Small Footprint Concatenative Text-to-Speech Synthesis System using Complex Spectral Envelope Modeling

Dan Chazan, Ron Hoory, Zvi Kons, Ariel Sagi, Slava Shechtman and Alexander Sorin

IBM Research Laboratory in Haifa, Israel
slava@il.ibm.com

Abstract

In this paper we present a method for speech modeling and its utilization in IBM's small footprint concatenative text-to-speech system. The method is based on frequency-domain, complex spectral envelope modeling, where the phase component plays a crucial role in attaining high quality speech synthesis. The modeling scheme presented enables low bit rate compression of the amplitude and phase information and low-complexity reconstruction of high quality speech with wide range pitch modification. Listening tests conducted for the overall text-to-speech system show a major improvement in MOS, compared to a previous, MFCC-based, system.

1. Introduction

In recent years there has been a growing demand for small footprint high quality concatenative text-to-speech (TTS) systems, a demand which comes mainly from the automotive and the mobile devices market. Typical requirements put a 10MB or smaller limit on the speech database size. Small footprint concatenative TTS requires sophisticated modeling of speech allowing deep compression, effective search of smoothly concatenating segments, as well as large range of pitch modification.

A small footprint TTS system is described in our earlier paper [1]. This system uses an MFCC (mel-frequency cepstral coefficients) speech representation. This model omits subtle details of the spectral envelope and lacks spectral phase information. Hence, the speech produced by this system sounds synthetic.

Several concatenative text-to-speech systems that use sinusoidal modeling, including phase information have been recently presented [4],[5]. However, these approaches usually have large footprint and complexity requirements.

In this paper we introduce an improved method for effective modeling of speech signals. The new parametric model is based on a harmonic sinusoidal speech representation and comprises separate harmonic amplitude and phase parameterization. It allows for low-complexity reconstruction of high quality speech with wide range pitch modification. A distance measure can be defined in the space of the model amplitude parameters, to enable searching for well matched successive segments. Such segments produce a continuous sound when concatenated and played back in a TTS system.

The paper is organized as follows. Section 2 describes the model and its parameter estimation or the *analysis* process. Section 3 describes the *synthesis* of the speech signal from the model parameters. In Section 4 the TTS listening tests results are reported.

2. Analysis process

The analysis process is based on approximation of the complex spectral envelope of the short-term (ST) speech signal. In this process, a frequency domain parametric model of the complex spectral envelope is derived from the sinusoidal representation or line spectrum [2] of the ST-signal.

The current modeling is based on the *harmonic* line-spectrum, for which the line frequencies are defined to be the fundamental frequency (pitch) multiples for voiced frames and correspond to the STFT (short time Fourier transform) frequency bins for unvoiced frames. The line spectrum, computed in the frequency (STFT) domain, is updated in a constant-frame manner (e.g. every 10 ms), rather than pitch-synchronously, as in [4],[5].

Computation of the line spectrum requires preliminary voicing classification and high-resolution pitch estimation. For these purposes we use a frequency domain peak analysis algorithm, previously reported in [3]. In this paper we deal with binary voicing decision only, though it is possible to use a continuous measure of voicing [1],[4]. We have found that incorporating a high-band random frequency dither for voiced frames, combined with careful modeling of unvoiced transients (which will be described later on) mostly compensates for the binary voicing decision and provides high quality speech synthesis from a compact set of model parameters.

The complex line spectrum for a voiced frame is obtained by the procedure referred to as *deconvolution* in the STFT domain. A vector of complex harmonic amplitudes associated with a set of line frequencies is computed so that its convolution with the Fourier transform of the windowing function best approximates the STFT in the least squares sense. This is done by solving a set of linear equations with a symmetric positive semi definite sparse matrix. The line frequencies are determined as STFT local maxima arguments, where the maxima are searched for in the vicinity of the pitch multiples. For the unvoiced frames we perform STFT and treat the result as the line spectrum.

The resulting line spectrum can be viewed as a sampling of the unknown complex spectral envelope at the line frequencies as follows:

$$H_k = S(f_k),\ \ k = 0,1,...,N-1, \qquad (1)$$

where:

N is the number of harmonics located inside the full frequency band corresponding to the sampling frequency (e.g. 0 - 11kHz for 22kHz sampling rate);

H_k are the harmonic complex amplitudes;

f_k are normalized line frequencies, (the Nyquist frequency is mapped to 0.5); for an unvoiced frame f_k=*k/LFFT* and *N=LFFT/2* where *LFFT* is the FFT length used (e.g. 512 for 22kHz sampling rate);

S(f) is the complex spectral envelope which is known at the line frequencies only.

Our goal is to determine a parametric model, approximating the complex spectral envelope $S(f),\ f \in [0,\ 0.5]$ which, when sampled at modified line frequencies (corresponding to some desired pitch value), produces naturally sounding speech. In addition, there should be a robust and computationally efficient method to estimate a spectral distance between modeled speech segments, which can be used for segment selection in the TTS system.

In order to accomplish the above requirements, we represent the complex spectral envelope in a polar form:

$$S(f) = A(f) \cdot e^{j \cdot \varphi(f)}, \tag{2}$$

where:

$$A(f) = |S(f)|, \quad \varphi(f) = \arg(S(f)), \tag{3}$$

and model the amplitude spectrum $A(f)$ and the phase spectrum $\varphi(f)$ separately. For the TTS segment selection, only spectral amplitude parameters will be used..

2.1. Amplitude spectrum modeling

Log-amplitude spectrum is modeled by a linear combination of basis functions $B_n(f),\ n = 1, 2, ..., L$:

$$\log(A(f)) = \sum_{n=1}^{L} c_n \cdot B_n(f). \tag{4}$$

Each basis function has a finite support called a frequency channel. In addition to the basis functions set, we define a frequency warping transform $\tilde{f} = F(f)$. All the frequency channels have the same width along the $\tilde{f}$ axis, and the channels corresponding to B_n and B_{n+1} half overlap each other in that scale. A well known Mel-frequency warping scale, which matches the human auditory perception, is used. Each basis function has a predefined shape. Triangular and truncated Gaussian basis functions have been tested and found to be similar in terms of the synthesized speech quality. A set of 32 basis functions allows synthesizing high quality speech at 22 kHz sampling rate. A set of four triangular basis functions defined in Mel-frequency scale is shown on Fig. 1 for illustration.

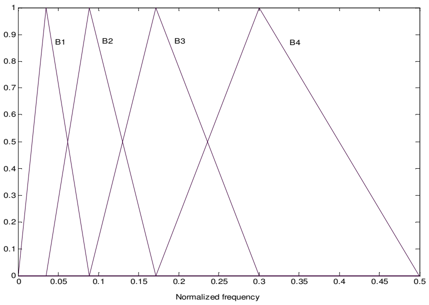

Figure1: A set of triangular basis functions for amplitude spectrum modeling

The model parameters c_n are determined by minimization of the expression:

$$\min_{\{c\}} \sum_{k=0}^{N-1} \left(\log|H_k| - \sum_{n=1}^{L} c_n \cdot B_n(f_k) \right)^2, \tag{5}$$

that is accomplished by a least squares solution of an over-determined set of linear equations. The number of parameters *L* is chosen so that even for high pitched female voice (characterized by a small number of harmonics) the number of harmonics (equations) is greater than the number of parameters (unknowns). However, the equation matrix might appear practically singular, since the frequency channel centers and the line frequencies are unevenly spaced along the transformed frequency axis. In order to overcome this problem we resample $\log|H_k|$ evenly in the warped frequency scale (e.g. Mel-scale) interpolating linearly along the warped frequency axis between the original harmonics.

Furthermore, the number of resampled harmonics may be modified to keep a predefined redundancy level of the observations. Thus, we generate *3L* harmonics which are evenly spaced in the Mel-frequency scale and use them in (5) instead of the original harmonics.

The energy *G* of the frame is combined together with the model parameters. A set of final amplitude spectrum parameters is computed as:

$$C_k = c_k - \frac{1}{L}\left(\sum_{i=1}^{L} c_i - \log G \right). \tag{6}$$

Thus, the energy is encoded by the sum of the amplitude parameters:

$$\log G = \sum_{k=1}^{L} C_k. \tag{7}$$

These amplitude parameters may be interpreted as break points in a piecewise-linear representation of the log-amplitude spectrum. Therefore, a perceptual distance measure in the TTS segment selection [1] can be computed on the parameter vectors. In the current TTS system we used a perceptually weighted distance between the *normalized* spectral parameter vectors, instead of MFCC vectors, used in [1].

2.2. Phase spectrum modeling

We model the phase spectrum only for voiced frames and for short unvoiced bursts or transient frames as discussed hereafter. For most of the unvoiced frames the phase spectrum is not modeled, and is randomly generated at the synthesis step.

2.2.1. Voiced phase spectrum modeling

Phase spectrum modeling is preceded by harmonic phases pre-processing. First we estimate and add a term that is linear in frequency, to the harmonic phases. This is equivalent to time-domain cyclic shift operation. Following [1] the linear term is computed so that the smoothness of the complex line spectrum is maximized. This step is followed by the well-known phase unwrapping operation. Finally, we compute a least squares approximation to the harmonic phases by a linear in frequency term and subtract this term from the harmonic phases.

Let $\phi_k,\ k = 0,1,...,N-1$ be the harmonic phases after the pre-processing. These harmonic phase values sample an unknown continuous phase spectrum $\varphi(f)$:

$$\phi_k = \varphi(f_k), \quad k = 0,1,...,N-1. \tag{8}$$

As was done in the log-amplitude spectrum modeling, the continuous phase spectrum is modeled by a linear combination of basis functions $P_n(f)$, *n=1, 2,…, M*:

$$\varphi(f)=\sum_{n=1}^{M} d_n \cdot P_n(f). \tag{9}$$

The basis functions are defined using a warped frequency scale $\tilde{f}=F(f)$ and have a predefined shape. The triangular functions, which were used for the amplitude spectrum modeling, provide good results for the voiced phase spectrum modeling as well. A set of sinusoidal basis functions $P_n(\tilde{f})=\sin(2\pi n\cdot\tilde{f})$ either in Mel-frequency or linear frequency $\tilde{f}=f$ also provides good results, but has been found to increase the computational complexity of the synthesis process.

The phase model parameters d_n are defined as follows. Both log-amplitudes $\log|H_k|$ and phases ϕ_k of the harmonics are re-sampled evenly in the transformed frequency scale by linear interpolation between their original values to obtain *K >> M* modified harmonics, e.g. *K=3M*. The re-sampling is done in order to guarantee the parameter estimation stability. The phase model parameters are obtained by minimization of the expression:

$$\min_{\{d\}} \sum_{k=0}^{K-1} |H_k|^{\alpha} \cdot \left(\phi_k - \sum_{n=1}^{M} d_n \cdot P_n(f_k)\right)^2, \tag{10}$$

where:

ϕ_k are the re-sampled harmonic amplitudes and phases and $\alpha > 0$ is a parameter controlling the influence of the spectral amplitudes on the phase approximation accuracy, e.g. $\alpha = 0.25$.

The minimization problem (10) is reduced to solving a set of linear equations with a symmetric positive definite matrix.

2.2.2. *Click detection and modeling*

The majority of unvoiced frames can be modeled by a Gaussian random process. The underlying speech production model is a white noise like excitation of the vocal tract generated by air flow. The vocal tract thus colors the white noise excitation by its frequency-amplitude characteristic. It follows from this model that the corresponding fragments of the speech signal are completely described by their power spectrum. Indeed, it is well known that good quality unvoiced speech can be obtained by generating a random phase spectrum [6].

However, there are some singular fragments of unvoiced speech where this model fails and phase information is essential for high quality speech synthesis. These fragments are characterized by essentially irregular excitation causing short audible bursts. Hereafter we call such fragments *clicks* and refer to other (usual) unvoiced frames as *regular* frames.

Typically, clicks appear within the sounds representing stop consonants like P, T, K, B, D and G. Sometimes clicks appear in other unvoiced sounds or even are a part of the background because of peculiarity of a specific speaker, recording equipment or environmental background. The "click-like" behavior characterizes also transient frames, classified as unvoiced and located on the voiced/unvoiced boundaries.

In any case an attempt to synthesize a click as usual unvoiced speech, i.e. using randomly generated phase, leads to smearing of its shape in time, and significantly deteriorates the auditory quality of the reconstructed speech signal.

Examples of regular unvoiced speech fragments and click-like fragments are shown on Fig. 2.

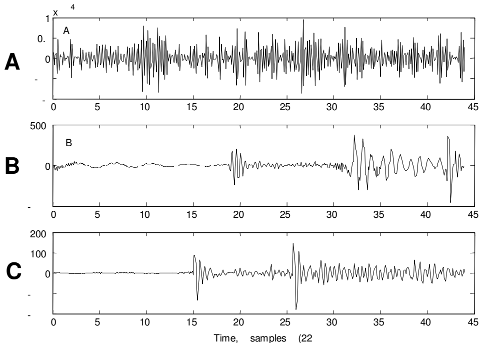

Figure 2: Regular unvoiced speech and clicks. A) regular frame, 'S' sound; B) unexpected click in a voiced-unvoiced transition; C) click in 'T'.

As illustrated by the examples of Fig. 2 the clicks might have very different waveform shape, e.g. white noise modulated by a step-wise envelope function, random impulse train, etc… What distinguishes the clicks from the regular unvoiced speech is their non-Gaussian behavior, which in particular means that the corresponding phases of the spectra contain information and therefore should be used during synthesis.

Our click detection method comprises measuring the departure of the unvoiced speech waveform within the analysis frame from a Gaussian process. Excess and entropy measures were tested, and the latter was found to produce the best detection results. It is well know that the Gaussian distribution has the highest entropy among all the distributions with the same variance.

Given an unvoiced segment of the time-domain waveform we compute a histogram of the waveform values. The values are binned into a predefined number of equally spaced containers spanning the dynamic range of the segment. The histogram is normalized by dividing each bin by the segment length. The normalized histogram gives an estimate of the discrete probability distribution function. Then the entropy estimate is computed as:

$$E=-\sum_{i=1}^{I} N_i \cdot \log_2(N_i), \tag{11}$$

where N_i are the values of the normalized histogram and *I* is the number of bins. The entropy estimate (11) is then compared against a predefined threshold. If the entropy estimate value is less than the threshold then the speech segment is declared a click. For example, the entropy estimate values obtained for the frames shown on Fig. 3 are: A – (regular) 4.04, B – (click) 2.66, C – (click) 2.57. In general the percentage of the click-frames among all the unvoiced frames does not exceed 10%.

Phase spectrum modeling is performed as described above for each click-frame. Typically click-frames have a flat amplitude spectrum. The model order (the number of basis functions) is increased beyond that of the voiced phase model because of a complex and rapidly varying nature of the click phase spectra. Thus we use 64 basis functions for 22 kHz speech.

In contrast to the voiced phase modeling we also accumulate the tangent of the linear in frequency terms subtracted from the harmonic phases during their preprocessing. This accumulated offset is an additional parameter which is stored together with the basis function coefficients and used by the synthesis process. It is done in order to prevent uncontrolled cyclical shift of the synthesized speech fragment which is acceptable for periodic voiced fragments but must be avoided in the click-like fragments, where correct waveform evolution in time is crucial.

2.3. Compression

The amplitude and phase spectra model parameters are compressed using a split vector quantization technique. The phase parameters are compressed with a variable bit rate, according to the frame categorization. Average bit-rate figures measured on a large speech database are 11kbs and 8kbs for 22kHz and 11kHz speech respectively, a bit-rate which falls in the low bit-rate coding category.

3. Synthesis process

The reconstructed speech signal is produced in a frame-by-frame fashion by synthesizing the frame spectrum, converting it by inverse FFT to time domain and overlap-adding with already synthesized part of the speech signal.

The model parameters associated with the current frame are passed to the synthesis process input together with the pitch and/or voicing decision. Note that in TTS synthesis the pitch comes from a prosody generation block (TTS front end), and can differ from the original pitch of the frame; but the original voicing decisions should be used in the synthesis process.

Normalized line frequencies are computed. For an unvoiced frame the harmonics are set to occupy each DFT point. For a voiced frame the frequencies are computed as multiples of the pitch frequency. For voiced frames, a random dither of line frequencies is introduced in the high frequency band. The high frequency harmonics are displaced from their initial locations by randomly generated shifts, whose variance gradually increases with the line frequency. This is done in order to achieve higher naturalness and pleasantness of the synthesized speech.

The harmonic amplitudes are computed by substitution of the line frequencies in formula (4), scaled to bring the frame energy to the level encoded by the model parameters and adjusted by the TTS gain control. For voiced and click frames the harmonic phases are computed by applying (9). For a regular unvoiced frame the phase values are generated randomly. A linear in frequency term (whose tangent is stored together with the phase parameters) is added to the harmonic phases of a click frame. If the current frame and the previous one are voiced then a linear phase term is computed, providing the best time-domain alignment between the frames, as proposed in [1]. Afterwards, an additional linear phase term is computed representing the frame shift in time (so that during the subsequent OLA operation the signal periods will coincide). Both linear terms are then added to the harmonic phase values.

The obtained line spectrum is convolved with the Fourier Transform of a windowing function (e.g. Hanning window), and the convolution result is sampled at the FFT points. Inverse FFT and OLA finalize the processing cycle.

4. Listening tests

A listening test was conducted with 3 sets of 25 synthesized sentences – the first using the old system (A), described in [1] the second using the new system (B), and the third with system B and an alternative speaker used for the concatenative voice database (both speakers are female speakers). A fourth set included reference natural speech samples.

The listeners were 40 native North American English, 20 male and 20 female.

Table 1: Mean Opinion Score listening test results

System	Speaker	MOS
A	1	2.29
B	1	2.90
B	2	3.14
Natural Speech	1	4.80

The tests show a significant MOS improvement, compared to the previous system – closing 24% of the gap between our system and natural speech (for speaker 1).

5. Conclusions and future work

The speech modeling approach presented in this paper has led to significant TTS quality improvement compared to the system [1]. The improvement is mostly due to the incorporation of phase information, which was lacking in the previous system. Another factor contributing to the improvement is a more accurate modeling of the amplitude spectrum.

In the future, we plan to further improve and enrich the current complex spectral envelope modeling and pitch modification scheme, in order to obtain better synthesis quality. Selective phase coding as well as difference/adaptive vector quantization techniques will be exploited to further reduce the system footprint. Furthermore, we intend to utilize this speech modeling method for high quality voice conversion.

6. References

[1] Chazan, D., Hoory, R., Kons, Z., Silberstein, D. and Sorin, A. "Reducing the footprint of the IBM trainable speech synthesis system", in *Proc ICSLP*, Denver, 2002

[2] 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.

[3] 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.

[4] 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.

[5] O'Brien, D. and Monaghan, A., "Concatenative synthesis based on a harmonic model", *IEEE Trans. Speech and Audio Processing*, vol. 9, no. 1, pp. 11-20, Jan. 2001.

[6] Deller J., Hansen J. and Proakis J., *Discrete-Time Processing of Speech Signals Processing*, IEEE Press, 1993.