N88-25752

SINUSOIDAL TRANSFORM CODING FROM 2400 to 9600 BPS

ROBERT J. McAULAY, Massachusetts Institute of Technology, Lincoln Laboratory, Speech Systems Technology, USA; THOMAS F. QUATIERI, Massachusetts Institute of Technology, Lincoln Laboratory, Speech Systems Technology, USA.

LINCOLN LABORATORY
Massachusetts Institute of Technology
244 Wood Street
Lexington, Massachusetts 02173-0073
United States

ABSTRACT

It has been shown that an analysis/synthesis system based on a sinusoidal representation of speech leads to synthetic speech that is essentially perceptually indistinguishable from the original [McAulay/Quatieri 1986]. Strategies for coding the amplitudes, frequencies and phases of the sine waves have been developed that have led to a multirate coder operating at rates from 2400 to 9600 bps. [McAulay/Quatieri 1987,1988]. The encoded speech is highly intelligible at all rates with a uniformly improving quality as the data rate is increased. A real-time fixed-point implementation has been developed using two ADSP2100 DSP chips. In this paper the methods used for coding and quantizing the sine-wave parameters for operation at the various frame rates are described.

INTRODUCTION

An analysis/synthesis system has been developed based on a sinusoidal representation of speech. The system leads to synthetic speech that is essentially perceptually indistinguishable from the original [McAulay/Quatieri 1986], offering the potential for a high-quality speech coding system. Since the parameters of the sinusoidal model are the amplitudes, frequencies and phases of the underlying sine waves, and since for a typical low-pitched speaker there can be as many as 80 sine waves in a 4 khz speech bandwidth, it is not possible to code all of the parameters directly.

In this paper the algorithms for encoding the amplitudes, frequencies and phases will be described that allow the Sinusoidal Transform Coder (STC) to operate at rates down to 2400 bps. The notable features of the resulting class of

coders are the intelligibility and naturalness of the synthetic speech, the preservation of speaker-dependent qualities so that talkers were easily recognizable, and the robustness against background acoustic noise and channel errors.

ESTIMATION OF THE PARAMETERS OF THE UNDERLYING SINE WAVES

The method for identifying the amplitude, frequencies and phases of the underlying sine waves is to first locate the peaks of the high-resolution spectrum of a frame of input speech. The latter is computed using the short-time Fourier transform applied to a speech buffer that is 2 1/2 times the average pitch period. The latter constraint is essential for correctly resolving all of the sinusoidal components. Setting the window therefore requires the computation of an average pitch period, and while any good pitch extractor could be used for this purpose, the STC system uses a new pitch estimation technique that has been derived by applying statistical decision theoretic methods to the sinusoidal model. The basic principle is to determine the fundamental frequency of a harmonic set of sine waves that provides a best fit to the measured set of sine waves. Therefore, as shown in Figure 1, the first step in the STC analysis system is to perform a high-resolution spectral analysis over the 0-1000 Hz frequency region to a speech buffer that is 2 1/2 times the lowest anticipated pitch period (25 ms). In the current implementation, a 512-point FFT is used over the entire speech bandwidth (0-4000 Hz), but computational efficiencies could be obtained by low-pass filtering, downsampling and the using a 128-point FFT over the 1000 Hz baseband region. In addition to the pitch, a voicing measure is determined and for those estimates corresponding to strongly voiced speech a short-term average pitch is computed, which is used to set the analysis window for the second analysis stage.

SINUSOIDAL ANALYSIS AND FINE-GRAINED PITCH ESTIMATION

The second analysis stage consists of another 512-point FFT over the entire 4000 Hz speech bandwidth. The magnitude of the resulting short-time Fourier transform (STFT) is examined for peaks, and now the amplitudes, frequencies and phases corresponding to the peaks are used to define the underlying speech model. The amplitudes and frequencies over the 0-1000 Hz region are supplied to the fine-grained pitch extractor which estimates the pitch using the same principles as that in the first analysis stage. Since the notion of "pitch" has no meaning in the case of unvoiced speech, it is better to think of the procedure as one of fitting a harmonic set of sine waves to the measured set of sine waves. It is a remarkable property of the STC system that this technique is capable of replicating both the voiced and unvoiced sounds as long as means are provided for properly accounting for the associated sine-wave phases. In fact it turns out that in the STC system all of the voicing information is embedded in the sine-wave phases. Preserving that information at low data rates requires suitable phase models that are tied to the voicing measure, a measure which is determined implicitly by the fine-grained pitch estimator.

AMPLITUDE AND PHASE CODING

Once the pitch has been estimated and coded (6-7 bits for pitch, 2-3 bits for voicing depending on the rate) it is used to determine a quasi-harmonic set of frequency bins which are applied to the original sine-wave set to assign one sine-wave per bin. A piece-wise linear envelope is fitted to the resulting set of amplitudes and frequencies and a piece-wise constant envelope is fitted to the resulting set of phases. If these envelopes are sampled at the harmonics of the coded pitch then very high quality, highly intelligible synthetic speech can be generated. The goal is to try to code these harmonic samples as efficiently as possible.

The first step in coding the amplitudes is to use the coded pitch to determine the number of sine-wave amplitudes that need be coded, which in turn determines the number of bits per amplitude. For a high-pitched speaker (170 Hz), all of the sine-wave amplitudes can be coded provided differential encoding is applied to the logarithmic values of the amplitudes of the neighbouring sine waves. The exact number of bits per differential is determined by the number of bits available at a given rate and the number of sine waves to be coded at that pitch. For low-pitched speakers it is not possible to encode all of the amplitudes, since for example a 50 Hz pitch corresponds to 80 sine waves which at even 1 bit per differential would require 4000 bps. which precludes operation at rates below 4800 bps. In the case of low-pitched speakers, therefore, only a certain number of the baseband sine waves are coded but then the frequency spacing between amplitude samples is increased logarithmically to exploit the critical band properties of the ear. The overall spectral level is preserved by coding the log-amplitude of the fundamental using 5 bits to cover a 90 dB dynamic range [McAulay/Quatieri, 1987]. The coding method is therefore similar to a channel vocoder [Holmes, 1980], for which the channel locations are pitch adaptive.

For rates above 4800 bps assigning more bits to the amplitudes does not lead to significantly better quality. At these rates it is better to begin to code the phases of the baseband sine waves. This is done using straightforward PCM encoding techniques with 5 bits assigned to each phase. As the rate is increased more of the baseband phases are coded to use up the remainder of the coding budget.

REGENERATION OF SINE-WAVE PARAMETERS AT THE RECEIVER

At the synthesizer (see block diagram in Figure 2), the first step is to decode the pitch and determine the bit allocation and coding tables that were used at the transmitter. If the logarithmic channel spacings were used, then an amplitude envelope is recreated which is sampled at the harmonics of the coded pitch to recover the sine-wave amplitudes. The phases are decoded and assigned to the appropriate number of baseband sine waves. For the sine waves for which no phases were coded, an artificial phase is regenerated by first phase-locking all of the sine waves to the phase of the fundamental, and then adding a random phase having a standard deviation that is determined by a nonlinear mapping of the the voicing measure.

At this stage a set of amplitudes, frequencies and phases have been regenerated at the coding frame rate, which typically is 50 Hz. for most applications.

COMPUTATIONALLY EFFICIENT SINE-WAVE SYNTHESIS

In the basic sinusoidal analysis/synthesis system speech reconstruction is done by matching the sine-wave parameters obtained on successive frames and applying linear interpolation to the amplitudes and cubic phase interpolation to account for the interaction of the phases and frequencies. While this model leads to high-quality synthetic speech, it is computationally expensive since it requires that every sine-wave be regenerated on a per sample basis. Studies have shown that this process itself could require as many as 3 ADSP2100 chips for 4000 Hz bandwidth speech. In order to achieve computational efficiency, the sine waves are most easily recreated using the inverse FFT overlap-add technique. Since the low-rate coders operate at a 50 Hz frame rate, the overlapping triangular windows are 40 ms wide and this leads to a roughness in the synthesized speech. This is due to the fact that the sine-wave parameters, which correspond to the vocal tract articulators, do not remain stationary for such a long interval of time. In fact in the development of the sine-wave model [McAulay/Quatieri 1986], it was found that the overlap-add method worked quite well provided the frame-rate was of the order of 100 Hz. In order to exploit the FFT method therefore, it is necessary to use the sine-wave parameters received every 20 ms to generate an interpolated set of parameters every 10 ms. This can be done by using the fact that the birth/death frequency matcher has established contiguous sets of sine-wave parameters at the 50 Hz rate. The amplitudes and frequencies can then be interpolated linearly, and the phases are then interpolated using a rule that requires that the mid- point phase lead to a sine wave that is a best fit to the trailing edge of the preceding sine wave and to the leading edge of the succeeding sine wave. With this new set of sine-wave parameters, the inverse FFT overlap-add method can then be applied at an effective 100 Hz frame rate, and as a consequence the roughness is eliminated, [McAulay/Quatieri 1988].

One of the fundamental objections to the STC system has been the fact that for speech in a very noisy background, the harmonic frequency coding has led to a synthetic background noise that has had a tonal quality. Although the coded noisy speech has been very intelligible, it has reduced the quality of the system since most listeners find the tonality of the synthetic noise to be very objectionable. To eliminate this effect, the receiver uses the voicing measure to up-date an average spectrum for the background noise during strongly unvoiced frames. This spectrum is then used to first suppress the harmonic representation of the noise [McAulay/Malpass 1980], and replace it with a wideband noise having the same spectral shape but with completely random amplitudes and phases, thereby totally eliminating the tonality. Moreover, since the randomization is done to the FFT coefficients before the inversion, there is little increase in computational complexity.

CONCLUSIONS AND DISCUSSION

The sinusoidal model has proven to be an effective parametric technique for the analysis and synthesis of speech. It has been used for modifying the time-scale and pitch- scale of speech [Quatieri/McAulay 1986] and for the enhancement of speech in AM radio broadcasting [Quatieri/McAulay 1988]. This paper has summarized the attempt to use the sinusoidal model for low-rate speech coding. The effort has led to a multi-rate speech coder that can operate effectively from 2400 to 9600 bps producing very intelligible speech with a quality that improves more or less uniformly as the data rate is increased. The coder is very robust in acoustic background noise and with 5 bits assigned for modest bit error detection and correction it can operate effectively over a channel with at least 1% bit errors. Furthermore, as a result of timing studies of the real-time fixed-point realization, computational choke-points have been located and the algorithm modified so that now the entire STC system has been realized using 2 ADSP2100 chips. The current effort is now focussed on developing more efficient coding strategies for improving the quality at 2400 bps and for operating the coder at even lower rates, perhaps down to 1200 bps.

ACKNOWLEDGEMENT

The authors would like to thank Marilyn Malpass of Lincoln Laboratory for her work on the architecture studies for the ADSP2100 implemetation of the STC system as this led to the identification of the computational choke-points. In addition all of the error protection work was her own. They would also like to acknowledge the interesting discussions with Dr. Michael Sabin of CYLINK, who has independently developed an implementation using two of the TMS320C25 DSP chips.

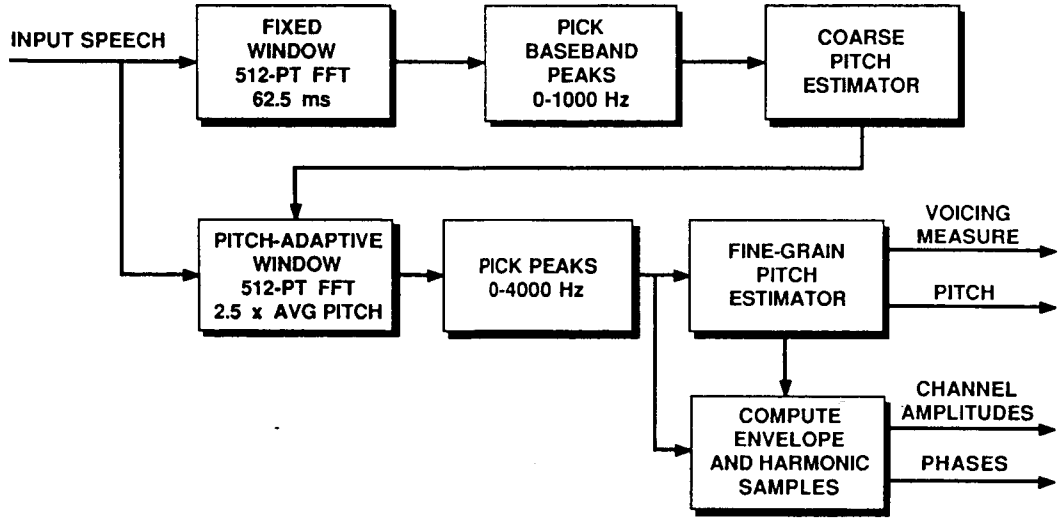

Fig. 1. STC Analysis.

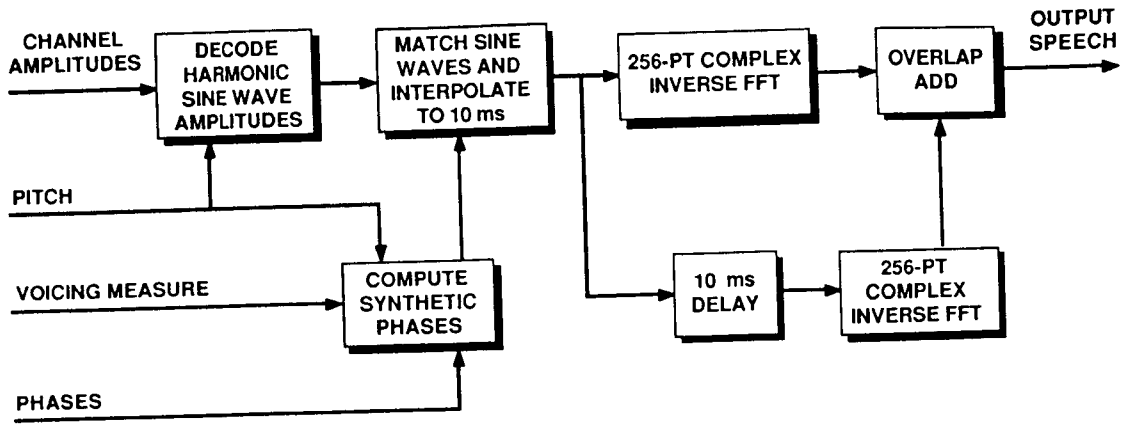

Fig. 2. STC Synthesis.

REFERENCES

McAulay, R.J., Quatieri, T.F. 1988.
Computationally efficient sine-wave synthesis and its application to sinusoidal transform coding. International Conference on Acoustics, Speech and Signal Processing, ICASSP '88, New York.

Quatieri, T.F. and McAulay, R.J. 1988.
Sinewave-based phase dispersion for audio preprocessing. International Conference on Acoustics, Speech and Signal Processing, ICASSP '88, New York, N.Y.

McAulay, R.J., Quatieri, T.F. 1987.
Multirate sinusoidal transform coding at rates from 2.4 Kbps. International Conference on Acoustics, Speech and Signal Processing, ICASSP '87, Dallas, TX.

McAulay, R.J., Quatieri, T.F. 1986.
Phase modelling and its application to sinusoidal transform coding. International Conference on Acoustics, Speech and Signal Processing, ICASSP '86, Tokyo, Japan.

McAulay, R.J., Quatieri, T.F. 1986.
Speech analysis/synthesis based on a sinusoidal representation. IEEE Trans. Acoustics, Speech and Signal Processing, ASSP-34, No. 4, pp. 744-754.

Quatieri, T.F. and McAulay, R.J. 1986.
Speech transformations based on a sinusoidal representation. IEEE Trans. Acoustics, Speech and Signal Processing, ASSP-34, No. 6, pp. 1449-1464.

McAulay R.J. and Malpass, M.L. 1980.
Speech enhancement using a soft-decision noise suppression filter. IEEE Trans. Acoustics, Speech and Signal Processing, ASSP-28, No. 2, pp. 137-145.

Holmes, J.N. 1980. The JSRU channel vocoder. In IEEE Proceedings Vol. 127, Pt. F, No. 1, pp. 53-60.