Theoretical Survey of the Formant Tracking Algorithm

Miss. Manisha H. Awatade
Electronics Department, Walchand Institute of Technology,
Solapur, Maharashtra, India

ABSTRACT
The formant is the important part of the phonetic characters, and reliable formant tracking algorithm is the base to study the phonetics. Based on the development course of the phonetic formant tracking algorithm, the linear prediction coding (LPC) and the model matching method are introduced emphatically, and there own advantages and disadvantages are analyzed, and the model matching method based on the hidden dynamic model will be the development direction of the future formant tracking technology.

General Terms
Speech recognition, speaker recognition.

Keywords
Formant, Tracking, Linear prediction coding, Model matching

1. INTRODUCTION
When pronouncing, the air currents pass the vocal tract, which will induce the resonance of the channel, and generate a group of resonance frequency that is called as the formant frequency, i.e. the formant for short. The formant is the important parameter to differentiate different vowels. The algorithm to position and mark the tracks of the change of the formant frequency with the time is called as the formant tracking algorithm. The formant tracking is the reflection of speaker's individual character. The acquirement of the formant parameter and the tracking algorithm have been widely used in the speaker recognition, the speech synthesis and the speech coding transfer, and they are the important research topics in the speech signal processing domain.

Formant is one of the major components of speech. The frequencies at which the resonant peaks occur are called the formant frequencies or simply formants. The formant of the signal can be obtained by analyzing the vocal tract frequency response

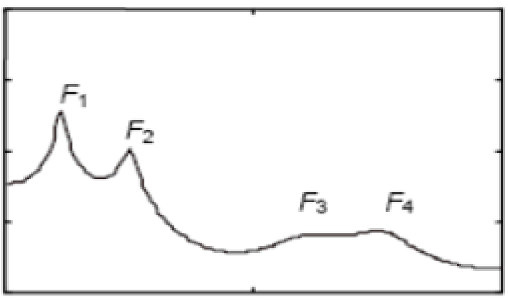

Fig 1: Vocal tract frequency response

Fig. 1 shows the vocal tract frequency response. The x-axis represents the frequency scale and the y-axis represents the magnitude of the signal. As it can be seen, the formants of the signals are classified as F1, F2, F3 and F4. Typically a voice signal will contain three to five formants. But in most voice signals, up to four formants can be detected. In order to obtain the formant of the voice signals, the LPC (Linear Predictive Coding) method is used. The LPC (Linear Predictive Coding) method is derived from the word linear prediction. In order to estimate the formants, coefficients of the LPC are needed. The coefficients are estimated by taking the mean square error between the predicted signal and the original signal. By minimizing the error, the coefficients are detected with a higher accuracy and the formants of the voice signal are obtained.

Based on the important meaning and the wide application foreground of the formant to the speech signal processing, many scholars have applied themselves to the study of the acquirement of the formant parameter and the tracking algorithm in recent years, and new algorithms are continually pushed. By the research and analysis of literatures, these algorithms can be classified as two sorts, i.e. the LPC method and the model matching method. These two methods have their own advantages and disadvantages. The separation of the linear prediction equation can exactly confirm the central frequency and bandwidth of the formant, but the ascriptions of pecks are difficult to be judged in LPC, which can be avoided by the model matching method, but the model needs to be trained by large numbers of data, and the training result depends on the quantity and kind of the training data. These two methods will be briefly introduced and analyzed as follows.

2. LINEAR PREDICTION CODING METHOD
First LPC method used is to acquire three front formant frequencies and extents in the phonetic fragment of vowel [1]. By the estimated characters of LPC spectrum, in the region that the energy of signals is strong, i.e. the region closing to the peak value of the spectrum, the LPC spectrum is closing to the signal spectrum, but in the region that the energy of signals is weak, i.e. the region closing to the vale of the spectrum, both spectrums are significantly different. So to check the peak values of the LPC spectrum can confirm the formant. In ideal situation, three front formants of the speech are three front formants of the LPC spectrum. An algorithm is presented which finds the frequency and amplitude of the first three formants during all vowel-like segments of continuous speech. It uses as input the peaks of the linear prediction spectra and a segmentation parameter to indicate energy and voicing. Ideally, the first three peaks are the first three formants. Frequently, however, two peaks merge, or spurious peaks appear, and the difficult part is to recognize such situations and deal with them. The general method is to fill formant slots with the available peaks at each frame, based on frequency position relative to an educated guess. Then, if a

peak is left over and/or a slot is unfilled, special routines are called to decide how to deal with them. Included is a formant enhancement technique, analogous to a similar technique which has been implemented via the chirp-z transform, which usually succeeds in separating two merged formants. Processing begins at the middle of each high volume voiced segment, where formants are most likely to be correct, and branches outward from there in both directions in time, using the most recently found formant frequencies as the educated guess for the current frame. The algorithm has been implemented at Lincoln Laboratory on the Univac 1219 and the Fast Digital Processor, a programmable processor, and has been tested on a large number of unrestricted sentences.

The LPC analysis algorithm is improved [2], and the new algorithm adopted the cepstral spectrum coefficient of LPC to acquire the parameters of formant. Comparing with [1] LPC spectrum estimation algorithm, the robustness of the improved algorithm was better when acquiring the formant of the fragment of vowel.

The Mel frequency scale according with human ear's hearing character is combined with the LPC analysis to estimate the first formant F1 and the second formant F2 of speech signals [3]. Traditional LPC algorithm needed to confirm the orders of the linear predictor according to the amount of the acquired formant, but this algorithm could acquire F1 and F2 by setting up a fixed order of the predictor, and needed not to change the orders of the linear predictor by changing the amount of the acquired formant.

An automatic tracking method of phonetic formant track adopted the dynamic programming method to realize the tracking of the formant by introducing the continual limited conditions of frequency [4]. First, acquire the candidate values of the formant frequency by seeking the roots of the linear predicted equation, then, establish a stationarity function as the limited condition of the frequency continuity, finally, by an improved Viterbi algorithm, compute the minimum value between the mapping value of the formant frequency in the current frame and the mapping value of the formant frequency in the last frame in the limitation of stationarity function to realize the track connection among frames of the formant. The key of this algorithm is to design a proper frequency continual limited stationarity function. However, the experiments showed that it was very difficult to design reasonable stationarity function [5].

A new formant-tracking algorithm using phoneme information is proposed [6]. Conventional formant-tracking algorithms obtain formant tracks by analyzing the acoustic speech signal using continuity constraints without any additional information. The formant-tracking error rate of the conventional methods is reportedly in the range of 10%–20%. In this paper, we show that if text or phoneme transcription of speech utterances is available, the error rate can be significantly reduced. The basic idea behind this approach is that given the phoneme identity, formant-tracking algorithms can have a better clue of where to look for formants. The algorithm consists of three phases: 1) analysis, 2) segmentation and alignment, and 3) formant tracking by the Viterbi searching algorithm. In the analysis phase, formant candidates are obtained for each analysis frame by solving the linear prediction polynomial. In the segmentation and alignment phase, the text corresponding to the input speech utterance is converted into a sequence of phoneme symbols. Then, the phoneme sequence is time aligned with the speech utterance. A hidden Markov model (HMM) based automatic segmentation algorithm is used for forced-time alignment. For each phoneme segment, nominal formant frequencies are assigned at the center of each phoneme segment. Then nominal formant tracks for the entire utterance are obtained by interpolating the nominal formant frequencies. In order to compensate for the coarticulation effect, different interpolation methods are used depending on the phonemic context. The interpolation process makes the formant-tracking algorithm robust to possible segmentation errors made by the HMM-based segmentation algorithm. As a result, the proposed formant-tracking algorithm does not require highly accurate alignment/segmentation. Finally, a set of formants is chosen from the formant candidates in such a way that the resulting formant tracks come close to the nominal formant tracks while satisfying the continuity constraints. The algorithm is tested using natural speech utterances and the performance is compared against formant tracks obtained by the conventional method using continuity constraints only. The new algorithm significantly reduces the formant-tracking error rate (5.0- 3% for male and 3.73% for female) over the conventional formant-tracking algorithm (13.00% for male and 15.82% for female).

In conclusion, because the orders of the coefficient predicted by the LPC method are confirmed beforehand, so the amount of the acquired complex conjugate peak pair is the half of the orders at most. Generally, the bandwidth of the extra peak is bigger than the bandwidth of the formant, so to acquire the formant means to judge the ascription of the peak. To compute simply, for the standard vowel signals, the LPC method can exactly confirm the central frequency of the formant and the bandwidth by separating the linear predicted equation. But if the voice signals are interfered by the noise source, the fake peak and the combined peak will occur on the frequency spectrum, which will bring large difficulty to judge whether formant the peak points belong to, and influence the tracing nicety of the formant track, that is the essential deficiency of the LPC method in the formant tracking analysis.

3. MODEL MATCHING METHOD

The model matching method avoids the problem that the LPC method is easy interfered by the fake peak and the combined peak, and it is the research hot in the acquirement of formant parameters and the tracking study in recent years. The model matching method experienced a development process from the HMM model to the HDM model.

In 1975, Baker put forward the idea to adopting the hidden Markov model (HMM) to trace the formant track [7]. but the experiment failed. Ten years ago, G. Kopec first successfully used the formant tracking method based on HMM [5]. He divided the formant tracing problem into two independent problems including checking and estimation, and the formant checking was to judge whether each frame speech signal had the formant or not, and the formant estimation is to endow certain frequency value for the checked formant. The checking and estimation of formant all adopted the Viterbi algorithm to search the optimal status sequence of HMM. This algorithm adopted the statistical method to realize the tracking of formant, but it could only realize the tracking of one formant.

After that, G. Kopc put forward the improved formant tracking algorithm based on HMM and the vector quantization (VQ) technology [5]. Comparing with the method in the literature two aspects were improved in the new algorithm. First, by set up two tracing modes, i.e. the single formant tracing mode and the multi-formant tracing mode, multiple formant tracks could be traced simultaneously, and the problem which could trace only one formant track was solved. Second, adopt the forward-backward algorithm (F-B algorithm) to replace the Viterbi algorithm to check and estimate the formant, because the Viterbi algorithm could generate one single status sequence, not a probability distribution, which would produce two problems. The first one is the problem that the formant checking is difficult, and the formant checking based on the Viterbi algorithm is to control the probability of the error checking and the peak value omission by setting up the thresholds, and it can not directly adjust the thresholds in the tracing process, so it is not flexible enough. And if the checking and the estimation of formant are implementing simultaneously, the formant checking performance will depend on the quantity which is used to denote the status of the formant parameters. With the increase of the status density of the frequency field space, the probability that the single status is checked will decrease. Therefore, with the increase of the status quantity, the probability that the real status is checked will gradually reduce. The HMM formant tracking algorithm based on F-B algorithm will avoid this problem. The second problem is that the formant track traced by the Viterbi algorithm is a group of discrete frequency values defined by the model status, but the formant track obtained by the F-B algorithm is the weighted average value in the status of discrete model, so the track obtained by the latter will be more smooth than the track obtained by the former.

The phonetic formant tracking method based on the Gaussian mixture model (GMM) is put forward [8], [9]. Because GMM is a continually distributed HMM which status is 1, this method still can be regarded as a formant tracking method based on the HMM model. There are two deficiencies to adopt the HMM model to solve the formant parameter tracking problem. First, when the algorithm is used to estimate the status of certain time of the formant track, it only takes the continuity of the track as the restriction condition to select the formant, which will easily induce the tracing error. Aiming at this deficiency, Minkyu Lee et al put forward an improved method [10]. By estimating the status of certain time of the formant track, they combined the phoneme information based on the speech signal text with the continuity of the track to be the restriction condition of selecting the formant, which could enhance the precision of the tracking. D. T. Toledaro et al also put forward similar improved method [11]. These methods could significantly enhance the tracing precision when they were used to trace the formant track of special people's sound which is related with the text, but for the tracking of non-special people's sound which is not related with the text, the had not obviously improved effects. Second, the algorithm needed large numbers of data to train the model, and the result of the final tracing was decided by the kind and the quantity of the training data. But in different using environments, whether the training data have sufficient representative quality or not could not be confirmed. This deficiency is instinctive for HMM. HMM is general statistical model which is widely used in many different domains. If it is applied in some special domain, special data will be needed to train the model. That is to say, the HMM is a data-driven model, and it doesn't involve any mechanism about the generation of data. So in the actual environment with noises and interferences, the formant tracking method based on HMM can not fulfill the actual requirements. To overcome this deficiency, the new established model should not only consider the observation data but also the pronunciation mechanism of the speech signals to describe the speech signals.

In the late of 1990s, L. Deng put forward a dynamic speech modeling method combining the metrics characters and the phonetic characters [12]. This method considered the conversion of the co-articulation and the neighboring phones in the pronunciation for the modeling process of the speech signals. It regards the pronunciation system of sound as a hidden dynamic system, and in which, each phone corresponds with one vector objective, i.e. when certain phone is pronounced, the muscles of the vocal cords and the track will approach certain objective status or shape according to the "program". This modeling method which is specially used for sound considers the generation mechanism of the speech signals, and gets rid of the modeling mechanism which only is driven by the data. Almost in the same term, Richards et al put forward similar speech modeling method which was named by the hidden dynamic model (HDM) [13]. To describe the dynamic structure of the sound, Richards mapped the hidden space on the phonetic character space by the nonlinear multiplayer perceptron (MLP), and trained the parameters of the model (target vector and the weight value of MLP) by the selected algorithm. HDM was successfully applied in the speech recognition [12], [14] and many new phonetic formant tracking algorithms were developed based on that. For example, I. Bazzi put forward a formant tracking algorithm based on the expectation maximization (EM) [16]. This algorithm is composed by two parts, and one is the acquirement of the mapping relationship between the formant and the phonetic observation information, which maps the formant parameters on the Mel frequency cepstral coefficient by a parameter-free nonlinear predictor, and establishes the prediction code text, and the other is the acquirement of the residual information of speech signals, which adopts the EM algorithm to train the residual coefficients of the speech signals and search the optimal format parameters in the prediction code text. Combining with the restriction condition with the target orientation, L. Deng et al put forward a nonlinear predictor which could be used in tracking of VTRs [16]. This nonlinear predictor maps the formant parameters on LPCC, not on MFCC, and because the LPCC has good separation character, so it can enhance the computation efficiency.

Above two algorithms all first quantifies the parameter space of formant, and maps the quantified formant parameters on MFCC or LPCC to form the prediction code text, and finally selects the optimal formant parameters by training the residual coefficients. To quantify the parameter space, the quantifying dimension should be selected, and too big quantifying dimension will produce large computation, and too small quantifying dimension will influence the tracing precision. Aiming at this problem, L. Deng also put forward an improved method. He regarded the formant parameters as the continual values of variables in the hidden status, and adopted the Kalman filtering and the smoothing technology to trace the track of VTRs, which could solve the problem induced by the quantification of frequency field space [15]. This method introduced extra prior information by the form of VTRs in the VTRs tracking process. Because the prior information can capture the timing character of VTRs track in the generation process of sound, so this method can exactly trace not only the VTRs with obvious frequency spectrum peak value, but also the VTRs phonetic segments (such as stops, spirants) without obvious formant structure. But this method has a deficiency, i.e. it needs to linearly process the nonlinear predictor, because the Kalman filtering is an implementation of Bayes filtering, and it is the optimal linear filter under the rule of the minimum square error, so it can not be used in the nonlinear occasions. The linear processing of nonlinear speech model will not only increase the computation, but the linearized model can not often represent real nonlinear model. To overcome the deficiency that the nonlinear predictor needs linear processing, foreign scholars applied the particle filtering technology in the formant tracking based on the HDM model in recent years. The particle filtering is another method to realize the Bayes filtering, and it adopts a group of randomly weighted particles to approach the posterior probability distribution, and because it is not limited by the linear Gaussian conditions, so it is widely applied in the control domain. Yanli Zheng et al first applied the particle filtering technology in the tracking of phonetic formant [17]. This method can process the nonlinear model, and needs not implement the linear processing to the nonlinear speech model, and it is easy to be implemented. But Yanli Zheng et al only offered a developmental idea, and the speech model they used was the simplified HDM without target orientation, so the tracing precision still needs to be enhanced.

4. SUMMARY

The formant tracking technology of speech signals is continually developing, and many foreign and domestic scholars are applying themselves to the research about it and have put forward many methods and algorithms. The formant tracking method developed from the LPC analysis method to the HMM model matching method and to the HDM model matching method. Now, the formant tracking method based on HDM is more and more emphasized by researchers. Of course, the HDM model matching method still has some deficiencies, for example, whether the established model can exactly describe the character of speech signals enough, and how to enhance the precision of the model tracing when simplifying the computation. As the research develops, these problems will be solved gradually, and the HDM model matching method will certainly exert important function in the domain of the speech signal formant tracking.

5. ACKNOWLEDGMENTS

Author likes to thank all the guidance and support obtained by our professors and colleagues in this work.

6. REFERENCES

] S.S.McCandless. 1974. An algorithm for automatic formant extraction using linear prediction spectra. IEEE Transactions on Acoustics, Speech, and Signal Processing. No.22(2). p.135-141.

[2] D.J.Broad, F.Clermont. 1989. Formant estimation by linear transformation of the LPC cepstrum. Journal of the Acoustical Society of America. No.86(5). p.2013-2017.

[3] A.M.De Lima Araujo, F.Violaro. (1998). Formant frequency estimation using a mel scale LPC algorithm. Telecommunications Symposium, ITS '98 Proceedings. Vol.1. p.207-212.

[4] D.Talkin. 1987. Speech formant trajectory estimation using dynamic programming with modulated transition costs. Journal Acoustical Society of America. No.82(S1). p.S55.

[5] G..Kopec. 1986. Formant Tracking Using Hidden Markov Models and Vector Quantization. IEEE Transactions on Acoustics, Speech, and Signal processing. No.34(4). p.709-729.

[6] Lee, M.; vanSanten, J.; Mobius, B.; Olive, J.; 2005 Formant tracking using context dependent phonemic information IEEE transaction on speech and audio processing vol.13.p.741-750.

[7] J.K.Baker. 1975. The Dragon System-An overview. IEEE Transactions of Acoustics, Speech, and Signal Processing. No.23(1). p.24-29..

[8] P.Zolfaghari, T.Robinson. 1996. Formant analysis using mixtures of Gaussians. Proc. ICSLP. Vol.2. p.1229-1232.

[9] J.Darch, B.Milner, S.Xu, S.Vaseghi and Y.Qin. 2005. Predicting formant frequencies from MFCC vectors. Proc. ICASSP. Vol.1. p.941-944.

[10] M.Lee, J.vanSanten, B.Mobius, J.Olive. 2005. Formant Tracking Using Context-Dependent Phonemic Information. IEEE Transactions on Speech and Audio Processing. No.13(5). p.741-750.

[11] D.T.Toledano, J.G..Villardebo, L.H.Gomez. 2006. Initialization, training, and context-dependency in HMM-based formant tracking. IEEE Transactions on Acoustics, Speech, and Signal processing. No.14(2). p.511-522.

[12] L.Deng. 1998. A dynamic, feature-based approach to the interface between phonology and phonetics for speech modeling and recognition. Speech Communication. No.24(4). p.299-323.

[13] H.B.Richards, J.S.Bridle. 1999. The HDM: A segmental hidden dynamic model of coarticulation. Proc. ICASSP. Vol.1. p.357-360.

[14] L.Deng, J. Ma. 2000. Spontaneous speech recognition using a statistical coarticulatory model for the vocal-tract-resonance dynamic. Journal of the Acoustical Society of America. No.108(6). p.3036-3048.

[15] L.Deng, L.J.Lee, H.Attias, A.Acero. 2007. Adaptive kalman filtering and smoothing for tracking vocal resonances using a continuous-valued hidden dynamic model. IEEE Transactions on Audio, Speech, and Language processing. No.15(1). p.13-23.

[16] L.Deng, A.Acero, I.Bazzi. 2006. Tracking vocal tract resonances using a quantized nonlinear function embedded in a temporal constraint. IEEE Transactions on Acoustics, Speech, and Language processing. No.14(2). p.425-434.

[17] Yanli.Zheng, M. Hasegawa-Johnson. 2004. Formant tracking by mixture state particle filter. Proc. ICASSP. Vol.1. p.565-568

.