International Journal of Innovation and Scientific Research
ISSN 2351-8014 Vol. 26 No. 2 Sep. 2016, pp. 384-389
© 2015 Innovative Space of Scientific Research Journals
http://www.ijisr.issr-journals.org/

Determining and comparing vocal tract length using vocal tract constrictions and linear predictive coding

Akankhya Sarmah and Kashyapi Kalita

Electronics and Telecommunication Engineering,
Assam Engineering College,
Guwahati, Assam, India

ABSTRACT: The length of the vocal tract is correlated with body size which can be determined by using formant frequencies in speech. The interconnections between vocal tract length and formant frequencies are explored here. Recorded and computer-synthesized vowel sounds are used to gauge the vocal tract length of a speaker under consideration. Vocal tract length assessment may play an important role in 'vocal tract normalization', which is crucial for speech perception and language acquisition. Vocal tract length is analyzed by applying two methods namely linear predictive coding (LPC) and vocal tract constriction (VTC). Speech signal is produced by the convolution of excitation source and time varying vocal tract system components. These excitation and vocal tract components are to be separated from the available speech signal to study these components independently. To reduce the complexity of deconvolving the given speech into excitation and vocal tract system components the 'Linear Prediction analysis' is developed. The VTC evidence gives a measure of the very low frequency component present in the signal and hence gives different range of frequency values for different types of sounds. Zero frequency filtering (ZFF) is used to give an approximate measure of vocal tract constriction in terms of the low frequency component present in the speech signal.

KEYWORDS: Vocal tract length, Formant frequencies, Speech signal, VTC, LPC.

1 INTRODUCTION

Different kinds of sound units have been studied in acoustic phonetics literature of which phones, phonemes, intonation, and the separation of words and syllables represent only those qualities of speech that are part of oral language. Sound is produced by the vocal cords or by air being forced through the mouth cavity with the tongue and lips shaped to accent a tone range. Of all the sound units, vowels are considered as voiced speech and consonants as unvoiced speech.

The frequency of vowels is used to determine the vocal tract length using the formula as given below:

1) L=c/4F1, where F1= the first formant (spectral peaks of the sound spectrum) frequency, c (34029cm/sec) = the speed of sound.

2) Length of back cavity (LB) is determined by using 2nd formant frequency and that of front cavity (LF) by 3rd formant frequency i.e. LB=c/2*F2, LF=c/2*F3 and total length of vocal tract is the addition of LB and LF i.e., L=LB+LF.

This vocal tract length will vary for each vowel as it's really difficult to achieve a completely neutral position of the vocal tract. So, a mid-central neutral vowel called 'schwa' having evenly spaced formants will be used.

To measure the formants accurately we will use two methods: Linear Predictive Coding (LPC) and Vocal Tract Constriction (VTC).

2 LINEAR PREDICTIVE CODING PROCEDURE

The redundancy in the speech signal is exploited in the LP analysis. The prediction of current sample as a linear combination of past p samples form the basis of linear prediction analysis where p is the order of prediction. The predicted sample ŝ (n) can be represented as follows,

$$\hat{\mathrm{s}}(\mathrm{n}) = \sum_{\mathrm{k}=1}^{\mathrm{p}} \mathrm{a_k}.\mathrm{s}(\mathrm{n}-\mathrm{k}) \quad (1)$$

Where, a_ks are the linear prediction coefficients and s(n) is the windowed speech sequence obtained by multiplying short time speech frame with a hamming or similar type of window which is given by,

$$s(n)=x(n).\omega\ (n) \quad (2)$$

where, ω(n) is the windowing sequence. The prediction error e (n) can be computed by the difference between actual sample s (n) and the predicted sample ŝ (n) which is given by,

$$\mathrm{e}(\mathrm{n}) = \mathrm{s}(\mathrm{n}) - \hat{\mathrm{s}}(\mathrm{n}) = \mathrm{s}(\mathrm{n}) + \sum_{\mathrm{k}=1}^{\mathrm{p}} \mathrm{a_k}.\mathrm{s}(\mathrm{n}-\mathrm{k}) \quad (3)$$

$$e(n)=s(n)- \hat{s}\ (n) \quad (4)$$

$$\mathrm{e}(\mathrm{n}) = \mathrm{s}(\mathrm{n}) + \sum_{\mathrm{k}=1}^{\mathrm{p}} \mathrm{a_k}.\mathrm{s}(\mathrm{n}-\mathrm{k}) \quad (5)$$

The primary objective of LP analysis is to compute the LP coefficients which minimized the prediction error e (n). The popular method for computing the LP coefficients by least squares autocorrelation method. This is achieved by minimizing the total prediction error. The total prediction error can be represented as follows,

$$\mathrm{E} = \sum_{\mathrm{n}=-\infty}^{\infty} e^2 \quad (6)$$

This can be expanded using the equation (5) as follows,

$$\mathrm{E} = \sum_{\mathrm{n}=-\infty}^{\infty}\left[\mathrm{s}(\mathrm{n}) + \sum_{\mathrm{k}=1}^{\mathrm{p}} a_k.s(n-k)\right]^2 \quad (7)$$

The values of a_ks which minimize the total prediction error E can be computed by finding

$$\partial E/\partial a_k$$

and equating to zero for k=0,1,2,...p.

$$\partial E/\partial a_k = 0$$

Each a_k gives p linear equations with p unknowns whose solution gives the LP coefficients. This can be represented as follows,

$$\frac{\partial \mathrm{E}}{\partial a_k} = \frac{\partial}{\partial a_k}.\left(\sum_{\mathrm{n}=-\infty}^{\infty}\left[\mathrm{s}(\mathrm{n}) + \sum_{\mathrm{k}=1}^{\mathrm{p}} a_k.s(n-k)\right]^2\right) = 0 \quad (8)$$

The differentiated expression can be written as,

$$\sum_{\mathrm{n}=-\infty}^{\infty} \mathrm{s}(\mathrm{n}-\mathrm{i}).\mathrm{s}(\mathrm{n}) = \sum_{\mathrm{k}=1}^{\mathrm{p}} a_k \sum_{\mathrm{n}=-\infty}^{\infty} s(n-i).s(n-k) \quad (9)$$

where, i=1, 2, 3...p. The equation (9) can be written in terms of autocorrelation sequence R (i) as follows,

$$\sum_{\mathrm{k}=1}^{\mathrm{p}} a_k.R(i-k) = R(i) \quad (10)$$

for i=1, 2, 3...p.

where the autocorrelation sequence used in equation (10) can be written as follows,

$$\mathrm{R}(\mathrm{i}) = \sum_{\mathrm{n}=\mathrm{i}}^{\mathrm{N}-1} s(n)s(n-i) \quad (11)$$

For i= 1, 2, 3...p and N is the length of the sequence. This can be represented in the matrix form as follows,

$$R.A=-r \quad (12)$$

where R is the p*p symmetric matrix of elements R(i, k) = R(|i-k|), (1<=i, k<=p), r is a column vector with elements (R(1),R(2), ...R(P)) and finally A is the column vector of LPC coefficients (a(1), a(2), ...a(p)). It can be shown that R is toeplitz matrix which can be represented as,

$$R = \begin{bmatrix} R(1), R(2), R(3), ..., R(P) \\ R(2), R(1), R(2), ..., R(P-1) \\ R(3), R(2), R(1), ..., R(P-2) \\ . \\ . \\ R(P), R(P-1), R(P-2), ..., R(1) \end{bmatrix} \qquad (13)$$

The LP coefficients can be computed as shown,

$$A = -R^{-1}.\mathrm{r} \qquad (14)$$

where R^{-1} is the inverse of the matrix R.

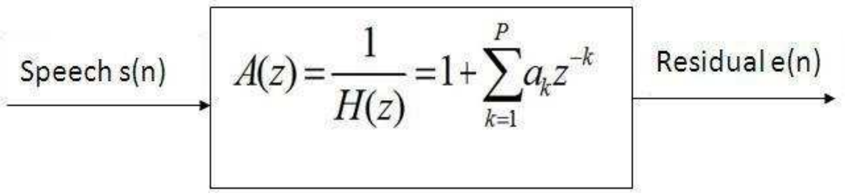

Figure 1: Computing the LP residual by inverse filtering

2.1 ESTIMATING VOCAL TRACT PARAMETERS BY LP ANALYSIS

LP analysis separates the given short term sequence of speech into its slowly varying vocal tract component represented by LP filter (H(z)) and fast varying excitation component given by the LP residual (e(n)). The LP filter (H(z)) induces the desired spectral shape for the shape on the flat spectrum (E(z)) of the noise like excitation sequence as given in equation (20). As the LP spectrum provides the vocal tract characteristics, the vocal tract resonances (formants) can be estimated from the LP spectrum. Various formant locations can be obtained by picking the peaks from the magnitude LP spectrum (|H (z)|). The figure 4 shows the first (F1), second (F2) and third formant (F3) frequencies estimated from the peaks in the LP magnitude spectrum.

S (z) =E (z).H (z) (20)

Where S (z) is the spectrum of the given short time speech sign

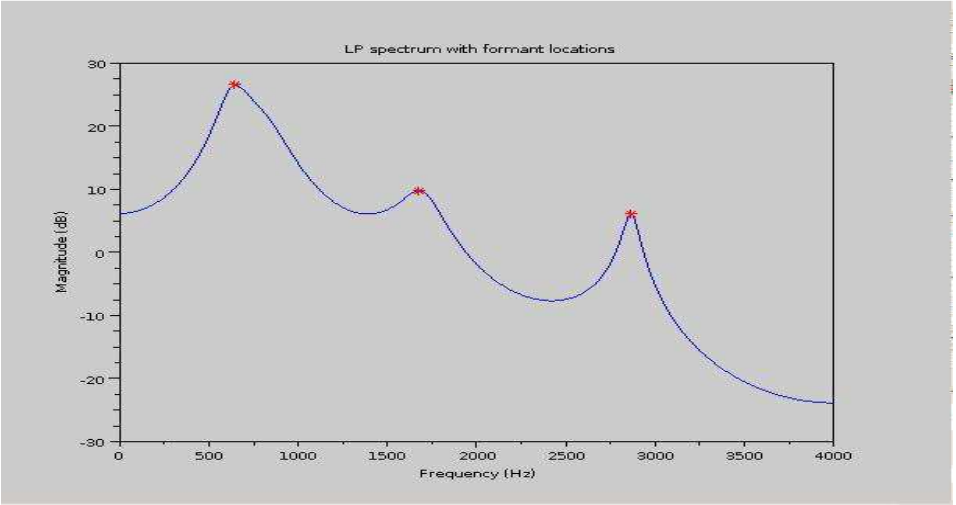

Figure 2: Formant locations corresponding to peaks in LP magnitude spectrum

Table 1

For practical purpose, 11 speakers both male and female were considered and following results were obtained:

SPEAKERS	**F1**	**L**	**F2**	**F3**	**L**
S1(male)	454.56	18.8920	2307.83	3170.17	12.7465
S2(female)	693.50	12.2673	3007.33	3263	10.8907
S3(female)	734.33	12.9979	3014.67	3478.66	10.5686
S4(female)	604.83	14.079	2823.83	3629.5	10.7155
S5(female)	616	13.8253	2995.66	3423.5	10.6974
S6(female)	506.2	16.8316	2679.83	3345.83	11.4805
S7(male)	558.2	15.3233	2115.17	2795.83	14.1392
S8(male)	548.83	15.5086	2509.17	3408.33	11.7746
S9(female)	479.33	17.7583	2976	3208	11.0164
S10(male)	593.83	15.0761	2487.33	3453.33	11.7708
S11(male)	522.17	16.0387	2252.83	2926.5	13.3766

3 VOCAL TRACT LENGTH CONSTRICTION METHOD

The evidence for low frequency dominant sounds is obtained by exploiting the sinusoidal nature of ZFF signal. The constriction of vocal tract dampens high frequency components in the resulting speech signal. For instance, in case of complete closure like voice bars, the resulting speech signal predominantly contains a low frequency component and looks like a sinusoidal signal. Accordingly, sounds containing dominant low frequency components exhibit high similarity in temporal domain with the sinusoidal like ZFFS, which is also low frequency dominant. This characteristic of ZFFS is exploited to obtain the proposed evidence about the VTC. The ZFFS can be computed from the speech signal in two steps. First, compute the output of a cascade of two ideal digital resonators at 0 Hz.

$$y(n)= -\sum_{k=1}^{4}(a_k)\, y(n-k) + x(n) \qquad (1)$$

where a1=4, a2=-6,a3=4, a4=-1 , and x(n) is the differenced speech signal. Then, remove the trend i.e.

$$y'(n)=y(n)-y(n)'' \qquad (2)$$

where y(n)''=(1/(2N+1)), $\sum_{n=-N}^{N} y(n)$ and 2N+1 correspond to the average pitch period computed over a longer segment of speech. The trend removed signal is the ZFFS. The positive zero crossings of the ZFFS will give the location of epochs.

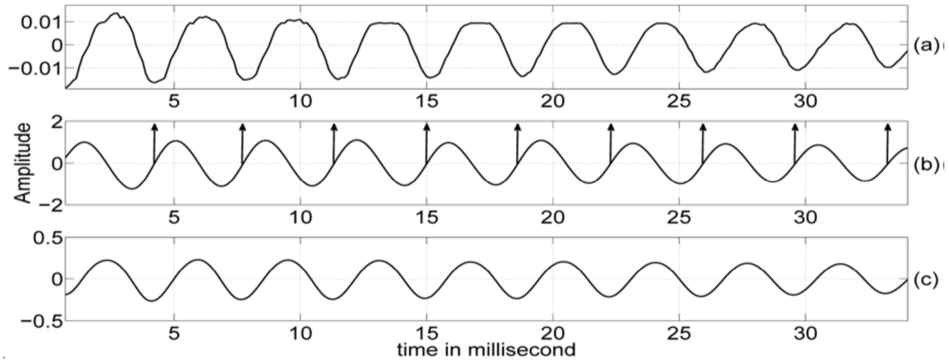

Fig. 3 (a) Voice bar region, (b) Zero-frequency filtered output of signal. Arrows show the epoch locations and (c) Inverted difference of ZFFS

Fig. 6(a) shows the voice bar portion of a speech signal and Fig 6(b) shows its ZFFS. The arrow markings show the epoch locations. The epoch location corresponds to zero crossing in ZFFS and to peak in the voice bar signal. To make both signals in same phase, difference of ZFF signal is computed and inverted. The obtained signal is shown in Fig. 6(c). To find the correlation between the two signals, an epoch based analysis is performed. Epoch interval is defined as the interval between successive epochs. In every epoch interval, ZFFS and speech signals are compared using the cosine kernel given by,

$$\frac{<x'(n)y'(n)>}{||x'(n)||||y'(n)||}$$

where, x' (n) and y' (n) are the speech signal and the ZFFS respectively between successive epochs. The cosine kernel value is proposed to be the measure of the match between the two. Fig. 7shows a portion of the speech signal for the utterance *she had your dark suit in greasy wash water all year* taken from TIMIT database. Fig. 7 also shows the cosine kernel evidence. To make this evidence equal to the length of speech, the cosine kernel value is computed at every epoch location and the value is duplicated until the next epoch location is reached. The figure shows the behaviour of evidence for different sounds. Evidence shows a high value for the voice bar regions (around 0.6 s and 0.9 s) and a very low value for the vowel regions (around 1 s).

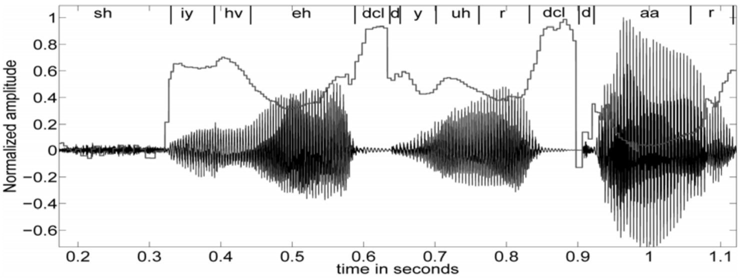

Fig.4 Speech signal with the proposed evidence.

The proposed evidence shows very high value for voice bar regions and very low value for low vowels.

For other sounds with relatively less constriction, the evidence shows an intermediate value.

From the typical values of F1,F2,F3 and VTC for the syllables “aa”, ”u” and “schwa” for both male and female, we found the standard ratio of “u” to “aa” as 48.73/51.26(male) and 49.96/50.03(female).

Now, using the formula, VTC_avg = [(H-L)*.5126] +L (male)

[(H-L)*.5003]+L (female)

where, H and L are VTC_avg of “aa” and “u” respectively.

In the following table, the VTC_avg for the 10 speakers are given.

Table 2

SPEAKERS	VTC_avg
S1(MALE)	0.449
S2(FEMALE)	0.5339
S3(FEMALE)	0.2495
S4(FEMALE)	0.5476
S5(FEMALE)	0.3991
S6(FEMALE)	0.2108
S7(MALE)	0.3227
S8(MALE)	0.4481
S9(MALE)	0.2982
S10(MALE)	0.5061
S11(MALE)	0.3186

A Look-Up table is constructed using a MATLAB code and top 5 values of the vocal tract length corresponding to the VTC_avg of all the speakers are taken.

Then, the length obtained from vocal tract constriction(VTC) method and linear prediction coding(LPC) method are compared as shown in the table below:

Table 3

SPEAKER	**Length Obtained using LPC Analysis**	**Length Obtained Using VTC Synthesis**
S1(FEMALE)	16.8316	16.3287
S2(FEMALE)	12.2673	13.7017
S3(FEMALE)	12.9979	12.48
S4(FEMALE)	14.079	16.0959
S5(FEMALE)	13.8253	18.8631
S6(MALE)	18.8920	19.0746
S7(MALE)	15.3233	17.6866
S8(MALE)	22.5086	22.9306
S9(MALE)	17.7583	16.6482
S10(MALE)	15.0761	20.9538
S11(MALE)	18.0387	18.0621

4 CONCLUSION

Vocal tract length for 11 speakers was obtained by both linear predictive coding and vocal tract constriction method. Out of the 11 speakers, results for 7 speakers obtained by linear predictive coding were approximately equal to the results obtained by vocal tract constriction method. Hence, we can come to a conclusion that vocal tract constriction method is another method for determining vocal tract length of a person.

ACKNOWLEDGEMENT

We are grateful to Prof. S.R.M Prasanna, Department of Electronics and Electrical Engineering, Indian Institute of Technology Guwahati under whose supervision we were able to complete our project. We would also like to thank Mr. Apurba Kalita, H.O.D, and ETE Department of Assam Engineering College for his guidance and encouragement all throughout.

REFERENCES

[1] K. N. Stevens*, Acoustics Phonetics*. Cambridge, MA, USA: MIT Press, 1999.

[2] G. Fant*, Acoustic theory of speech production*. The Hague, The Netherlands: Mouton, 1960.

[3] T. Gay, L.-J. Boe, and P. Perrier, *Acoustic and perceptual effects of changes in vocal tract constrictions for vowels*, *J. Acoust. Soc. Amer.*, vol. 92, pp. 1301–1309, 1992.

[4] S. K. Hong and S. W. Yoon, *Effect on vowel production of constriction in the vocal tract*, *J. Korean Phys. Soc.*, vol. 46, pp. 840–847, 2005.