Inverse Filtering Based Harmonic plus Noise Excitation Model for HMM-based Speech Synthesis

Zhengqi Wen[1], *Jianhua Tao*[2]

[1, 2] National Laboratory of Pattern Recognition,
Institute of Automation, Chinese Academy of Sciences, 100190, Beijing, China
zqwen,jhtao@nlpr.ia.ac.cn

Abstract

In this paper, a new Voicing Cut-Off Frequency (VCO) estimation method based on inverse filtering is presented. The spectrum of residual signal got from inverse filtering is split into sub-bands which are clustered into two classes by using K-means algorithm. And then, the Viterbi algorithm is used to search a smoothed VCO contour. Based on this new VCO estimation method, an adaptation of Harmonic Noise Model is also proposed to reconstruct the residual signal with both harmonic and noise components. The proposed excitation model can reduce the buzziness of speech generated by normal vocoders using simple pulse train, and has been integrated into a HMM-based speech synthesis system (HTS). The listening test showed that the HTS with our new method gives better quality of synthesized speech than the traditional HTS which only uses simple pulse train excitation model.

Index Terms: speech synthesis, excitation model, inverse filtering, voicing cut-off frequency

1. Introduction

Statistical parametric speech synthesis systems, especially Hidden Markov Models (HMM)-based speech synthesis systems (HTS), have shown their ability to produce understandable and natural-sounding voices [1]. Compared to the traditional unit-selection systems [2] which are highly depending on the quality of the recorded database, HTS takes a lot of advantages which are related to its flexibility due to the statistical modeling process. In [1], Zen et al. summed up these advantages and disadvantages over unit-selection synthesis systems. The main disadvantage of HTS is the quality of synthesized speech which sounds buzzy. It is a typically problem encountered in normal vocoders that excitation model used is either a pulse train or a white Gaussian noise during voiced and unvoiced segments respectively [3].

To reduce the buzziness of speech generated by HTS using the traditional excitation model, some approaches have been proposed in literature. In [4], Yoshimura et al. integrated the mixed excitation model into HTS. In [5], Cabral et al. used the Liljencrants-Fant waveform [6] to model glottal source because it has a decaying spectrum at high-frequency region which is similar to the real glottal source. In [7], Drugman et al. proposed an adaptation of the Deterministic plus Stochastic Model (DSM) for residual signal. In [8], Raitio et al. used Iterative Adaptive Inverse Filtering (IAIF) [9] to decompose voiced speech into the vocal tract transfer function, voice source spectrum and glottal waveform. All these methods which focus on excitation modeling have showed their ability to improve the quality of synthesized speech.

This paper describes an excitation model based on inverse filtering technique. In source-filter theory, speech signal is modeled by an excitation signal and a filter which contains the effects of the source, vocal tract and lip radiation. In spectral analysis stage, all these effects are incorporated together in the spectral coefficients and so residual signal got from inverse filtering has a flat spectrum as showed in Figure 1. We take an adaptation of Harmonic plus Noise Model [10] and treat the harmonic region as a sum of a number of harmonically related sinusoids and the noise region as a high-passed white Gaussian noise. In order to split the spectrum, a new Voicing Cut-Off Frequency (VCO) estimation method is proposed which split the spectrum of residual signal into sub-bands which are clustered into two classes by using K-means algorithm. And then Viterbi algorithm is used to search a smoothed VCO contour. Listening tests showed that the quality of synthesized speech of HTS using the proposed excitation model is better than traditional excitation model with the same configurations and databases.

The remaining part of this paper is organized as follows. Section 2 gives a detail description of the VCO estimation method we proposed in this paper. Section 3 illustrates a vocoder based on the Harmonic plus Noise Excitation Model. In Section 4 we integrate the excitation model into the HTS. Then a listening test is conducted in Section 5. Finally in Section 6, conclusions and future work are summarized.

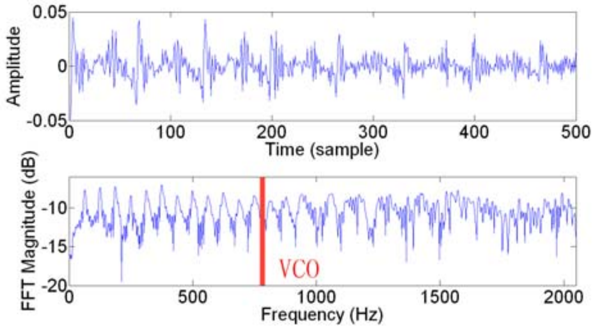

Figure 1: *Residual and Corresponding FFT Spectrum.*

2. Voicing Cut-Off Frequency

The phenomenon that most speech frames have harmonic structure in low-frequency region and noise structure in high-frequency region as showed in Figure 1 had caught researchers' attentions. They assumed that the spectrum of the speech frames could be split into two distinct regions: a low-frequency harmonic region and a high-frequency noise region [11]. The split frequency is called Voicing Cut-Off frequency (VCO) as marked in Figure 1.

There are numerous methods to estimate the VCO. In [12], Hermus et al. summarized these methods into three categories: analysis-by-synthesis methods, spectral domain methods and time domain methods, and figured out the advantages and disadvantages of these methods and then proposed a new method which combines the ideas from [13] and [14]. In Hermus's method, a speech frame of two pitch period is used to calculate a discrete Fourier transform (DFT) and define the peakiness based on the fact that the odd lines and the even lines of DFT contain harmonic components and noise components, respectively. The effective of this method mostly depends on the accuracy of pitch estimation.

To alleviate the dependency on the accuracy of pitch estimation, in this paper we proposed a new VCO estimation method based on statistical method. Our estimation method is motivated by [12] and [13]. In our method, we split spectrum of residual signal into sub-bands based on the estimated frequency. These sub-bands are aligned with the max value before clustering. K-means algorithm is used to cluster these sub-bands into two classes and then the Viterbi algorithm is adopted to do time smoothing to get a smoothed VCO contour.

2.1. Spectral Estimation and Splitting

Let $s(k), k = 1,2,\cdots,K$ be a frame of residual signal with corresponding discrete Fourier transform (DFT) $S(n), n = 1,2,\cdots,N$ by a rectangle windowing. In this paper, we suppose N is odd and only consider the first $(N+1)/2$ DFT coefficients, i.e. the "positive" frequencies of the DFT. Then we split the spectrum into sub-bands based on the frequency of this frame and these sub-bands are aligned with the max value. The number and the length of these sub-bands are calculated by:

$$length = \lfloor f * N / Fs \rfloor \tag{1}$$

$$number = \lfloor Fs / f \rfloor \tag{2}$$

where Fs is sampling frequency and f is the frequency of a frame signal.

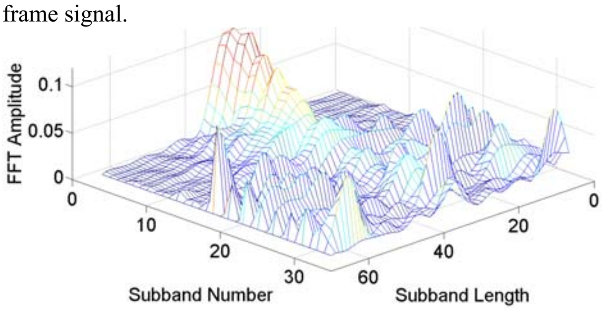

Figure 2: *Splitting the spectrum of residual signal into sub-bands.*

As showed in Figure 2, sub-bands from low-frequency region have similar shape but irregular form from high-frequency region. So it is possible to use K-means algorithm to cluster these sub-bands into two classes and split the spectrum into two regions: low-frequency harmonic region and high-frequency noisy region.

2.2. K-means Clustering

We treat every sub-band as a one-dimension vector and use a definition to measure the distance between every two sub-bands. In this paper, we adopt the angle between two sub-bands defined in Eq. 3 as distance measure.

$$d(x_i, x_j) = \cos^{-1}\left(\frac{\sum_{k=1}^{M} x_i(k) \cdot x_j(k)}{\sqrt{\sum_{k=1}^{M}(x_i(k))^2} \cdot \sqrt{\sum_{k=1}^{M}(x_j(k))^2}}\right) \tag{3}$$

where x_i is a sub-band and M is the length of a sub-band.

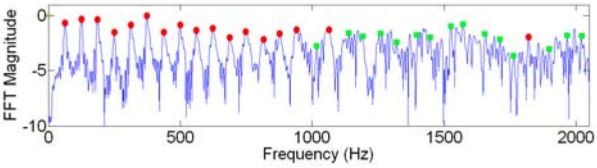

Figure 3: *K-means Clustering result using angle as distance measure.*

Sub-bands are divided into two classes which are labeled by different marks and different colors in Figure 3 and we assume that one is harmonic and the other is noise. We define VCO candidate by the gradient of clustering results as Eq. 4 expressed. As in Figure 3 showed, we can find there is not only one VCO candidate and thus need a time smoothing procedure to get a smoothed VCO contour.

$$gradient_{i,k} = |mark_k - mark_{k+1}| \tag{4}$$

where $mark_k$ and $mark_{k+1}$ are the clustering results of k^{th} and $(k+1)^{th}$ sub-band.

2.3. Time Smoothing

Different from smoothing algorithms which only considered the adjacent frames, we use dynamic programming to search a smoothed path through concatenating target function which is defined in Eq. 5.

$$d(i_m, (i+1)_n) = 1 / \left(|m-n| + l\right) \tag{5}$$

where i_m is m^{th} sub-band of i^{th} residual signal frame and $(i+1)_n$ is n^{th} sub-band of $(i+1)^{th}$ residual signal frame, l is a coefficient which is used to control the smoothness of the VCO contour.

We adopt the Viterbi algorithm showed in Eq. 6 to maximum the target score to find smoothed VCO contour.

$$\begin{aligned} V_{0,k} &= P(x_0 \mid k) \cdot gradient_{0,k} \\ V_{t,k} &= gradient_{t,k} \cdot \max_{m \in M_t}(d((t-1)_m, t_k) \cdot V_{t-1,m}) \\ y_T &= \arg\max x_{m \in M_T}(V_{T,m}) \\ y_{t-1} &= Ptr(y_t, t) \end{aligned} \tag{6}$$

where T is the total frame number, M_t is the number of sub-band of t^{th} residual signal frame, $V_{t,k}$ is the best VCO contour from t^{th} residual signal frame to k^{th} residual signal frame.

The smoothed VCO contour is showed in Figure 4. In this figure, the frequency range is normalized to $0 \sim \pi$ and the frame shift is 0.05s.

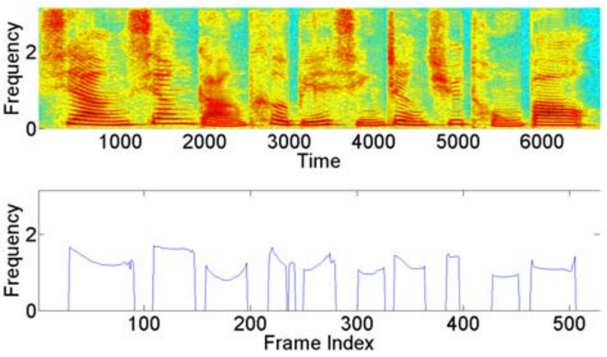

Figure 4: *A waveform spectrum and corresponding smoothed VCO contour.*

3. Vocoder Based on Harmonic plus Noise Excitation Model

Traditional excitation model in LPC-like vocoders is either a pulse train or a white Gaussian noise during voiced and unvoiced segments respectively. The synthesized speech of these vocoders sounds buzzy for the reason that this excitation model in voiced segments has regular harmonic structure in high-frequency region. In order to attenuate this problem, we proposed an adaptation of Harmonic plus Noise model and use the VCO estimated in previous section to divide the spectrum into low-frequency harmonic region and high-frequency noisy region.

3.1. Harmonic Region

Harmonic Region is represented as a sum of a number of harmonically related sinusoids in Eq. 7. And the number of sinusoids is calculated by Eq. 8.

$$s_h[t]=\sum_{n=1}^{k}\sin(2\pi nft+\varphi(n)) \qquad (7)$$

$$k=\left\lfloor \frac{VCO}{f} \right\rfloor \qquad (8)$$

where f is the frequency of a frame of residual signal, $\varphi(n)$ is the initial phase of n^{th} sinusoid.

3.2. Noisy Region

In [15], Pantazis et al. study Triangular envelope, Hilbert envelope and Energy envelope to model the temporal characteristics of noise, and listening test showed a clear preference on the Energy envelope and Hilbert envelope. In this paper, we adopt the Energy envelope in modeling the noisy region. Based on the calculated VCO, a white Gaussian noise is high-passed and its temporal structure is modeled by Energy envelope.

As Figure 5 showed, the residual signal is reconstructed by adding high-frequency noise and low-frequency harmonic together and then filtered by spectral coefficients to synthesis speech.

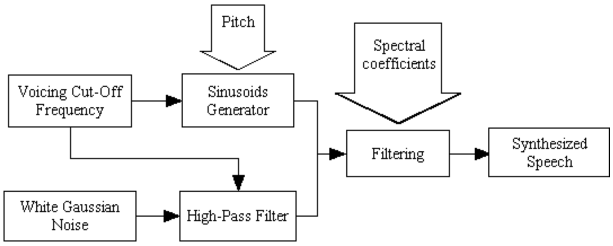

Figure 5: *The flowchart of a vocoder based on Harmonic plus Noise excitation model.*

4. Integration into the HMM-based Speech Synthesis System

Our implementation of HMM-based Speech Synthesis System is based on the HTS toolkit available in [16] and some modifications are needed while integrating the Harmonic plus Noise excitation model. Mel-cepstral coefficients are used as spectral parameters. The excitation parameters including pitch and VCO should treat differently between voiced and unvoiced region and in our HTS we adopt a multi-space distribution (MSD).

Context dependent HMMs are trained and clustered using a tree-based context clustering technique based on minimum description length principle. Details of the contextual factors are described in [3].

In synthesis stage, input text is transited to a context dependent label sequence which is used to construct a sentence HMM. Parameter generation algorithm [17] is adopted to generate the spectral and excitation parameters from the sentence HMM. The excitation is generated and speech is synthesized from obtained excitation and spectral coefficients as described in Section 3.

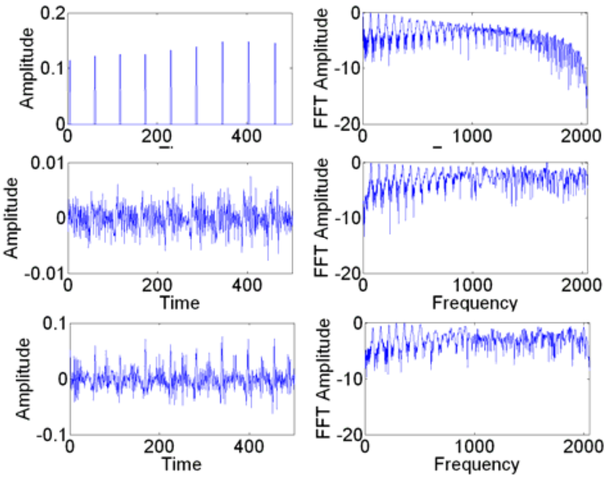

Figure 6: *A comparison between residual signal and reconstructed residual signal. From top to down are Pulse Excitation signal, Harmonic plus Noise Excitation signal and residual signal.*

5. Experiments and Results

5.1. Excitation Generation

The harmonic plus noise excitation generated by using the pitch and VCO parameters is compared to the corresponding residual and traditional pulse plus noise excitation. In figure 6, waveform and amplitude spectrum are considered in this comparison. It is showed that the harmonic plus noise excitation model is better than traditional pulse trains or noise excitation model in reconstructing the residual signal, especially in the high-frequency region.

5.2. Listening Test

CMU ARCTIC databases [18] named SLT and BDL which are uttered by a US English female speaker and a US English male speaker respectively are used to train the HTS. In order to evaluate the effectiveness of the proposed harmonic plus noise excitation model, HTS using the traditional excitation model with the same configurations and databases is also trained. A comparative evaluation is conducted on the quality of synthesized speech of these two typed HTS. In this evaluation, a preference test is conducted by 10 participants. They are asked to listen to both versions of 3 groups of sentences and to attribute which one is better. These 3 groups of sentences have different length and contain different rhythm information.

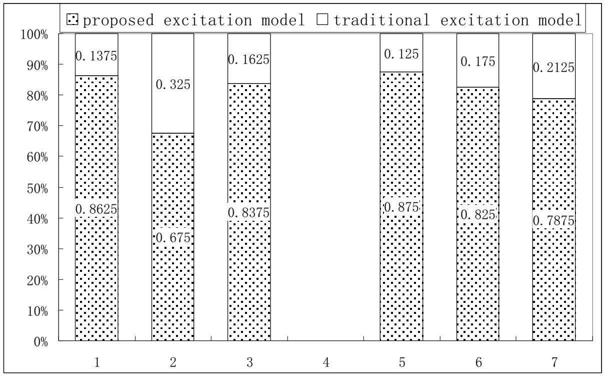

Figure 7: *Listening test results of the comparison of traditional excitation model and harmonic plus noise excitation model. The left three groups are based on BDL database and the right three groups are based on SLT database.*

In Figure 7, listening test results show that the harmonic plus noise excitation model proposed in this paper have significantly improved the quality of synthesized speech. The average preference score of all voices is about 81.04% and BDL is 79.17%, SLT is 82.92%. It means that female voices are better than male voices. The main reason for this is that female voices have higher frequency and more noise in high-frequency region than male voices. Higher frequency means larger gaps between harmonics. Female voices need more high-frequency noise to sound naturalness.

However, participants in this listening test noticed that some sentences sound too noisy. It is mostly attributed to the sharp change from the harmonic region to the noise region. So in next step we will model both the harmonic and the noise in the whole frequency band.

6. Conclusion and Future works

In this paper, a new VCO estimation method based on inverse filtering is proposed. Based on this VCO contour, we proposed an adaptation of Harmonic plus Noise Model to reconstructed residual signal and then integrated the excitation model into HTS. A listening test is conducted to compare the quality of synthesized speech of HTS based on traditional excitation model and the proposed excitation model. The results showed a significant improvement of the quality of synthesized speech.

However, the proposed excitation model only considered the harmonic part and the noise part separately, in future work we will focus our attentions on both of them modeling in the whole frequency band and using of other high-precision excitation models.

7. Acknowledgements

The work was supported by the National Science Foundation of China (No. 60873160 and No.90820303) and China-Singapore Institute of Digital Media (CSIDM).

8. References

[1] Heiga, Z., Keiichi, T. and Alan, W.B., “Statistical parametric speech synthesis”, Speech Communication, 51(11):1039-1064, 2009.

[2] Hunt, A.J. and Black, A.W., “Unit selection in a concatenative speech synthesis system using a large speech database”, IEEE ICASSP, 373-376, 1996.

[3] Yoshimura, T., Tokuda, K., Masuko, T., Kobayashi, T. and Kitamura, T., “Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis”, EUROSPEECH, 2347-2350, 1999.

[4] Yoshimura, T., Tokuda, K., Masuko, T. and Kitamura, T., “Mixed excitation for HMM-based speech synthesis”, EUROSPEECH, 2259-2262, 2001.

[5] Cabral, J., Renals, S., Richmond, K. and Yamagishi, J., “Towards an Improved Modeling of the Glottal Source in Statistical Parametric Speech Synthesis”, ISCA SSW6, 2007.

[6] Fant, G., Liljencrants, J. and Lin, Q., “A four parameter model of glottal flow”, STL-QPSR4, 1-13, 1985.

[7] Drugman, T., Wilfart, G. and Dutoit, T., “A Deterministic plus Stochastic Model of the Residual Signal for Improved Parametric Speech Synthesis”, Interspeech, 2009.

[8] Raitio, T., Suni, A., Pulakka, H., Vainio, M. and Alku, P., “HMM-based Finnish text-to-speech system utilizing glottal inverse filtering”, Interspeech, 1881–1884, 2008.

[9] Plumpe, M., Quatieri, T. and Reynolds, D., “Modeling of the glottal flow derivative waveform with application to speaker identification,” IEEE Trans. Speech Audio Processing, 7(5):569–585, 1999.

[10] Stylianou, Y., Laroche, J. and Moulines, E., “High-quality speech modification based on a harmonic + noise model,” EUROSPEECH, 451–454, 1995.

[11] Makhoul, J., Viswanathan, R., Schwartz, R. and Huggins, A., “A mixed source model for speech compression and synthesis,” IEEE ICASSP, 163–166, 1978.

[12] Hermus, K., Van hamme, H. and Irhimeh, S., “Estimation of the voicing cut-off frequency contour based on a cumulative harmonicity score,” *IEEE SP Letters*, 14(11):820–823, 2007.

[13] Stylianou, Y. “Applying the harmonic plus noise model in concatenative speech synthesis”, *IEEE Trans. Speech Audio Process.*, 9(1):21–29, 2001.

[14] Kim, E.-K., Han, W.-J. and Oh, Y.-H., “A new band-splitting method for two-band speech model,” *IEEE Signal Process. Lett.*, 8(12):317–320, 2001.

[15] Pantazis, Y. and Stylianou, Y., “Improving the modeling of the noise part in the Harmonic plus Noise Model of speech”, IEEE ICASSP, 2008.

[16] [Online], “HMM-based Speech Synthesis System (HTS)”, http://hts.sp.nitech.ac.jp/.

[17] Tokuda, K., Yoshimura, T., Masuko, T., Kobayashi, T. and Kitamura, T., “Speech parameter generation algorithms for HMM-based speech synthesis”, IEEE ICASSP, 1315–1318, 2000.

[18] [Online], “CMU_ARCTIC speech synthesis databases”, http://festvox.org/cmu_arctic/index.html.