Proceedings of the 5th WSEAS International Conference on Signal Processing, Istanbul, Turkey, May 27-29, 2006 (pp15-19)

A Comparative Study of Formant Frequencies Estimation Techniques

DORRA GARGOURI, Med ALI KAMMOUN and AHMED BEN HAMIDA
Unité de traitement de l'information et électronique médicale, ENIS
University of Sfax
TUNISIA

Abstract: This paper presents two techniques of formants estimation based on LPC and cepstral analysis. These methods are implemented with Matlab and applied to the problem of accurate measurement of formant frequencies.
The first algorithm estimate formant frequencies from the all pole model of the vocal tract transfer function. The approach relies on the source – filter model supposing that the speech signal can be considered to be the output of a linear system. The spectral peaks in the spectrum are the resonances of the vocal tract and are commonly referred to as formants.
The cepstral algorithm picks formant frequencies from the smoothed spectrum. The approach relies on decomposing the speech signal by homomorphic deconvolution into two components: the first component presents the excitation, while the second component is intended to present vocal tract resonances. The result, called cepstrum, is then used to estimate the smoothed spectrum. Formant picking is achieved by localizing the spectral maxima from the envelope.
Results show the efficiency of LP based technique and the limitation of the cepstral technique in the estimation of formants of high frequencies.

Keywords: LPC, Cepstre, formant, cepstrum, vowel

1 Introduction

The problem of formant extraction has received considerable attention in speech analysis and recognition [8]. The use of formant frequencies is appealing in principle due to the important role in determining the phonetic content as well as the close relation to the vocal tract geometry. Unfortunately, reliable formant frequencies are very difficult to extract from the speech wave. However, several studies have shown that there exist approximately linear relationships between formant frequencies and other spectral representations [8, 9]
Although in the long run automatic formant analysis of speech has received considerable attention and a variety of approaches have been developed, the calculation of accurate formant features from the speech signal is still considered a non-trivial problem.
In this sense, we present in this paper two techniques of speech treatment based on Cepstral analysis and LPC for the estimation of the first four formants frequencies. These techniques are applied to vowels pronounced by different speakers.
The outline of this paper is as follow. In the next section, we describe formant frequencies estimation technique based on LPC analysis. Section 3 deals with formant extraction technique based on cepstral analysis. Next, we present the experimental evaluations and comparisons of the results. And finally, the conclusion of this study is stated.

2 LPC Based Formants Estimation Technique

The vocal tract can be modeled as a linear filter with resonances. The resonance frequencies of the vocal tract are called formant frequencies. Graphically, the peaks of the vocal tract response correspond roughly to its formant frequencies. Therefore, if the vocal tract is modeled as a time-invariant, all-pole linear system, then each of the conjugate pair of poles corresponds to a formant frequency (resonance frequency).
The peaks of the vocal tract response in each configuration correspond roughly to its formant frequencies [1, 2, 3, 6]. For voiced and periodic speech (as in sustained vowels) the vocal tract can be modeled by a stable all-pole model. The resonances or peaks of the vocal tract transfer function (poles of the H(z) transfer function) correspond roughly to the formant frequencies of a particular sound.
The linear prediction theory is well-documented in the literature [1, 2, 3, 4, 5, 6] so, here we will briefly

review the mechanics of computing a linear prediction model [2], and then discuss the implications in formant frequencies estimation.
In fact, the speech signal can be defined as:

$$s(n) = -\sum_{i=1}^{N_{LP}} a_{LP}(i).s(n-i) + e(n) \quad (1)$$

Where N_{LP}, a_{LP} and e(n) represent, respectively, the number of coefficients in the model, the linear prediction coefficients and the error in the model.
Equation (1) can be written in Z-transform notation as a linear filtering operation:

$$E(z) = H_{LP}(z).S(z) \quad (2)$$

E(z) and S(z) represent, respectively, the Z-transform of the error signal and the speech signal. $H_{LP}(z)$ is defined as a linear prediction inverse filter:

$$H_{LP}(z) = \sum_{i=0}^{N_{LP}} a_{LP}(i) z^{-i} \quad (3)$$

Formant frequencies can be estimated from the LP smoothed spectrum. From this spectrum, local maxima are found and those of small bandwidths are related to formants [3]. Then, Peak-picking can be used to estimate formants, but this method provides a significant improvement over the accuracy that would be expected from an attempt to pick peaks from the unprocessed speech spectrum.
However, we will use another way to estimate formant frequencies based on the relationship between formant and poles of the vocal tract filter [7].
The denominator of the transfer function may be factored:

$$1 + \sum_{i=1}^{N_{LP}} a_{LP}(i) z^{-i} = \prod_{k=1}^{N_{LP}} (1 - c_k.z^{-1}) \quad (4)$$

Where C_k are a set of complex numbers, with each complex conjugate pair of poles representing a resonance at frequency:

$$\hat{F}_k = \left(\frac{F_s}{2\pi}\right) \tan^{-1}\left[\frac{\mathrm{Im}(c_k)}{\mathrm{Re}(c_k)}\right] \quad (5)$$

And bandwidth:

$$\hat{B}_k = -\left(\frac{F_s}{\pi}\right) \ln |C_k| \quad (6)$$

If the pole is close to the unit circle then the root represents a formant:

$$r_k = \sqrt{\mathrm{Im}(c_k)^2 + \mathrm{Re}(c_k)^2} \geq 0.7 \quad (7)$$

3 Cepstrum Based Formants Estimation Technique

The vocal tract shape can be considered as a "filter" that filters the excitation to produce the speech signal. The frequency response of the filter has different spectral characteristics depending on the shape of the vocal tract. The spectral peaks in the spectrum are the resonances of the vocal tract and are commonly referred to as formants.
A feature that is common to nearly all spectral shape models is the derivation of the spectral envelope through some kind of smoothing operation. Smoothing is intended to remove the irrelevant harmonic detail.
The homomorphic decomposition is designed to separate convolved signal components.
Let

$$S(t) = g(t) \otimes h(t) \quad (8)$$

Where $\otimes$ denotes convolution, g(t) and h(t) are respectively the contribution of the excitation and vocal tract.
This kind of method represents the spectral envelope by computing the power spectrum using the Fourier Transform, and performing an inverse Fourier transform of the logarithm of that power spectrum. Low-order coefficients (8 to 16) of this inverse are retained [10]. Formants are finally estimated from the smoothed spectral envelope using constraints on formant frequency ranges and relative levels of spectral peaks at the formant frequencies.
By inverse Fourier transform of the log spectrum, the cepstrum is computed. The expression of the cepstrum is:

$$ç(n) = FFT^{-1}(Log(FFT(s(n)))) \quad (9)$$

At this state, the excitation (g(n)) and the vocal tract shape (h(n)) are superimposed, and can be separated using conventional signal processing such as a temporal filtering (liftring). In fact, the low order terms of the cepstrum contain the information relative to the vocal tract. This contribution becomes unimportant from a sample n_0 (n_0 corresponds to the fundamental frequency F_0). The visible periodic peaks beyond n_0

reflect the impulses of the source. Theses two contributions can be separated by a simple temporal windowing F

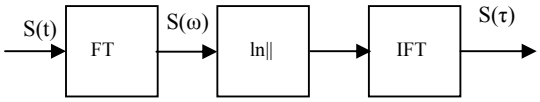

Fig 1: The homomorophic decomposition.

The first cepstral coefficients contain essentially the contribution of the vocal tract and that the periodic "peaks" visible on the suite c_n ($n \geq n_0$ (n_0 corresponding to F_0)) reflect the impulses of the source [1].
The smoothed cepstral envelope of the vocal tract can be obtained easily by the following schema:

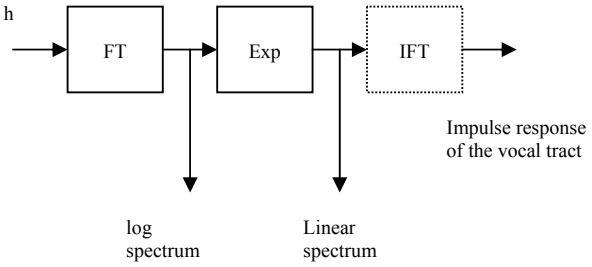

Fig 2: Transformation of Cesptrum to smoothed spectrum.

After calculating the smoothed spectrum, we can afterward extract amplitudes corresponding to the vocal tract resonances. This can be easily obtained by localizing the spectral maxima from frequency bands corresponding to the first four formants (200-900 Hz for (F1), 1600-2800 Hz for (F2), 1400-3800 Hz for (F3), and 3700-4600 Hz for (F4)) [8].
We can also extract fundamental frequency by localizing the order of the cepstral maxima corresponding to n_0 .

4 Experiments and results

The speech data (16 kHz sampling frequency) used in this study pertains to the TIMIT speech corpus.

For our experiments, we used ten different subjects from each sex. All speakers read the same text (sa1.wav). From the 22 different English vowels and diphthongs present in TIMIT database we have selected six vowels. These vowels are [ih, ix, aa, ux, iy, y].
All Cepstral and LP coefficients (12 coefficients) have been computed from pre-emphasized speech signal using 512 points Hamming windowed speech frames.
For LPC based technique, Formants frequencies candidates are calculated by solving the prediction polynomial using Levinson-Durbin algorithm. Only poles agreeing with equation 7 are considered as formant candidates.
Various experiments have been carried out on a set of wav files selected from the TIMIT corpus. We tested the formant estimation algorithms on different male and female subjects. For each vowel pronounced by each speaker, we extracted the first four formant frequencies by the two techniques described above.
The mean values of formant frequencies estimated by LPC method are summarized in Table 1and those estimated by Cepstral method are summarized in Table 3.
As can be seen there is considerable subject to subject variability in the measurements of formant frequencies.
In comparison with Cesptrum smoothed spectrum based algorithm, the formant estimation algorithm, based on LP coefficients, proves more accuracy in the measurement of formants F3 and F4
In order to compare the efficiency of these algorithms, the standard deviation of formant frequencies estimated by LPC and Cepstral based techniques has been computed. The results are summarized in Tables 2 and 4.
From these results, we can notice that the standard deviation is more important for Cepstral algorithm. However, there is a narrow range in the estimated values of formant frequencies estimated by LPC method. Also, we remark that the standard deviation increases with the order of the formant.
The previous results allow us to collect an important explanation about Cepstral and LPC method; in comparison with the Cepstral technique, the LP algorithm is a practical way to estimate formants of the speech signal especially at high frequencies.

	F1		F2		F3		F4	
	male	female	male	female	male	female	male	female
ih	412.33	466.54	1835.39	2329.84	2878.92	3056.34	4158.17	4157.20
ix	648.09	732.29	1894.44	2070.54	2636.81	2925.48	3918.56	4176.05
aa	700.92	724.81	1443.75	1646.01	2523.83	2834.31	3698.52	4124.73
ux	366.23	417.65	1493.98	1985.36	2692.42	2907.70	3606.63	4039.29
iy	336.87	407.80	1984.94	2313.68	2997.67	3181.75	3681.36	4150.69
y	253.39	356.26	2205.18	2478.26	3206.33	3375.80	4264.30	4234.32

Table 1: Mean values of formant frequencies (in Hz) estimated by LPC based technique

	F1		F2		F3		F4	
	male	female	male	female	male	female	male	female
ih	4.03	15.16	16.56	25.88	19.40	37.09	43.27	43.56
ix	19.46	23.80	12.22	23.00	55.35	35.50	25.90	43.75
aa	7.50	23.55	30.70	18.28	15.57	34.39	37.55	43.22
ux	2.87	13.57	19.23	22.05	49.27	35.28	98.29	42.32
iy	4.21	13.25	14.18	25.70	26.15	38.61	26.32	43.49
y	4.67	11.58	13.56	27.53	40.09	40.96	36.87	44.36

Table 2: Standard deviation of formant frequencies estimated by LPC based technique

	F1		F2		F3		F4	
	male	female	male	female	male	female	male	female
ih	390.6	431.25	1888	2434	3178.2	2912.5	4041	4034.4
ix	559.4	721.88	1772	2056	2718.8	2865.7	3969	4128.14
aa	637.5	703.13	1431	1475	3025	2984.4	4050	3990.66
ux	359.4	390.63	1650	1950	3256.3	2906.3	3909	3987.53
iy	309.4	365.63	1984	2284	3175	2900	3872	4021.89
y	340.6	431.25	2069	2434	2984.4	2912.5	4003	4034.4

Table 3: Mean values of formant frequencies (in Hz) estimated by Cepstral analysis based technique.

	F1		F2		F3		F4	
	male	female	male	female	male	female	male	female
ih	42.31	112	251.38	316.47	178.037	236.08	306.9	219.72
ix	78.58	132.95	112.24	291.61	193.215	179.24	288.29	177.65
aa	116.2	114.35	279.43	275.51	209.264	430.06	318.75	309.36
ux	42.31	66.291	328.99	379.66	298.588	406.4	156.21	268.48
iy	47.62	70.726	156.78	255.37	325.179	365.85	198.98	209.9
y	95.98	79.944	391.36	387.2	308.055	204.22	200.06	247.84

Table 4. Standard deviation of Formant frequencies estimated by Cepstral analysis based technique

5 Conclusion

We presented in this paper two techniques of formants extraction based on cepstral analysis and linear prediction coefficients.

The LP model was generated using the autocorrelation method based on the Levinson-Durbin recursion. The cepstral envelope is generated using the homomorphic deconvolution based on the separation of the excitation and vocal tract contribution.

Vowels pronounced by 10 speakers from each sex, were analyzed using the cepstral and LP methods in order to estimate formant frequencies (vocal tract resonances). Significant variations among the speakers were observed for all the acoustic measures.

To compare the accuracy of these algorithms, the standard deviation has been computed. Results show the existence of a narrow range in the values of formant frequencies estimated by LP based technique. This range will be wider for formant frequencies estimated by cepstral technique especially for formants of high frequencies. These results confirm that LP based technique are more efficient for the estimation of formants frequencies.

References

[1] Calliope, "*La parole et son traitement automatique*", 1989, pp. 283-3090.

[2] Joseph Picone, "Signal Modeling Techniques in Speech Recognition ", *Proceedings of IEEE*, final copy, June 1993.

[3] Bernard Gold and Lawrence L. Rabiner, "Analysis of Digital and Analog Formant synthesizers", *IEEE Transactions on Audio and Electroacoustics,* VOL. AU-16, NO. 1 March 1968.

[4] Lawrence L. Rabiner, James W. Cooley, Howard D. Helms, Leland B. Jackson, James F. Kaiser, Charles M. Rader, Ronald W. Schafer, Kenneth Steiglitz, and Clifford J. Weinstein "Terminology in Digital Signal Processing", *IEEE Transactions on Audio and Electroacoustics,* VOL. AU-20, NO. 5 March 1972.

[5] Lawrence L. Rabiner, Bishnu S. Atal, and Marvin R. Sambur, "LPC Prediction Error—Analysis of Its Variation with the Position of the Analysis Frame", *IEEE Transactions on Acoustics, Speech, and Signal Processing*, VOL. ASSP-25, NO. 5, October 1977.

[6] Paavo Alku and Susanna Varho, "A Frequency Domain Method to Improve Modeling of Formants in Speech Coding Applications of Linear Prediction", Helsinki, University of Technology.

[7] Yanli Zheng and Mark Hasegawa-Johnson, "Formant Tracking by Mixture State Particle Filter", *Icassp* 2004.

[8] Issam Bazzi, Alex Acero and Li Deng, "An expectation maximization approach for formant tracking using a parameter-free non-linear predictor", *Microsoft Research, One Microsoft Way*, Redmond, WA, USA.

[9] Jesper Högberg, "Prediction of formant frequencies from linear combinations of filterbank and cepstral coefficients", *TMH-Q*, April 1997.

[10] R. W. Schafer & L. R. Rabiner, “System for automatic formant analysis of voiced speech,” *Journal of the Acoustic Society of America*, Vol.47, N°2, pp.634-648, 1970.