# VOCAL TRACT RESONANCES TRACKING BASED ON VOICED AND UNVOICED SPEECH CLASSIFICATION USING DYNAMIC PROGRAMMING AND FIXED INTERVAL KALMAN SMOOTHER

*İ. Yücel Özbek, Mübeccel Demirekler*

Department of Electrical and Electronics Engineering
Middle East Technical University, Turkey
*email:{iozbek, demirek}@metu.edu.tr*
*web:www.eee.metu.edu.tr/~yozbek/*

## ABSTRACT

This paper presents a systematic framework for accurate estimation of vocal tract resonances (formants) using neither training data nor a phonetic transcription. In the proposed method, the speech signal is segmented in voiced and unvoiced parts and the resonance frequencies of the vocal tract are estimated by dynamic programming and further processed by using Kalman filtering/smoothing for each part. The performance of the proposed method is compared with three different methods which are baseline, WaveSurfer [10] and MSR [5]. The proposed method reduces the overall vocal tract resonances (for F1, F2 and F3) estimation error rate by 35%, 39.6% and 2.74% over the baseline, WaveSurfer and MSR methods respectively.

***Index Terms—*** formant tracking, vocal tract resonances, VTR, Kalman filtering/smoothing, voiced and unvoiced speech classification

## 1. INTRODUCTION

Vocal tract resonances (VTRs) contain very useful information about uttered speech and speaker. They are used in many speech applications (i.e. speech recognition, synthesis, accent classification etc). Hence, reliable estimation of formants is important in order to improve performance of these applications. Recently, numerous methods are proposed to track formants that use Kalman filtering (KF) [5, 7, 11, 14], dynamic programming (DP) [1, 2, 4], HMM [15], GMM [6] or combination of them [3, 13]. In this work, we combine Kalman filtering/smoothing and dynamic programming algorithm to track and estimate formant frequencies accurately. Doing this combination, we consider formant tracking process as a kind of multi-target tracking process. In multi-target tracking applications, there are two important issues; data association (that is, which measurement belongs to which target), and position estimation. Using a similar idea, we consider formant candidates form LPC analysis stage as measurements from targets that correspond to formant frequencies. DP is considered as a data association stage, in which labeling of the formant candidates are handled. Estimation of formant location is done in KF stage. The proposed method is explained in Figure-2 in detail. From our point of view, without using the KF stage, the tracker has lack of main estimation stage. Figure-2 indicates that the formant tracking procedure applied to voiced and unvoiced parts of the speech are not the same. Indeed this is one of the factors that improve the performance of the system. The reason for this differentiation is the basic observation that for voiced regions formant candidates given by LPC is much more reliable compared to the unvoiced regions. The direct implication of this observation is the differentiation of parameters of trackers two cases. For the voiced speech, nominal formant frequencies (independent of phone) are used as additional information in DP part with relatively low importance. Furthermore, the formant measurements (output of the DP stage) contain "low noise" so the model generated for KF part has a small measurement noise covariance. For the unvoiced speech, the line connecting formants (similar to [2, 12, and 16]) of the proceeding and succeeding voiced regions are used as nominal formant frequencies which are called "estimated nominal VTRs". They are quite effective in DP stage where LPC outputs are not reliable. KF parameters are selected according to a re-examination of the voicing decision. The measurement covariance parameter in KF is relatively high for unvoiced part due to less reliable LPC outputs

## 2. BASELINE METHOD

Before explaining the proposed method, we introduce our baseline system that can be seen in Figure-1. The baseline method is conventional formant tracking algorithm based on dynamic programming [1, 2, and 4]. The sub-blocks of the baseline system are explained in Section 3.3.

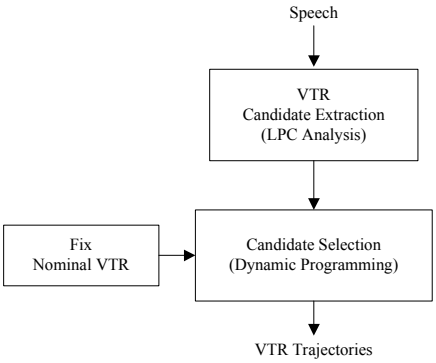


Figure-1 General scheme of baseline formant estimation procedure

## 3. PROPOSED METHOD

The general scheme of proposed VTR estimation procedure can be seen in Figure-2. The sub-blocks are explained as follows.

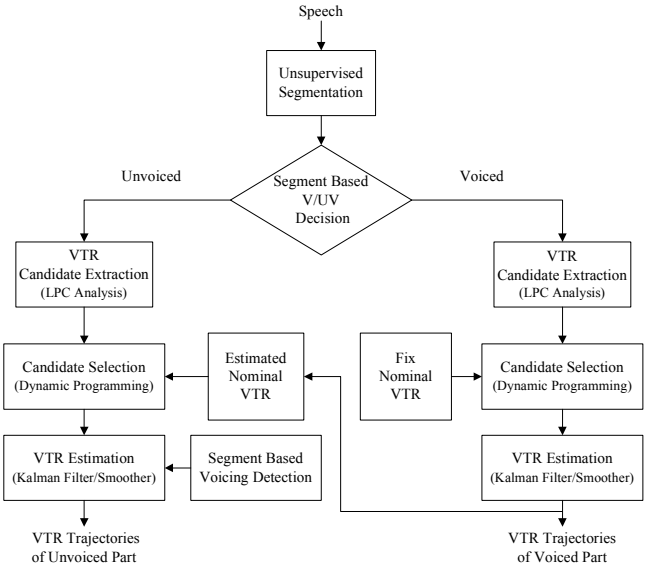


Figure-2 General scheme of proposed VTR estimation procedure

### 3.1. Unsupervised Speech Segmentation and Segment Based Classification: Voiced vs. Unvoiced

In this work, we use Level building dynamic programming (LBDP) algorithm in order to segment speech signal into homogenous units [8]. The number of segments is L and is chosen as L = 40.T , where T (sec) is the total duration of the speech utterance. After segmentation phase, each segment is classified as voiced or unvoiced by using two energy thresholds, which are the average energy of the segment in dB and the energy ratio of the low frequency band (100-900 Hz) to the high frequency band (3700-5000 Hz) in dB.

### 3.2. Vocal tract Resonance Candidates Based on LPC

After pre-emphasis stage, speech signal is divided into frames. For each frame, the frequencies and the bandwidths of formant candidates are calculated as

$$R_k = \frac{\omega_k}{2\pi T_s} \text{ and } B_k = -\frac{\ln(|c_k|)}{\pi T_s}$$

where, $c_k$, $R_k$ and $B_k$ are the $k^{th}$ complex root of denominator polynomial of LPC filter (real roots are discarded), frequency of resonance candidates and its bandwidth respectively. $T_s$ is the corresponding sampling frequency. In this work, the frame length and frame rate are 40 and 6 msec respectively. Also, we choose sampling frequency to be 10 KHz and the LPC order to be 12.

### 3.3. Estimation of Vocal Tract Resonances

In this work, the estimation of the vocal tract resonances is handled by Kalman smoothing. For each resonance, we use one Kalman filter. The critical point in this method is to choose correct measurement (resonance) candidate to update Kalman filter. For this purpose it is necessary to associate the resonance candidates with formant tracks. There are some methods in the literature to solve this problem [7, 11].In this work, we use dynamic programming (DP).

#### *3.3.1. VTR Candidate Classification (Selection)*

We use dynamic programming (DP) to find resonance candidates for VTR estimation phase. The states of the DP are all possible formant track/candidate associations. As an example, for 4 track and 6 candidates, the number of states is $N_s$=15. From the definition it is obvious that $N_s$ may change for each frame. Incremental costs related with DP are $D_L$(Local cost) and $D_T$(Transitional cost). Definition of them is similar to [1, 2, 4] and are given below. The local cost $D_L(.)$ is related to our knowledge about VTR without using any temporal context and it is defined as

$$D_L(S_k = m) = \sum_{i=1}^{N} \left( \alpha B_{im} + \Gamma \eta_i \frac{\left| R_{im} - R_i^n \right|}{R_i^n} \right)$$

Here, $S_k$ denotes the state at frame k. N is the number of VTR, B is the bandwidth of the resonance which is weighted by α and is independent VTR index. R and $R^n$ are the VTR candidate and nominal VTR values respectively. The normalized mean distance between the candidate and nominal VTR is weighted by $\eta_i$ and Γ.

The transitional cost $D_T(.)$ which forces the resonance candidates to be continuous is defined as:

$$D_T(S_k = m \mid S_{k-1} = p) = \sum_{i=1}^{N} \varphi_i \left( \frac{R_{im}(k) - R_{ip}(k-1)}{R_{im}(k) + R_{ip}(k-1)} \right)^2$$

Where R(k) and R(k-1) are the resonance candidates at frame k and k-1. $\varphi_i$ is the weight which is VTR dependent. Hence, the total cost at $k^{th}$ frame for $S_k = m$ is

$$D(S_k = m) = D_L(S_k = m) + \min_p \{D_T(S_k = m \mid S_{k-1} = p) + D(S_{k-1} = p)\}$$

The backtracking procedure of DP gives the best resonance frequencies that means the VTR candidates are classified into VTR index and they are ready for final VTR estimation phase. On the contrary to, the baseline method, the Γ parameter of DP is set for voiced and unvoiced parts differently.

*3.3.2. VTR Estimation*

In the VTR estimation phase Kalman smoothing is used. The state-space representation of the dynamic system model is given as follows;

$$x_k = Ax_{k-1} + Gw_{k-1} \text{ and } y_k = Hx_k + v_k$$

The $w_k$ and $v_k$ are Gaussian random processes with known covariance Q and R, which are defined as [11,9]

$$w_k \sim N(0;Q) \text{ and } v_k \sim N(0;R)$$

We choose the following model parameters and the state;

$$A = \begin{bmatrix} 1 & T \\ 0 & 1 \end{bmatrix}, G = \begin{bmatrix} T^2/2 \\ T \end{bmatrix}, H = \begin{bmatrix} 1 & 0 \end{bmatrix}, \; x_k = \begin{bmatrix} F_k & \dot{F}_k \end{bmatrix}^T$$

where, T is the time difference of consecutive frames which is constant. $F_k$ is corresponding to the resonance frequency that is estimated and $\dot{F}_k$ is it's time derivative. $y_k$ is the classified resonance candidate (measurement). Standard Kalman filtering / smoothing techniques are applied for the state estimation [9].

## 4. EXPERIMENTAL RESULTS

In this section we show experimental results and compare the performance of the proposed method with our baseline system, WaveSurfer [10] and MSR methods [5]. The experiments were carried out using hand labeled VTR database [5] which has been introduced recently. The database contains 516 utterances (sentences) and it is publicly available[1]. The experimental results of this work cover all 516 sentences of the database. VTR estimation errors (in Hz) are measured by averaging absolute VTR differences between the estimated and hand labeled reference values over all frames, which is defined as follows.

$$E_i = \frac{1}{N_c} \sum_{i=1}^{N_c} \left| \hat{F}_i - F_i^r \right| \quad \text{i=1,2,3}$$

where, $E_i$ is the estimation error of $i^{th}$ VTR, $\hat{F}_i$ and $F_i^r$ are the corresponding estimated and hand labeled reference VTR's respectively and $N_c$ is the total number of frames. The hand labeled database has 10 KHz sampling frequency. For error calculation, the hand labeled data is up-sampled so that it has same sampling rate as the proposed system. For a more detailed examination, we measure VTR estimation error for broad phonetic classes as well.

The comparison of the proposed and the baseline system is given Table-1.

Table-1 The error produced by the proposed and baseline methods for broad phonetic class

| Classes | Proposed | | | Baseline | | |
|---|---|---|---|---|---|---|
| | **E1** | **E2** | **E3** | **E1** | **E2** | **E3** |
| **Vowels** | 53 | 73 | 98 | 56 | 74 | 108 |
| **Semivowels** | 65 | 84 | 139 | 71 | 105 | 176 |
| **Nasals** | 93 | 194 | 156 | 108 | 236 | 178 |
| **Fricatives** | 119 | 126 | 156 | 224 | 185 | 227 |
| **Affricatives** | 144 | 150 | 167 | 243 | 197 | 186 |
| **Stops** | 120 | 135 | 168 | 208 | 216 | 249 |
| **AVERAGE** | 83 | 105 | 131 | 122 | 137 | 169 |

The comparison of MSR and WaveSurfer is given in [5] (Although 538 sentences are used in [5], we use 516 of them since only 516 sentences are publicly available) and repeated here for over all evaluation of our method

Table-2 The error produced by the MSR and WaveSurfer methods for broad phonetic classes (This table is taken from [5] for comparison purpose)

| Classes | MSR | | | WaveSurfer | | |
|---|---|---|---|---|---|---|
| | **E1** | **E2** | **E3** | **E1** | **E2** | **E3** |
| **Vowels** | 64 | 105 | 125 | 70 | 94 | 154 |
| **Semivowels** | 83 | 122 | 154 | 89 | 126 | 222 |
| **Nasals** | 67 | 120 | 112 | 96 | 229 | 239 |
| **Fricatives** | 129 | 108 | 131 | 209 | 263 | 439 |
| **Affricatives** | 141 | 129 | 149 | 292 | 407 | 390 |
| **Stops** | 130 | 113 | 119 | 168 | 210 | 286 |

The comparison of the proposed method with WaveSurfer and MSR's method [5] can be seen in Table-1, Table-2, Figure-3, 4, 5.

## 5. DISCUSSION AND CONCLUSIONS

The experimental results show that the proposed method is significantly better than both the baseline system and WaveSufer. The method also has a significantly better

[1] In [5], 538 sentences are introduced; however, 516 of them are publicly available.

performance compared to MSR's method [5] in vowel and semi-vowel phonetic classes where the VTRs are well-defined. This result can be seen in Figure-4, Table-1 and Table-2. On the other hand, it is comparable to the MSR's method for the remaining phonetic classes. The overall performance (for F1, F2 and F3) of the proposed method is slightly better than MSR's method, which can be seen in Figure-5.

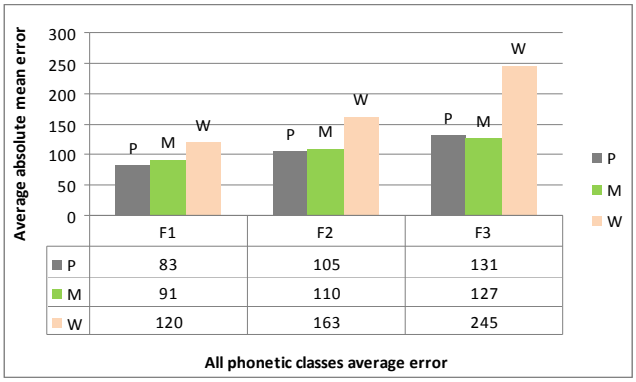


Figure-3 The error produced by the proposed (P), MSR (M) and WaveSurfer (W) for all phonetic classes.
(Note: MSR and WaveSurfer 's results are calculated using Table-2)

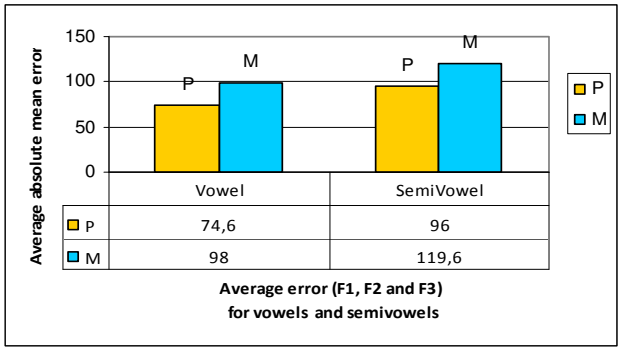


Figure-4 The error produced by the proposed (P), MSR (M) for vowels and semivowels phonetic classes

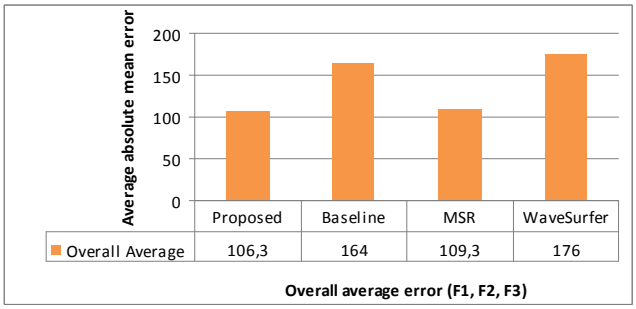


Figure-5 The error produced by the proposed, baseline, MSR and WaveSurfer for overall average (F1, F2, F3)
(Note: MSR and WaveSurfer 's results are calculated using Table-2)

Examination of Table-1 shows that the performance of the proposed method for nasal phonetic class is relatively low. The reason for this is that the resonance candidates of proposed method are obtained using LPC analysis which chooses spectral peaks as VTRs. In hand labeled database, however, spectral valleys are chosen as VTRs for some nasal consonants as explained in [5]. We are currently studying on different types of VRT candidate extraction methods in order to further increase the performance of nasal and unvoiced speech parts.

## 6. REFERENCES


[1] D. Talkin. "Speech formant trajectory estimation using dynamic programming with modulated transition costs" *J. Acoust. Soc. Am.*pp. S55. S1, 1987

[2] Lee, M., VanSanten, J., Mobius, B., Olive, J., "Formant tracking using context-dependent phonemic information". *IEEE Trans. Speech Audio Process*. pp. 741–750, 2005.

[3] Q. Yan, et.al "Formant tracking linear prediction model using HMMs and Kalman filters for noisy speech processing" *Computer Speech and Language.* Elsevier pp.543-561. 2007

[4] K. Xia, C. Espy-Wilson, "A new formant tracking algorithm based on dynamic programming,". in ICSLP, pp.55-58. 2000

[5] Li Deng. et al "A database of vocal tract resonance trajectories for research in speech processing"*ICASSP*,Vol.1.pp.369-372 2006

[6] J. Darch. et al "MAP prediction of formant frequencies and voicing class from MFCC vectors in noise **"** *Speech Comm***.** Elsevier V 48, I. 11, pp 1556-1572. 2006

[7] Li Deng. et al "Adaptive Kalman Filtering and Smoothing for Tracking Vocal Tract Resonances Using a Continuous-Valued Hidden Dynamic Model" . *IEEE Trans. Speech Audi. Pro.*. 2007

[8] Sharma, M.; Mammone, R**.;** "Blind speech segmentation: Automatic segmentation of speech without linguistic knowledge" *ICSLP* pp.1237 – 1240. 1996

[9] Y. Bar-Shalom and T. E. Fortmann. *Tracking and Data Association*. New York: Academic, 1988.

[10] The speech tool, WaveSufer: http://www.speech.kth.se/

[11] Özbek İ. Yücel, Mübeccel Demirekler. "Tracking of Visible Vocal Tract Resonances (VVTR) Based on Kalman Filtering" (*INTERSPEECH'2006*),September17-21 Pittsburgh, Pennsylvania

[12] S.A. Fattah,W.-P. Zhu, M.0.Ahmad."An approach to formant frequency estimation at low signal-to-noise ratio". *ICASSP 2007.*

[13] C. Glaisert, M. Heckmann, E Joublin, C. Goerick "Joint Estimation of formant trajectories via specto-temporal smoothing and Bayesian techniques". *ICASSP 2007.*

[14] D. Rudoy, D.Spendley and P. J. Wolfe. "Conditionally linear Gaussian models for tracking of vocal tract resonances". (*INTERSPEECH'2007*),

[15] Toledano, D.T. et. al "Initialization, training, and context-dependency in HMM-based formant tracking" *IEEE Trans. Speech Audi. Pro.*. 2006 Volume 14, pp: 511 – 523

[16] Mustafa, K.; Bruce, I.C.; "Robust formant tracking for continuous speech with speaker variability" *IEEE Trans. Speech Audi. Pro.*. 2006 Page(s):435 - 444