ISSN: 2278 – 909X
International Journal of Advanced Research in Electronics and Communication Engineering (IJARECE)
Volume 4, Issue 12, December 2015

A Comparative Study of Formant Estimation

Safya Bhore, Milind S. Shah

***Abstract*— The primary aim of speech analysis is to extract speech parameters or features that represent important characteristics of a waveform. Formant frequencies are one of the important features of the speech production and vocal tract model and therefore its estimation has always received considerable attention. This paper presents a comparison of three formant extraction techniques namely Cepstral analysis, Linear Prediction based Cepstral (LPCC) technique and Burg's method. These three methods have been implemented in MATLAB for estimation of lowest three formants of various speech samples, which is then compared with the values of formants obtained from Praat software. These results have been tabulated. It was observed that, in general, the Burg algorithm is more accurate than Cepstral analysis and LPCC techniques.**

***Index Terms*— Formant frequencies, Cepstral analysis, Linear prediction based cepstral analysis (LPCC) and Burg method.**

I. INTRODUCTION

Speech is inevitable to humans. Due to its importance in human lives it has always remained one of the most widely researched areas. Speech processing and its study help us to understand the nature of speech recognition, synthesis, coding etc. Among the various speech processing parameters, formant is an important parameter. According to the source filter theory of speech production, vocal tract can be modeled as a tube [1]. This tube may have resonances upon vibration. These spectral resonances of the vocal tract are called formants [2].

Formant estimation and modeling of speech in terms of formants is necessary in various speech processing domains, since formants very efficiently describe the essential aspects of speech using very limited set of parameters [2]. Hence, the behavior of the first three to four formants is crucial and sufficient in many applications. They help in perception of speech sounds, determining the phonetic content of speech, and used widely in recognition systems [1-2]. For example, the spacing between F2 and F3 help to distinguish between glides in the syllable initial position [2]. In fact, some studies also relate these resonance frequencies to the age of speakers and even their heights which can be helpful for forensic purposes [3-4]. The set of formants are unique to every phoneme. Also formant values for the same voiced sound may vary, but limited, from person to person.

The aim of this paper is to estimate formant frequencies of speech samples (vowels and VCV syllables) of three male and female speakers using Cepstral analysis, Linear Prediction based Cepstral analysis (LPCC) and Burg algorithm techniques and compare their results. These techniques are implemented in MATLAB software. The formant values estimated by each of the three algorithms are compared with the estimated formant values by Praat software [5] for the same speech samples. Thus the implemented techniques are compared for their performance in terms of accuracy of the formant estimation by calculating Root Mean Square Error (RMSE). Praat software is used as a reference software here because when synthesized vowels with specific formant values were generated and given to Praat software as well as all the three implemented algorithms, the results of Praat gave the least RMSE. Also, since Praat resamples any speech signal given as input to 10 kHz, for comparison purpose, resampling is an additional step carried out in each of the algorithms before processing.

The outline of this paper is as follows. Section II gives the implementation details of the three methods. Section III presents the results in terms of RMSE values and the compassion of all the three methods is discussed. Section IV gives the summary and conclusion

II. IMPLEMENTATION

A. Cepstrum based Formant Estimation

According to the source filter theory of speech production, speech $s(t)$ is composed of the excitation signal $e(t)$ and the vocal tract components $h(t)$ [2]. Cepstral analysis aims to make use of this fact for separating the signal into its components in a simplistic manner. So signal is viewed as a linear combination of these two components [6]. For this purpose, it is required to transform signal into frequency domain $S(\omega)$ and then perform the log magnitude as given in (1) and (2). A transformation back does not lead to the time domain but to what is called as the cepstral domain and the resultant spectrum obtained is called as the cepstrum $C(n)$

$$|S(\omega)| = |E(\omega)||H(\omega)| \quad (1)$$

And,

$$C(n) = IDFT(\log|S(\omega)|) \quad (2)$$

Where,

$|S(\omega)|$ = speech spectrum

$|E(\omega)|$ = excitation signal spectrum

$|H(\omega)|$ = vocal tract transfer function

$C(n)$ = cepstral coefficients

The process of estimating formants by Cepstral analysis

Manuscript received Dec, 2015.

***Safya Bhore**, Electronics and Telecommunication, F.C.R.I.T. , Navi Mumbai, India.*

***Milind S. Shah,** Electronics and Telecommunication, F.C.R.I.T., Navi Mumbai, India.*

technique is given in Fig. 1. Firstly, the signal is resampled to 10 kHz and then pre-emphasis is carried out. The pre-emphasized signal is then windowed and framed. For our implementation, a Hamming window of 20 msec duration is used. Next, the cepstrum is calculated. As the lower part of the cepstrum corresponds to the vocal tract information, this part of the cepstrum is retained with a low time liftering window whose cut off frequency is chosen to be 30 samples. The liftering operation is explained in [7]. Cepstral coefficients are estimated for each frame. The smoothed spectrum is calculated to obtain the vocal tract signal [6]. The smoothening is done by discrete Fourier transform operation performed on the signal. The formants which correspond to peaks in the smoothed spectrum are detected by a peak picking algorithm [7]. The first three peaks in the smoothed spectrum are identified as the first three formant frequencies [6].

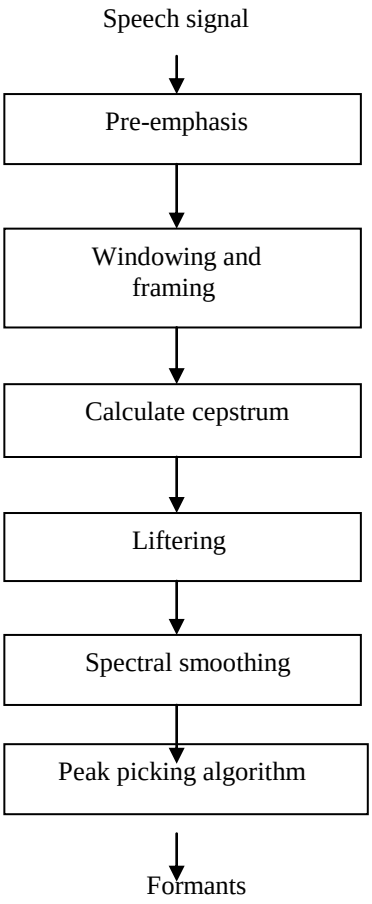

Fig. 1. Algorithm for formant estimation using Cepstral Analysis [4]

B. LPCC

The implemented algorithm is shown in Fig. 2. As in the previous method, resampling, pre-emphasis and windowing is carried out on the speech signal. The prediction parameters are computed for every frame. LP filter order p of 12 is chosen. Next, LP parameters are converted to cepstral coefficients according to the following set of equations where $a_0, a_1, \ldots, a_p$ represent the LP parameters, p is the order of the LP filter and c_m are the cepstral coefficients. Cepstral coefficients from LPC are calculated using (3). This method of cepstral coefficients computation avoids taking the Fourier transforms [8].

$$C_m = \begin{cases} \log G & n = 0 \\ -a_m + \frac{1}{m}\sum_{k=1}^{m-1} -(m-k)\, a_k\, c_{m-k} & 1 \le m \le p \\ \sum_{k=1}^{p} -\frac{m-k}{m}\, a_k\, c_{m-k} & p < m < n \end{cases} \quad (3)$$

Where G is the gain of LPC filter.

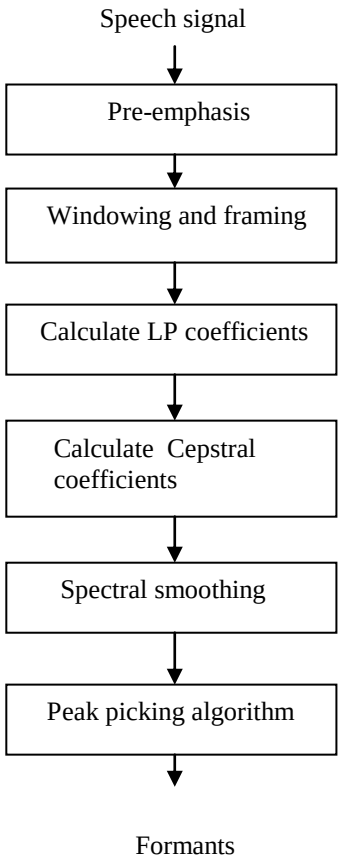

Fig. 2. Algorithm for formant estimation using Cepstral Analysis [4]

C. Burg algorithm

The process of estimating formants by Burg's method is given in Figure 3. Speech signal is first resampled to 10 kHz. After this, pre-emphasis, windowing and framing using a 20 msec Hamming window is done. These steps remain the same as that in the previous two techniques. Next the autoregressive parameters are obtained via the Burg method. For this purpose, LP filter order of 10 is used for the computation for male and female sound samples. This allows for the estimation of approximately 4 formant values in each frame. Roots of the prediction polynomial are then obtained. A lower LPC order is sufficient for females and increasing it results in spurious peaks in the spectrum. Because the LPC coefficients are real-valued, the roots occur in complex conjugate pairs so only the roots with positive imaginary part are retained and their corresponding angles are determined by (4).

$$angz = \sum \tan^{-1}\left(\frac{\text{Im } r_k}{\text{Re } r_k}\right) \quad (4)$$

In the above equation, r_k is the retained root of a polynomial. The angular frequencies in rad/sample are converted to Hz and the bandwidths of the formants are calculated by the following equations [6].

$$F_k = \left(\frac{F_s}{2\pi}\right) \tan^{-1}\left(\frac{\text{Im } r_k}{\text{Re } r_k}\right) \quad (5)$$

$$B_k = -\left(\frac{-F_s}{2\pi}\right) \ln |r_k| \quad (6)$$

These formant values are sorted in ascending order and he first three values are chosen as the first three formants. These values are found for every frame.

III. Results and Discussion

A. Speech material

Natural speech recordings of vowel sounds (/a/, /i/, /u/) and VCV syllables (/aba/, /afa/, and /aya/) were recorded for 3 male and 3 female speakers respectively at a sampling rate of 16 kHz and used as speech material

B. Results

Formant values are estimated for every frame. RMSE is then calculated between formant values obtained by all the three algorithms respectively and Praat output, for vowels (/a/, /i/, /u/) and VCV syllables (/aba/, /afa/, /aya/) of three male and three female speakers. Average RMSE across three male and female speakers are calculated for vowels and VCV syllables. Table 4.1 and 4.2 shows the average RMSE results for male and female vowels and VCV by Cepstral analysis, tables 4.3 to 4.4 shows the results for LPCC method and table 4.5 and 4.6 for Burg method.

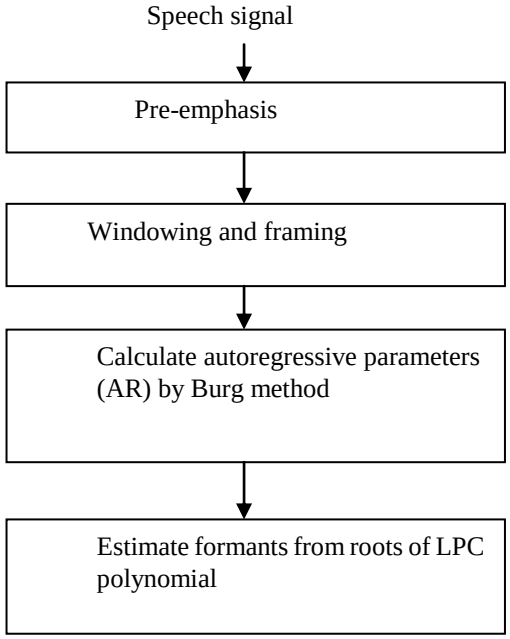

Fig. 3. Algorithm for formant estimation based on Burg algorithm

Table.1: Average RMSE of F1, F2, F3 formants of vowels by Cepstral technique

	RMSE for F1		RMSE for F2		RMSE for F3	
	Male	Female	Male	Female	Male	Female
/a/	99.58	205.21	58.37	131.33	147.27	634.87
/i/	45.58	100.92	58.37	741.56	269.27	441.69
/u/	197.79	253.46	752.49	1046.72	788.71	988.10

Table 2: Average RMSE of F1, F2, F3 formants via Cepstral analysis technique for natural VCV syllables

	RMSE for F1		RMSE for F2		RMSE for F3	
	Male	Female	Male	Female	Male	Female
/aba/	173.47	150.73	421.62	569.07	473.1	448.9
/afa/	233.14	350.25	380.01	454.55	446.02	506.48
/aya/	160.71	239.77	577.79	418.25	455.75	507.43

Table 4.3: Average RMSE of F1, F2, F3 formants of vowels by LPCC technique

	RMSE for F1		RMSE for F2		RMSE for F3	
	Male	Female	Male	Female	Male	Female

/a/	92.31	113.93	486.76	415.54	371.44	493.73
/i/	28.63	214.85	261.72	682.28	352.69	676.77
/u/	35.26	62.91	216.45	971.74	503.1	879.51

Table 4.4: RMSE of F1, F2, F3 formants via LPCC method for natural VCV syllables

	RMSE for F1		RMSE for F2		RMSE for F3	
	Male	Female	Male	Female	Male	Female
/aba/	175.59	216.38	641.22	369.29	608.13	477.38
/afa/	178.83	184.74	551.53	476.52	569.58	587.35
/aya/	153.09	107.58	351.87	319.13	521.1	466.26

Table 4.5: Average RMSE values of formants values of vowels by Burg algorithm

	RMSE for F1		RMSE for F2		RMSE for F3	
	Male	Female	Male	Female	Male	Female
/a/	5.77	32.88	3.18	43.64	8.042	546.58
/i/	3.732	50.97	4.50	482.40	9.567	112.73
/u/	4.83	54.90	4.932	180.67	17.19	592.37

Table 4.17: RMSE of F1, F2, F3 formants via Burg algorithm for natural VCV syllables

	RMSE for F1		RMSE for F2		RMSE for F3	
	Male	Female	Male	Female	Male	Female
/aba/	11.831	143.46	50.071	162.67	15.432	333.39
/afa/	24.26	153.99	30.52	232.01	28.46	573.34
/aya/	10.56	71.94	24.36	332.91	12.83	607.85

From the tables it can be observed that in general, for all the three algorithms, there was an increased error with increasing formant order. RMSE for F1 for the vowel /a/ is highest in case of Cepstral analysis method. This suggests that by Cepstral analysis method it is difficult to resolve nearby formants. It is also observed that RMSE for VCV syllables is higher than that of vowels. This is because there are many voiced / unvoiced sections in the syllables and the algorithm assumes some random values during the unvoiced/consonant sections."

IV. CONCLUSION

Three techniques for the estimation of speech formant frequencies based on cepstral analysis, linear prediction based cepstral analysis and Burg algorithm were compared. We compared the formant frequencies of natural vowels and VCV syllables for three male and three female speakers,

estimated by the three implemented methods with the values obtained from PRAAT software. Burg algorithm implementation resulted in considerably less error than the other two methods. Out of the remaining two, LPCC technique gave better results than Cepstral technique.

REFERENCES

[1] R.W. Schafer and L.R. Rabiner, "System for automatic formant analysis of voiced speech," *J. Acoust. Soc. Am.*, vol. 47, no. 2, pp. 635-648, 1969.
[2] Douglas O' Shaughnessy, *Speech Communications: Human and Machine*, 2nd ed., Hoboken: New Jersey, John Wiley & Sons, 2000.
[3] P. Busby and P. L. Plant, "Formant frequency values produced by pre-adolescent boys and girls*," J. Acoust. Soc. Am.*, vol. 97, no. 4, pp. 2603-2606, 1995.
[4] H. Cao, Y. Wang and J. Kong, "Correlations between body heights and formant frequencies in young male speakers," *ISCSLP*, pp. 536-540, 2014.
[5] B. Paul and D. Weenik (2015). *Praat: doing phonetics by computer Version 5.4.22* [Online]. Available: from http://www.praat.org/
[6] G. Gargouri, M. A. Kamoun, M. A. Zerzri and A. B. Hamida, "Cepstrum vs LPC: A comparative study for speech formant frequencies estimation," *GESTS Trans. Intl. Comm. Signal Process.*, vol. 9, no. 1, pp. 87-102, 2006.
[7] G. Gargouri, M. A. Kamoun M. A. Zerzri and A. B. Hamida, "Cepstral method evaluation in speech formant frequencies estimation*," ICIT*, vol. 3, pp. 1612-1616, 2004.
[8] G. Gargouri, M. A. Kamoun, M. A. Zerzri and A. B. Hamida, "Formant estimation techniques for speech analysis," *International Conference on Machine Intelligence*, pp. 96-100, 2005.