

Tracking Formant Trajectory of Tracheoesophageal Speech Using Hidden Dynamic Model

Feng Xue

Teaching Research Office of Zhangjiagang Education Bureau, Zhangjiagang City, 215600, China

Tel: 86-139-6227-5277 E-mail: zjgjyszb@szedu.com

Gang Lv (Corresponding author)

School of Electronic Information, Soochow University, Suzhou 215006, China

Tel: 86-512-6219-0898 E-mail: lvgang@suda.edu.cn

The research was financed by Canadian Center of Science and Education (No. B2009-122). (Sponsoring information)

Abstract

In this study, a method of tracking the formant trajectory of Tracheoesophageal (TE) speech using the Hidden Dynamic Model (HDM) with dynamic target orientation was put forward based on the characteristics of TE speech. At first, a PIF-LPC algorithm was used to extract the formant parameters of TE speech. The parameters as a weighted dynamic component were then introduced in HDM; afterwards, HDM was solved using the particle filtering algorithm based on the prior distribution of this parameter to track formant trajectory. It was shown in simulation tests that the method was of high precision and good robustness owing to the orientation effect of dynamic targets. With this method, not only the interferences of spurious peaks and merged peaks in TE speech were overcome, but also continuous TE speech signals were tracked.

Keywords: Tracheoesophageal speech, Hidden dynamic model, Formant tracking

1. Introduction

Laryngeal cancer was the common malignant tumors in head and neck of human beings. Its morbidity in China was 30-50/100 thousand. Among them, 50%-60% laryngeal cancer patients required total laryngectomy. After the surgery, tracheoesophageal (TE) speech is one alternative for voice restoration (Jeffrey, 2002, PP. 282-294).

TE speech is characterized by a generally lowered frequency, near normal intensity, and because of access to the large volume of pulmonary air, generally normal temporal features when compared to normal speakers (Rob, 2008, PP. 4517-4520). Therefore, formants as the most fundamental parameter to characterize TE speech signals play important roles in realization of TE speech conversion by extracting and tracking precisely the parameters of formant trajectories.

Hidden Dynamic Model (HDM) is an acoustic modeling method integrating the characteristics of Prosody with that of Phonetics (Richards, 1999, PP. 357-360). In HDM, the speech signal system is regarded as a hidden dynamic model. In this hidden dynamic space, each sound corresponds to a vector target, i.e., the muscles of vocal cords and vocal track approach a certain target status according to a 'programs' when a certain sound is made (Deng, 2000, PP. 3036-3048). Thus, the problem of tracking the formant trajectory of TE speech is converted to the problem of solving the target status in a successive time series. It was proved that, with HDM, not only formant trajectory of the vowel section in continuous speech signals is tracked effectively, but also formant trajectory 'disappeared' at the contoid section in continuous speech signals is revealed (Deng, 2007, PP. 13-23).

2. HDM and its solution

2.1 State space model of speech

Although speech signals have the characteristic of time-variance, the formant parameters change a little within a short term (10~30ms). Thus, the state equation was described as the following formula (Zheng, 2004, PP. 565-568):

$$X_t = X_{t-1} + V_{t-1} \quad (1)$$

Where, V_{t-1} was random noise of system, and X_t was the system status at moment t and was composed of the K -dimension formant frequency vector F and bandwidth vector B , i.e.:

$$X = (F, B) = (f_1, f_2 \cdots f_K, b_1, b_2 \cdots b_k) \quad (2)$$

All-pole model for the transfer function of the speech signal track was expressed as follows:

$$H(z) = G \prod_{i=1}^K \frac{1}{(1 - z_i z^{-1})(1 - z_i^* z^{-1})} \quad (3)$$

Where, $z_i = e^{-\pi \frac{b_k}{f_s} + j2\pi \frac{f_k}{f_s}}$, $z_i^* = e^{-\pi \frac{b_k}{f_s} - j2\pi \frac{f_k}{f_s}}$, and f_s was the sampling frequency of signals.

With inverse Z-transform performed on the above formula, cepstral coefficient of the n th LPC was obtained as follows:

$$C_n = \sum_{k=1}^K \frac{2}{n} e^{-\pi \frac{b_k}{f_s}} \cos(2n\pi \frac{f_k}{f_s}) \quad (4)$$

In the above formula, the transformation from the formant parameters to LPCC was achieved. Thus, the observed output was indicated as follows:

$$Y_t = C(X_t) + W_t \quad (5)$$

Where, W_t was the observed noise of system and obeyed the mean value of zero. The variance was the normal distribution of σ_x^2 .

Status equation (1) and observation equation (5) constituted the HDM.

In order to gain better tracking effects, Li et al. put forward an improved HDM (Deng, 2004, PP. 557-560). According to the acoustical characteristic of speech, F1, F2, F3, ..., of speech signals are arranged from low to high in frequency domain. For instance, generally F1 concentrated at low frequencies around 500Hz, while F3 concentrated at high frequencies around 3500Hz. According to this characteristic, Li used a static vector (known as the orientation target) to characterize the prior acoustical information of speech formants and believed that formant frequency would approach the static vector as the time tended to the infinity. Thus, formula (1) was converted as:

$$X_t = [1 - \Phi]X_{t-1} + \Phi U + V_{t-1} \quad (6)$$

Where, U was the orientation target, and Φ was the weighted value.

2.2 Particle filtering

The state space equation could be solved through particle filtering. Particle filtering expressed by the probability density of particle is a sequential Monte Carlo simulation method based on Bayesian Theorem (Fredrik, 2002, PP. 425-437). The idea of particle filtering is to indicate the needed posterior probability density through the weighted sum of a series of random samples, so as to obtain the estimated value of current state. Since the posterior probability function $p(x_t | y_t)$ of the function is generally unknown probability distribution, it could not be sampled directly. Therefore, particle x_t^i ($i=1, 2, \dots, n$) is generally sampled through an importance density function $q(x_t | y_t)$ with the probability density distribution known and same to $p(x_t | y_t)$.

Suppose

$$w_t(x_t) = \frac{p(x_t | y_t)p(y_t)}{q(x_t | y_t)} \quad (7)$$

And the expectation of an arbitrary function $f(x_t)$ was as follows:

$$E(f(x_t)) = \frac{\int f(x_t)w_t(x_t)q(x_t | y_t)dx_t}{\int w_t(x_t)q(x_t | y_t)dx_t} \quad (8)$$

Accordingly, particle x_t^i was collected from importance density function $q(x_t | y_t)$ to obtain the discrete approximation of expectation $E(f(x_t))$ as follows:

$$\overline{E(f(x_t))} = \frac{\frac{1}{n} \sum_{i=1}^n f(x_t^i) w_t(x_t^i)}{\frac{1}{n} \sum_{i=1}^n w_t(x_t^i)} = \sum_{i=1}^n f(x_t^i) \overline{w_t^i} \quad (9)$$

$$\text{Where } \overline{w_t^i} = \overline{w_t(x_t^i)} = \frac{w_t(x_t^i)}{\sum_{i=1}^n w_t(x_t^i)} \quad (10)$$

was the normalized particle weight, and n was the total number of particles. As the number of particles was large, such estimation would be converged to the actual posterior probability density.

If importance density function $q(x_t | y_t)$ was decomposed as:

$$q(x_t | y_t) = \frac{q(x_0)}{\prod_{j=1}^t q(x_{j-1}, y_j)} \quad (11)$$

Then (7) was written as:

$$w_t^i = w_{t-1}^i \frac{p(y_t | x_t^i) p(x_t^i | x_{t-1}^i)}{q(x_t^i | x_{t-1}^i, y_t)} \quad (12)$$

Thus, the particles sample set $\{x_t^i, \overline{w_t^i}\}_{i=1}^n$ characterizing posterior probability density function $p(x_t | y_t)$ was obtained. In the process above, the basic algorithm of Sequential Importance Sampling (SIS) was illustrated. SIS had the problem of weight degradation, i.e., after multiple iterations, all particles but one had minute weights. It suggested that a lot of work had been wasted on the particle upgrade for $p(x_t | y_t) = 0$. Accordingly, a resampling method was introduced. The idea of this method was to remove particles with small weights and to preserve and reproduce particles with larger weights. This method reduced the effect of particle degradation to a certain extent. However, since particles with small weights were eliminated, the particle set after resampled lost the variety; accordingly, the real probability distribution was not reflected with the problem of sample impoverishment emerging. Auxiliary Particle Filtering (APF) was an improved algorithm to solve this problem (Pitt, 1999, PP. 590-599). Based on SIS, an importance density function was introduced in APF. It satisfied:

$$q(x_t, v_t | x_{0:t-1}, y_{1:t}) \propto p(y_t | v_t) p(x_t | x_{t-1}) p(x_{t-1} | y_{1:t-1}) \quad (13)$$

Where, v_t was an auxiliary variable. Particles were extracted according to the above formula, and their weights were solved according to the following formula:

$$w_t^i = w_{t-1}^i \frac{p(y_t | v_t) p(y_t | x_t^i) p(x_t^i | x_{t-1}^i)}{q(x_t^i | x_{t-1}^i, y_t)} \quad (14)$$

The auxiliary sampling of APF was conducted at moment $t-1$, and the measurement information y_t at moment t was taken into account. Thus, the variety of particles was improved, and the particles sampled were more close to the real state at moment t .

3. HDM based on dynamic target orientation

In Section 2, an HDM based on static target orientation was introduced. In practice, two problems were noticed in this model. Firstly, for continuous speech, orientation target U was supposed to be not a static vector but a dynamic vector varying with time. Secondly, weighted value Φ also was supposed to be not a constant but a variable adjusting itself with actual condition. Accordingly, HDM based on dynamic target orientation was put forward.

First, state equation (6) was transduced as:

$$X_t = [1 - \Phi_t] X_{t-1} + \Phi_t v_t + V_{t-1} \quad (15)$$

Where, v_t was the dynamic target orientation at moment t , it was solved by PIF-LPC algorithm (Gang, 2009, PP. 127-135). Φ_t was the weighted variable at moment t . It was solved according to the following formula:

$$\Phi_t = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{\|C_t - C(v_t)\|^2}{2\sigma^2}} \quad (16)$$

Where, C_t was the coefficient vector of LPCC derived from the state space at moment t ; $C(v_t)$ was nonlinear mapping of the auxiliary variable from the formant parameter to the LPCC coefficient according to formula (4) at moment t ; σ was an user-defined coefficient. According to formula (16), when the observational results at moment t were close to the value of the prior auxiliary variable with the weight approaching 1, the current state output was dominated by the value of the prior auxiliary variable; whereas the weight approached zero, the current state output was dominated by the transfer results of the state space.

The algorithm for solving the formant trajectory of TE speech with HDM based on dynamic target orientation included the following steps:

- (1) Initialization, $t=0$, $x_0^i \sim p(x_0)$, $w_0^i = n^{-1}$.
- (2) Auxiliary variable v_t was solved with PIF-LPC algorithm.
- (3) Auxiliary particle filtering.
 - 1) Particles were extracted according to the importance density function in formula (13);
 - 2) The weights of particles were calculated according to formula (14);
 - 3) Particle weights were normalized according to formula (10);
 - 4) Particle set x_t^i was resampled;
- (4) Weight Φ_t was calculated according to formula (16).

- (5) The state output of HDM at moment t was calculated according to formula (15).

The steps from (2) to (5) were repeated, till the tracking of all data frames was finished.

4. Test results and analysis

TE sound observation group consisted of 17 patients with the treatment of speech rehabilitation after total laryngectomy. Among them, there were 9 males and 8 females, with the average age of 64.7. The group was tested in the environment with the yawp less than 45dB, taking comfortable sitting position, and sent out sustained and stable vowels. The sampling frequency of speech signal was 8kHz, and the quantization precision was 16bits; Hamming window was employed with 256 sampling points in each frame, and the frame shift was 1/4 of the frame length. For the convenience of comparing the accuracy of test results, the solved formant trajectories were labeled on their corresponding spectrograms.

Fig. 1 shows the tracking effect of TE speech /shang hai/ with HDM based on static target orientation and HDM based on dynamic target orientation separately. In the figure, the static target orientation U was set as [600,2000,3000,50,100,150], and the weighted value Φ was set as 0.3. The formant trajectory solved with HDM had the continuity, and all formant trajectories always kept respectively within their rational frequency-band ranges, not only in vowel sections, but also in contoid and transitional sections. F1, F2 and F3 increased successively from low frequencies to high frequencies without mutual superposition. The reason was that HDM based on target orientation brought continuity constraints to the formant trajectories in contoid and transitional sections to avoid random leaps. Therefore, HDM was more suitable for tracking TE speeches. However, the weighted oriented targets in HDM would affect the distribution of particles, and the effect was controlled by the weight. When particles distributed around the real formant frequencies, i.e., the real format state could be covered by the particles extracted from the suggested distribution, the oriented targets would bring positive effects on the tracking. If the orienting targets were not arranged rationally, however, some adverse influences would be brought to the tracking. Especially, when the real formant trajectory of speeches had large span in the frequency domain, static orientation targets brought negative effects on the tracking usually. As shown in graph, from section /sh/ to section /ang/, F2 sharply decreased from 2700Hz at 0.2s to 1800Hz at 0.22s with the frequency span up to 900Hz. In this case, static orientation targets cause the improper particle distribution with tracking performance degraded. Correspondingly, the correct tracking results were obtained with HDM based on dynamic target orientation. This was because that HDM based on dynamic target orientation not only integrated the prior acoustical characteristics solved with PIF-LPC in real time, but also adjusted the proportion of weight dynamically according to formula (16). Consequently, the weight approached the auxiliary variable including the acoustical characteristics at the speech sections with the significant formant structures

(e.g., vowel sections) under the condition that the output satisfied the continuity constraint.

5. Conclusions

TE speech is characterized by poor intelligibility and voice quality. Acoustic analysis of TE speech has the potential of quantifying the voice quality and assisting the speech pathologist in determining and monitoring the therapy process. In this paper, a sort of HDM based on dynamic target orientation was put forward. The formant parameters of TE speech were solved with PIF-LPC before being integrated into HDM as the dynamic target orientation. The weights of oriented targets were adjusted in real time through the comparison with actual observations. Finally, the formant parameters as auxiliary variables were added in particle filtering to improve the variety of particles in the course of resampling. In HDM, with particle impoverishment avoided, the accurate tracking of formant trajectory was achieved finally. Simulation tests proved that, with HDM based on dynamic target orientation, not only the interferences of spurious peaks and merged peaks to the conventional LPC algorithm at vowel sections were avoided effectively, but also the formant trajectories disappeared at the contour and transitional sections were tracked. Therefore, HDM based on dynamic target orientation was an approach of good robustness and high precision to track formant trajectory of TE speech.

References

- Deng L., Ma J. (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):3036-3048.
- Deng L., Lee L.J., Attias H., Acero A. (2004). A structured speech model with continuous hidden dynamics and prediction-residual training for tracking vocal tract resonances. *Proceeding of the ICASSP 2004*,557-560.
- Deng L., Lee L.J., Attias H., Acero A. (2007). Adaptive kalman filtering and smoothing for tracking vocal resonances using a continuous-valued hidden dynamic model. *IEEE Transactions Speech Audio Process*, No. 15(1):13-23.
- Fredrik G., Niclas B., Urban F. (2002). Particle filters for positioning, navigation and tracking. *IEEE Transactions on signal processing*, No. 50(2): 425-437.
- Gang L., Heming Z. (2009). Formant frequency estimations of whispered speech in Chinese. *Archives of Acoustics*, No. 34(2):127-135.
- Jeffrey P.S., Mary A.C. (2002). Acoustic cues to the voicing feature in tracheoesophageal speech. *Journal of speech, language, and hearing research*, No.45:282-294.
- Pitt M.K., Shephard N. (1999). Filtering via simulation: Auxiliary particle filters. *American Statistical Association*, No. 94(446) 590-599.
- Ray M.C., Vij P., Philip D. (2008). On the prediction of speech quality ratings of tracheoesophageal speech using an auditory model. *Proceeding of the ICASSP 2008*, 4517-4520.
- Richards H. B., Bridle J. S. (1999). The HDM: A segmental hidden dynamic model of coarticulation. *Proceeding of the ICASSP 1999*, 357-360.
- Zheng Y., Hasegawa-Johnson M. (2004). Formant tracking by mixture state particle filter. *Proceeding of the ICASSP 2004*, 565-568.

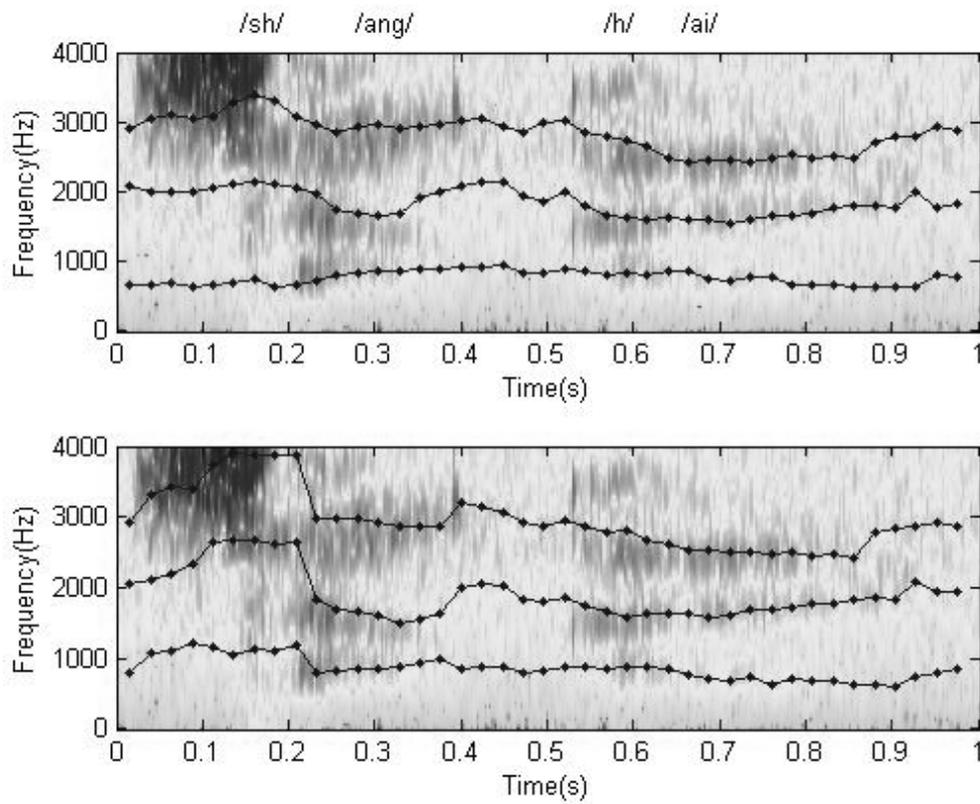


Figure 1. The upper graph shows the tracking effect of TE speech /shang hai/ with HDM based on static target orientation; the lower graph shows the tracking effect with HDM based on dynamic target orientation