Mixture Model Based Approach for Formant Tracking

Parimala K
ee15metech11024@iith.ac.in

Annapurna Kala
ee13b1014@iith.ac.in

***Abstract*— The problem statement is to track the formants in continuous speech robustly. The approach consists on two major steps.**
1) Computing the Pyknogram.
2) Trying to fit the pyknogram in a mixture model.
Gaussian Mixture Model and Student's-t Mixture Model are implemented. Experimental results on TIMIT database are shown for both the mixture models.

I. Introduction

A formant is an acoustic resonance of the human vocal tract system. Phonemes are detected based on the spectro-temporal information in formants. They are alo used in analysis, synthesis and enhancement. Tracking formants is difficult considering the fact that peaks are often missed due to windowed voice source impulse train. And noise further increases the problem.
Previously Linear Prediction based formant tracking was implemented. But the basic assumption of LP Model is that the Vocal Tract System (VTS) contributes to only spectral peaks. This leads to inaccurate predictions for spectral zeros in case on nasals.
The nonparametric approaches are based on the observation that the formants are characterized by local energy maxima and also correspond to frequency modulations in the spectrum of speech. A combination of local amplitude/frequency-demodulation is used for formant tracking. Multiband Demodulation Algorithm (MDA) of Potamianos and Maragos[1] is used for this purpose. The main output of the MDA algorithm is the pyknogram, which is effectively a two-dimensional representation of the density of the instantaneous frequencies of signal components present in various frequency bands. Pruning the output of MDA could a solution, but its stil not reliable.

II. Approach

We used a hybrid approach to tackle the above mentioned problems. Its hybrid as it uses both non-parametric approach (MDA) as well as parametirc approach (statistical modelling of pruned pyknogram) to estimate formants. Since the pyknogram density is multimodal, mixture models are used. Initially speech signal is divided into frames of 10 msec and the following procedure is applied on each speech frame.

A. Computing Pyknogram

A speech signal of each frame can be divided into N AM-FM signals where each signal corresponds to a particular frequency band. A Gabor Filterbank is used for this purpose.

Gabor Filter

Each passband signal, say r(t), comprises of a combined amplitude modulation AM and frequency modulation FM structure.

$$r(t) = a(t)cos(2\pi(f_c(t) + \int q(\tau)d\tau) + \theta) \tag{1}$$

where r(t) is frequency-modulating signal and f_c(t) is the central frequency corresponding to the bandpass component.A filtering scheme is needed to isolate the resonance signal r(t) from the speech signal before demodulation can be performed. A real Gabor bandpass filter is used for this purpose with impulse response h(t) and frequency response H(f).

$$h(t) = e^{-(\alpha t)^2} \tag{2}$$

$$H(f) = \frac{(\pi)^{1/2}}{(2\alpha)}(e^{-\frac{(\pi(f-\nu))^2}{(\alpha)^2}} + e^{-\frac{(\pi(f+\nu))^2}{(\alpha)^2}}) \tag{3}$$

where ν is the center frequency of the filter chosen equal to the formant frequency F, and α is the bandwidth parameter. The center frequencies range form 300 Hz to 4 KHz with an increment of 10 Hz. Bandwidth of each filter is 100 Hz.

ESA

The energy separation algorithm ESA is used to demodulate a speech resonance r(t) into amplitude envelope a(t) and instantaneous frequency f(t) signals. The ESA is based on an energy-tracking operator invented by Teager.

$$\psi(x[n]) = x(n)^2 - x(n-1)x(n+1) \tag{4}$$

$$f(t) = \frac{1}{2\pi}\frac{\psi\hat{s}(t)}{\psi s(t)} \tag{5}$$

$$||a(t)|| = \frac{1}{2\pi}\frac{\psi s(t)}{\psi\hat{s}(t)} \tag{6}$$

where $\hat{s}$(t)=$\frac{ds}{dt}$

Spectral Moments

The short-time estimates of the bandpass signal F is obtained using the first spectral moment computed about the spectral centroid.

$$F(t,\eta_k) = \frac{\int_t^{t+T} f_k(t)||a_k(t)||^2}{\int_t^{t+T}||a_k(t)||^2} \tag{7}$$

where T denotes the duration over which the weighted average is computed and is η_k the k^{th} center frequency of the filter in the filterbank. T is 10 msec. A scatterplot of spectral moments $F(t,\eta_k)^s$ gives the Pyknogram.

Pruning

Potamianos et al. [1] proposed a heuristic approach to estimate the raw formants from the pyknogram as follows:
$RF = \eta : (F(t,\eta) = \eta)\&(F(t,\eta+1) - F(t,\eta) < 0.7)\&(\eta < 50$

B. Student's-t Mixture Model

In a pruned pyknogram, the raw formant estimates are densely populated in the formant regions. We fit a multi-modal mixture model to these formant estimates.

Distribution

The pdf of Student's-t Mixture Model is a follows:

$$p(x(t),\psi^t) = \sum_{i=1}^{4} \pi_i^t \, p_X(x(t)/\mu_i^t, \Sigma_i^t, \nu_i^t) \tag{8}$$

where

$$p_X(x) = \frac{\tau(\frac{\nu+1}{2})||\Sigma||^{\frac{-1}{2}}}{(\pi\nu)^{\frac{1}{2}}\tau(\frac{\nu}{2})(1+(\frac{\delta(x;\mu,\Sigma)}{\nu})^{\frac{\nu+1}{2}})} \tag{9}$$

where

$$\delta(x;\mu,\Sigma) = (x-\mu)^T\Sigma^{-1}(x-\mu) \tag{10}$$

The heaviness in the tail of a t-distribution can be controlled using the degrees-of-freedom parameter ν.

Expectation Maximization

The parameters of the t-distribution are to be estimated using expectation maximization.

1) M-Step: Posterior probabilities are τ and u

$$\tau_{ij}^(t,k+1) = \frac{\pi_i^{(t,k)} p(x_j(t)/\mu_i^{(t,k)}, \Sigma_i^{(t,k)}, \nu_i^{(t,k)})}{\sum_{i=0}^{4} \pi_i^{(t,k+1)} p(x_j(t)/\mu_i^{(t,k)}, \Sigma_i^{(t,k)}, \nu_i^{(t,k)})} \tag{11}$$

$$u_{ij}^{(t,k+1)} = \frac{\nu_i^{(t,k)}+1}{\nu_i^{(t,k)} + \delta(x_j(t)/\mu_i^{(t,k)}, \Sigma_i^{(t,k)})} \tag{12}$$

2) E-Step: If N(t) is the number of raw formant estimates obtained at t

$$\pi_i^{(t,k+1)} = \sum_{j=1}^{N(t)} \frac{\tau_{ij}^{(t,k+1)}}{N(t)} \tag{13}$$

$$\mu_i^{(t,k+1)} = \frac{\sum_{j=1}^{N(t)} \tau_{ij}^{(t,k+1)} u_{ij}^{(t,k+1)} x_j(t)}{\sum_{j=1}^{N(t)} \tau_{ij}^{(t,k+1)}} \tag{14}$$

$$\Sigma_i^{(t,k+1)} = \frac{\sum_{j=1}^{N(t)} \tau_{ij}^{(t,k+1)} u_{ij}^{(t,k+1)} \beta_{ij}^{(t,k+1)}}{\sum_{j=1}^{N(t)} \tau_{ij}^{(t,k+1)}} \tag{15}$$

where

$$\beta_{ij}^{(t,k+1)} = (x_j(t) - \mu_j^{(t,k+1)})(x_j(t) - \mu_j^{(t,k+1)})^T \tag{16}$$

and $\nu_i^{(t,k+1)}$ would be the solution of

$$-\psi(\frac{\nu_i^{(t,k)}}{2}) + log(\frac{\nu_i^{(t,k)}}{2}) + 1 + \frac{1}{n_i^{(t,k+1)}} \sum_{j=1}^{N(t)} \tau^{(t,k+1)}(logu_{ij}^{(t,k+1)} - u_i($$

Here ψ is digamma function and is defined as $\psi(x) = (d/dx)log\tau(x)$ and $n_i^{(t,k+1)}$=$\sum_{i=0}^{N(t)} \tau_{ij}^{(t,k+1)}$. The four formant frequencies in each frame correspond to the μ^s of the multimodal mixture model.

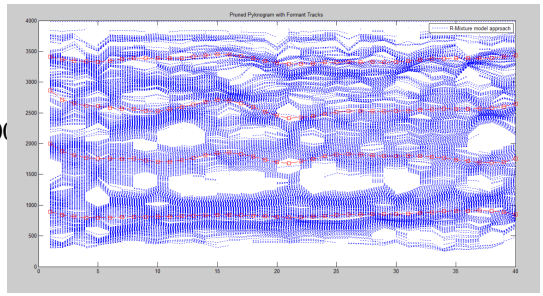

Fig. 1. Formants tracked for the phonim ba

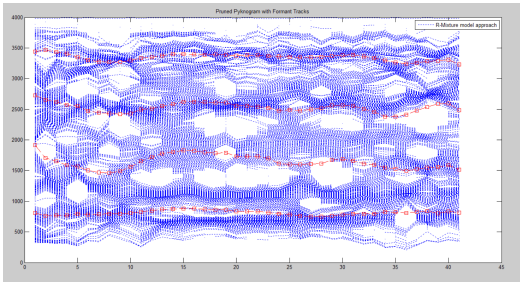

Fig. 2. Formants tracked for the phonim pa

Initialization

Initialization is done by applying K-Means clustering in each frame and smoothing over all frames. Initialization for K-means is done randomly one for every 1000 Hz.

III. Results

The approach is applied on distinct phonims as well as TIMIT speech frames. For distinct phonims, the formants are tracked properly Where as in case of speech frames, first two formants are tracked reliably but error is observed for third and fourth formant. The above mention is could be observed in the figures.The percentrage deviation($Fig5$) and the deviation($Fig6$) are shown below.

IV. Conclusion

The reason why TMM modeling of Pyknogram is might be the fact that we use Expectation Maximization for model fitting where we assume the speech data is independent where as its not. Using a HMM based approach might give relatively better results.

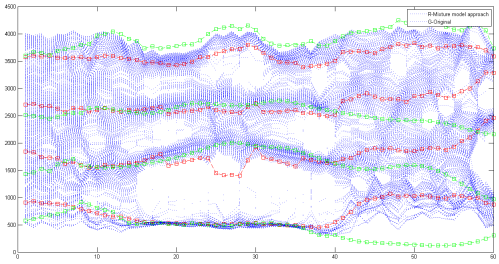

Fig. 3. Formants tracked for a TIMIT speech frame compared against the manual formant tracks

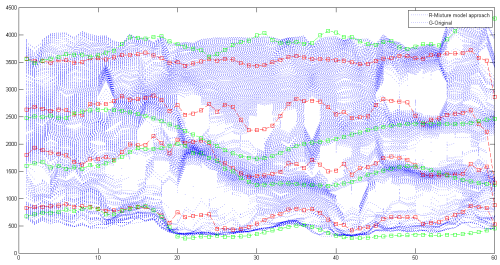

Fig. 4. Formants tracked for a TIMIT speech frame compared against the manual formant tracks

F1	F2	F3	F4
10.2018	6.6943	5.8441	6.4053
16.7272	6.7807	3.1973	6.8096
17.5933	6.2059	4.7712	7.0685

Fig. 5. Percentage deviation for four formants for several speech signal

F1	F2	F3	F4
54.31	105.90	177.71	246.48
64.95	167.59	83.106	211.46
89.15	115.67	97.25	365.069

Fig. 6. Deviation for four formants for several speech signal

REFERENCES

[1] A. Potamianos and P. Maragos, Speech formant frequency and bandwidth tracking using multiband energy demodulation, J. Acoust. Soc. Amer., vol. 99, no. 6, pp. 37953806, 1996

[2] A Mixture Model Approach for Formant Tracking and the Robustness of Students-t Distribution Harshavardhan Sundar, Student Member, IEEE, Chandra Sekhar Seelamantula, Member, IEEE, and Thippur V. Sreenivas, Senior Member, IEEE