Low Bit-Rate Speech Coder Based on a Long-Term Model

Orit Lev (Fellah) and David Malah
Department of Electrical Engineering
Technion, Israel Institute of Technology

1. Introduction

Low Bit Rate Speech Coders, for transmission at rates below 4Kbps, have received much attention in recent years. Existing low bit-rate speech coders, such as LPC10, CELP, MELP and WI (Waveform Interpolation [1]), are based on models that can adequately represent only short speech segment and therefore use short analysis frames (20-40ms).

In this paper, we present a low bit rate speech coder based on a Long-Term Model (LTM) for voiced speech, proposed in [2]-[4], and on the WI coder. In the LTM, a periodic input signal undergoes a time-varying spectral shaping representing the evolution of the pitch-cycle waveform. The resulting signal, which has a fixed pitch period but a time-varying pitch-cycle waveform, is multiplied by a time-varying gain function that represents the variation in speech loudness. The resulting signal then undergoes a time-axis warping, which represents the evolution of the pitch period, yielding the output speech signal. Since the LTM allows evolution of the pitch period and the pitch-cycle waveform in the analysis frame, it facilitates long analysis frames (100-160msec). The spectral shaping in the proposed coder is based on WI. In WI, speech (or LPC residual) is observed as a continuously evolving sequence of pitch cycle waveforms. A subset of these waveforms is extracted and coded. In the decoder, after inverse quantization, missing waveforms are synthesized by interpolation. The extracted waveforms are normalized to a fixed length and sequentially aligned using a cyclical shift. Then, a two-dimensional surface, called Prototype Waveform Surface or Characteristic Waveform (CW) $u(t,\phi)$ is produced from these waveforms, where t represents time and ϕ represents the phase (location) in the waveform. The CW is transformed to the frequency domain, using DFT along the ϕ-axis, to benefit from the different perceptual significance of the magnitude and phase. The CW is then Low-Pass filtered along the t-axis, to produce the SEW (Slowly Evolving Waveform) surface which contains the voiced component of the speech. The complementary High-Pass (unvoiced) component is called REW (Rapidly Evolving Waveform). The SEW magnitude is coded at high spectral resolution and low temporal resolution. The REW magnitude is coded at low spectral resolution and high temporal resolution. Typical WI coders [1] don't code the phase, and use a random phase for REW and fixed phase for SEW. Recently, Gotessman et. al. [6],[7] proposed to code the SEW phase by VQ, using analysis-by-synthesis, where the selected phase minimizes the distortion between the SEW, and the combination of the reconstructed SEW magnitude and the candidate phase. The proposed coder, described next, is a modification of the WI coder, taking advantage of the LTM for voiced frames. For unvoiced frames, standard LPC-10 coding is used. The average bit rate of the proposed coder is 2 Kbps.

2. Proposed Coder

The proposed coder first performs adaptive segmentation of the speech signal into 64-160ms voiced frames, 32ms unvoiced segments and 32ms low-energy (silence) segments [5]. The segmentation is based on energy level, signal periodicity, zero crossings count and on the ratio between energy at low frequencies and the overall energy. Unvoiced segments are coded by LPC-10 techniques. 'Silence' segments are represented by low-level white Gaussian noise ('comfort noise') generated at the decoder, so that only their energy level needs to be coded and transmitted. The segmentation attempts to produce long voiced frames for efficient coding. Voiced frames lengths are determined by limiting the Log Spectral Distance (LSD) between sections at the beginning and end of the frame. This limitation restricts the change in the pitch-cycle waveform in the frame, keeping the coding error low. The lengths are also restricted by delay constraints of the coder. A block diagram of the voiced-frames coder is described in Fig. 1.

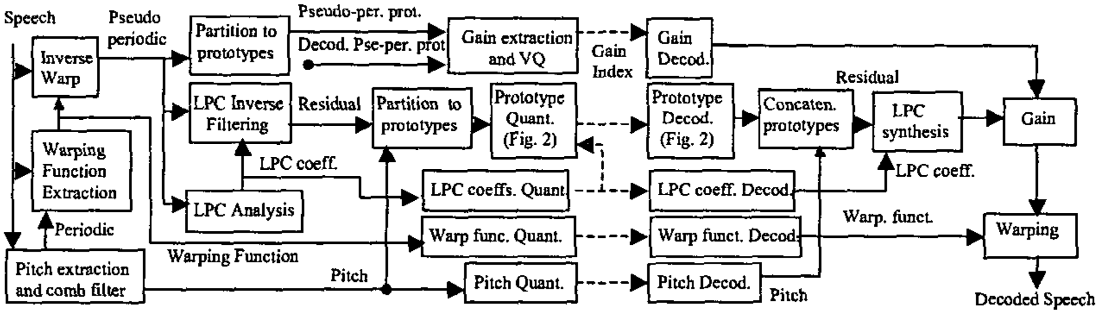

Fig. 1: Voiced-frames coder. Left- Analysis, right- Synthesis

First, the inverse time-warping function is estimated as in [2], by calculating a piecewise linear mapping from a periodic signal (produced by a comb filter) to the speech signal. The warping function is coded by quantizing its initial value and slopes. The warping function is used to convert the input speech frame into a pseudo-periodic signal with a constant pitch period but a time-varying pitch-cycle waveform. LPC analysis is performed on the pseudo periodic signal yielding a residual signal. The CW is generated from the residual signal in a simpler way than in the WI case, since the residual signal here has a constant period. The waveforms (prototypes) are extracted at a rate of one prototype per pitch cycle, by simply partitioning the residual signal according to the known pitch period, and the CW is generated by arranging the equal-length prototypes in a two dimensional array. Prototype waveform alignment is thus not needed. The LPC coefficients are transformed to LSF (Line Spectral Frequencies) and quantized. The prototypes gain is extracted using analysis-by-synthesis, by comparing the pseudo-periodic prototypes to the pseudo-periodic decoded prototypes, and coded by a 10 bit VQ. In the synthesis, the LPC residual signal is generated by simply concatenating the equal-length reconstructed prototypes. The residual signal undergoes LPC synthesis, gain and time warping to produce the decoded speech signal. The coding of the prototype waveform surface, which is the heart of the coder, is described in Fig. 2.

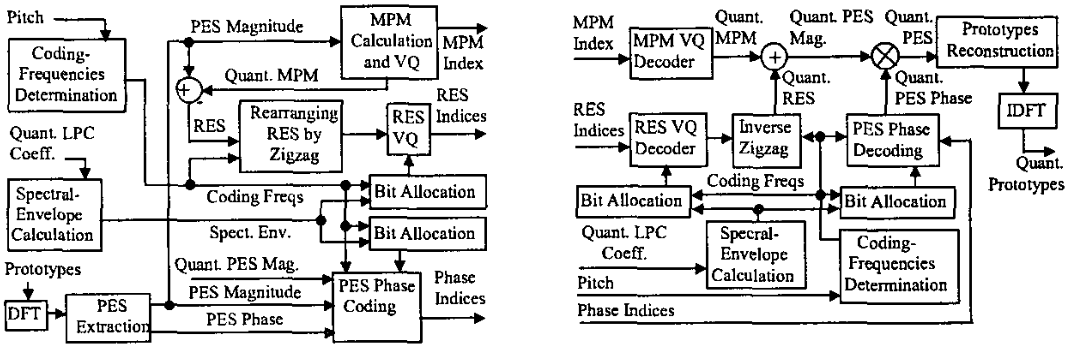

Fig. 2: CW Surface coder for voiced frames. Left- Analysis, right- Synthesis

To efficiently code the CW we propose a technique by which the prototype waveform spectrum evolution along the time axis is coded instead of the WI approach of coding the spectrum of each prototype along the phase axis. This is not possible in WI because of its much shorter analysis frame. The CW is transformed to the frequency domain by DFT along the ϕ-axis, as in WI. For each frequency, the shape obtained when moving on the surface along the t-axis represents the time evolution of the prototype spectra at this frequency. These shapes are denoted PES (Prototype Evolution Shapes). Since the PES are generally smoother than the prototype spectra, they are simpler to code, saving bits. Further bit reduction is achieved by assigning the bits for PES coding according to their significance to speech perception. Initially, there is a PES vector for each frequency. Then, a subset of frequencies used for coding, denoted Coding Frequencies (CF) is selected, as described in the next section. Following that, the spectral envelope is calculated from the LPC coefficients (after quantization and reconstruction). The assignment of bits to the CF is done as in [8], using the spectral envelope as the weighting function, such that frequency having a higher spectral envelope value is assigned more bits. The PES is decomposed into magnitude and phase surfaces, and a different bit assignment is used for each surface. The PES magnitude surface is decomposed into a RES (Rapidly Evolving Shapes) surface and a MPM (Mean Prototype Magnitude) vector. The MPM is the vector of PES means. It is coded using a 10-bit VQ. Since the MPM dimension is pitch dependent, 4 codebooks are used for 4 representative dimensions (18,27,33,50). The appropriate codebook is selected according to the pitch value. The reconstructed MPM is subtracted from the PES, yielding the RES surface. The RES are rearranged using zigzag scanning (ZS) as described in the next section. The rearranged RES are coded using a VQ tree. 4 VQ trees are used for 4 representative RES dimensions (14,20,24,40). The appropriate VQ tree is selected according to the dimension of the rearranged RES and the appropriate codebook in the VQ tree is selected according to the bit assignment. The PES phase is coded by a method similar to SEW phase coding in [6],[7]. The PES are also rearranged by ZS. The phase is coded using a VQ tree. For each coding frequency, the representative phase vector is the vector that minimizes, when combined with the rearranged PES magnitude, the MSE with respect to the rearranged PES. The final coded phase results from applying an inverse ZS to the representative phase vectors. As in the RES magnitude case, 4 VQ trees are used for 4 representative dimensions (14,20,24,40).

3. Coding Frequencies Selection and Zigzag Scanning

Zigzag scanning (ZS) is shown below to overcome the problem of different RES surface dimensions for speakers with different pitch frequencies. The problem is described in Fig. 3. Fig. 3a shows the RES surface for a female's speech frame representing a speaker with high pitch frequency. The number of frequencies is small, due to the short pitch period, and therefore the number of prototypes is large resulting in a large RES dimension. Fig. 3b shows the RES surface for an equal length male's speech frame representing a speaker with low pitch frequency. The number of frequencies is large, due to the long pitch period, and therefore the number of prototypes is small, resulting in a small RES dimension. Using VQ, high pitch frequency speakers require a small number of accesses to a large dimension VQ. For a given total bit allocation, a large number of bits will be assigned to each codebook, resulting in efficient coding. Low pitch frequency speakers, however, require large number of accesses to a small dimension VQ. For the same total bit allocation, a small number of bits will be assigned to each codebook resulting in codebooks that are not rich enough.

In this case it is better to combine several RES together and receive longer vectors. This will result in a small number of large dimension vectors and in efficient coding as in the case of high pitch frequency speakers. The proposed ZS method is based on the above reasoning and thus combines neighboring RES to obtain a fixed number of longer shapes. The method is described in Fig. 4.

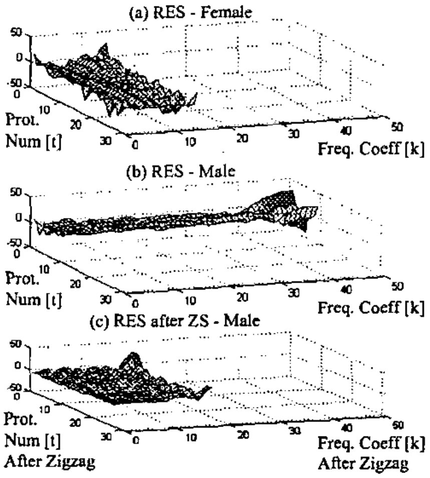

Fig. 3: RES surfaces

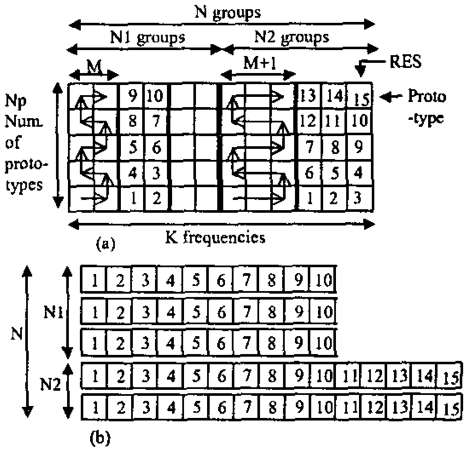

Fig. 4: Zigzag Scanning method: (a) Original RES. (b) Rearranged RES.

The K RES, corresponding to K frequencies, are rearranged to form N new shapes. N1 new shapes are created by combining M RES, and N2 new shapes are created by combining M+1 RES. Each of the N coding frequencies is the lowest frequency in the corresponding RES group. The combination of neighboring RES is done by ZS as shown in Fig. 4. The idea behind scanning in a zigzag is that every step causes only a small change in the RES surface value. Horizontal changes are small due to similar prototype magnitude values at adjacent frequencies. Vertical changes are small since adjacent prototypes have similar values at a given frequency, since adjacent prototypes are similar in voiced speech. Fig. 3c shows the male RES surface after they were rearrangement by ZS. As can be seen, the RES surface dimensions are now similar for the two demonstrated speakers.

5. Results

The number of bits for the voiced-frames coder, assuming a mean voiced frame of 768 samples (6 sections of 128 samples) is 186. The bit assignment is: Warping function- 31 (offset-7, slopes- 6x4), LPC coefficients- 24 (LSF split-VQ), RES- 60 (VQ tree, 0,...,6 bits), MPM- 10 (VQ), PES Phase- 40 (VQ tree, 0,...,6 bits), Gain- 10 (VQ), average pitch- 7 (scalar quantizer), and framing decision- 4 (Frame type-1, Voiced frame length-3). The effective length of the frame is 708 samples because of a 64 samples overlap between adjacent speech frames. The resulting rate of the voiced-frames coder is 2.11 Kbps. The unvoiced frames are coded by LPC-10 at 2.4Kbps and the 'silence' frames are coded at 0.25Kbps (8 bits to code energy in a 32ms frame). To estimate the mean overall rate, we assume that 65% of speech duration is voiced, 25% unvoiced, and 10% silence, resulting in overall rate of 2Kbps, on average. For female speech, the quality of the decoded speech is good. For male speech, the decoded speech has some hoarseness at this rate, so that some further work is needed.

References

[1] W. B. Kleijn, J. Haagen, "Waveform Interpolation for Coding and Synthesis", Chapter 5 in W. B. Kleijn and K. K. Paliwal Eds., *Speech Coding and Synthesis*", Elsevier Science B.V., 1995.

[2] Y. Stettiner, "Long-Term Model For Voiced Speech with Application to Co-Channel Separation", PhD Dissertation, Technion, 1995.

[3] Y. Stettiner, D. Malah and D. Chazan, "Estimation of the Parameters of a Long Term Model for Accurate Representation of Voiced Speech", Proc. IEEE, ICASSP-1993, pp. 534-537, 1993.

[4] Y. Stettiner, D. Malah and D. Chazan, "Dynamic Time Warping with Path Control and Non-Local Cost", Proc. IEEE, Intl. Conf. on Pattern Recognition, ICPR, 1994.

[5] O. Fellah, "Low Bit-Rate Speech Coding Based on a Long-Term Model", M.Sc, thesis, Technion, 2000 [in Hebrew].

[6] O. Gottesman, "Dispersion Phase Vector Quantization For Enhancement of Waveform Interpolation Coder", ICASSP-1999, Vol. 1, pp. 269-272, 1999.

[7] O. Gottesman and A. Gersho, "Enhanced Waveform Interpolative Coding at Low Bit-Rate", IEEE Trans. On Acoustics, Speech, and Signal Processing, Vol. 9, No. 8, pp. 786-798, 2001.

[8] R. V. Cox, R. E. Crochiere, "Real-Time Simulation of Adaptive Transform Coding", IEEE Trans. on Acoustics, Speech, and Signal Processing, Vol. ASSP-29, No. 2, pp. 147-154, 1981.