INTERNATIONAL JOURNAL FOR TRENDS IN ENGINEERING & TECHNOLOGY
VOLUME 3 ISSUE 3 – MARCH 2015 – ISSN: 2349 – 9303

Speech Analysis and synthesis using Vocoder

Kala A[1]
[1]SNS College of Technology,
ME ECE, Department of ECE,
Coimbatore-641035
kalaalwar@gmail.com

Vanitha S[2]
[2]SNS College of Technology,
Assistant Professor, Department of ECE,
Coimbatore-641035
vanitharajanneel@gmail.com

Abstract— In this paper, I proposed a speech analysis and synthesis using a vocoder. Voice conversion systems do not create new speech signals, but just transform existing one. The proposed speech vocoding is different from speech coding. To analyze the speech signal and represent it with less number of bits, so that bandwidth efficiency can be increased. The Synthesis of speech signal from the received bits of information. In this paper three aspects of analysis have been discussed: pitch refinement, spectral envelope estimation and maximum voiced frequency estimation. A Quasi-harmonic analysis model can be used to implement a pitch refinement algorithm which improves the accuracy of the spectral estimation. Harmonic plus noise model to reconstruct the speech signal from parameter. Finally to achieve the highest possible resynthesis quality using the lowest possible number of bits to transmit the speech signal. Future work aims at incorporating the phase information into the analysis and modeling process and also synthesis these three aspects in different pitch period.

Index Terms— Frequency Cepstral Coefficient, Pitch Detection, Spectral Envelope Estimation, Maximum Voiced Frequency, Harmonic plus Noise Model.

1 INTRODUCTION

Today, the synthesis quality and recognition rate are so that commercial applications are following from them. So, for speech synthesis, I can cite Telecommunications, Multimedia and Automobile with for example for Telecommunications, the vocalization of SMS, the reading of mails, the phone access to fax and e-mails, the consulting of databases, automatic answer phones (ex : Chrono Post) ; for Multimedia, the speech interface between man and machine, the help for teaching reading or / new languages (educational tools and / software), the help for teaching reading to blind people, bureaucratic tools ; and at last for the Automobile, the alert and video surveillance systems, the Internet access among others for the mail reading. Some companies as Nuance, Scan soft and Acapela-Group are present on the business market.

It aims at the development of implementing a speech processing model called *"Harmonics Plus Noise Model"* (HNM). It's in fact a hybrid model since it decomposes the speech frames into a harmonic part and a noise part. It normally has to produce a high quality of artificial speech.

In voice conversion systems do not create new speech signals, but just transform existing ones. This is the reason why this paper has been focused on synthesis. Understood in this context, speech vocoding is different from speech coding. The main goal of speech coding is to achieve the highest possible resynthesis quality using the lowest possible number of bits to transmit the speech signal. Real time performance during analysis and reconstruction is also one of its typical requirements. In the statistical parametric frameworks mentioned above, vocoders must have not only these high resynthesis capabilities, but also provide parameters that are adequate to statistically model the underlying structure of speech, while information compression is not a priority.

Vocoders are a class of speech coding systems that analyze the voice signal at the transmitter, transmit parameters derived from the analysis, and then synthesize the voice at the analysis, and then synthesize the voice at the receiver using those parameters. All vocoder attempts to model the speech generation process as a dynamic system and try to quantify certain physical constraints of the system. These physical constraints are used to provide a parsimonious description of the speech signal [4]. Vocoders are, in general, much more complex than the waveform coders and achieve very high economy in transmission bit rate. However, they are less robust, and their performance tends to be talker dependent. The most popular among the vocoding systems is the *linear predictive coder* (LPC). The other vocoding schemes include the channel vocoder, formant vocoder, cepstrum vocoder and voice excited vocoder.

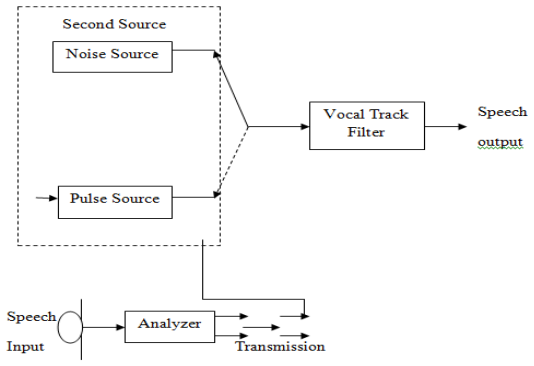

Fig. 1 Speech Generation Model

Fig 1. Shows the traditional speech generation model that is the basis of all vocoding systems. The sound generating mechanism forms the source and is linearly separated from the intelligence modulating vocal tract filter which forms the system. The speech signal is assumed to be of two types: voiced and unvoiced sound ("m", "n", "v" pronunciations) are a result of quasiperiodic vibrations of the

vocal chord and unvoiced sounds (“f”, “s”, “sh” pronunciations) are fricatives produced by turbulent air flow through a construction. The parameters associated with this model are the voice pitch, the pole frequencies of the modulating filter, and the corresponding amplitude parameters. The pitch frequency for most speakers is below 300 Hz, and extracting this information from the signal is very difficult. The pole frequencies correspond to the resonant frequencies of the vocal tract and are often called the formants of the speech signal. For adult speakers, the formants are centered around 500 Hz, 1500 Hz, 2500 Hz and 3500 Hz. By meticulously adjusting the parameters of the speech generation model, good quality speech can be synthesized [2].

MFCC

Wrapping of signals in the frequency domain using 24 filter banks are done. This filter is developed based on the behavior of human ear’s perception, or each tone of a voice signal with an actual frequency f, measured in Hz, it can also be found as a subjective pitch in Mel frequency scale [10]. The Mel frequency scale is determined to have a linear frequency relationship below 1000 Hz and a logarithmic relationship higher than 1000Hz. The Mel frequency higher than 1000 Hz is,

$$\text{Mel (f)} = 2595 * log_{10}\left(1 + \frac{f}{700}\right) \qquad (1.1)$$

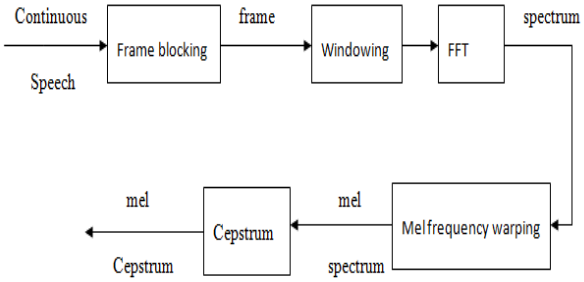

Fig. 2 MFCC Processor

In this final step, convert the log Mel spectrum returns to time. The result is called the Mel frequency Cepstrum coefficients (MFCC).

2 HNM ANALYSES

The analysis consists of estimating the harmonic and noise parameters. By the decomposition into two independent parts, these are estimated separately. First, I have to separate the voiced frames from the unvoiced frames and then compute the parameters used for the synthesis. The Proposed system is shown in figure 2.1.

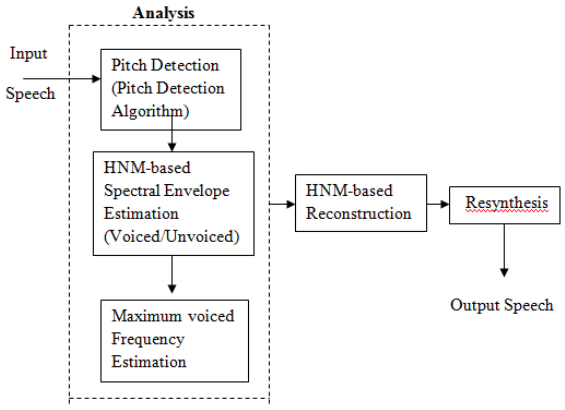

Fig. 2.1 Block diagram of Harmonic plus Noise Model System

The steps of the analysis process are the following:

- First Estimation of fundamental frequency
- Voiced/Unvoiced decision for each frame
- Spectral Envelope Estimation
- Estimation of the maximum voiced frequency FM
- Refinement of the fundamental frequency

2.1 FIRST ESTIMATION OF THE FUNDAMENTAL FREQUENCY (PITCH):

The first step consists of the estimation of the pitch f0. This parameter is estimated every 10ms. I will see that the length of the analysis window will depend on this local pitch. One method co-developed with the HNM algorithm is explained here. This method is based on an autocorrelation approach and is obtained by fitting the original signal with another defined by a pitch (sum of harmonics) in the frequency domain:

$$\varepsilon = \frac{\int_{-\frac{1}{2}}^{\frac{1}{2}} \left[|S_w(f)| - |\hat{S}_w(f)|\right]^2 df}{\int_{-1/2}^{1/2} |S_w(f)|^2 df} \qquad (2.1)$$

Where $S_w(f)$ are the Short Term Fourier Transform of the speech segment S (t) (Blackman weighted segment whose length is equal to 3 times the maximum fundamental period T_0^{max}) and$\hat{S}_w(f)$, the Short Term Fourier Transform of a purely harmonic signal obtained from a fundamental frequency F_0.

To avoid some pitch errors, a “peak tracking” method is needed. This has to look at two frames forward and backward from the current one. The minimum error path is found and by this way the pitch is associated.

2.2 VOICED/UNVOICED DECISION

The frames extracted every 10 ms (whose length is always 3 times T_{0max}) have then to be classified as "voiced" or "unvoiced". We apply the Short-Term Fourier Transform (STFT) with a number of points NFFT equal to 4096 (with zero padding) to the current frame and we call it S (f).

From this first STFT, we can evaluate the first four amplitudes of the harmonics (the first of which is the fundamental). We note S^ (f), which is a set of amplitudes of harmonics (of f0).
The following criterion is applied:

$$E = \frac{\int_{0.7f_0}^{4.3f_0}(|S(f)| - |\hat{S}(f)|)^2 df}{\int_{0.7f_0}^{4.3f_0}(|S(f)|)^2 df} \qquad (2.2)$$

The frame is declared "voiced" if E is less than the threshold of -15dB, and "unvoiced" otherwise.

.

2.3 SPECTRAL ENVELOPE ESTIMATION:

Assuming a simplified speech production model in which a pulse-or-noise excitation passes through a shaping filter, the term *spectral envelope* denotes the amplitude response of this Filter in frequency. Such an envelope contains not only the contribution of the vocal tract, but also the contribution of the glottal source. In unvoiced frames, the spectrum of the noise-like excitation is flat, which means that the response of the filter coincides with the spectrum of the signal itself (except for a scaling factor). In voiced frames, the spectrum of the pulse-like excitation has the form of an impulse train with constant amplitude and linear-in-frequency phase placed at multiples of, Therefore, the spectrum of the signal shows a series of peaks that result from multiplying the impulses of the excitation by uniformly spaced spectral samples of the filter response. Assuming local stationarity, full-band harmonic analysis returns these discrete samples of the spectral envelope. Then, a continuous envelope can be estimated via interpolation.

2.4 ESTIMATION OF MAXIMUM VOCED FREQUENCY

This parameter is also estimated every 10ms. In the beginning, I work in the Interval $\left[\frac{f_0}{2}, \ldots, \frac{3f_0}{2}\right]$ of the absolute spectrum. I look for the greatest amplitude and the corresponding voiced frequency in this interval, which I denote A_m and f_c, respectively. I also compute the sum of the amplitudes (called the cumulative amplitude A_{mc}) located between the two minima around the greatest voiced frequency. The other peaks in the band are also considered (occurring at frequencies denoted by f_i) in the same interval, with the two types of amplitudes $A_m(f_i)$ and $A_{mc}(f_i)$. I compute the mean of these cumulative amplitudes, denoted by $\overline{Amc\ (fi)}$.

Then apply the following test to the greatest frequency f_c:

If . $\frac{A_{mc}(f_c)}{A_{mc}(f_i)} > 2$ and $A_{mc}(f_c)$-max{A_m(i)}>13dB

L being the number of the nearest harmonic of f_c.

Then the frequency f_c is declared "voiced" and the next $\left[\frac{3f_0}{2}, \ldots, \frac{5f_0}{2}\right]$ is considered and the same criterion is applied. The highest voiced frequency found will correspond to the maximum voiced frequency F_M. However, to avoid mistakes, a 3 points, median smoothing filter is applied.

As well, the frequency F_M can vary greatly from one frame to the next. In order to reduce abrupt jumps I can also use another median filter on this time-varying frequency. Five points are in general used here.

2.5 REESTIMATION OF THE FUNDAMENTAL FREQUENCY

Using the frequencies (fi) declared as voiced in the previous step, I try to minimize the following function: $E(\hat{f}_0) = \sum_{i=1}^{L(1)} |f_i - i\hat{f}_0|$ with L (i) representing the number of voiced frequencies and f_0 the initial estimation of the pitch. The minimum is reached for the new estimation of the pitch.

3 RESULTS AND DISCUSSION

In this paper, the enrollment of the user a data record was maintained in the database with different text information. This database contains 53 different voices (25 female, 28 male). The voices were taken from different speech synthesis and recognition databases in English. The specific utterances representing each voice were chosen randomly among candidates with a suitable duration (around 5 s). Although the recording conditions were database dependent, in all cases the sampling frequency was 16 kHz and the signal-to-noise ratio was checked to be high enough for analysis-synthesis purposes.

Table 1 Information about the Speakers

NUMBER OF MALE SPEAKERS	NUMBER OF FEMALE SPEAKERS	AVERAGE AGE	LANGUAGE
25	28	24	ENGLISH

The speech signal used in the training phase for a particular speaker is shown in Table 1.

The first analysis step to be performed is pitch detection. Pitch detection algorithms (PDA) used to exhibiting very good performance when applied clean signals (that the signals involved in speech synthesis usually show high signal-to-noise ratio). The vocoder presented in this paper includes an implementation of the Autocorrelation-based algorithm.

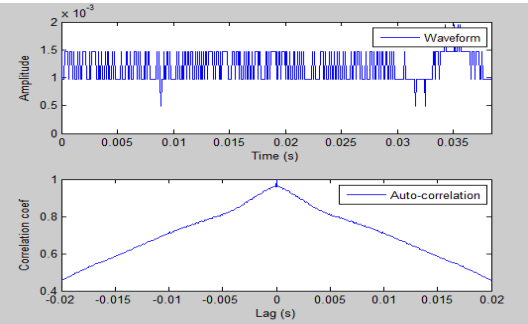

Fig 3.1 Autocorrelation of a Signal

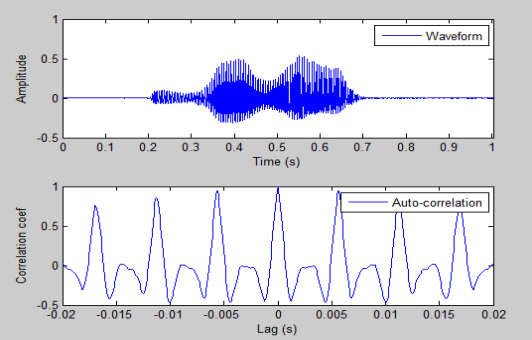

Fig 3.2 Pitch Estimation of a Signal

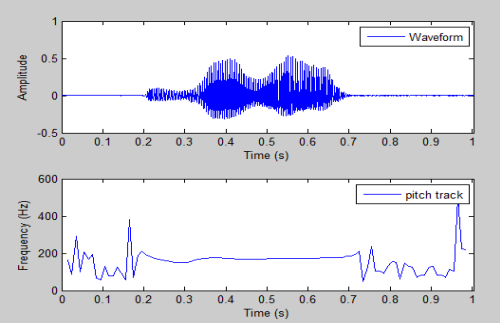

Fig 3.3 Pitch Tracking of a Signal

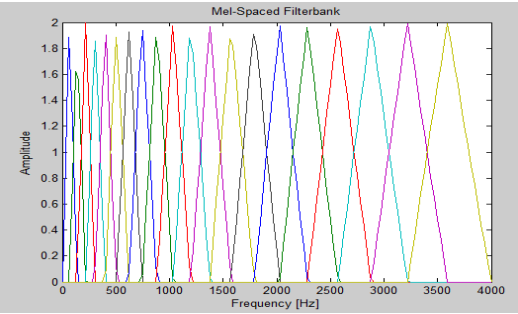

Fig 3.4 Mel scale Filter Bank

In fig 3.1 shows the Autocorrelation output of a speech signal. Here the correlation coefficient of a signal have the length is 2*maxlag+1. The correlation coefficient has the maximum value is '1'.

After autocorrelation the pitch will be estimated from each and every frame using correlation coefficients. It is shown in fig 3.2.

In fig 3.3 shows that the computation time, which should be minimized to remove the delay. Determination of voiced segment from voice, and the capable to remove the silence from sound. It provides good resolution as well as to avoid gross errors.

When I sample a spoken syllable, we will be having many samples. Then I try to extract features from these sampled values. Cepstral coefficient calculation is one of such methods. Here I initially derive Short Term Fourier Transform of sampled values, then take their absolute value (they can be complex) and calculate log of these absolute values. There after I, go for converting back them to time domain using Discrete Cosine Transform (DCT). I have done it for five users and first ten DCT coefficients are Cepstral coefficients.

It takes into account physiological behavior of perception of human ear, which follows linear scale up to 1000 Hz and then follows the log scale. Hence I convert frequency to Mel domain using a number of filters. Then I take its absolute value, apply log function and convert back them into time domain using dct. For each user I had feature vectors having 20 MFCC coefficients each. For visualization purposes I only show few feature vectors and their MFCC. In the above figure I have only chosen few feature vectors. Each column refers to a feature vector. The element of each column and the corresponding MFCCs. As I had chosen the first 24 DCT coefficients, hence each column will be having 24 elements.

In this paper the envelope of the signal estimates by using spectral envelope estimation. Fig 3.5 gives the spectral envelope estimation of speech signal.

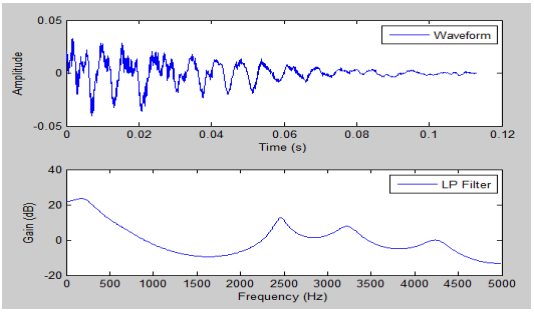

Fig 3.5 Spectral Envelope Estimation

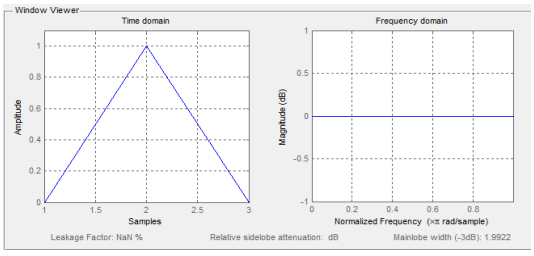

Fig 3.6 3-period Hanning Window

Maximum Voiced Frequency (MVF) is used in various speech models as the spectral boundary separating periodic and a periodic components during the production of voiced sounds. Windowing is essential as it determines the harmonicity properties of

the resulting spectra. In all cases, the window length should be proportional to the pitch period. In this work, I have used a 3 period-long Hanning window as I found it to be suited for the amplitude spectra to exhibit a good peak-to-valley structure.

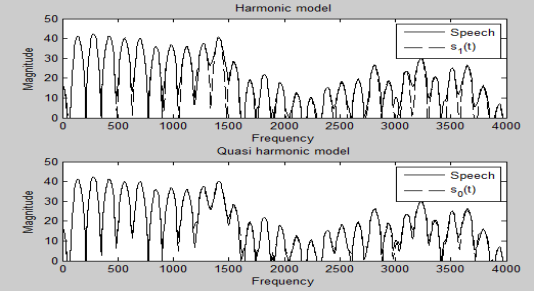

Fig 3.7 HMvsQHM

From this figure the quasi harmonic model provides smooth signal than the Harmonic model.

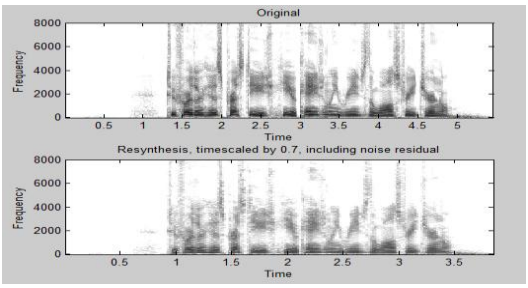

Fig 3.8 Resynthesized Speech Signal

Periodic signals can be approximated by a sum of sinusoids whose frequencies are integer multiples of the fundamental frequency and whose magnitudes and phases can be uniquely determined to match the signal - so-called Fourier analysis. One manifestation of this is the spectrogram, which shows the short-time Fourier transform magnitude as a function of time. A narrowband spectrogram (i.e. one produced with a short-time window longer than the fundamental period of the sound) will reveal a series of nearly-horizontal, uniformly-spaced energy ridges, corresponding to the sinusoidal Fourier components or harmonics that are an equivalent representation of the sound waveform. Below is a spectrogram of a brief clarinet Melody; the harmonics are clearly defined.

The key idea behind maximum voiced frequency is to represent each one of those ridges explicitly and separately as a set of frequency and magnitude values. The resulting analysis can be resynthesized by using low bit rate of information.

4 CONCLUSION

In this paper three aspects of analysis have been discussed: pitch refinement, spectral envelope and maximum voiced frequency estimation. The proposed a vocoder system was analyzed under different windowing schemes and by varying the length of frames with different overlaps. From the analysis of the three techniques, the fundamental frequency (173.611 Hz) and the formant frequencies (5 different values) are obtained. The Quasi harmonic analysis model is used to implement a pitch refinement algorithm which improves the accuracy of the subsequent spectral envelope estimation. Harmonic plus noise model is also used to reconstruct the speech signals from parameters. Therefore the harmonic model yields more spectral envelopes than sinusoidal analysis. Future work aims at incorporating the phase information into the analysis and modeling process and also synthesis these three aspects in different pitch period.

REFERENCES

[1] H. Zen, K. Tokuda, and A. W. Black, "Statistical parametric speech synthesis," *Speech Commun.*, vol. 51, no. 11, pp. 1039–1064, 2009.

[2] Y. Stylianou, "Harmonic plus noise models for speech, Combined with statistical methods, for speech and speaker modification," Ph.D. dissertations, École Nationale Supèrieure de Télécommunications, Paris, France, 1996.

[3] A.Kain, "High resolution voice transformation," Ph.D. dissertation, OGI School of Sci. and Eng. at OHSU, Portland , OR, 2001.

[4] T. Toda, A. W. Black, and K. Tokuda, "Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 8, pp. 2222–2235, Nov. 2007.

[5] J. L. Flanagan, "Parametric representation of speech signals," *IEEE Signal Process. Mag.*, vol. 27, no. 3, pp. 141–145, 2010.

[6] HMM-Based Speech Synthesis System (HTS), [Online]. Available: http://hts.sp.nitech.ac.jp/

[7] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Simultaneous modeling of spectrum, pitch and duration in HMMbased speech synthesis," in *Proc. Eurospeech*, 1999, pp. 2347–2350.

[8] K. Tokuda, T. Kobayashi, T. Masuko, and S. Imai, "Mel-generalized cepstral analysis—A unified approach to speech spectral estimation," *Proc. Int. Conf. Spoken Lang. Process.*, vol. 3, pp. 1043–1046, 1994.

[9] G.Senthil Raja, Dr.S.Dandapat, "Performance of Selective Speech Features for Speaker Identification", Journal of the Institution of Engineers (India), Vol. 89, May 29, 2008

[10] Md.Rashidul Hasan, Mustafa Jamil Md.Golam Rabbani,Md.Saifur Rahman, "Speaker Identification using Mel Frequency Cepstral Coefficients", 3rd International conference on Electrical and computer engineering ICECE 2004, Dec 2004

[11] Sandipan Chakroborty, Goutam Saha, "Improved Text-Independent Speaker Identification using Fused MFCC & IMFCC Feature Sets based on Gaussian Filter" , International Journal of Signal Processing 5:1, 2009

[12] I.Saratxaga, I. Hernaez, M. Pucher, E. Navas, and I. Sainz, "Perceptual importance of the phase related information in speech," in *Proc. Interspeech*, 2012.

[13] I.Sainz, D. Erro, E. Navas, I. Hernaez, J. Sanchez, I. Saratxaga, I.Odriozola, and I. Luengo, "Aholab Speech Synthesizers for Albayzin 2010," in *Proc. FALA*, 2010, pp. 343–347.

[14] H. Zen, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Hidden semi-Markov model based speech synthesis," in *Proc. ICSLP*, 2004, vol. II, pp. 1397–1400.