method. The estimated formant data are compared with the specified formant data used in the synthesis of vowels in order to compute the error in the estimation. The influence of following factors on the error is studied; the duration of analysis interval, location of analysis interval relative to the excitation (epochal) instant and the fundamental frequency F0. The last factor, viz., F0 has the most significant influence on the error. The error shows a systematic trend with respect to F0 that may be utilized to improve the accuracy of estimation.

Linear prediction (LP) Formulations: LP is one of the most powerful automatic technique for the estimation of formant frequencies and bandwidths [6, 7]. There are two formulations of LP technique; the covariance based method [1] and the autocorrelation based method [2].

When analysis is carried out over the closed glottis interval (CGI) of a vowel sound, covariance method gives a better accuracy [8]. Since the analysis interval is less than a pitch period, the accuracy of estimation is not influenced by F0 or the voice source pulse shape. However, identification of CGI of a vowel sound is an extremely difficult task. Further this approach assumes zero glottal flow during CGI, i.e., no leakage.

For an analysis interval much larger than a pitch period, say of three to four pitch periods (30-40 msec), there is not much of a significant difference in the results obtained by the two formulations. The autocorrelation method is relatively better in terms of computational efficiency and numerical stability. Hence, in this study we limit ourselves to the application of autocorrelation formulation of LP.

LP analysis: LP technique estimates a set of coefficients, {a0, a1,...aM}, where M is the order of LP. The z-transform of the sequence, A(z), corresponds to the reciprocal of an optimal all-pole digital filter. Typically M is twice the expected number of formants (complex conjugate pole pairs) within the folding frequency.

Estimation of formant data using LP technique: There are two approaches for estimating the formant data after obtaining LPCs. In one approach, the log magnitude squared spectrum of 1/A(z) is computed and the peak locations are identified to obtain formant frequencies. When two formants are close to each other they merge into a single broad peak in the spectrum. In order to resolve such peaks, peaks in the negative derivative of phase spectrum of 1/A(z) are identified to obtain the formant frequencies. The derivative of phase spectrum shows sharper peaks [9,10]. Improved resolution is obtained at the cost of emphasizing the spurious pole usually present in the LP analysis. The second derivative of log-spectrum [11] gives a much better resolution than the derivative of phase spectrum.

In the second approach, adopted in this work, the roots of the polynomial of A(z) are solved. The roots correspond to the locations of formants within the unit circle expressed in polar coordinates. Accordingly, the angle subtended by a vector from the origin to a root determines the formant frequency with 2π radians corresponding to the sampling frequency. The radial distance from the origin to the root is related to the bandwidth. The estimated formant data are later arranged in an ascending order as the root solving may pick up the roots in no specific order.

II. Experimental Setup

II.A. Synthesis of vowels: In the present study, a cascaded four formant filters is used in the synthesis. Formant filters are set to known values of formant data of a chosen vowel. Four choices of vowels are considered: /a/, Schwa (denoted as /E/), /u/ and /i/. Formant frequencies used for these vowels are as shown in Table-I. Bandwidths for all the formants and for all vowels, BW(i), i = 1 to 4, is set to be the same and equal to 50 or 100 or 150 or 200 Hz in the various experiments. Synthesis is done with a sampling frequency, Fs, of 8000 Hz.

Table-I
Formant Frequencies of Vowel Sounds
Used for the Experiments

Sl No	Vowel	F(1)	F(2)	F(3)	F(4)
1	/a/	730	1090	2500	3500
2	/E/	500	1500	2500	3500
3	/u/	300	900	2500	3500
4	/i/	270	2200	2500	3500

A steady vowel is synthesized with a constant F0 and constant amplitude of excitation. Experiments are conducted for a range of F0 from 70 to 210 Hz, in steps of 1 Hz. However, see Sec.III.B. There are two options for the source or excitation function: (a) Impulse or (b) voice source pulse. In this article the study is based on option (a). Study based on option (b) is planned for a volume of PRAVS relating to Voice Source.

II.B. Analysis: Synthesized vowel is of 100 ms duration, A frame of samples of 40 ms around the mid-part of the vowel is used for analysis. No pre-emphasis is applied since the source is an impulse function. However, windowing has been applied. An 8th order LP analysis is performed.

The error in the estimation of formant frequency is defined as

$$\text{EFF(i)} = [\ \text{Estimated F(i) - Acutal F(i)}\ ]$$

The maximum absolute relative error in per cent (MAE) is defined as:

$$\text{MAE in F(i)} = \text{Max of } [100\ |\text{EFF(i)}|\ /\ \text{F(i)}\ ]\text{, over all F0}$$

Similarly, the error in the estimation of bandwidth is defined as

$$\text{EBW(i)} = [\ \text{Estimated BW(i) - Actual BW(i)}\ ]$$

The maximum absolute relative error in per cent (MAE) is defined as:

$$\text{MAE in BW(i)} = \text{Max of } [100\ |\text{EBW(i)}|\ /\ \text{BW(i)}\ ]\text{, over all F0}$$

III. Results

III.A. Results on the influence of analysis interval

To study the influence of analysis interval on the accuracy of estimation, we set BW(i)=100 for i=1,4. MAE only for the first formant is illustrated since the error is much larger for the first formant compared to other formants. Analysis interval in the range of 10 to 50 ms in steps of 5 ms is considered. MAE in F(1) and MAE in BW(1) are plotted against the analysis interval for F0 variation in the range of 70 to 180 Hz in Fig.1.

It is seen that the MAE in F(1) as well as MAE in BW(1) sharply decrease and remain the same for analysis interval equal to or greater than 25 ms. Although the bandwidth BW(1) is the same (100 Hz) for all vowels, the maximum error in the estimation of BW(1) is higher for low F(1) cases.

One may wonder if there is any advantage of making the analysis interval equal to an integral multiple of pitch period? If this is so, then we expect the error to show valleys when analysis interval is an integer multiple of pitch period. Analysis results for the choice of integer as well as fractional values of pitch period are shown in Figs. 2a and 2b for F(1) and BW(1) respectively. MAE decreases smoothly as the analysis interval in increased up to about 3 pitch periods. MAE saturates for analysis interval beyond 3 pitch periods. In other words, there is no advantage in making analysis interval to be integer multiple of pitch period. Some researchers have proposed pitch synchronous analysis. In case of shorter analysis interval, covariance method is to be preferred as it doesn't require windowing. The presence of voice source pulse affects the covariance method. In the autocorrelation method of LP, windowing has to be done. For short analysis interval, windowing distorts the signal. Thus an analysis interval of one pitch period gives rise to larger errors. In another experiment (not illustrated) it is found that the location of excitation instant relative to the mid location of window function has no significant effect on the error.

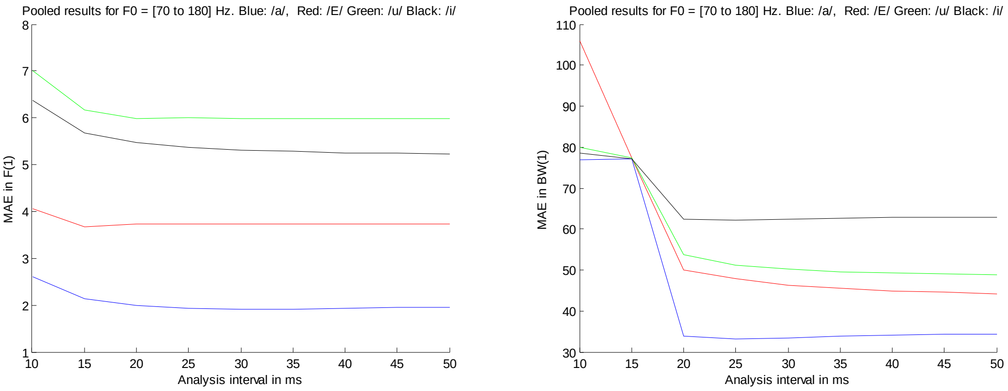


Fig. 1. The maximum absolute error MAE in (a) F(1) and (b) BW(1) as a function of analysis interval. *MAE is computed across F0 = 70 to 180 Hz. Vowel /a/ (Blue),vowel /E/ (Red), vowel /u/ (Green, vowel /i/ (Black).*

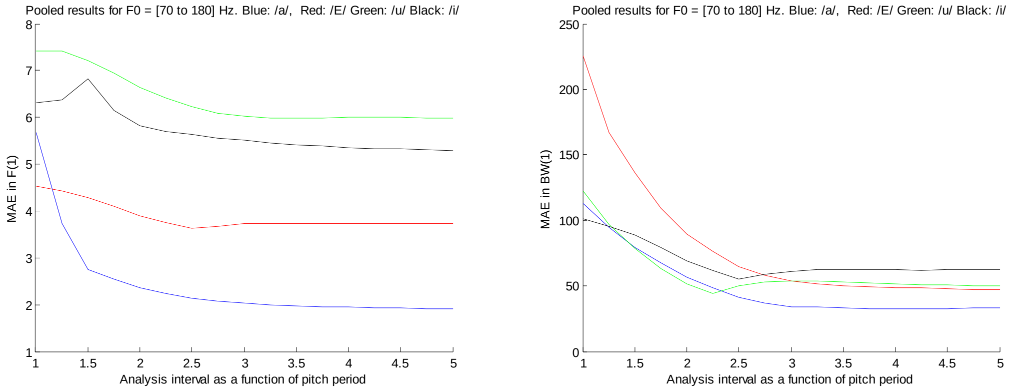


Fig. 2. The maximum absolute error MAE in (a) F(1) and (b) BW(1) as a function of analysis interval as related to pitch period. *MAE is computed across F0 = 70 to 180 Hz. Vowel /a/ (Blue),vowel /E/ (Red), vowel /u/ (Green, vowel /i/ (Black).*

A larger analysis interval (30 to 40 ms) appears to be a preferred choice. This is not a major issue in the analysis of a steady vowel. However, for analysis of vocalic segments during CV or VC transitions where there occur rapid changes in formant frequencies, the estimated formant data would then correspond to the average values over the analysis interval. A discussion on this issue is not a subject matter for the present study.

III. B. Results on the influence of F0 on the estimation of formant frequencies

An analysis interval of 40 ms has been used to study the influence of F0 over the range of 70 to 210 Hz in steps of 1 Hz. Synthesis is done in the time domain. Pitch period has to be integral number of samples. The pitch period in number of samples (NT0) is Fs/F0. For F0=70 Hz (or 210 Hz), 8000/70=114.28 (or 8000/210=38.09) which has to be converted to 114 (38) samples. For F0=144 and 145, NT0 happens to be the same, viz., 55 samples. In the experimental studies reported henceforth, NT0 is varied between 114 to 38 in steps of -1. However, the results (plots) are presented with respect to F0 (computed as Fs/NT0).

Relative Error: Figure 3a shows MAE Vs formant frequency across the entire range of F0 for all the four synthetic vowels for the choice of bandwidth=100 Hz. In the case of first formant, MAE varies from about 4% (for the highest F(1) of /a/) to 10% (for the lowest F(1) of /i/); In the case of second formant MAE varies from about 1.5% (for the highest F(2) of /i/) to about 4% (for the lowest F(2) of /u/); MAE is about 1.2% for F(3) and about 0.6% for F(4). MAE decreases almost exponentially with the increase in the value of formant frequency. Results on the dependence of error in the estimation of formant frequencies as a function of bandwidth is presented in Sec. III.D.

In order to bring out the dependence of the error on F0, only the error in F(1) is illustrated in Fig.3b. EFF(1) in percent Vs F0 Hz shows an oscillatory behaviour. The oscillatory behaviour of the error is discussed further below. For any given F0 (in the vertical direction), EFF(1) increases with a decrease in F(1). The largest error occurs for vowel /i/ with the lowest F(1) and the smallest error occurs for vowel /a/ with the highest F(1). By this logic, EFF(i) for i>1 is expected to decrease with increasing values of F(2) to F(4). This seems to suggest that closer the spacing between F0 and F(1), larger is the error. If this is indeed so then an analysis of 'mirror spectrum' should give better results. See Appendix-A on testing of this hypothesis.

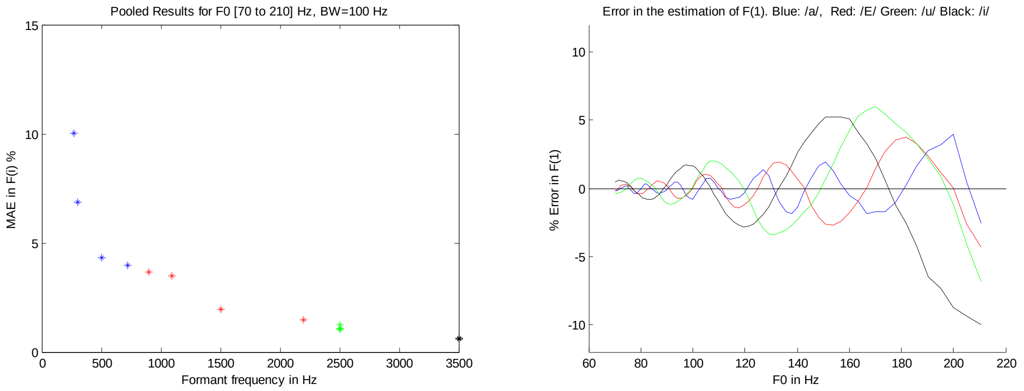


Fig. 3. (a) Dependence of the Maximum absolute relative error (MAE) on the values of formant frequencies. Results are pooled for all the four synthetic vowels. (b) Error in the estimation of F(1) in percent Vs F0 in Hz for four synthetic vowels. BW=100 Hz.

Magnitude of error: A relative error of 10% in the estimation of F(1) for vowel /i/ may appear high, but the magnitude of the error is 27 Hz. Similarly, relative error of 4% in F(1) of vowel /a/ may appear low but the magnitude of the error of is 28.8 Hz, which is nearly the same as that of F(1) for vowel /i/. An error of 3.66% for F(2)=900 Hz and 1.46% for F(2)=2200 Hz correspond to about 32 Hz. An error of 1.26% in F(3) is about 32 Hz and an error of 0.6% in F(4) is about 21 Hz. Hence, it is better to consider the magnitude of error in Hz rather than the relative error in percent.

Oscillatory behaviour of EFF(1): In order to bring out the dependence of the error on F0, only the error in F(1) is illustrated. In this case the bandwidth is chosen as

100 Hz for all the formants and for all vowels. A plot of EFF(1) in Hz Vs F0 in Hz (Fig.4a) shows an oscillatory behaviour. A probable reason for the oscillatory behaviour of error is discussed in Appendix-B. The number of oscillatory cycles in EFF(1) increases with increasing F(1). That is, vowel /a/ has the largest and vowel /i/ has the least number of cycles of EFF(1).

There are certain values of F0 for which the error is zero. The values of F0 where the error reaches zero depends on the value of F(1). Hence the locations of zero-crossing in EFF(1) don't coincide for different vowels. Further, for a given vowel, the spacing between two successive zero-crossings in EFF(1) is not uniform and the spacing increases as F0 increases.

As F0 increases, the peak-to-valley excursion increases. The magnitude of the maximum negative error of -27 Hz for vowel /i/ is nearly the same as the magnitude of the maximum positive error of +28.4 Hz for vowel /a/. EFF(1) at successive peaks and valleys increases as F0 increases. The spacing between successive peaks or successive valleys increases with increase of F0. The locations of peaks or valleys in EFF(1) don't coincide for the different vowels.

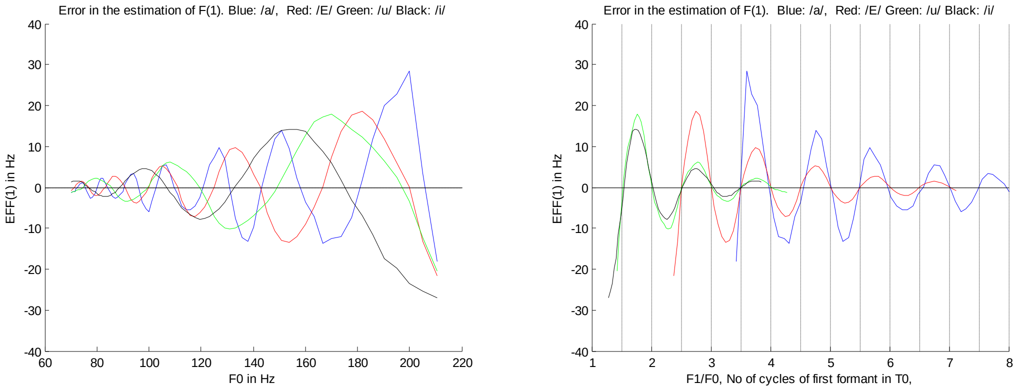


Fig. 4. (a) Error in the estimation of F(1) in Hz Vs F0. (b) Error in the estimation of F(1) in Hz Vs F1/F0. Vowel /a/ (Blue),vowel /E/ (Red), vowel /u/ (Green, vowel /i/ (Black). Bandwidth=100 Hz.

Harmonicity: A better insight on the behaviour of error with respect to F0 is obtained by plotting EFF(1) in Hz Vs F(1)/F0. See Fig.4b. The ratio F(1)/F0 is henceforth denoted as H(1), which represents the number of cycles of the first formant within one pitch period in the time domain. For example, for vowel /a/, when F0=180 Hz, H(1)=4 implying that there are four full cycles of F(1) in one pitch period in the time domain. In the spectrum this implies that the fourth harmonic coincides with F(1). Depending on the relative values of F(1) and F0 there could be a whole number of cycles plus a fractional part of cycle of F(1) within a pitch period in the time domain. The lower and upper limits of H(1) is different for different vowels for the same range of F0 considered. For vowel /i/ with the lowest F(1), the range of H(1) is 1.22 (270/220) to 3.85 (270/70) whereas for vowel /a/ with the highest F(1), the range is from 3.22 (720/220) to 10.28 (720/70). In Fig.4b, the x-axis is shown for a range of (1 to 8).

The EFF(1) Vs H(1) also shows an oscillatory behaviour. Unlike, the plot of EFF(1) Vs F0 (Fig.4a), in the plot of EFF(1) Vs H(1) (Fig.4b) the zero-crossings, peaks and valleys for different vowels align over the overlapping values along x-axis. Error in F(1) is zero for all vowels for H(1) = 2, 3, 4, 5 (dashed vertical lines) etc. This is to be expected since in such cases the spectral peak at F(1) coincides with one of the harmonics in the log spectrum. The error in F(1) is zero for all vowels also for H(1) close to 2.5, 3.5, 4.5, 5.5 (dashed vertical lines) etc, which is surprising. This may arise since the harmonics are equally spaced on either side from the spectral peak at F(1).

Peaks of positive EFF(1) occur for all vowels near H(1) = 1.75, 2.75, 3.75 etc. Valleys of negative EFF(1) occur for all vowels near H(1) = 2.25, 3.25, 4.25 etc. For a given vowel, the maximum error occurs near the lower end of H(1). The error at successive positive (or negative) peaks decreases exponentially. This can be seen clearly for the case of vowel /a/. Can such a systematic trend in the EFF(1) Vs H(1) be captured by a rule? This is discussed in Sec. IV.

Influence of bandwidth on the estimation of formant frequencies: The above results have been derived empirically for a choice of BW(i)=100 Hz, i=1 to 4. The maximum error in the estimation of formant frequencies depends also on the value of bandwidth in addition to F0. Fig.5 shows the results for four different choices of the bandwidth, 50, 100, 150 and 200 Hz. MAE decreases sharply as the bandwidth increases as seen along a vertical line for a given F(1). For example, for the first formant of vowel /i/, MAE decreases from 12% (for BW(1)=50 Hz) to 5.4% (for

BW(1)=200 Hz). For the first formant of vowel /a/, MAE decreases from 4.6% to 2.3%. For an increase in bandwidth from 50 Hz to 200 Hz, MAE in F(1) decreases almost by a factor of 2.

EFF(1) in percent Vs H(1) for four different choices of bandwidth for synthetic vowel /a/ shows an oscillatory behaviour (Fig.5b). The frequency of oscillation is the same for the different choices of bandwidth. This result has been observed for other vowels as well but not illustrated here. For any given choice of H(1) (along a vertical line), the error decreases with increasing bandwidth. Also, the error at successive peaks (or valleys) decrease exponentially.

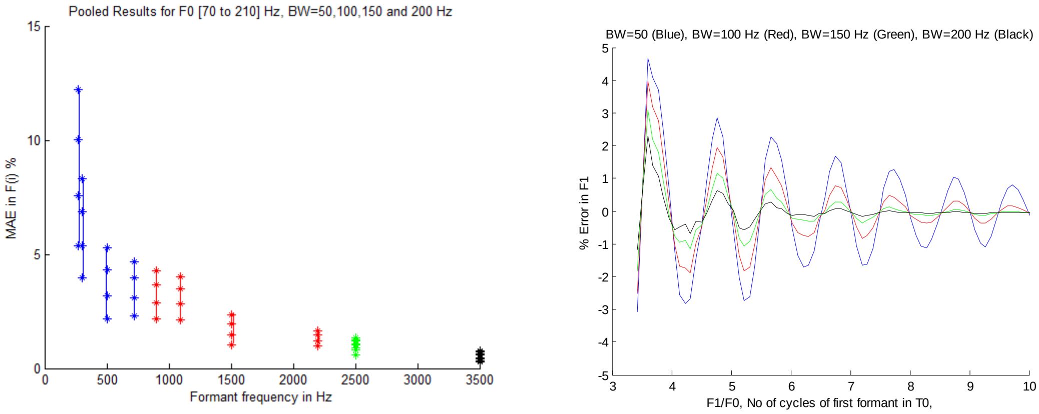


Fig. 5. (a) Influence of bandwidth on MAE (b) Error in F(1) Vs F0 for four choices of Bandwidth for vowel /a/

III.C. Influence of F0 on the Estimation of Bandwidths

Four test bandwidths are used: 50, 100, 150 and 200 Hz for all the four formants and four vowels. Vowels are synthesized over the entire F0 range of [70 to 210] Hz. Pooled results for all vowels and for the entire range of F0 in shown in Fig.6. MAE in BW(1) in percent Vs test bandwidth and MAE in BW(1) in Hz Vs test bandwidth are shown in Fig.6a and Fig.6b respectively. The range of error (not the estimated bandwidth) is (a) [150 to 100] Hz for test BW=50 Hz; (b) [100 to 70] Hz for test BW=100 Hz; (c) [70 to 45] Hz for test BW=150 Hz; and (d) [30 to 50] Hz for test BW=200 Hz.

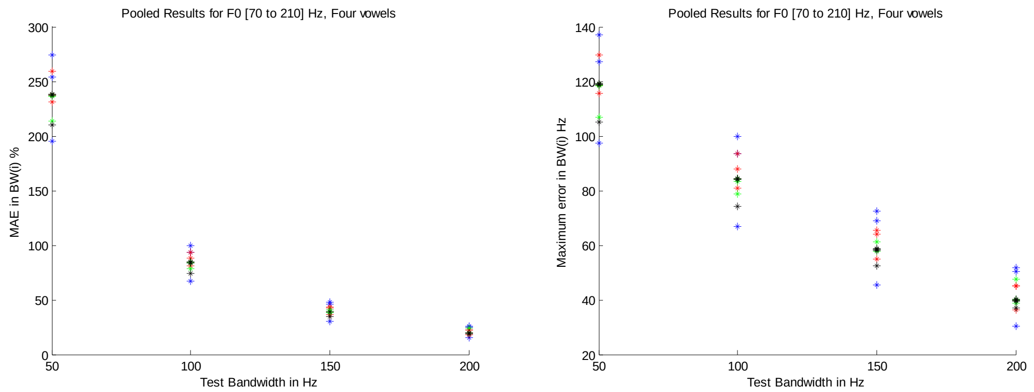


Fig.6. MAE in BW(1) Vs test bandwidth (a) in percent (b) in Hz for four different choices of bandwidth. The spread in the results arises owing to four formants of four different vowels.

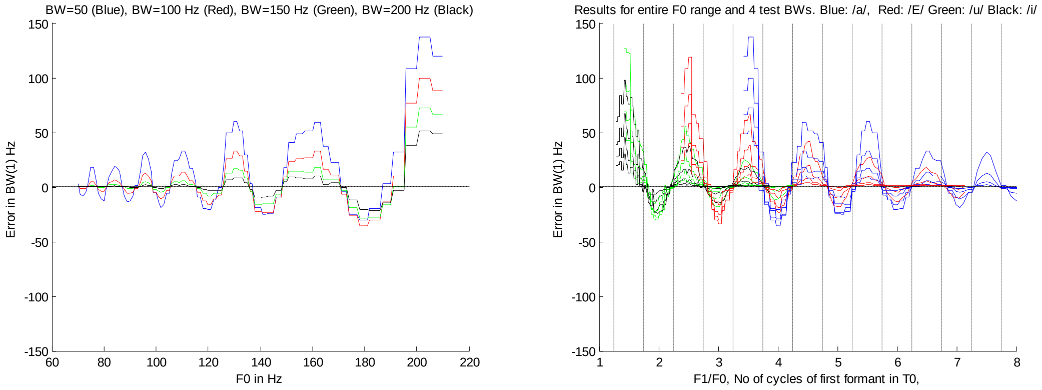


Fig.7. (a) EBW(1) Vs F0 for vowel /a/ and for four different choices of test bandwidth. (b) EBW(1) in Hz Vs H(1) for four vowels and four different choices of bandwidth.

Oscillatory behaviour of EBW(1): In order to bring out the dependence of the error on F0, the error in BW(1) is illustrated in Fig.7a for vowel /a/ as a function of F0 and for four different choices of test bandwidth. The error shows an oscillatory behaviour. For any given F0 (along the vertical direction), EBW(1) increases with decrease in BW(1). The maxima and minima in the error increases with increase in F0. Generally there is a larger over estimation (peak positive error) rather than an under estimation (peak negative error) in BW(1). The spacing between successive zeros or peaks or valleys in EBW(1) is not the same. The location of zero error differ for different test bandwidths even for the same vowel. Similar results are seen for other vowels but not illustrated for the sake of clarity as the results overlap heavily. As seen earlier in Fig.4a for EFF(1), even for EBW(1), the locations of zero-crossings, peaks and valleys don't coincide for the different vowels.

A better insight can be obtained by plotting EBW(1) in Hz against H(1) as shown in Fig.7b. The oscillatory behaviour is similar for all the choices of test bandwidth. For any given choice of H(1) (along a vertical line), the error decreases with increasing bandwidth. The locations where the error reaches zero or a peak or a valley is the same for all vowels and the spacing between successive zero-crossings or peaks or valleys is the same for all vowels over the applicable range of H(1). Also, for the same vowel, the error at successive peaks or valleys decrease exponentially. Generally there is a larger over estimation (positive error) rather than an under estimation (negative error) in BW(1).

III.D. Influence of Other Formants on EFF(1) and EBW(1)

In the above sections we have reported results on the influence of F0 and influence of bandwidth on the estimation of F(1). How much is the influence of other formants, F(2) to F(4) on the estimation of F(1)? To address this issue, two experiments are conducted: (i) The influence on the error for four different choices of F(2) is studied, keeping F(1), F(3) and F(4) the same (ii) EFF(1) for a signal synthesized with four formants is compared against EFF(1) obtained for a signal synthesized with a single formant.

III.D.1. Influence of F(2) on EFF(1) and EBW(1)

A four formant vowel sound is synthesized with F(1)=720 Hz and for four different choices of F(2) = 900, 1000, 1200 and 1500 Hz. F(3) and F(4) are set to be 2500 and 3500 Hz, respectively. Bandwidth for all four formants is set to be 100 Hz. F(1)=720 has been chosen since it gives the widest range of H(1). Fig. 8a shows the plot of EFF(1) Vs H(1). Fig.8b shows the plot of EBW(1) Vs H(1).

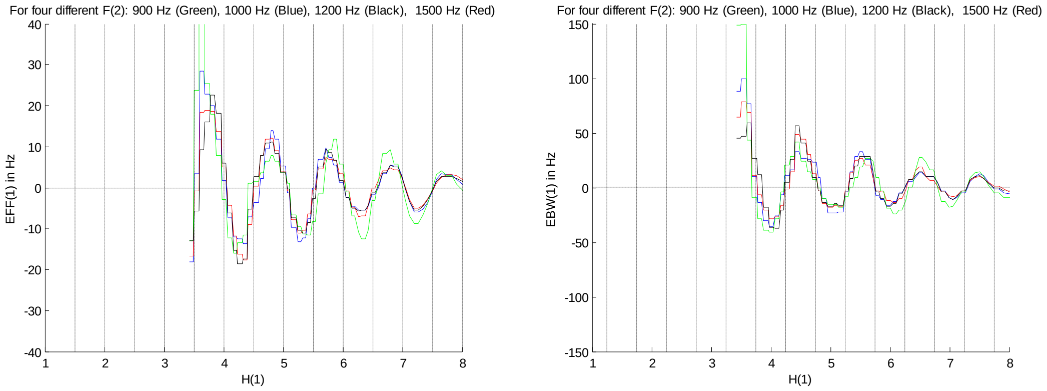


Fig. 8. Four formant vowel. F1=720 Hz and four different choices of F(2), F3=2500 Hz, F(4)=3500 Hz. (a) EFF(1) Vs H(1). (b) EBW(1) Vs H(1)

EFF(1) is nearly independent of F(2), except for the case of F(2)=900 Hz. Similarly, EBW(1) is nearly independent of F(2) except for F(2)=900 Hz. There is a very large error for the case of F(2)=900 since in this case there is a cross-over of the 3-dB down bandwidth frequencies of the first two formangs: F(1)+100=820 Hz, F(2)-100=800 Hz.

So, it can be said that as long as two formants don't overlap, the behaviour of EFF(1) and EBW(1) is independent of F(2). There is only a slight difference in the values of maxima for cases other than F(2)=900 Hz around H(1)=3.75.

III.D.2. Error for four formants case and error for a single formant case

Assuming non-overlapping of formants, since EFF(1) and EBW(1) are nearly independent of F(2), we expect the results to be independent of third and fourth formants as well since F(3) and F(4) are much greater than F(2). To verify this, results of a four formant vowel /a/ is compared against a single formant case with F(1)=720 Hz. The former signal is analyzed with M=8 and the latter signal is analyzed with M=2. The results are shown in Fig.9. The results for the two cases overlap for the entire range of F0, except for a very slight difference in the value of maximum error.

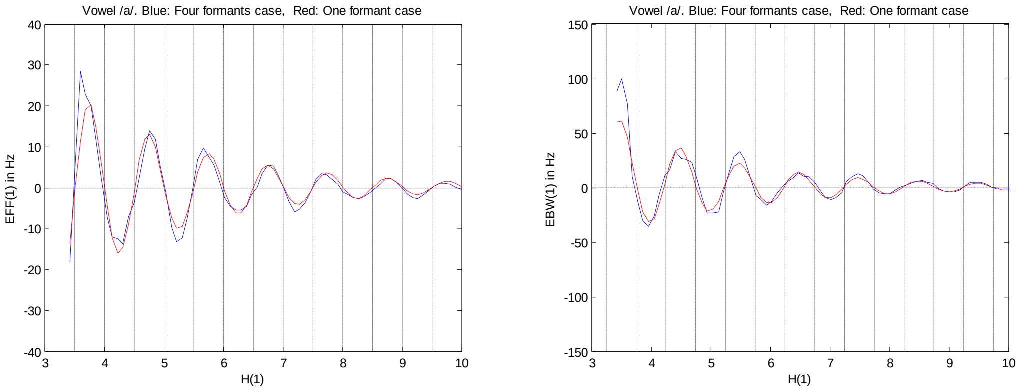


Fig. 9. Four formant case (in Blue) and One Formant case (in Red) for vowel /a/. Bandwidth=100 Hz. (a) Error in the estimated formant frequency (b) Error in the estimated bandwidth

The important consequence of this experiment is that EFF(1) Vs H(1) of a vowel with four formants, F(1) to F(4), is nearly the same as EFF(1) Vs H(1) of a signal synthesized with only one formant, F(1). It demonstrates that estimated errors in formants don't affect one another. Formants are considered to be nearly orthogonal as per LP analysis technique.

III.D.3. Modelling the Oscillatory Behaviour of EFF(1) and EBW(1)

Fig.10a shows EFF(1) Vs H(1) for a single formant case. Two examples are shown, one with F(1)=400 Hz and the other with F(1)=700 Hz. For both examples, BW(1)=100 Hz. Fig.10b shows EBW(1) Vs H(1) for a single formant case for the same two examples considered above.

When x-axis is chosen as H(1), the spacing between successive zeros is uniform. EFF(1) can be modelled by an exponentially damped sinusoid of the form y = A0 exp(-α(x-x0)) cos(2π (x-xf0)), where xf0 is the location of the first positive peak, A0 is the value of the error at xf0. Let the value at the next positive peak be A1. Then alpha=ln(A0/A1). Similarly, EBW(1) can be modelled by an exponentially damped sinusoid of the form y=B0exp(-β(x-xb0)) cos(2π(x-xb0))

The modelled (rule based) curves are shown in Red in Figs.11a and 11b. There is a very good match between the EFF(1) and the modelled curve. For EBW(1) match is good for x>xb0 with a slight mismatch for x<xb0. A family of such curves can be

generated for different choices of bandwidth with different (A0, xf0, alpha) and (B0, xb0, beta).

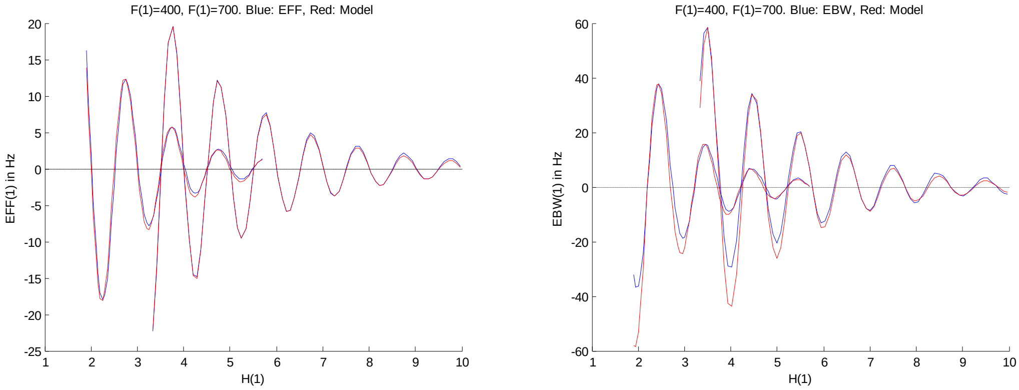


Fig.10. (a) EFF(1) Vs H(1) (in Blue) and modelled curve (in Red) (b) EBW(1) Vs H(1) (in Blue) and modelled curve (in Red)

IV. Improving the accuracy of estimation of F(1)

It is possible to capture the systematic variation in EFF(1) Vs H(1) or EBW(1) Vs H(1) by a rule as seen above. Assume that a rule for EFF(1) Vs H(1) or EBW(1) Vs H(1) can be formed. Can such a rule be used for correcting the formant data estimated by LP technique? In the experiments above H(1) is computed as F(1)/F0, where F(1) is the actual first formant frequency. During analysis only the estimate of F(1) and estimate of BW(1) are available and not the actual values. Let H'(1) be defined as [Estimated F(1)]/F0. EFF(1) Vs H(1) and EFF(1) VS H'(1) are to be compared to see if a derived rule for the former case can be carried over to the latter case. Similarly, EBW(1) Vs H(1) and EBW(1) Vs H'(1) are to be compared to see if a derived rule for the former case can be carried over to the latter case.

IV.A. EFF(1) Vs H'(1)

Plot of EFF(1) Vs H(1) (in Blue) as well as the plot of EFF(1) Vs H'(1) (in Red) are shown in Fig.11a for vowel /a/. Plot of EBW(1) Vs H(1) (in Blue) as well as the plot of EBW(1) Vs H'(1) (in Red) are shown in Fig.11b for vowel /a/. It is seen that the two plots are slightly shifted relative to each other suggesting that one can make use of H'(1) in place of H(1).

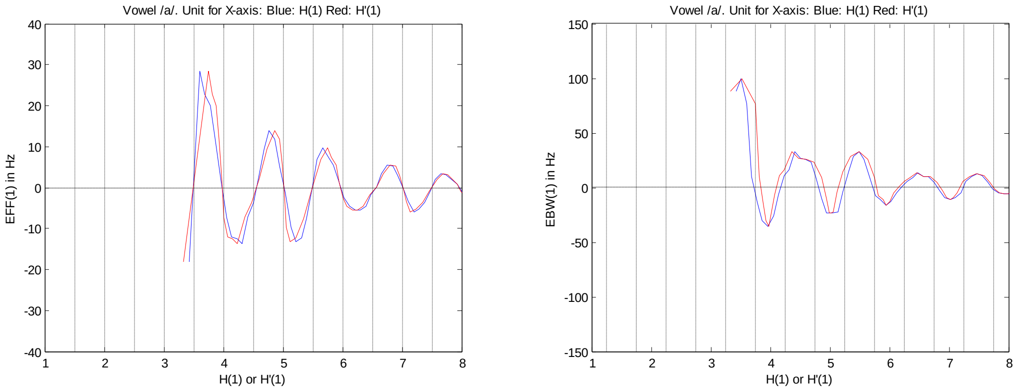


Fig.11. (a) EFF(1) Vs H(1) and EFF(1) Vs H'(1) for vowel /a/ (b) EBW(1) Vs H(1) and EBW(1) Vs H'(1) for vowel /a/

IV. B. Improving the accuracy of estimated formant data

There are two approaches for improving the accuracy in the initial estimates of formant data: (a) By Table look-up (b) By Analysis-synthesis approach. The procedure of re-estimation (or correction) can be considered to be an inversion process.

Based on the findings reported earlier, re-estimation of formants can be done one formant at a time. We illustrate the procedure of these two approaches with respect to improving the accuracy of estimation of only the first formant. The same procedure may be used for improving the accuracy of estimation of other formants as well, if required. Some preliminary steps are to be followed that are common to both the approaches and these preliminary steps are given below.

IV.C. Look-up Table approach

Generating the Tables

(i) A single formant signal is synthesized with the input variables F(1), BW(1) and F0. Let us refer to these input values as calibration data.

(ii) A wide range of values is used for each of the input variables and with a step size equal to the desired accuracy. As an example, for synthesis, a range of [270 to 720] Hz in steps of 5 Hz may be used for calibration-F(1). An appropriate range may be chosen if the interest is to correct error in the estimation of F(2). For calibration-BW(1) a range of [50 to 200] Hz in steps of 20 Hz may be used. A range of [70 to

220] Hz in steps of 1 Hz may be used for calibration-F0. In practice, user may select the range and step size as desired.

(iii) For each triad of input calibration data, a single formant signal is synthesized. Synthesis is done with a sampling frequency, Fs, which has to be the same as that of natural vowels required to be analyzed.

(iv) Subsequently, analyze synthesized signal using LP technique with order M=2. Formant frequency and bandwidth are estimated from LPCs of order 2. For a second order LP polynomial the estimated formant frequency and estimated bandwidth can be obtained using closed form expression instead of root-solving. Let [1, b1, b2] be the estimated LPCs. Let ff1 and bw1 be the estimated values of formant frequency and bandwidth, respectively for the single formant case. The relations are derived as follows:

Since $b2 = \exp(-2\pi\, bw1/Fs)$, $bw1 = (Fs/2\pi) \ln(1/b2)$

since $b1 = 2 \cos(2\pi\, ff1/Fs) \exp(-\pi\, bw1/Fs) = 2 \cos(2\pi\, ff1/Fs)\, (b2)^{0.5}$

$$ff1 = (Fs/2\pi) \cos^{-1}[\, - (b1/2)\, (b2)^{-0.5}\,]$$

Phase-I: Fill-up a Table-FF with columns representing F0 and rows representing estimated formant frequency, ff1. Each cell in Table-FF points to yet another Table, viz., Table-BW whose columns are F0 and rows are the estimated bw1. The cell in Table-BW has a pair of entries, viz. the calibration [F(1), BW(1)] used for preparing the Tables.

Phase-II: Given a natural vowel, apply LP analysis and obtain the initial estimates for the first formant frequency and bandwidth. Identify the cell in Table-FF and subsequently the cell in Table-BW and assign the calibration data to be the re-estimates of F(1) and BW(1).

Example-1: In Peterson and Barney (P&B) formant data [12], consider male speaker 1, vowel '/uh' with measured data F0=158 Hz. F(1)=660 Hz and F(2)=980 Hz. These measured data are assumed to be initial estimates obtained by LP technique. H(1) for the measured data corresponds to H'(1)=4.17. One can expect a very low error in F(1) and a large error in BW(1) for this value of H'(1). In P&B Data, bandwidths are not published. Assume that these initial estimates are to be corrected. Use a range of calibration input data of NT0=[52 to 47] Hz in steps of -1 which corresponds to a range of F0=[153.8 170.2] Hz. Use the range for Calibration-F(1)=[620 to 700] Hz in steps of 10 Hz and BW(1)=[50 to 200] Hz in steps of 25 Hz. The results are sorted with respect to estimated ff(1) and part of the results where estimated F(1) is close to 660 Hz are shown in the Table below. The last column labelled 'Err1' will be explained in Section IV.D.

Sl No	F0	Est F1	Est B1	Calib F1	Calib B1	Err1
1	163.265	659.588	191.163	660	200	0.241655
2	156.863	659.609	132.022	670	125	0.097158
3	166.667	659.988	41.622	640	50	0.134939
4	166.667	660.088	77.15	650	100	0.056023
5	160	660.346	119.806	670	125	0.072722
6	170.213	660.376	80.212	640	75	0.060571
7	**156.863**	**661.048**	**91.466**	**680**	**50**	**0.007332**
8	163.265	661.182	45.1	680	50	0.116976
9	153.846	661.604	141.859	670	125	0.129117
10	163.265	661.705	80.42	670	100	0.038725
11	160	661.852	85.803	680	75	0.016498
12	166.667	662.007	189.676	660	200	0.249842

Entries corresponding to the calibration data that are closest to the measured data (158, 660, ?) are shown in Red. There are four choices (rows shown in Red) which are closest to the measured F0 and measured F(1). The ambiguity can be resolved if estimated bandwidth is also available. The final choice also depends on the relative confidence in the measurement or the relative importance of the three parameters. See the next Section.

Example.2: We make use of formant data published by Peterson and Barney [12]. These are measured from spectrograms. However, let us assume them to be initial estimates obtained by applying LP technique in order to illustrate the procedure. In Peterson and Barney (P&B) formant data [12], consider female speaker 38, vowel '/i/' with measured formant data F0=187 Hz. F(1)=330 Hz. H(1) for the measured data corresponds to H'(1)=1.76. One can expect a large error in F(1) and very low error in BW(1) for this value of H'(1). In P&B Data, bandwidths are not published. Assume that these initial estimates are to be corrected. Use a range of calibration input data of NT0=[46 to 40] Hz in steps of -1 which corresponds to a range of F0=[173.91 200] Hz. Let the range for Calibration-F(1)=[290 to 370] Hz in steps of 10 Hz and

BW(1)=[50 to 200] Hz in steps of 25 Hz. The results are sorted with respect to measured F0 since expected error in the estimation of F(1) is large (H'(1)=1.76). A part of the results are shown below.

SL No	NTO	F0	Est F1	Est B1	Calib F1	Calib B1
1	43	186.047	331.794	94.277	310	50
2	43	186.047	329.033	106.127	310	75
3	43	186.047	339.506	108.278	320	100
4	43	186.047	336.006	128.644	320	125
5	43	186.047	332.851	150.906	320	150
6	43	186.047	330.16	174.398	320	175
7	43	186.047	327.952	198.653	320	200

Entries closest to F0=186 Hz are shown in Red in the above Table. If the estimated bandwidth is low then the re-estimated F(1) is 310 Hz and if the estimated bandwidth is large then the re-estimated F(1) is 320 Hz. If the re-estimated bandwidth is available then the ambiguity may be resolved. Also, re-estimated or corrected BW(1) can also be determined. Since H'(1) is close to the location of maximum error, F(1)=310 may chosen as the corrected value.

Comments on Look-up Table Approach: Look-up Table approach requires a huge memory but computational requirement is very low. Look-up Table approach is quiet general and may be utilized even when the error doesn't show any systematic trend or when there is a mutual dependence amongst formants. In general, a multi-dimensional Table can be prepared mapping one set of actual values for the input variables (calibration data) to another set of values corresponding to measurable parameters. Given the measured parameters, by inversion the calibration data may be found, which are more accurate estimates. For example, Atal et al [13] report a mapping from 4-dimensional articulatory space (vocal tract area function model) to a 3-D acoustic measurement space (first three formant frequencies) and present an inversion algorithm that can be used to go back from the measurement space to the input variables space. If there are multiple solutions (one-to-many mapping), additional constraints have to be used to resolve the ambiguity. Although it is well known that F0 strongly influences the estimated formant data, it is surprising that Look-up Table approach has not been reported in the literature to correct the estimated formant data.

It would be an interesting exercise to re-estimate the formant data published by P&B using the above procedure.

IV.D. Analysis-by-Synthesis Approach

The above procedure when automated using a suitable distance measure leads to the Analysis-by-Synthesis approach. Different weights may be assigned to the three measurements depending the relative reliability in the estimation. Usually, F0 estimation is far more reliable than the bandwidth estimation. The procedure is explained below:

a) The initial estimates obtained using LP technique or Spectrographic measurement will be denoted by Given_F0, Given_F1, Given_BW1. These estimates are to be improved upon since it is known that F0 introduces errors in the estimates of F(1) and BW(1).

b) Synthesize a single formant vowel with Calibration-F0, Calbration-F1, Calbiraton-BW1. In the implementation a range of values would be used for these variables. The range to be used depending on the expected error magnitudes in the estimates.

c) Estimate the formant data by applying LP technique on a frame of synthesized vowel. Let us refer to these as Estimated_F0, Estimated_F1 and Estimated_BW1.

c) Minimize the Error measure that is defined as

$$\begin{aligned} \text{Err1} = {} & \text{W(F0)*(Given\_F0 - Estimated\_F0)/Given\_F0} \\ & + \text{W(F1)*(Given\_F1 - Estimated\_F1)/Given\_F1} \\ & + \text{W(BW)*(Given\_BW - Estimated\_BW1)/Given\_BW} \end{aligned}$$

where W(F0), W(F1) and W(B1) are the relative weights associated with F0, F(1) and BW(1), respectively.

d) When the Given data matches with the estimated data, i.e., when Err1 is lowest, the data used for calibration would then correspond to the re-estimated data.

Example.1: Given F0=158 Hz, F(1)=660 Hz, BW(1)=90 Hz. This is same as the Example-1 considered in the previous Section except that the bandwidth is also specified here.

The computed ERR1 for the choice of W(F0)=0.5, W(F1)=0.3 and W(BW)=0.2 is shown in Table-1. The global minimum based on AbS approach gives the global error of 0.007 and re-estimated values as F0=156.86 Hz, F(1)=680 Hz, BW(1)=50 Hz. This particular choice is shown in bold face in the Table-1 above.

The results will be different with a different set of weights. Thus for weights of 0.4, 0.3, 0.3 for W(F0), W(F1), W(BW1), the results are 156.86 Hz, 670 Hz and 75 Hz for F0, F(1) and BW(1) with the global minimum error being 0.004. A better way of assigning the weights is to make use of the systematic error seen earlier. Since for the given data, H(1)=660/158=4.17, we expect the accuracy in the estimation of F(1) to be better and that of BW(1) to be worse and hence W(F1) can be lower and W(BW1) can be higher. Thus with W(F0)=0.5, W(F1)=0.2 and W(B1)=0.3 the results are F0=156.86 Hz, F(1)=670 Hz and BW(1)=75 Hz with the global minimum error of 0.007.

Example-2: Given F0=187 Hz, F(1)=330 Hz, BW(1)=100 Hz. This is same as the Example-2 considered in the previous Section except that the bandwidth is also specified here.

For the choice of W(F0)=0.5, W(F1)=0.3 and W(BW)=0.2, the global minimum based on AbS approach gives Err1 as 0.006 and re-estimated values as F0=186 Hz, F(1)=310 Hz, BW(1)=60 Hz. This particular choice is NOT seen in Table-2 but it is closest to Row-2 in the Table-2. The choice of W(F0)=0.4, W(F1)=0.4 and W(BW)=0.2 also gave the same results. This suggests that F(1)=310 Hz is very likely to be the re-estimated value.

V. Conclusion

Error in the estimation of formant data decreases with increasing analysis interval and preferably the analysis interval must be more than 30 ms or 3 pitch periods, whichever is longer; Error in the estimation of formant frequency sharply increases as F1 decreases below 500 Hz and for F(1)/F0 below 3; higher formants (with formant frequency above 1500 Hz) show a low error of the order of 2%; Error shows an oscillatory behaviour implying that both over estimation and under estimation do occur; Error in the estimation of F(1) is zero for (F(1)/F0) equal to 2,3,4 etc and also for 2.5, 3.5, 4.5, 5.5 etc; Error in the estimation of F(1) shows negative peaks for

F(1)/F0=2.25, 3.25 etc and positive peaks for F(1)/F0=2.75, 3.75 etc. Formants with larger bandwidths have a lower error;

Error in the estimation of bandwidth decreases with increasing analysis interval; Error in the estimation of bandwidth is much greater (of the order of 40 to 100%) compared to the error in the estimation of formant frequency (4 to 10%); Larger the bandwidth, lower the error; Error in bandwidth estimation also shows an oscillatory behaviour; Estimated error in BW(1) is zero at 1.25, 1.75, 2.25, 2.75 etc whereas peaks of error in BW(1) occur near (F1/F0) = 2, 2.5, 3, 3.5, etc. This trend in error in the estimation of BW(1) is exactly opposite of the trend of errors in F(1).

One probable explanation for the oscillatory behaviour is presented in Appendix-B. A rigorous explanation is found wanting for the trends seen in the estimated error functions

A major finding of the study is that the error in the estimation of F(1) and BW(1) are not influenced by other formants significantly, especially when formants don't overlap. Similar remarks apply to other formants. This means that the error in the estimation of F(1) and BW(1) for synthesized signals of a four formant filter and single formant filter are about the same.

Because of the influence of F0, the estimated formant data differ from the specified formant data even for a single formant signal. Conversely, given the estimated formant data and F0 of a natural vowel, by an inverse process we can find out the actual values. In other words, we can re-estimate or correct the errors in the estimated formant data of vowel sounds. This approach can be used to correct the published data. This approach of improving the accuracy of initial estimation by using calibration data appears to be quiet general and the approach may be extended to other estimation problems as well.

Throughout this study we have considered a periodic impulse train as the source to synthesize vowels. Hence, this article is only of theoretical interest. A more appropriate method would be to use a voice source pulse with or without interaction to synthesize a vowel and study the influence of not only F0 but also the voice source parameters on the estimation of formant data. This is planned for future work.

References

[1] Atal B. S. and Hanauer S. L., “Speech analysis and synthesis by linear prediction of speech wave”, J. Acoust. Soc. Amer., vol. 50, no. 2, pp. 637-655, 1971.

[2] Markel J. D., and Gray A. H., "A linear prediction vocoder based upon the autocorrelation method," IEEE Trans. Acoust. Speech and Sig. Proc., vol. ASSP-22, Apr. 1974, pp. 124-134.

[3] Makhoul J., "Linear prediction: A tutorial review", Proc. IEEE, vol.63, Apr 1975, pp. 561-580.

[4] Atal B. S., "Influence of pitch on formant frequencies and bandwidths obtained by linear prediction analysis", J. Acoust. Soc. Am., vol.55, S81, 1974.

[5] Monsen R. B., and Engebretson A. M., "The accuracy of formant frequency measurements,", Vol.26, Issue-1, JSHR, March 1983, pp.89-97.

[6] McCandless S. S., "An algorithm for automatic formant extraction using linear prediction spectra," IEEE Trans. Acoust. Speech and Sig. Proc., vol. ASSP-22, pp.135-141, Apr 1974.

[7] Markel J. D., "Digital inverse filtering - A new tool for formant trajectory estimaiton," IEEE Trans. Aud. and El. Acost., AU-20, June 1972, pp.129-137.

[8] Ananthapamanabha T. V. and Yegnanarayana B., "Epoch extraction from linear prediction residual for identification of closed glottis interval," IEEE Trans. Acoust. Speech and Sig. Proc., vol. ASSP-27, No.4, Aug. 1979, pp. 309-319.

[9] Huggins W. H., "A phase principle for complex-frequency analysis and its implications in auditory theory," J. Acoust, Soc. Am., vol.24, 1952, pp.582-589.

[8] Yegnanarayana B., "Formant extraction from linear prediction phase spectrum", J. Acoust. Soc. Am., 1978, vo1.63(1), pp.1638-1640.

[11] Christensen R., et al., “A comparison of three methods of extracting resonance information from predictor-coefficient coded speech’” IEEE Trans. Ac. Speech and Sig. Proc., vol. ASSP-24, Feb. 1976.

[12] Peterson G. E., and Barney H. L., "Control methods used in a study of the vowels," J. Acoust. Soc. Am., vol.24, March 1952, pp.175-184.

[13] Atal B. S., Chang J. J., Mathews M. V. and Tukey J. W., "Inversion of rticulatory-to-acoustic transformation in the vocal tract by a computer-sorting technique," J. Acoust. Soc. Am., Vol.63 (5), May 1978, pp.1535-1555.

APPENDIX-A

Mirror-Image Spectrum

A simple transformation of a speech signal {s(n)} by the equation {s1(n)} = { (-1) ^n {s(n) } (i.e., multiplying alternate samples by +1 and -1), produces a mirror image spectrum with respect to the folding frequency (Fs/2), where Fs is the sampling frequency. In the short-time spectrum of {s1(n)}, the spectral peak at F(1) gets transformed to (Fs/2)-F(1) whereas there is no change in F0 and the higher formants gets transformed downwards with F(4) in the transformed signal being is closer to f=0. Thus the spacing between F0 and the transposed first formant is increased considerably. Does this transformation result in an improved accuracy in the estimation of F(1) and decreased accuracy in the estimation of F(4)?

For the transformed signal {s1(n)}, let the estimated formant frequencies be G(1), G(2),...G(4). The new estimates of the formant frequencies are (Fs/2)-G(i).

Let {R(k)} and {R1(k)} be the autocorrelation coefficients of {s(n)} and {s1(n)}, respectively. Let {a(k)}, {a1(k)} be the estimated LPCs for {s(n)} and {s1(n)}, respectively. Then it can be shown analytically that {R1(k)}={ (-1)^k R(k)} and {a1(k)}= {(-1)^k a(k)} and that the roots of the polynomial A1(z) in z-plane are rotated versions of the roots of the polynomial A(z), rotated by 180 degrees, which leads to the solution that the estimates of formant frequencies of the transposed signal {s1(n)}, G(i) = (Fs/2)-Estimates of F(i) of {s(n)}. In other words, there is no improvement in the accuracy of estimation of formant data by analyzing the transformed signal. This has also been confirmed by the experimental data. Hence, the lower accuracy in the estimation of low F(1), high F0 vowels is not due to the spacing between F0 and F(1) but due to some other reason yet to be investigated. Also, the better accuracy of higher formants doesn't arise because of the relatively more number of cycles in the analysis interval or within a pitch period.

APPENDIX-B

A Probable Reason for the Oscillatory Behaviour of EFF(1) and EBW(1)

Consider a single formant signal. Time domain property of the first formant frequency seems to provide some insight into oscillatory behaviour in error. Depending on the relative duration of the first formant cycle and pitch period as well as the bandwidth, at the immediate next excitation instant, first formant frequency oscillation may have a low value (nearly a zero-crossing) or a large positive or negative peak. If more number of first formant cycles have elapsed between two successive excitations then the peak amplitude decreases to a larger extent due to damping and the magnitude of discontinuity in the amplitude at the next excitation instant will be lower. We hypothesize that the error in the estimation is dependent on the magnitude of discontinuity at the immediate next excitation instant. In order to test this hypothesis, we analyse the impulse response of a single formant vowel (no windowing) with LP of order 2. We vary the analysis interval. The end point of the analysis interval has differing amplitudes. Figure. B1a shows the impulse response and the end points of the analysis intervals (dashed vertical lines). In Figure. B1b, the relative amplitude (scaled ten times) at the end of the analysis interval is shown by Black dashed line. The variation in the amplitude with the varying analysis interval clearly shows an oscillatory behaviour as expected.

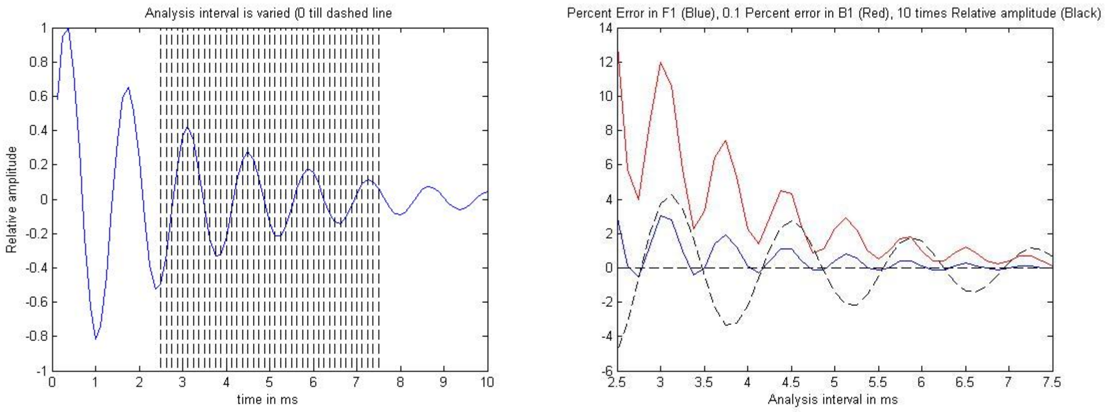


Fig. B1. (a) Impulse response of a single formant (F1=720 Hz, B1=100 Hz) signal. Analysis interval extends from the origin till the dashed line. (b) Error in the estimation of Formant frequency F1 (Blue) and (1/10 of) error in the estimation of bandwidth B1 (in Red) and the relative amplitude (10 times) at the ending instant of analysis interval (Black).

The error in the estimation of F(1) and the (1/10th) error in the estimation of B1 Vs the analysis interval are also shown in Fig. B1b. The error in F(1) Vs analysis interval as well as the error in B(1) Vs the analysis interval show an oscillatory behaviour and the peaks in the error decreases as the analysis interval is increased. The peaks and valleys in the two error graphs coincide.