Tracking of Visible Vocal Tract Resonances (VVTR) Based on Kalman Filtering

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

ABSTRACT

This paper analyzes vocal tract resonance (VTR) frequency trajectories and their relationship to formants from a new point of view. Considering abrupt/continuous changes in the physical geometry of vocal tract, VTR may change in number, suddenly change their positions or may leak to some regions where they usually do not exist. We define the visible VTR (VVTR) as VTR that can be seen from the spectrogram. So we propose an algorithm, based on Kalman filtering, that can handle all these changes in VVTR. The suggested properties of VVTR trajectories and the performance of the algorithm are demonstrated on several examples.

Index Terms: formant tracking, vocal tract resonance, Kalman filtering,

1. INTRODUCTION

The two terms, vocal tract resonance (VTR) and formant frequencies can be used interchangeably in the literature. VTR frequencies are related to physical system and defined as resonance frequencies of air path of articulatory system. VTRs are independent of existence of air (excitation) in that articulatory system. Previously (VTR) frequencies are called formant frequencies and defined only for vowel-like sounds in the literature. However, scientists from different areas such as phonetics [5, 6] considered formant frequencies as spectral prominence and they used them as distinguishing features for some obstruent sounds such as plosives and fricatives. Therefore, nowadays, in general, formants are considered as frequencies that occur due to vocal tract resonances and are defined in acoustic domain with evidence of spectral prominence. Similarly, tracking resonance frequencies in speech utterance is handled in different perspective. Kopec [1] tracks formant frequencies only for vowel-like sounds and also labels some regions that formant frequencies do not exist. Lee et al. [2] tracks formant frequencies for speech utterances that contain unvoiced consonants but tracking performance is evaluated only for vowel-like regions. Their reason to track formant frequencies in unvoiced regions is to increase the tracking performance of vowel-like regions. Deng et al. [3, 4], extends formant frequency tracking into unvoiced regions including unvoiced closure and call it tracking VTRs. They consider formant and VTR as same for voiced regions and assume that in non-sonorant regions VTRs are hidden (unobservable) but they should be tracked somehow to maintain the continuity of articulatory movement which is independent of the existence of air (excitation).

Our perspectives are somehow different from previous studies:

- Since articulators change continuously, some resonance frequencies may have continuous trajectories (including sonorant and obstruent region) according to physical geometry of air path. However, some resonance frequencies can completely disappear or newly appear, especially when physical geometry of air path changes abruptly as in the case of nasal and plosive sounds.
- Number and frequency of a VVTR changes according to context and may change for the same phone. As an example resonance frequencies of unvoiced closure (leakage formant) can occasionally be visible in the spectrogram. Also an unexpected resonance frequency may appear in a nasalized vowel.
- It is difficult to decide on the number of resonance frequencies that exists in a speech utterance. The generally excepted procedure of tracking 5 or 3 formant frequencies seems to give some wrong indications about VTRs.

Considering all these observations we introduce the concept of tracking visible vocal tract resonance (VVTR) frequencies. VVTR includes well-known formant frequencies defined for vowel-like sounds, extra formants due to nasalization, leakage formant in obstruent regions etc, when they exist.

Therefore, our tracking method is different from previous studies in the following ways:

- The number of visible resonance frequencies that we want to track depends on the utterance and is not known a priori. It may change in time for a given speech utterance. Therefore, the tracking algorithm should have the capability of tracking different numbers of resonance frequencies along the speech utterance.

- The tracking algorithm should have the capability of initiating new trajectories and ending the already existing ones.

In this study we present a new strategy to track VVTR trajectories in a fully automatic manner without using any phonemic information. In section 2, we will describe the tracking algorithm. Experimental results are given in section 3, and section 4 is devoted to discussion and conclusions.

2. VVTR TRACKING ALGORITHM

Our proposed VVTR tracking algorithm is mainly based on multi-target tracking algorithms which are widely used in tracking literature. Here we tailored the existing methods to suit VVTR tracking in speech utterance. Our VVTR tracking algorithm operates in four phases; 1) Speech analysis, 2) Track start/end decisions 3) Gating and association and 4) Tracking, as shown in Figure 1. In the analysis phase, resonance candidates are obtained by solving the linear prediction polynomial obtained from LPC analysis. In the track start/end decision phase new trajectories start if there are consistent track candidates or old trajectories end if no more consistent candidates are available. In the gating and association phase, each trajectory is associated to a suitable resonance candidate if such a candidate exists in a sufficiently close neighborhood of it. Finally, in the tracking phase, the tracks are generated by Kalman filtering. Following sections explain each block in more detail.

2.1 Speech analysis phase

Analysis phase generates candidates for resonance frequencies by finding the roots of the linear prediction polynomial obtained from LPC analysis. Common procedures are applied to find LPC coefficients [9]. The prediction polynomial can be written as follows;

$$A(z) = 1 - \sum_{k=1}^{P} a_k z^{-1} = \sum_{k=1}^{P/2} (1 - c_k z^{-1})(1 - c_k^* z^{-1})$$

where, P is the order of the LPC filter. Resonance frequencies are found using only complex roots of prediction polynomial.

$c_k = |c_k| e^{jw_k}$ is k^{th} complex roots of $A(z)$.

Resonance frequency and its bandwidth can be expressed as;

$$f_k = \frac{w_k}{2\pi T_s} \quad ; k^{th} \text{ resonance frequency candidate}$$

$$B_k = -\frac{\ln|c_k|}{\pi T_s} \quad ; \text{estimated bandwith}$$

where, T_s is sampling period

The sampling frequency can be chosen according to the interest of highest resonance frequency band. Bandwidth of the resonance frequencies are used as a threshold in the selection of resonance candidates. That is the roots of the LPC polynomial with large bandwidths may not be used as a resonance candidate.

2.2 Track Start/End Decision phase

Track Start/End Decision phase provides the tracking algorithm with three decisions about trajectories (tracks); start a new VVTR trajectory, confirm VVTR trajectory and end an existing VVTR trajectory. The state diagram of this phase is given in Fig.2. State 1 is for track initiation. Any point obtained as a resonant frequency from the LPC analysis that is not in the gate of any existing track is considered as a possible future track in State 1. At the following frame the state of the candidate track will be changed either as 0 or 2. The change in the state is done according to the resonant candidates of the next frame. If there is a measurement, i.e. a candidate resonant frequency which is close to the track candidate generated at State 1 (in the gate of the track considered) than this track goes to state 2. Otherwise it goes to state 0. States 0 and 2 are waiting states and waiting times are design parameters. In State 2 Kalman filters are initiated for the candidates. The trajectories which do not have measurements in their gates go to State 0 where they wait a certain time for a measurement. If they can not receive measurement in the gate of the trajectory, they go to State -1 and are deleted. However in the case of consistent measurements they go to State 2. The track candidates in State 2 go to State 100 and declared as tracks if they receive consistent measurements for a certain number of times. State 100 is the tracking state and all trajectories are tracked by the tracking algorithm if they take measurements. If they do not take measurements in certain time interval that the state of this track is set to State -100 and is ended.

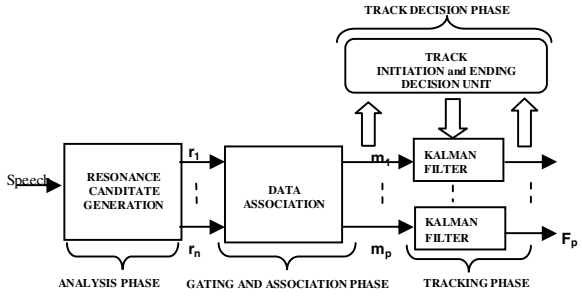


Figure 1: General VVTR tracking method

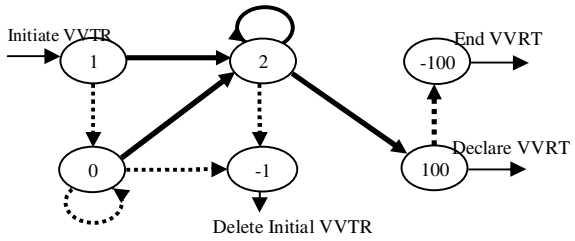


Figure 2: State flow diagram of the track decision phase (The solid lines denote trajectories that take consistent measurement The dashed lines denote trajectories that do not take consistent measurement)

2.3 Gating and association phase

Gating and association phase is mainly related to Track Start/End Decision phase of the algorithm. The gate of a trajectory determines upper and lower limits for the candidate resonance frequencies. Trajectories consider a measurement as consistent if it is in between these upper and lower limits. The VVTR tracking algorithm use two types of gates: constant, and variable. The upper and lower frequency values of a constant gate is constant. As an example, if we define constant gate value as 150 Hz and current trajectory value is 500 Hz than the trajectory considers resonance candidates as

consistent, if they are in between 500-150=350Hz and 500+150=650 Hz. The constant gate is used for those trajectories whose states are at State 1, State 0 or State 2. The variable gate is used for the trajectories in State 100, i.e. after VVTR is declared. The variable gate changes as time and it is defined as:

$$\text{Gate}_k = \sqrt{\text{Th}.S_k}$$

where, S_k and Th are measurement covariance matrix and predefined threshold respectively. Then, the upper and lower limits of gate are defined as follows.

$$\text{Limits}_k = \hat{y}_k \mp \sqrt{\text{Th}S_k}$$

In this expression $\hat{y}_k$ is the predicted measurement at time instant *k*, which is obtained using prediction step of the Kalman filter. This expression indicates that the gate value is changed according to the measurement covariance. As the theory suggests, if a consistent measurement does not come, the gate value increases. If more than one candidate falls in the gate of any trajectory, the nearest-neighbor procedure is applied to associate the candidate with the trajectory.

2.4 Tracking phase

In the tracking phase Kalman filter is used to track resonance frequencies. The state-space representation of the dynamic system model is given as follows;

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

According to our observation, we decide that trajectories of resonance frequencies can be modeled approximately as linear functions of time in short time intervals. The nonlinear changes of the trajectory (rapid change in trajectories) are modeled via process noise of the above given dynamic system. Therefore, we choose following model;

$$A = \begin{bmatrix} 1 & T \\ 0 & 1 \end{bmatrix}, G = \begin{bmatrix} T^2/2 \\ T \end{bmatrix}, H = \begin{bmatrix} 1 & 0 \end{bmatrix}$$

where, T is the measurement interval which is constant. This model is also known as constant velocity model in target-tracking applications [8]. The state vector x_k is defined as;

$$x_k = \begin{bmatrix} F_k & \dot{F}_k \end{bmatrix}^T$$

where, F_k is corresponding to the resonance frequency that is tracked and $\dot{F}_k$ is it's time derivative. y_k is the measurement vector obtained from resonance candidates. w_k and v_k are known as process and measurement noises that are assumed to be Gaussian and independent.

Tracking algorithm uses standard Kalman filtering equations [8] where measurements are the associated frequencies obtained at the association phase. First components of the filtered states give the related VVTR track.

3. EXPERIMENTAL RESULTS

In this section, we will give spectrograms and VVTR frequency tracker outputs of some utterances that are taken from Continuous Speech Turkish Database [10]. Phonetic transcriptions (with SAMPA alphabet of Turkish [13]) of the fragments are given in the corresponding figures. We use WaveSurfer speech tool [11] to show results of VVTR tracker.

We also use formant tracking results of WaveSurfer for comparison purposes. The first spectrogram is for the Turkish word *'fakat'*. The spectrogram of *'fakat'* is given in Fig. 3. In this figure, the indicated regions denote leakages of resonance frequencies that appear in the closure before the plosives 'k' and't'. The solid lines are results of proposed VVTR tracker. It is clear that algorithm is successful in tracking and ending leakage resonance frequency tracks as well as the normal resonance frequencies of vowels.

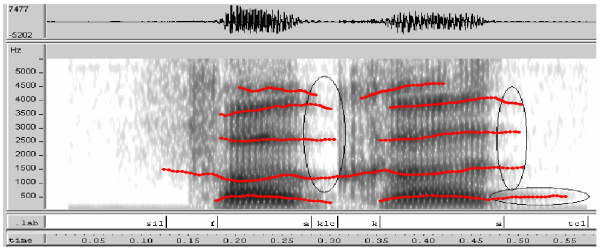


Figure 3: VVTR tracking result of the word 'fakat' with transcription (f-a-kcl-k-a-tcl-t)

Figure 4 shows spectrogram of a fragment '*O hantaldı'*.This speech utterance contains special phenomena about vowel nasalization which is on *'a'* that appears before the nasal sound *'n'*. Since the proposed VVTR tracker is capable of tracking different number of resonance frequencies it tracks extra resonance frequency due to nasalization as well the resonances of *'a'* as shown in Figure 4. Also, the leakage resonance frequencies in the closure part of *'d'* are successfully tracked and these tracks are connected to resonance frequencies of the next vowel *'l'*. Note that tracking the leakages generates the required continuity of the vocal tract resonances when such continuity exists. On the other hand it is clear that whenever the vocal tract shape change causes an abrupt change in its resonance frequency, tracks are ended and new ones are started as explained above.

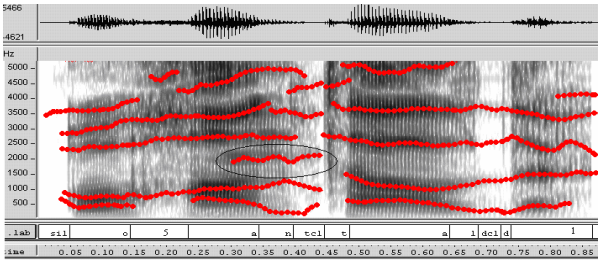


Figure 4: VVTR tracking result of the fragment 'O hantaldı' with transcription 'o- 5-a-n-tcl-t-a-l-dcl-d-1'

Figure 5 shows VVTR tracker results of the statement *'sürücügillerden'*. This sound also contains other special phenomena which is called "velar pinch", that occurs at velar consonant *('k', 'g', 'N')*. At these phones two resonance frequencies merge together. The encircled region in the figure shows that the VVTR algorithm is also successful in detecting and tracking this phenomenon.

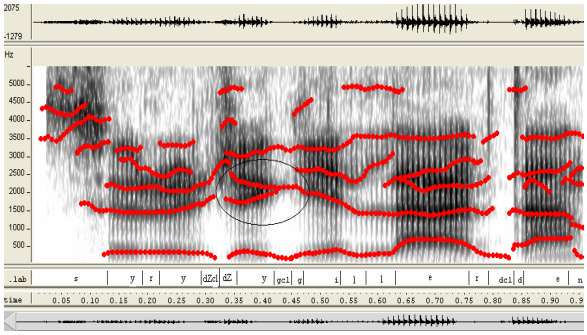
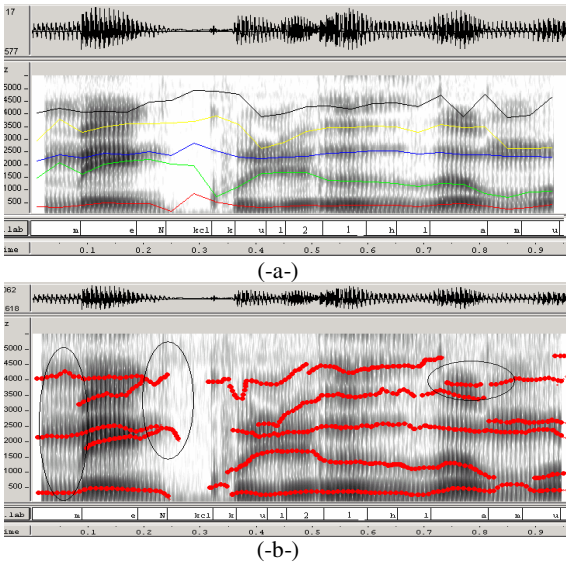


Figure 5: VVTR tracking result of the word 'sürücügillerden' with transcription 's-y-r-y-dZcl-dZ-y-gcl-g-i-l-l-e-r-dcl-d-e-n'

In Figure 6, the tracking results of the proposed VVTR tracker and WaveSurfer speech tool are compared for the same fragment *'menkulü ıhlamur'*. WaveSurfer has fixed number of formants which is set to 5. A close examination of this utterance shows that the first nasal 'm', shown as first encircled region in the figure, has 3 VVTR tracks however the number increases to 5 at the vowel region of 'e'. It can be seen that the VVTR tracks are again reduced to 3 corresponding to "velar pinch" before *'N'* which denotes phonetic velar *'n'* as the second encircled region indicates. The last marked region of Fig. 6 shows the difference between the WaveSurfer tracks and our tracker for the last 'a'. Our tracking system gives a much more satisfactory result for the 5th VVTR compared to WaveSurfer.



(-a-)

(-b-)

Figure 6: WaveSurfer and VVTR tracking result of word 'menkulü *ı*hlamu' ('m-e-n-kcl-k-u-l-y- 1-5-l-a-m-u') in part-a, part-b respectively

4. DISCUSSION AND CONCLUSIONS

In this study we have analyzed the vocal tract resonances and their visibility in the spectrogram. In the literature it is usually assumed that the vocal tract resonances are continuous due to the continuity of the movements of the articulators, however this continuity is not visible in the spectrogram of the speech. In our analysis we observed that vocal tract resonances may not be continuous especially at the nasals, but not restricted only to them, where vocal tract changes structurally. Based on our observations of spectrograms of several sentences we proposed a tracking algorithm that can start, end and merge tracks of vocal tract resonances. The tracking algorithm is based on Kalman filtering and is inspired by multiple target tracking methods.

Visibility of the vocal tract resonances in the spectrogram is another issue of this paper. In the examples we showed that at some regions where vocal tract resonances exist due to 'leakage', continuity of the tracks is maintained by the algorithm.

The experimental results are very promising and show that the approach given here can be used in many speech applications such as speech synthesis, recognition, segmentation etc. The VVTR frequencies can be considered as "*anchor formant*" frequencies hence the result also can be used for classical formant frequency tracking applications such as given by Xia et al. [12].

5. REFERENCES

[1] G. Kopec, "Formant Tracking using Hidden Markov models and Vector Quantization", *IEEE Trans. Acoustics, Speech and Sig. Proc.*, ASSP-34, August 1986, pp. 709-729.

[2] M. Lee, J. van Santen, B. Möbius, and J. Olive, "Formant tracking using segmental phonemic information," *IEEE Trans. Acoustics, Speech and Sig. Proc.*, 2005.

[3] L. Deng , A. Acero, I. Bazzi, "Tracking Vocal Tract Resonances Using a Quantized Nonlinear Function Embedded inatemporalConstraint" *IEEE Trans. Acoustics, Speech and Sig. Proc.*, 2006.

[4] L. Deng and Z. Ma, "Spontaneous speech recognition using a statistical co articulation model for the hidden vocal-tract resonance dynamics," *J.Acoust. Soc. Amer.*, vol. 108, no. 6, pp. 3036–3048, 2000.

[5]Yanli Zheng, Hasegawa-Johnson, M "Particle filtering approach to Bayesian formanttracking" *Statistical Signal Processing, 2003 IEEE Workshop*

[6] Harvey M. Sussman, Helen A. McCaffrey, and Sandra A.Manhews. An investigation of locus equations as a source of relational invariance for stop place categorization. *J. Acoust. Soc. Am,* 90(3):1309-1325, September 1991.

[7] K.Stevens, AcousticPhonetics.Cambridge, MIT Press,1998.

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

[9] Rabiner, L.R. and Schafer, R.W. *Digital processing of speech signals*. Prentice Hall, Englewood Cliffs, 1978.

[10] Özgül Salor, BryanPellom, Mübeccel Demirekler "On developing new text and audio corpora and speech recognition tools for the Turkish language*"ICSLP-2002, Colorado USA*

[11] WaveSufer1.8: http://www.speech.kth.se/wavesurfer/

[12] Xia, Kun / Espy-Wilson, Carol (2000): "A new strategy of formant tracking based on dynamic programming", In *ICSLP-2000*, vol.3, 55-58

[13]SAMPA(Turkish) http://www.phon.ucl.ac.uk/home/sampa/