Proceedings of the Institute of Acoustics

# FORMANT TRACKING : A COMPARISON OF SEVERAL PARAMETRIC METHODS

F. Plante, W.A. Ainsworth

Communication and Neuroscience Dept., Keele University, Keele, Staffs ST5 5BG

## 1. INTRODUCTION

Formants are one of the oldest representations for speech analysis and production. It is well established that formants are less powerful than the cepstral representation for recognition task but they allow the vocal tract configuration to be described. In order to improve models of production and to study speech gestures, it is important to have an accurate estimation of formants and their transitions.

Formant tracking is an old and as yet unsolved problem. With increasing computational power, new sophisticated methods are proposed which take a priori knowledge of glottal wave into account. But, if from a theoretical point of view they appear more efficient, in practice, the results are quite far from those expected. In this paper we compare several parametric methods of formant tracking in order to determine the contribution of each hypothesis assumed by these methods.

The main problem in comparing methods, is that we do not know the true values of formant frequencies and bandwidth. Until now a qualitative visual comparison has been employed. In this paper we try to establish some quantitative criteria in order to achieve a more objective comparison.

After reviewing the general scheme of formant tracking, and studying the relation between the methods, some results obtained for the four first formants with Vowel-Vowel transitions for male and female speakers are presented.

## 2. METHODS

### 2.1 Scheme of formant tracking

In order to better understand the part of analysis we review the general scheme of formant tracking in figure 1. This can be subdivided into three steps.

The first is to obtain a spectrum that represents the frequency characteristics of vocal tract from a frame of the signal. For this, we need to remove the contribution of the glottal wave from the signal. Two main approaches can be used [1]. Either a spectrum obtained with cepstral filtering or a spectrum obtained by a parametric approach. The second approach requires knowledge of the production of the signal.

The next is to extract the formants from this spectrum. Three methods are generally used. Either using a peak-picking method on the spectrum [2] or based on the derivative of its phase [3] or, in case of parametric approach, using the roots of the predictor polynomial [4]. The main problem at this stage, is to define criteria in order to differentiate formant and a spectral shaping poles. Limitations on the value of the bandwidth and frequency range for each formant are often used.

The last step is to post-process these values by considering all or a large part of the signal. It is in this stage that we can introduce criteria of smoothing and continuity. This last step will not be examined in this paper.

For an accurate analysis of formant transitions, it is important to have a small time window. In this case, the parametric approach is well known for its better frequency resolution [5]. For the parametric approach, extracting formants from roots of predictor polynomial gives better results [6].

FORMANT TRACKING WITH PARAMETRIC METHODS

As criteria for defining a formant, we used the following frequency range (in Hertz) : 200-1000 for F1; 600-2900 for F2; 1800-3700 for F3; and 2800-4500 for F4. Ranges are large to include formants of normal adult voices. In the limitation of the formant bandwith, we tested 500 and 900 Hz.

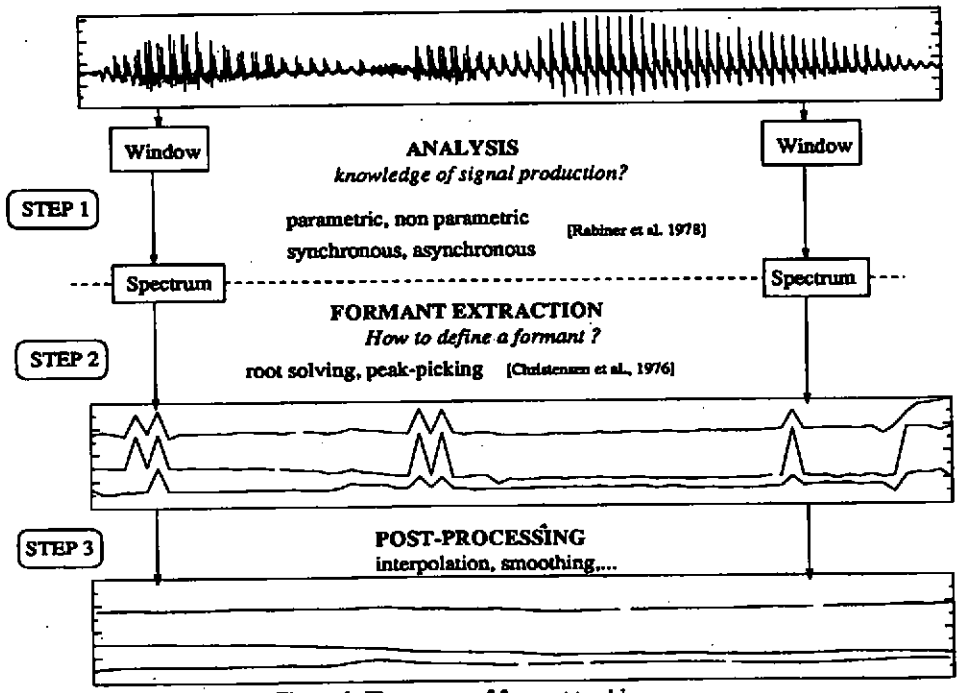


Figure 1: Three steps of formant tracking.

2.2 ARMA analysis and its derivatives

The principle of ARMA analysis is that the sample $n$ of signal $s$ is predictable from the linear combination of the P previous samples of the output ($s_{n-1}...s_{n-P}$) and the M previous samples of the input $u$ (eq.1). P is the order of the autoregressive part (AR) and M the order of the moving average part (MA).

$$s_n = -\sum_{i=1}^{P} a_i s_{n-i} + \sum_{j=0}^{M} b_j u_{n-j} \qquad \text{eq.1}$$

Using the z transform, we obtain a formulation in the frequency domain. If we define H(z) as the transfer function of the system we have :

$$H(z) = \frac{S(z)}{U(z)} = \sum_{j=0}^{M} b_j z^{-j} \Big/ \left(1 + \sum_{i=1}^{P} a_i z^{-i}\right) \qquad \text{eq.2}$$

The error or residual $e_n$ is defined as :

$$e_n = s_n - \hat{s}_n \qquad \text{eq.3}$$

If the signal $s$ is created by a true ARMA process, the residual is zero. In other cases, it carries information about the difference between the model and the reality.
From values of $M$, $P$ and the form of $u$, we obtain different types of analysis, based on different hypotheses. Figure 2 shows the links between these different analyses. There are two main ways of simplifing the solving of equation 1.

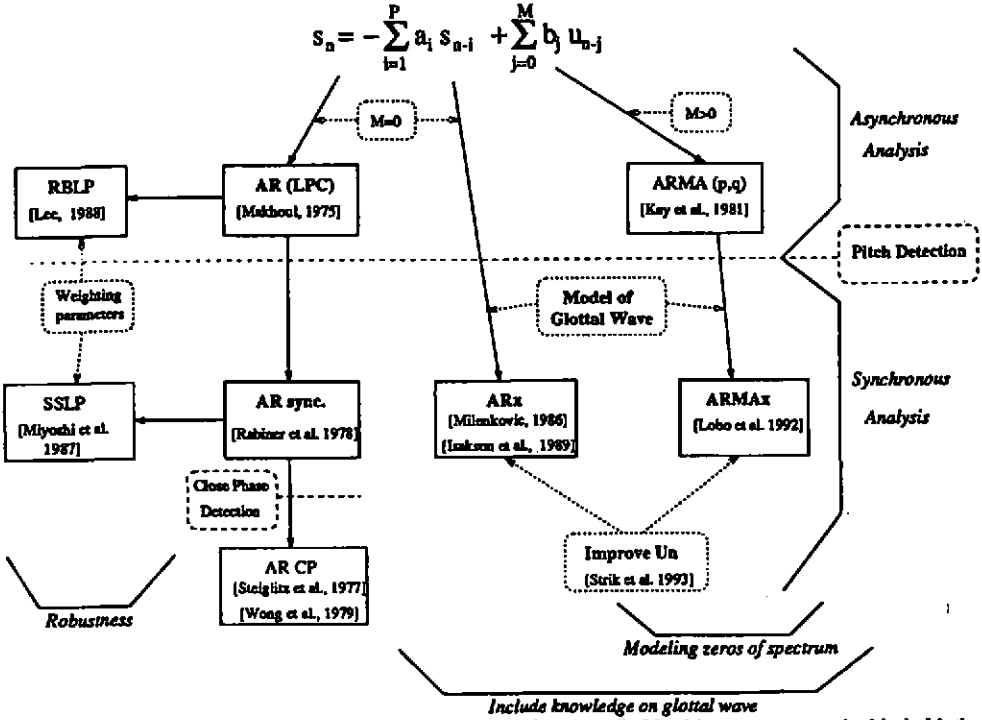


Fig 2 : Link between several parametric methods. The analyses studied in this paper are marked in bold, the aims of the analyses are marked in italic. Hypotheses or constraints are given in dotted box.

2.2.1 Linear Prediction and derivatives. The oldest way is to take M and $u$ equal to 0. We obtain AR analysis or Linear Prediction [7]. Equation 1 is now modified as shown in equation 4. In this method the glottal wave excitation is consider to be unknown .

$$\hat{s}_n = -\sum_{i=1}^{P} a_i s_{n-i} \qquad \text{eq. 4}$$

The transfer function becomes : $H(z)=1/(1+\sum_{i=1}^{P} a_i z^{-i})$ eq.5

and the residual : $e_n = s_n - \hat{s}_n = s_n - \sum_{i=0}^{N} a_i s_{n-i}$ eq.6

Comparing eq.2 and eq.5, it emerges that the transfer function had only poles in its spectrum. This error is important only in the case of nasal sounds, when the nasal cavity is coupled with the vocal tract. Thus the residual $e_n$ contains information about this error and information about the excitation (eq.3 and eq.6). This is interesting from the point of view of pitch analysis [8], but it is disadvantageous for formant tracking. In effect, in the model we assume that the spectrum of the residual is flat (minimisation in the sense of least squares). If the residual contains excitation information, it has a slope at -6dB/octave. To improve the analysis, we could be employ pitch synchronous analysis. Analysis performed on one pitch period gives better results [1]. Nevertheless, in this case, we are always in disagreement with the theoretical model. In effect as with the pitch period, the glottal wave does not have a flat spectrum. In order to agree with the hypothesis of the LP, we need to perform the analysis only on the closed phase period [9, 10]. The main problem of this method is to determine the closed phase. If the laryngograph signal is recorded simultaneously this task can be quite easy [11], but with only the speech signal it is an open problem [12].
Another approach to improve the result of LP analysis is to weight the prediction coefficients [13]. This approach is mainly used to increase the robustness against the noise[14]. We do not consider these methods in this paper.

2.2.2 Introducing knowledge about the excitation. Derivatives of LP analysis try to minimise the influence of the error introduced by not taking the excitation into account. So it seems useful to introduce information about the glottal wave in the parametric model [15,16,17]. For this we replace $u_n$ by $g_n$, a model of glottal wave. We used a modified version of Lilijencrants-Fant model [18]. We distinguish four parts in the pitch period described by 6 parameters. This implies that the analysis is performed on one pitch period.
Either we considered that the all-pole model is sufficient and in this case with put M=0, (ARx analysis) or we take M>0 to take the zeroes in the spectrum into account (ARMAx analysis). We study the cases M=0 and M=6.
In both cases, the problem of this analysis is to fit the model of the glottal wave with reality (GAR and GARMA analyses). Some non-linear optimization methods exist [19], the more important is to define some criteria for stopping the iteration. We can look at the difference between the residual and the glottal model, or the difference between resynthesis speech and real speech taking either the filter given by the analysis or a filter with "good" formants. No comparison was performed between these different ways. We decided to choose the first for computational time aspect and used the Levenberg-Marquardt method [20].
This approach seems to give good results, but few comparisons have been made with other analyses [21]. We study the exact improvement brought about by this method in comparison of asynchronous or synchronous LP.

2.3 Criteria of comparison
The main problem for comparing different formant tracking methods is that we have no method of knowing the real formant values. In pitch detection for example, the laryngographic signal gives the true value of pitch. From this it is possible to establish criteria and achieve quantitative comparison [22].

FORMANT TRACKING WITH PARAMETRIC METHODS

At present, in formant tracking the comparison is achieved by visual comparison with the spectrogram. The best method is selected based on aspect of continuity and smoothing. The problem is that these criteria could be obtained from any methods with good post-processing. It is for this reason that we do not do this in order to compare the methods.
We can introduce criteria which give indication on the continuity and stability of formant varaition. First we computed the percentage of missing formant values (MFV). For a formant, the analysis not provides value contains in frequency range of formant. Secondly, we computed the frequency stability of the formant k using Perturbation Quotient (PQ) eq.7. It is the sum on all the formant transition of the difference between a value and local mean value. This parameter has been used for compute the stability of the pitch [23].

$$PQ(k) = \frac{100}{N-2} \sum_{i=2}^{N-1} \frac{\left| \frac{F_{i+1}^k + F_i^k + F_{i-1}^k}{3} - F_i^k \right|}{F_i^k} \qquad \text{eq.7}$$

These parameters enable the continuity and smoothing of the formants to be described, as in a visual comparisons. Nevertheless, these parameters do not allow to know if and where an error occued.
We used test of signes to compare results obtained with two values of bandwidth, of between pitch synchronous analyses. A classical test of comparison of percentage is used to compare LP with other analyses.

## 3. RESULTS

### 3.1 Database and parameters

The comparison of methods is achieved with a database of vowel transitions in $V_1V_2$ context produced by one male and one female speakers. The six vowels are /a/, /i/, /u/, /o/, /ɛ/ and /ɜ/.
The signal was sampled at 16kHz. The detection of Glottal Closure Instant is achieved with inverse filtering method [24]. The classical linear prediction is performed with a window of 32 ms, with an overlap of 50%.
All analyses are based on the method of covariance and use the QR transformation to resolve the system of equations. In the case of non stability of the filter, roots outside the unit circle were reflected inside [4].

### 3.2 Quantitative comparison.

Results for male and female speaker are given respectively in table 1 and 2. For both speakers, classical Linear Prediction (LPC) gives best results, in terms of smoothing (smaller PQ values). This is expected because the analysis is performed on a quite large window (32 ms), which corresponds to 3 to 6 pitch periods. However, the percentage of MFV is significatively higher for the LPC analysis.
We can see that using a limitation of bandwith at 900 Hz decrease the percentage of MFV significantly which coincides with an increase of the perturbation quotient that is significant for the female speaker.
There are no significant differences between the different analyses, either the missing or the perturbation quotient.

### 3.2 Visual comparison

In figure 3, we show formant tracking obtained with different methods with bandwith limited to 500 Hz for the utterance /ɛa/ spoken by the female speaker.

It can be seen that LPC gives the more regular values, but does not allow the formant trajectories to be accurately followed during the transitions. This error will be aggravated in the case of faster transitions.
As with MFV and PQ parameters, there are not many differences between methods. The errors are situated in the same areas for all the methods.

| | | Bandwidth<500 | | | | | Bandwidth<900 | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| male speaker | | F1 | F2 | F3 | F4 | Total | F1 | F2 | F3 | F4 | Total |
| LPC | *MFV* | *2.7* | *2.8* | *18.4* | *25* | *12.2* | *1.9* | *0.6* | *2.9* | *5.9* | *2.8* |
| | PQ | 1.44 | 2.45 | 1.41 | 1.02 | 6.32 | 1.56 | 2.34 | 1.36 | 2.4 | 7.66 |
| AR | *MFV* | *3.5* | *3.9* | *18.9* | *24.6* | *12.7* | *2.2* | *0.6* | *2.1* | *7.1* | *3* |
| sync | PQ | 3.7 | 6 | 3.29 | 3.05 | 16.04 | 4.02 | 5.02 | 3.31 | 4.3 | 16.65 |
| ARx | *MFV* | *3.6* | *3.3* | *17.7* | *23.6* | *12.1* | *2.1* | *0.3* | *2.2* | *6.4* | *2.7* |
| | PQ | 3.36 | 5.52 | 2.98 | 3.11 | 14.97 | 3.45 | 4.86 | 3.14 | 4.34 | 15.79 |
| GAR | *MFV* | *3.2* | *3.2* | *17.5* | *23.7* | *11.9* | *1.9* | *0.3* | *2* | *6.4* | *2.7* |
| | PQ | 3.48 | 5.56 | 3.07 | 3.07 | 15.17 | 3.55 | 4.85 | 3.07 | 4.25 | 15.73 |
| ARMAx | *MFV* | *3.2* | *3.5* | *16.6* | *23.3* | *11.7* | *2* | *0.5* | *2.1* | *6.3* | *2.7* |
| | PQ | 3.28 | 5.3 | 3.15 | 3.36 | 15.09 | 3.38 | 4.77 | 3.29 | 4.4 | 15.84 |
| GARMA | *MFV* | *2.9* | *3.4* | *16.7* | *23.5* | *11.6* | *2.1* | *0.5* | *2* | *5.9* | *2.6* |
| | PQ | 3.33 | 5.47 | 3.24 | 3.41 | 15.45 | 3.46 | 4.89 | 3.29 | 4.39 | 16.03 |

Table 1 : Percentage of missing and frequency pertubation for male speaker.

| | | Bandwidth<500 | | | | | Bandwidth<900 | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| female speaker | | F1 | F2 | F3 | F4 | Total | F1 | F2 | F3 | F4 | Total |
| LPC | *MFV* | *0.7* | *4.8* | *27.6* | *28.1* | *15.3* | *0.4* | *1.3* | *10.3* | *12.4* | *6.1* |
| | PQ | 0.89 | 0.85 | 1.11 | 0.79 | 3.64 | 0.98 | 1.65 | 2.19 | 2.91 | 7.73 |
| AR | *MFV* | *1.3* | *2.3* | *14.9* | *21.9* | *10.1* | *0.5* | *0.3* | *5.7* | *8.3* | *3.7* |
| sync | PQ | 3.88 | 4.35 | 3.04 | 3.85 | 15.2 | 3.73 | 4.58 | 4.08 | 5.65 | 18.04 |
| ARx | *MFV* | *1.9* | *3.4* | *16.1* | *22.1* | *10.9* | *0.8* | *0.5* | *6.1* | *8.5* | *4* |
| | PQ | 4 | 4.43 | 2.81 | 3.81 | 15.04 | 3.99 | 5.36 | 4 | 5.48 | 18.83 |
| GAR | *MFV* | *1.2* | *3.5* | *15.7* | *22.5* | *10.7* | *0.3* | *0.7* | *6* | *8.6* | *3.9* |
| | PQ | 3.9 | 4.83 | 3.05 | 3.91 | 15.69 | 3.98 | 5.56 | 4.23 | 5.58 | 19.36 |
| ARMAx | *MFV* | *1* | *3.2* | *14.9* | *21.2* | *10.1* | *0.5* | *0.6* | *6.4* | *7.5* | *3.7* |
| | PQ | 3.09 | 4.92 | 3.34 | 4.14 | 15.49 | 3.17 | 6.02 | 4.68 | 5.95 | 19.83 |
| GARMA | *MFV* | *0.9* | *3.5* | *14.9* | *21.4* | *10.2* | *0.4* | *0.6* | *6.1* | *7.4* | *3.6* |
| | PQ | 3.32 | 4.98 | 3.51 | 4.32 | 16.12 | 3.39 | 5.98 | 4.64 | 5.85 | 19.87 |

Table 2: Percentage of missing and frequency perturbation for the female speaker.

## 5. CONCLUSION

In this paper, we compared some parametric approaches for extracting formants. Contrary to most studies of this kind, we used objective criteria in order to compare the methods.
We did not observe significant differences between analyses excepted for the case of asynchrounous linear prediction. The choice of the limitation in the bandwith of formant seems more important, especially for the female speaker.
Nevertheless, we cannot decide which is the best method according to the criteria of MFV and PQ or visual comparison for all the database. To obtain the best formant tracking, it will necessary to adapt the choice of the analysis with the utterance and speaker. For this, using of objective parameters could be a great help.

An effort must be made in the way to establish reference parameters, as in pitch analysis for example. In this work only continuity and smoothing criteria were used. For a complete and objective comparison of methods, introduce a measure from the spectrogram will be necessary, in order to reproduce criteria of visual comparison and allow a localisation of errors.
Another limitation of this study, is that with increasingly sophisticated analysis methods increaing numbers of criteria are needed. It is difficult to establish wheter the optimum parameter set has been found.
Formant tracking is as yet an unresolved problem. It is utopian to want a method that will work for every type of speech. We need to adapt the analysis to the signal. For this we must establish objective criteria which could help us in this adaptation. It in this direction that futher work will be performed.

## 6. REFERENCES

[1] L.R. RABINER R.W. SCHAFER "Digital processing of speech signals"Prentice-Hall, New Jersey, (1978).
[2] S S McCANDLESS " An algorithm for automatic formant extraction using linear prediction spectra" IEEE ASSP Vol.22 135-141, (1974).
[3] B YEGNANARAYANA "Formant extraction from linear prediction phase spectra" JASA Vol.63 1638-1640 (1978).
[4] B S ATAL S L HANAEUR "Speech analysis and synthesis by linear prediction of the speech wave" JASA Vol.50 637-655, (1971).
[5] S M KAY, S L MARPLE "Spectrum analysis- A modern perspective" Proc. IEEE Vol.69, pp1380-1419 (1981).
[6] R L CHRISTENSEN, W J STRONG, P PALMER "A comparison of three methods of extracting resonance information from prediction coefficient coded speech" IEEE ASSP Vol 24 8-14 (1976).
[7] J MAKHOUL "Linear Prediction : A tutorial review" Proc. IEEE Vol.63 561-580 (1975).
[8] J.D MARKEL, A. H. GRAY "Linear Prediction of speech" Springer-Verlag, Berlin, (1976)
[9] K STEIGLITZ, B DICKINSON "The use of time domain selection for improved linear prediction" IEEE ASSP Vol.25 34-39 (1977).
[10] D Y WONG, J D MARKEL, A H GRAY "Least squares glottal inverse filtering from the acoustic speech waveform" IEEE ASSP Vol.27 350-355, (1979).
[11] A K KRISHNAMURPHY, D CHILDERS "Two channel speech analysis" IEEE ASSP Vol.34 730-743 (1986).
[12] W. HESS "Pitch determination of speech signals : Algorithms and devices" Springer Verlag, Berlin, 1983.
[13] Y MIYOSHI, K YAMATO, R MIZOGUCHI, M YANAGIDA, O KAKUSHO "Analysis of speech signals of short pitch period by a sample selective linear prediction" IEEE ASSP Vol.35 1233-1240 (1987).
[14] C H LEE "On robust linear prediction of speech" IEEE ASSP Vol. 36 642-650 (1988).
[15] Y M CHENG D O'SHAUGHNESSY "Parameter sensitivity and robust estimation in an ARX model with glottal excitation" ICASSP 89 Vol.1 564-567 (1989).
[16] A ISAKSON M MILLNERT "Inverse glottal filtering using a parametrized input model" Signal Processing Vol.18 435-445 (1989).
[17] P MILENKOVIC "Glottal inverse filtering by joint estimation of an AR system with a Linear input model" IEEE ASSP Vol.34 28-42 (1986).
[18] A LOBO "Estimation of the voice source and its modeling in speech synthesis" PhD Keele University (1992).
[19] H STRIK, B CRANEN, L BOVES "Fitting a Lilijencrants-Fant model to inverse filter signals" EUROSPEECH'93 Vol.1 103-106 (1993)
[20] W. H. PRESS, S. A TEUKOLSKY W. T. VETTERLING B. P. FLANNERY "Numerical recipes in C" Cambridge University Press 683-688, 1992.
[21] A LOBO, W A AINSWORTH "Evaluation of a glottal ARMA model of speech production" ICASSP 92 (1992).
[22] L R RABINER B S ATAL M R SAMBUR "LPC prediction error analysis of its variation with the position of the analysis frame." IEEE ASSP Vol.25 434-442, (1977).
[23] Y. KOIKE "Application of some acoustic measures for the evaluation of laryngeal dysfunction" Studia Phonologica, Vol.7 17-23, (1973).
[24] F. PLANTE, G. MEYER, W. A. AINSWORTH "Pitch analysis : Auditory model versus inverse filtering". Proc Inst. Acoust. (1994).

ACKNOWLEDGMENT

This work was supported by Contract SCI-CT92-0786 of the EC Science Programme.

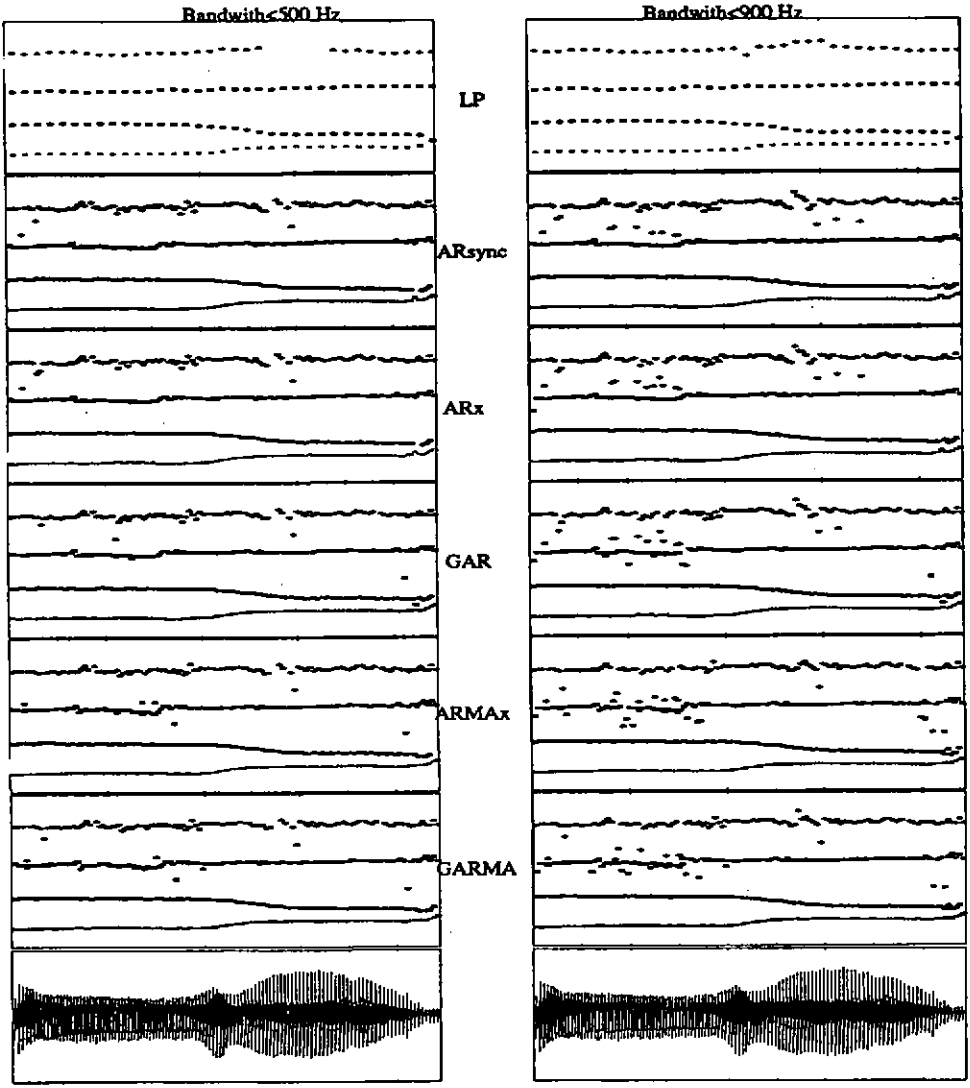


Figure 3: Formant tracking obtains by analyses with bandwith limited to 500 and 900 Hz for the utterance /ea/ spoken by the female speaker.