Tracking Formant Trajectory of Continuous Chinese Whispered Speech with Hidden Dynamic Model Based on Dynamic Target Orientation

Gang Lv, Heming Zhao
School of Electronic Information, Soochow University, E-mail: lvgang@suda.edu.cn
doi:10.4156/jcit.vol5. issue9.23

Abstract

Aimed at the characteristics of Chinese whispered speech formants, i.e., migrating to high-frequency, increased bandwidth, and increased spurious peaks and merged peaks, a method of tracking the formant trajectory of continuous Chinese whispered speech using the Hidden Dynamic Model (HDM) with dynamic target orientation was put forward in this study. The calculation proceeded as follows: firstly, the PIF-LPC algorithm was used to evaluate the formant parameters of whispered speech (PIF-LPC is an improved LPC algorithm. In PIF-LPC, pole interaction factors are used to correct the formant bandwidth of residual poles, to reduce the effect of pole intersection and to improve the accuracy of formant parameters); then, the extracted formant parameters as dynamic target orientation were introduced in HDM and compared with the actual observation results for real-time adjustment of the weight of dynamic target orientation; finally, HDM was solved through auxiliary particle filtering (APF), so as to realize the tracking of the formant trajectory of whispered speech. It was shown in the experimental results that the interferences of spurious peaks and merged peaks were avoided when the formant trajectory of continuous whispered speech was tracked by this method.

Keywords*: Whispered Speech, Hidden Dynamic Model, Formant Tracking*

1. Introduction

The whispered speech is a special articulatory model. When people make whispered speech, their vocal cord does not vibrate with no F0 in speech signal [1]. Therefore, formants as the most fundamental parameter to characterize whispered speech signals play important roles in realization of whispered speech conversion and speaker recognition by extracting and tracking precisely the parameters of formant trajectories.

The first and second formants of Chinese whispered speech migrate to high frequency relative to that of normal speech. The migration amplitude of the first formant is significantly larger than that of the second; meanwhile, the bandwidths of all formants increase [2]. It generates the problem of pole intersection and produces more spurious peaks and merged peaks on the spectrogram. Therefore, formant trajectory of whispered speech is not tracked accurately with the conventional LPC algorithm.

Hidden Dynamic Model (HDM) is an acoustic modeling method integrating the characteristics of Prosody with that of Phonetics [3]. 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 [4]. Thus, the problem of tracking the formant trajectory of whispered speech is converted to the problem of solving the target status in a successive time series. In Article [5], 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.

In this study, a method of tracking the formant trajectory of continuous whispered speech using the HDM with dynamic target orientation was put forward based on HDM and the characteristics of Chinese whispered speech formants. In this method, an improved PIF-LPC algorithm was used to extract the formant parameters of whispered 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 whispered speech were overcome, but also continuous whispered speech signals were tracked.

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 [6]:

$$X_t = X_{t-1} + V_{t-1} \tag{1}$$

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.:

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

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

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

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 nth LPC was obtained as follows:

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

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

$$Y_t = C(X_t) + W_t \tag{5}$$

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 [7]. 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:

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

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 [8]. 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 $p(x_t \mid y_t)$ of the function is generally unknown probability distribution, it could not be sampled directly. Therefore, particle x_t^i (i=*1,2…,n*) is generally sampled through an importance density function $q(x_t \mid y_t)$ with the probability density distribution known and same to $p(x_t \mid y_t)$.

Suppose

$$w_t(x_t) = \frac{p(x_t \mid y_t)p(y_t)}{q(x_t \mid y_t)} \tag{7}$$

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

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

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

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

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

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 $q(x_t \mid y_t)$ was decomposed as:

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

Then (7) was written as:

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

Thus, the particles sample set $\{x_t^i, \overline{w_t^i}\}_{i=1}^n$ characterizing posterior probability density function $p(x_t \mid y_t)$ 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 $p(x_t \mid y_t)$ =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 [9]. Based on SIS, an importance density function was introduced in APF. It satisfied:

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

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:

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

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. PIF-PLC algorithm

The auxiliary variable v_t mentioned in the introduction of particle filtering in the above section could be acquired directly with LPC. Due to the distinctive acoustical characteristics of whispered speech, however, the formant parameter extracted with LPC was interfered by spurious peaks and merged peaks with the results full of considerable mistakes. Thus, the causes of such mistakes were analyzed in this section, and an improved algorithm was introduced to extract auxiliary variable particles.

In the frequency domain, if the sampling frequency is F_s, the formant F_i and the 3dB bandwidth B_i from the LPC algorithm can be converted to the pole with the angle of ϕ_i and the radius of r_i in the z domain according to the following formulas:

$$\text{The radiation angle of the pole: } \phi_i = 2\pi \frac{F_i}{F_s} \tag{15}$$

$$\text{The radius of the pole: } r_i = e^{-\frac{B_i * \pi}{F_s}} \tag{16}$$

The power spectrum of the pole z_i in the z domain is

$$\left| H(e^{j\theta}) \right|^2 = \prod_{i=1}^{n} \frac{1}{1 - 2r_i \cos(\theta - \phi_i) + r_i^2} \tag{17}$$

For the convenience of this discussion, we first suppose two poles z_1 and z_2, so the power $\left|H(e^{j\phi_1})\right|^2$ at the radiation angle ϕ_1 is

$$\frac{1}{(1-r_1)^2}\cdot\frac{1}{1-2r_2\cos(\phi_1-\phi_2)+r_2^2}=\frac{1}{(1-r_1)^2}\cdot\Delta|H| \tag{18}$$

Where $\Delta|H|$ is defined as the pole interaction factor (PIF) of pole z_2 with pole z_1 and can be used to measure the interaction effect [10].

In the z domain, when poles z_1 and z_2 gradually converge, the differences in their radiation angles would be reduced, according to formula (18), the angle difference between these two poles is diminished, the PIF is increased, and the corresponding spectral peak at the angle ϕ_1 is raised. Otherwise, the PIF is decreased and the corresponding spectral peak is lowered. This brings up the pole interaction problem that then leads to the result that one or more real roots are regarded as spurious and deleted from the original LP polynomial. To reduce the degradation of pole interactions, we proposed an improved algorithm. In the new design, the radii of the remaining poles are modified to make the spectral energy of the formant polynomial equal to that of the original whispered LP polynomial at the formant frequencies.

Suppose the reserved formant pole with the angle of ϕ_i and the radius of r_i is z_i, and the deleted formant pole with the angle of ϕ_j and the radius of r_j is z_j. According to formula (18), the power at the angle of ϕ_i is

$$\left|H(e^{j\phi_i})\right|^2=\frac{1}{(1-r_i)^2}\times\frac{1}{1-2r_j\cos(\phi_i-\phi_j)+r_j^2}=\frac{1}{(1-r_i^{'})^2} \tag{19}$$

Here, $r_i^{'}$ denotes the corresponding new pole radius when deleting pole z_j and reserving the unchanged pole energy, and M denotes the deleted pole amount.

In addition, we must consider the influence on other reserved poles when changing the radius of a pole. So, the formula (19) could be extended as

$$\begin{aligned}&\frac{1}{(1-r_i)^2}\prod_{k\neq i}^{N}\frac{1}{1-2r_k\cos(\phi_i-\phi_k)+r_k^2}\times\frac{1}{1-2r_j\cos(\phi_i-\phi_k)+r_j^2}\\&=\frac{1}{(1-r_i^{'})^2}\prod_{k\neq i}^{N}\frac{1}{1-2r_k^{'}\cos(\phi_i-\phi_k)+r_k^{'2}}\end{aligned} \tag{20}$$

Here, r_k are radii of other reserved poles: $r_k^{'}$ is the corresponding pole radius after modification, and N is the amount of reserved linear predictive multinomial pole.

The algorithm is realized in the following six ways:

(1) Create a whisper spectrum and determine LP roots from the LP polynomial,
(2) Calculate the bandwidth for each root by using formula (16),
(3) Descend the bandwidths and classify the root with the largest bandwidth as the deleted pole,
(4) Start with the root whose bandwidth is the smallest and obtain the new root by using formula (19),
(5) Use formula (20) to modify the rest of the formant roots; and,
(6) Continue this process from Step 3 to Step 5 until only the all saved formants are obtained.

4. 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:

$$X_t = [1-\Phi_t]X_{t-1} + \Phi_t v_t + V_{t-1} \tag{21}$$

Where, v_t was the dynamic target orientation, namely, the formant parameter solved with PIF-LPC at moment t; Φ_t was the weighted variable at moment t. It was solved according to the following formula:

$$\Phi_t = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{\|C_t - C(v_t)\|^2}{2\sigma^2}} \tag{22}$$

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 (22), 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 continuous Chinese whispered 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 was solved with PIF-LPC.

(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;
5) The state of output filtering $x_t = \sum_{i=1}^{n} \overline{w_t^i} x_t^i$.

(4) Weight Φ_t was calculated according to formula (22).

(5) The state output of HDM at moment t was calculated according to formula (21).

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

5. Test results and analysis

The sample database of test speeches was composed of 1632 section of continuous whispered speech recorded by 100 testers with the duration between 0.8s and 4.7s. 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.

In Fig.1, it is shown the tracking effect of continuous Chinese whispered speech /su zhou da xue/ using LPC and PIF-LPC separately. It can be seen that both LPC and PIF-LPC tracking the significant formant correctly on spectrograms, such as section /a/ and section /ue/. However, LPC brought more mistakes comparing with PIF-LPC. For instance, with LPC, there were mistakes in the extraction of F1,

F2 and F3 at 1.25s, F2 and F3 at 1.5s, and F2 at 1.9s. Corresponding to these moments, correct values of formant were extracted with PIF-LPC. At such contoid sections as /s/, /zh/ and /x/, such problems as the peak merging (F3 merged with F4) and the incompleteness of formant structures (F1 and F2 being unconspicuous) caused mistakes of tracking with LPC. In transitional speech sections, the formant peaks solved with two algorithms were in random states. Therefore, neither of the two algorithms tracked formant trajectory of continuous whispered speech accurately.

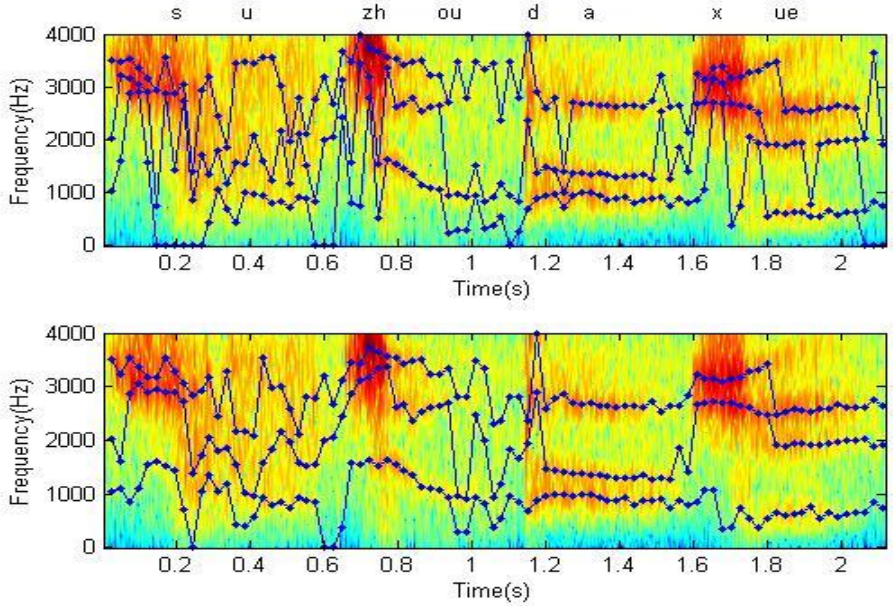

Figure 1. The tracking effect of continuous Chinese whispered speech /su zhou da xue/ with LPC was shown in the upper graph; the tracking effect with PIF-LPC was shown in the lower graph

Fig.2 shows the tracking effect of continuous Chinese whispered speech /su zhou da xue/ 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 [500,2000,3000,50,100,150], and the weighted value Φ was set as 0.3. Compared with LPC and PIF-LPC, 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 continuous 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, F1 at section /ou/ decreased from 1500Hz at 0.8s to 700Hz at 1.1s with the frequency span up to 800Hz, and F2 at section /a/ from 1.2s to 1.55s tracked falsely. In this case, static orientation targets might 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 (22). 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.

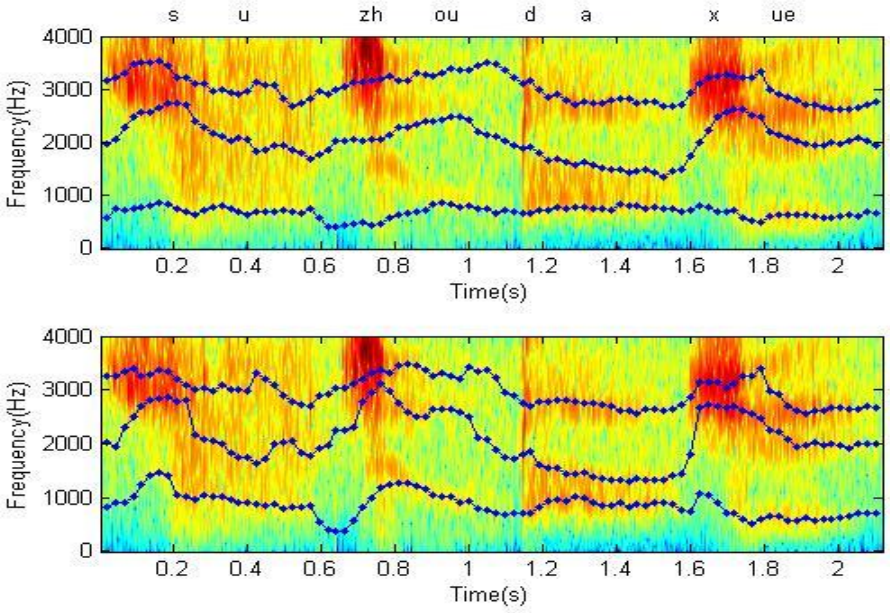

Figure 2. The upper graph shows the tracking effect of continuous Chinese whispered speech /su zhou da xue/ with HDM based on static target orientation; the lower graph shows the tracking effect with HDM based on dynamic target orientation

6. Conclusions

A sort of HDM based on dynamic target orientation was put forward in this study. The formant parameters of whispered 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 contoid 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 continuous Chinese whispered speech.

7. Acknowledgements

The authors would like to express thanks to Yebin. W. for his assistance in providing the earlier research data. This work is supported by the National Natural Science Foundation of China, under Grant No. 61071215 and Canadian Center of Science and Education under Grant No. B2009-122.

8. References

[1] Morris R.W., Clements M.A., “Reconstruction of speech from whispers”, Medical Engineering and Physics, vol. 24, no. 7, pp. 515-520, 2002.
[2] Itoh T., Takeda K., Itakura F., “Analysis and recognition of whispered speech”, Speech Communication, vol. 45, no. 2, pp. 139-152, 2005.

[3] Richards H. B., Bridle J. S., "The HDM: A segmental hidden dynamic model of coarticulation", In Proceeding of the ICASSP, pp. 357-360, 1999.
[4] Deng L., Ma J., "Spontaneous speech recognition using a statistical coarticulatory model for the vocal-tract-resonance dynamic", Journal of the Acoustical Society of America, vol. 108, no. 6, pp. 3036-3048, 2000.
[5] Deng L., Lee L.J., Attias H., Acero A., "Adaptive kalman filtering and smoothing for tracking vocal resonances using a continuous-valued hidden dynamic model", IEEE Transactions Speech Audio Process, vol. 15, no. 1, pp. 13-23, 2007.
[6] Zheng Y., Hasegawa-Johnson M., "Formant tracking by mixture state particle filter", In Proceeding of the ICASSP, pp. 565-568, 2004.
[7] Deng L., Lee L.J., Attias H., Acero A., "A structured speech model with continuous hidden dynamics and prediction-residual training for tracking vocal tract resonances", In Proceeding of the ICASSP, pp. 557-560, 2004.
[8] Fredrik G., Niclas B., Urban F., "Particle filters for positioning, navigation and tracking", IEEE Transactions on signal processing, vol. 50, no. 2, pp. 425-437, 2002.
[9] Pitt M.K., Shephard N., "Filtering via simulation: Auxiliary particle filters", American Statistical Association, vol. 94, no. 446, pp. 590-599, 1999.
[10] Gang L., Heming Z., "Formant frequency estimations of whispered speech in chinese", Archives of Acoustics, vol. 34, no. 2, pp. 127-135, 2009.