HIGH QUALITY SINUSOIDAL MODELING OF WIDEBAND SPEECH FOR THE PURPOSES OF SPEECH SYNTHESIS AND MODIFICATION

Dan Chazan[1], Ron Hoory[1], Ariel Sagi[1], Slava Shechtman[1], Alex Sorin[1], Zhi Wei Shuang[2], Raimo Bakis[3]

[1]IBM Research Laboratory in Haifa, Israel,
[2]IBM China Research Lab,
[3]IBM T.J. Watson Research Center

slava@il.ibm.com

ABSTRACT

This paper describes an efficient sinusoidal modeling framework for high quality wide band (WB) speech synthesis and modification. This technique may serve as a basis for speech compression in the context of small footprint concatenative Text to Speech systems. In addition, it is a useful representation for voice transformation and morphing purposes, e.g., simultaneous pitch modification and spectral envelope warping. The conventional sinusoidal modeling is enhanced with an adaptive frequency dithering mechanism, based on a degree of voicing analysis. Considerable reduction of the amount of model parameters is achieved by high band phase extension. The proposed model is evaluated and compared to the alternative STRAIGHT framework [1]. Being simpler and considerably more efficient than STRAIGHT, it outperforms it in speech quality for both speech reconstruction and transformation.

1. INTRODUCTION

In the framework of Concatenative Text to Speech (CTTS) the acoustic units may be represented by some speech model rather then by an exact or compressed waveform. It is desired that such a model will, on one hand, provide almost transparent speech reconstruction and, on the other hand, enable spectral domain speech manipulations (i.e. spectral smoothing, vocal tract morphing, pitch modification etc.). This results in improved unit concatenation, especially in the case of small footprint CTTS systems [4]. Such models are also highly beneficial for voice morphing purposes (i.e. changing the personality of a speaker while preserving the spoken context). In this paper an enhanced complex envelope model for speech, based on sinusoidal speech representation, is introduced. A brief description of a well-known reference speech analysis/synthesis model, called STRAIGHT [1], will be given first. Recently, it has been used as a speech modeling engine for a wide range of applications, starting from speech coding to CTTS [2] and voice morphing [3]. It is followed by the description of a computationally efficient and precise constant-frame-rate sinusoidal model estimation algorithm, which is the subject of this work. A scheme for model enhancement and considerable reduction of the amount of parameters is described. Finally, a technique of complex envelope resampling for the purposes of voice morphing and/or pitch modification is depicted. It is demonstrated, that both spectral envelope amplitude and phase may be interpolated and resampled during speech manipulations, in order to attain a high quality morphed speech. Quality and naturalness of both morphed and reconstructed signal are compared to STRAIGHT-generated signals.

2. SPEECH ANALYSIS/SYNTHESIS MODELS

2.1. Analysis/synthesis by STRAIGHT system

The STRAIGHT research software package uses a speech model especially designed for real-time speech manipulations. It represents a speech spectrum by a spectral envelope which is smoothed in frequency and in time and then densely sampled along the time (each 1 ms) and frequency (1024 point FFT analysis) axes [1]. That smooth and over-sampled representation allows a wide range of real-time speech manipulations during the reconstruction stage though it involves an analysis process of very high computational complexity [1].

STRAIGHT's analysis algorithm does not extract phase information. Rather, its reconstruction algorithm adopts the minimum phase assumption for the spectral envelope and further applies all-pass filters in order to reduce the buzz timbre of the reconstructed signal [1]. During reconstruction the spectral envelope surface is resampled at the desired time and frequency points (according to the target spectrum/pitch/duration modification). Pitch cycle waveforms are generated and combined together by the PSOLA procedure [2].

2.2. Complex spectral envelope modeling

2.2.1. Model parameters

Many speech production models represent voiced speech by a sum of quasi periodic (i.e. harmonic) and noise-like signals. Stationary sinusoidal modeling is widely used to describe the harmonic part of a windowed speech frame s_w, due to its simplicity and accuracy [7]:

$$s_w(n) \cong \hat{s}_w(n) = w(n)\sum_{k=0}^{L} A_k \sin(\theta_k n + \varphi_k), \quad (1)$$

where $w(n)$ is a symmetric window, e.g. Hamming or Hann, and θ_k is the k-th harmonic frequency which will be addressed in more detail later on. (1) is equivalent to the frequency domain equation: (2)

$$S_w(\theta) \cong \hat{S}_w(\theta) = \sum_{k=-L}^{L} C_k W(\theta - \theta_k) = \sum_{k=0}^{L} C_k W(\theta - \theta_k) + \overline{C}_k W(\theta + \theta_k),$$

where S_w is a short time spectrum, C_k is a complex amplitude of the k-th harmonic and $W(\theta)$ is the Fourier transform of $w(n)$. We define the harmonic frequency θ_k as the position of the highest local maximum found on the short time amplitude spectrum $\|S_w(\theta)\|$ in a close vicinity of $\theta_0 k$, i.e. the k-th multiple of the basic angular pitch frequency θ_0. The sequence $\{C_k\}_{k=0}^{L}$ is referred to as *line spectrum*. The line spectrum can be interpreted as a result of sampling of a continuous complex spectral envelope at the harmonic frequencies. We consider the line spectrum as a means of the complex spectral envelope parameterization. Hence, the harmonic modeling will be further referred to as complex spectral envelope extraction. Once the line spectrum estimate is obtained, it is possible to calculate the complex spectral envelope at any frequency using an appropriate interpolative model.

Computation of the line spectrum requires high-resolution pitch estimation. For this purpose we use the frequency domain pitch detector, proposed in [5].

Once the harmonic frequencies are determined, the line spectrum estimate can be obtained by minimizing a spectrum approximation error:

$$E = \sum_{m=0}^{N-1} \left\| S_w(\theta_m) - \hat{S}_w(\theta_m) \right\|^2, \theta_m = \pi m / N, \quad (2)$$

where N is half of the FFT length. The minimization is accomplished by solving an over-determined set of linear equations. It was shown in [6], that a solution of a similar problem, when using a window length of two pitch cycles is equivalent to a solution of a Toeplitz set of linear equations. Below, we'll show that it may be accurately approximated by a solution of a sparse set of linear equations, when using a long enough analysis window, compared to the FFT length. This solution is more efficient than the one proposed in [6]. Model equation (2) may be rewritten in a matrix form:

$$\hat{S}_w = \mathbf{W}_{\mathrm{Re}}\mathbf{c}_{\mathrm{Re}} + \mathbf{W}_{\mathrm{Im}}\mathbf{c}_{\mathrm{Im}}, \quad (3)$$

where $\mathbf{c}_{\mathrm{Re}}$ and $\mathbf{c}_{\mathrm{Im}}$ are respectively the real and imaginary parts of the line spectrum. $\mathbf{W}_{\mathrm{Re}}$ and $\mathbf{W}_{\mathrm{Im}}$ are matrices containing shifted replicas of the window Fourier transform as their columns:

$$\begin{cases} w_{\mathrm{Re}}(i,k) = W(\frac{i\pi}{N} - \theta_k) + W(\frac{i\pi}{N} + \theta_k) \\ w_{\mathrm{Im}}(i,k) = W(\frac{i\pi}{N} - \theta_k) - W(\frac{i\pi}{N} + \theta_k) \end{cases}, 0 \le i \le N, 0 \le k \le L. \quad (4)$$

Utilizing smooth windowing functions makes it possible to use a truncated version of the window Fourier transform. For example, setting $N = 256$ and using a 440 samples long Hamming window results in the effective window spectrum bandwidth of about 7 samples. Thus, the $\mathbf{W}_{\mathrm{Re}}$ and $\mathbf{W}_{\mathrm{Im}}$ matrices are sparse. These matrices differ from each other by only the first and last several columns. The spectral modeling error (2) in matrix notation is given by:

$$E = \left\| \mathbf{S} - \mathbf{W}_{\mathrm{Re}}\mathbf{c}_{\mathrm{Re}} - j\mathbf{W}_{\mathrm{Im}}\mathbf{c}_{\mathrm{Im}} \right\|^2 \quad (5)$$

The minimization of (5) with respect to the line spectrum $\mathbf{c}$ is accomplished by solving two sets of sparse linear equations:

$$\begin{cases} (\mathbf{W}_{\mathrm{Re}}{}^T\mathbf{W}_{\mathrm{Re}})\mathbf{c}_{\mathrm{Re}} = \mathbf{W}_{\mathrm{Re}}{}^T\mathbf{S}_{\mathrm{Re}} \\ (\mathbf{W}_{\mathrm{Im}}{}^T\mathbf{W}_{\mathrm{Im}})\mathbf{c}_{\mathrm{Im}} = \mathbf{W}_{\mathrm{Im}}{}^T\mathbf{S}_{\mathrm{Im}} \end{cases} \quad (6)$$

An efficient algorithm for solution of (6) can be found in [9].

In Figure 1 an example for amplitude spectral envelope estimation is presented. It can be observed that the envelope obtained by the proposed method (solid line) is very close to the one obtained by the STRAIGHT algorithm (dashed line) within the perceptually important frequency band whereas the proposed analysis method has significantly lower complexity.

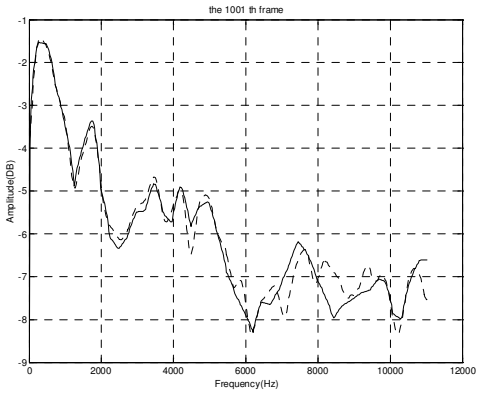

Figure 1. *Instantaneous amplitude spectral envelopes of phoneme 'i' obtained by STRAIGHT (dashed line) and by the harmonic analysis (solid line).*

The line spectrum estimation procedure described above is carried out for voiced and mixed-voiced frames only. For pure unvoiced frames the short-time Fourier transform (STFT) is used as the line spectrum estimate. The analysis procedure may be performed either at a constant frame update rate (e.g. each 5ms or 10ms) or pitch-synchronously, i.e., the analysis window includes one pitch period and is centered on each glottal closure instant (GCI). The results reported in this work have been obtained by using constant frame update rate.

In general, each voiced frame is described by the complex spectral envelope parameters $\{C_k\}_{k=0}^{L}$ and the set of quasi harmonic frequencies $\{\theta_k\}_1^L$. However we do not store

all the harmonic frequencies, but only the pitch frequency. Thus the number of the model parameters is reduced considerably. At the reconstruction step, the harmonic frequencies are generated as multiples $\tilde{\theta}_k = \theta_0 k$ of the pitch frequency. Then a random frequency dither is applied to $\tilde{\theta}_k$ in order to improve the reconstructed speech naturalness as described in section 2.2.3 below.

It is convenient for our purposes to represent the line spectrum (and the underlying complex spectral envelope) in a polar form: $C_k = A_k e^{j\varphi_k}$ and to consider harmonic amplitudes A_k and phases φ_k. When speech modification is required, the amplitudes and phases are interpolated separately, as depicted in section 3. Then a phase alignment correction is applied, as described in section 2.2.4. Finally, a short time spectrum is reconstructed according to (2), converted to time domain and overlap-added with the already reconstructed part of the speech signal.

2.2.2. Artificial High-Band phase generation

In the current speech modeling scheme High-Band (HB) phase may be artificially generated from the Narrow-Band (NB) complex spectral envelope. The proposed method is inspired by the techniques described in [8]. As a first stage of HB phase generation the complex spectral envelope is generated below a predefined cut-off frequency θ_C by using the NB harmonics, while HB harmonics are set to zero. Then the complex spectral envelope is converted to time domain. Similarly to [8] a non-linear operation such as wave rectification or sign operation is applied to the NB signal in order to generate a HB component. The modified signal is transformed back to the frequency domain and the phases, generated at the HB harmonic frequencies, are used in combination with the original HB line spectrum amplitudes. Informal listening tests demonstrated that replacement of the original harmonic phases by the artificial HB phases carried out for WB speech is unperceivable with $\theta_C = 1.1398$ [rad] which is equivalent to 4 kHz at sampling rate of 22.05 kHz. This technique makes it possible to eliminate 64% of the harmonic phase parameters, while the reconstructed speech quality is practically unaffected.

2.2.3. Harmonic model enhancement

The raw harmonic model omits the noise component of the speech, which is dominant at high frequencies of wide band speech signals. It has been found, that incorporation of a random harmonic frequency dither improves the perceptual quality of the reconstructed speech and makes it perceptually close to the original speech. The dither is applied above a threshold frequency of about 3 kHz and gradually increases towards high frequencies. Further improvement is reached when the dither parameters (i.e. the starting frequency and the growth rate) are determined dynamically as a function of a voicing degree parameter estimated at the analysis step. This parameter is computed as a scaled multiple of the spectral modeling error derived from (2) and a frequency error defined by a weighted standard deviation of the actual harmonic frequencies θ_k from the corresponding multiples $\theta_0 k$ of the basic pitch angular frequency.

2.2.4. Frame alignment at analysis and reconstruction

The line spectrum estimate reflects the waveform of a representative pitch cycle within the frame. Shifting the analysis window along the time axis results in a cyclic shift of the representative pitch period waveform. Spectral envelope modification (e.g., warping or resampling), performed during the reconstruction are not invariant with respect to the cyclic offset. It turns out that the best modification results are obtained when the original representative pitch period waveform has most of its energy concentrated near the origin. We explore two techniques that enable us to achieve this goal. The first is based on a time-domain pitch mark extraction. We define the pitch marks as the position of high peaks of the time-domain speech signal located at approximately one pitch period distance from each other. The analysis window center originally defined by a given constant inter-frame offset is then shifted so that it coincides with the closest pitch mark. A similar approach is used in a pitch synchronous analysis technique. An alternative technique was introduced in [4]. First, the complex line spectrum is estimated with a constant offset between analysis windows. Then, a linear in-frequency term of the unwrapped phase of the complex line spectrum is estimated, and subtracted from the phase (the estimation is carried out by a weighted-by-amplitudes regression). This operation is equivalent to a cyclic rotation in time domain so that most of the waveform energy is concentrated near the beginning of the representative pitch period.

The coarse alignment performed at the analysis step does not guarantee sufficient mutual alignment of the consecutive reconstructed frames, especially when morphing and/or pitch modification is involved. This is why we perform a *relative* alignment as a part of the reconstruction procedure. The relative alignment algorithm finds a linear in frequency additive phase term which maximizes the cross-correlation between the pitch cycle waveforms associated with the current and the previously reconstructed frames. Finally, an additional linear term is applied, which accounts for the constant offset between the consecutive reconstructed frames. As a result, the OLA produces a waveform which smoothly evolves along the time axis.

3. SPEECH MANIPULATIONS

With the underlying assumption that the spectral envelope is continuous in time and in frequency, basic speech manipulation procedures can be performed by spectrum interpolation and resampling in frequency (spectrum warping, pitch modification) or in time (duration modification). In the case of STRAIGHT, interpolation (during analysis) and resampling (during synthesis) are preformed for the amplitude information only. In the

proposed method the complex spectral envelope, containing both amplitude and phase information, can be interpolated and resampled. These operations are applied separately to the harmonic amplitudes and phases.

Good quality of the manipulated speech is obtained when performing the following interpolation scheme. For the amplitudes, a log-linear interpolation is used. Phase interpolation is carried out by a linear interpolation on the complex line spectrum values, followed by phase extraction.

4. EXPERIMENTAL RESULTS

The quality of reconstructed speech was evaluated. A MOS listening test was conducted with 3 sets of 10 samples: a set of 22KHz PCM records of two professional US English female speakers; a set of signals, reconstructed by STRAIGHT and a set of signals, reconstructed by the proposed system (with 5 ms update period and phase extrapolation starting from 4 kHz). 24 non-professional listeners heard the samples in random order and rated them on a 1-5 scale. The results are presented in Table 1. It can be seen that the proposed model outperforms the STRAIGHT system and gets very close to the PCM score.

Table 1. MOS results for natural and reconstructed speech.

System	PCM	Proposed	STRAIGHT
MOS	4.54	4.23	3.92

To evaluate the quality of morphed female speech the proposed model was compared with the reference STRAIGHT modeling. The speech morphing process was conducted by passing the voiced sections of the source signal through a spectral warping transformation and a pitch modification transformation as described in [3]. In practice, the two transformations were carried out as a single step, by resampling the original spectrum at $\tilde{\theta}_k = F^{-1}(\theta_{0,tar} k)$, where $F(\theta)$ is a frequency warping function, and $\theta_{0,tar}$ is the desired pitch frequency. This operation was performed on the amplitude spectrum in the case of STRAIGHT and on the complex envelope, in the case of the proposed modeling. Two different transformations were used on the two US English female speakers. In both, the warping function was a piecewise linear function with slopes varying from 0.6 to 1. The average pitch was modified by 15% in one transformation and by 30% in the other. The morphing transformations were performed on 10 sentences from each speaker. An A-B preference test was conducted with 24 non-professional listeners. The results are summarized in Table 2.

Table 2. A-B preference test for morphed speech

Pref.	No pref.	Proposed		STRAIGHT	
		Any pref.	Strong pref.	Any pref.	Strong pref.
%	37.9	34.4	6.7	27.7	5.4

The test results show preference of the proposed system based morphing over the STRAIGHT based one (10% preference among distinguishable comparison pairs). The quality improves mainly due to a better phase modeling: less metallic quality, less buzz, better performance in transients. It should be noted, however, that more listeners than expected were not quite sensitive to the artifacts caused by the usage of artificial phase.

5. SUMMARY

This paper has introduced a computationally efficient constant frame-rate sinusoidal modeling technique. It has been shown that high quality, close to transparent speech may be reconstructed from the model parameters. In spite of being more computationally efficient and having less model parameters, the proposed algorithm outperforms the STRAIGHT algorithm in both reconstructed and morphed speech quality.

11. REFERENCES

[1] Kawahara H., "Speech representation and transformation using adaptive interpolation of weighted spectrum: VOCODER revisited", in *ICASSP 97*, vol.2, pp.1303-1306 (1997.4).

[2] H. Zen, T. Toda, "An Overview of Nitech HMM-based Speech Synthesis System for Blizzard Challenge 2005", in *INTERSPEECH-2005*, 93-96.

[3] W. Hamza, R. Bakis, Z. W. Shuang, H. Zen, "On Building a Concatenative Speech Synthesis System from the Blizzard Challenge Speech Databases", in *INTERSPEECH-2005,* 97-100.

[4] D. Chazan, R. Hoory, Z. Kons, A. Sagi, S. Shechtman, A. Sorin, "Small footprint concatenative text-to-speech synthesis system using complex spectral envelope modeling", in *INTERSPEECH-2005*, 2569-2572.

[5] Chazan, D., Zibulski, M., Hoory, R. and Cohen, G. "Efficient periodicity extraction based on sine-wave representation and its application to pitch determination of speech signals", in Proc Eurospeech 2001, 2427-2430.

[6] Stylianou, Y., "Applying the harmonic plus noise model in concatenative speech synthesis", IEEE Trans. Speech and Audio Processing, vol. 9, no. 1, pp. 21-29, Jan. 2001.

[7] McAulay, R., Quatiery, T. "Speech analysis/synthesis based on a sinusoidal representation", IEEE Trans. Acoust., Speech & Signal Processing, vol. 34, no. 4, pp. 744-754, Aug. 1986.

[8] Makhoul J. and Berouti M., "High-frequency regeneration in speech coding systems", in *Proceedings of the ICASSP*, Washington, DC, pp. 428-431, 1979.

[9] Davis, T.A., "UMFPACK Version 4.0 User Guide", Dept. of Comp. and Inf. Sc. and Eng., Univ. of Florida, Gainesville, 2002.