Speech coding at low and very low bit rates

Geneviève Baudoin
Signal processing and Telecommunications department ESIEE
BP 99 93162 Noisy Le Grand CEDEX FRANCE
baudoing@esiee.fr

Abstract

In this paper, a review of the main algorithms for speech coding at low and very low bit rates from typically 50 bps to less than 4000bps is done. Then a new segmental method with automatically derived units for very low bit rate coding is presented.

1 Introduction

In classical wired telephone systems speech is digitized at 64 Kbps. Many algorithms [2, 3] have been proposed to reduce this bit rate. Three different ranges are classically distinguished:
- High bit rates above 16 Kbps, corresponding to waveform coding algorithms not specific to speech,
- Medium bit rates from 4 Kbps to 16 Kbps, corresponding to hybrid waveforms coding techniques taking into account the specific features of speech or of human auditive perception. the main representant of this class is the CELP method [?].
- Low and very low bit rates from $\sim$ 50 bps to 4 Kbps, corresponding to vocoders (Voice coders).

This paper will focuss on the last category of low and very low bit rates coders.

The theoretical minimum bit rate for a speech coder is about 60 bps, corresponding to $\sim$ 50 phonemes and an average phone rate of 10 phones/s.

Applications of low bit rate speech coding include increasing of the capacity of telephone networks, secure telephony with encryption, pagers, internet telephony, answering machines, HF communications, and adaptive multi-rate communications where the source coding and the channel codings are adapted to the quality of the channel.

The evaluation of the low and very low bit rate coders is done through subjective tests such as mean opinion score (MOS) and through understanding tests such diagnostic Rythme Tests (DRT) or Diagnostic acceptability tests (DAM) under different condition of ambiant noise or channel error rates.

A speech coding system is made of a coder (analysis) and a decoder (synthesis). The coder analyses the speech signal to extract a reduced set of pertinent parameters that are coded on a limited number of bits. The decoder uses these parameters to reconstruct a synthetic speech.

Most of the algorithms make use of a simplified linear model of speech production. The speech signal is modelised by the "excitation-filter model" as the output of a linear all pole filter (synthesis filter) the tranfer function of which represents the spectral envelope (formants), excited by an input signal accounting for the fine structure of the speech spectra. Speech coders segment the signal in quasi-stationnary frames of $\sim$ 20ms. On each fame, it extracts parameters representing the spectral envelope and parameters characterizing the excitation such as energy, voicing and fundamental frequency F_0.

2 Low bit rate speech coders, vocoders

For low bit rate speech coding, typically from 800 bps to 4000 bps, it is no longer possible to use waveform coding. The coder must eliminate perceptually irrelevant information.

2.1 Classical 2-states vocoders

In the 1^{rst} vocoders (channel, formant or LPC vocoders) the different frames of the speech signal were classified as voiced (V) or unvoiced (UV). These classical vocoders applied the excitation-filter model. The synthesis of the decoded speech used a reconstructed excitation signal made of white noise for UV frames and of a periodical pulse signal at F_0 for voiced frames.

In channel vocoders introduced by Dudley in 1939 The analysis evaluates the energy, the voicing, the fundamental frequency F_0 and the relative power of the signal in different adjacent frequency bands (typically 10 bands). In the decoder, speech is synthesized by passing a synthetic excitation signal through a bank of bandpass filters, the output of these filters being scaled to fit the relative powers of the original signal. Outputs of the filters are summed and the result is scaled to fit the original energy.

In formant vocoders, the analysis determines the positions, amplitudes and bandwidth of the 3 first formants, as well as the energy of the frame, the voicing and F_0. In the decoder, the excitation signal is filtered by 3 tuned filters at the formant frequencies. The resulting synthetic speech signal is scaled to the frame energy. This technique gives a comprehensible speech at 1200 bps, but formant estimation is a very difficult task.

In Linear Prediction LPC coders [5], the spectral envelope of speech is estimated by the amplitude transfer function of an all pole filter $(1/A(z))$. A linear prediction of the speech signal is applied to obtain the coefficients a_i of the filter. the number of coefficients is between 8 and 16. The coder calculates and transmits spectral coefficients, energy, voicing and F_0. The synthetic speech is generated by filtering the reconstructed excitation by the synthesis filter $1/A(z)$ and scaling the ouput to fit the original energy.

In these 3 coders, excitation is too succinctly represented. In a 2400 bps coder, there are approximately 2000 bits/s to code the spectral envelope and only 500 bps for the excitation. The 2-state approach is not adapted to mixed sounds (such as voiced fricatives). It cannot represent sounds that are voiced until a frequency f_{max} and unvoiced after. Plosive sounds cannot be correctly modelled by white noise. The decoded signal sounds noisy, lacks of clarty and presents tonal distortions.

2.2 Recent Vocoder algorithms

The new algorithms have in common a better representation of the voiced parts of speech and of the evolution of voicing parameters. Spectral parameters are efficiently coded by vector quantization.

MBE MultiBand Excited Coders

In MBE coders [6], speech is analysed in many frequency bands and is declared voiced or unvoiced in each band. The number of bands is of the order of the number of fundamental harmonics. The spectral envelope $H(f)$ and the fine structure $E(f)$ of the Short Term Fourier Transform $X(f)$ of the frame signal are separately approximated by $\hat{E}(f)$ and $\hat{H}(f)$. The synthetic signal $\hat{x}_n$ is obtained in the frequency domain by $\hat{X}(f) = \hat{E}(f)\hat{H}(f)$. The transmitted parameters are F_0, voicing information for each band, and parameters describing the spectral envelope.

Improved MBE coding (IMBE) have been standarding at 4150 Kbps for the Inmarsat-M system.

STC Sinusoïdal Transform Coders

McAulay and Quatieri [7] with many others have developed coders based on sinusoïdal models. Speech is modeled by a sum of sinusoïds with time varying amplitudes, frequencies and phases. For voiced part of speech, the frequencies are related to the harmonics of the fundamentals frequency and evolve slowly with time. The peaks of the Short Term fourier Transform can be used to determine the parameters of the sinusoïds. For UV speech, the frequencies are uniformely spaced.

A multirate sinusoïdal coder was developed at MIT Lincoln Labs [7] with rates from 1.8 to 8 Kbps. For lower bit rates the phase information is not transmited.

Prototype Waveform Interpolation WI coders

In WI coders [9], spectral parameters are obtained by linear prediction. The residual signal is calculated. The pitch period is estimated. Then a characteristic waveform (CW) is extracted from the residual signal at regular intervals (typically at a rate of 480 Hz). The length of the CW, for voiced speech, is equal to a pitch period $p(t_m)$ at the calculation time t_m and is arbitrary for unvoiced sounds. The CW calculated at time t_m is normalized to a length of 2π, giving the CW $u(t_m, \tau)$. At each time instant t is associated a periodical waveform $u(t, \tau)$ of period 2π, represented by its Fourier series coefficients. and obtained by linear interpolation between 2 successive CW at times t_m and t_{m+1}. The denormalized length $p(t)$ of this waveform is obtained by linear interpolation on the pitch $p(t_m)$. The excitation signal can be obtained from the interpolated waveforms $u(t, \tau)$. For voiced speech, the CW evolves slowly while for unvoiced speech it evolves rapidly. These 2 components of speech are separated by filtering by a high-pass and a low-pass filter applied to $u(t, \Phi)$ along the t axis. The two 2-dimensional resulting components are quantized separately. It is perceptually unrelevant to code precisely the REW component. A rough representation of the amplitude spectrum is sufficient. But as it evolves rapidly it has to be transmitted at a sufficient rate (typically 240 Hz). On the contrary the SEW component has to be coded precisely, but transmission of the CW can be done at a low rate (40 Hz). The quantization of the SEW is done on its Fourier series coefficients by vector quantization.

In the synthesizer, the different parameters are linearly interpolated. The 2 components SEW and REW are reconstructed from Fourier coefficients. The total reconstructed excitation $et(t)$ is obtained by adding the Fourier coefficients of reconstructed REW and SEW. It is filtered by the LPC synthesis filter and at last by enhancement post-filter.

A 2400 bps WI coder has been realized [9]. It gives better perceptual results as the 4800 bps FS1016 standard.

MELP Mixed Excitation Linear Prediction Coders

The recent DOD standard at 2400 bps [10] is a MELP coder. It presents a good communication quality. It uses a multiband mixing model with frequency dependent voicing strengths. The mixed excitation is modelled as the sum of a noisy and a pulse component.

The transmited parameters are Pitch, aperiodic flag, 10 first Fourier magnitudes of voiced residuals, bandpass vocing strengths, vector quantized spectral coefficients and 2 gains.

5 voicing strengths (Quantized to 0 or 1) are determined by analysing the signal on 5 frequency bands. The aperiodic flag is set by this analysis to indicate that the pulse component of he excitation should be aperiodic rather than periodic. The analysis also calculates the magnitudes of the first 10 pitch harmonics of the residual for voiced speech. These magnitudes are vector quantized.

In the synthesis, all the MELP parameters are interpolated pitch-synchronously. The pulse component of the excitation is calculated by inverse Fourier transform of one pitch period in length applied on Fourier magnitudes. For UV sounds or when the aperiodic flag is set, a jitter is added to the pitch value. The pulse and noise components are filtered and added. The pulse filter is determined by the sum of all band-pass coefficients for the voiced frequency bands and the noise filter is determined in the same way for the unvoiced bands. Excitation is filtered by an adaptive spectral enhancement filter and by the LPC synthesis filter. Then it is scaled and filtered by a pulse dispersion filter to spread the excitation energy on a pitch period.

HSX Harmonis Stochastic coders

Harmonic Stochastic coding [8] is quite similar to MELP. The synthetic excitation is the sum of an harmonic plus a stochastic component. Typically, excitation is harmonic until a certain frequency and stochastic after. The reconstructed excitation is filtered by the LPC synthesis filter and scaled to the frame energy. A post-filter is applied to enhance formants.

The analysis determines pitch, energy, LPC parameters and the level of voicing in 4 subbands, the level of voicing being forced to be a decreasing function.

3 Very low bit rate speech coder

For coding at bit rate under a few hundred of bps, it is no longer possible to work frame by frame, a variable length segmental approach is necessary [4, 12, 11, 1, 13].

Very low bit rate coders use a speech recognizer in the coder and a symbol to speech synthesizer in the decoder. The coder does a symbolic transcription of the speech using a codebook of elementary variable length units that can be phonetic units (such as phonemes or syllables) or automatically derived data driven acoustics units. For each segment, the coder transmits the symbol corresponding to the recognized unit plus auxiliairy parameters such as fundamental frequency, energy, length of the unit. Generally the synthesis is done by concatenation of representants of the elementary units.

The bit rate to code the sequence of recognized units is between 50 and 150 bps The bit rate for coding of auxiliairy parameters is of the same order.

Phonetic vocoders require the phonetic transcription of the training data base which is a heavy human task proned to errors and that has to be repeated for each new language. So the automatic determination of characteristic speech units based uniquely on raw speech data is an interesting approach. We will now present a new segmental vocoder based on a set of automatically derived data driven units.

4 A new segmental vocoder based on data-driven units

The above mentioned problems led us to define such a set of units *automatically* in an unsupervized mode. For this purpose, we dispose of number of tools which proved their efficacity in automatic speech processing: temporal decomposition (TD), non-supervised clustering, multigrams (MGS), Hidden Markov Models (HMM) and others.

The set of basic units was initialized using the Temporal Decomposition of linear prediction cepstral coefficients, vector quantization (VQ) and eventually multigrams (MG). It was further refined by HMMs.

The Temporal Decomposition (Atal) detects quasi-stationnary parts in the parametric representation of speech. It approximates the trajectories of the speech parameters as a succession of target values and interpolation functions. This 1^{rst} stage determines the boundaries of segments which are then clustered by Vector Quantization (size 64).

HMM are used in the training step to refine segments in the dictionnary, to model them and to detect these segments in the input speech. HMM parameters are initialized using context free and context dependent Baum-Welch training with TD+VQ or TD+VQ+MG transcriptions. The refinement consists of iterative steps of corpus segmentation (using previously trained HMMs), and model parameters re-estimation of those models using the new segmentations and labellings. Refinement improves the coherence of models with data (increasing likelihood) and

the perceptual coherence of acoustic segments in the different classes.

Multigrams (MG) may serve for finding characteristic sequences of quantized TD events or of segments determined by HMM. The method is based on finding optimal segmentation of symbol string into variable length sequences (multigrams) by maximizing the likelihood of segmentation for a given multigrams codebook.

Figure 1 represents the different steps of the derivation of the speech units.

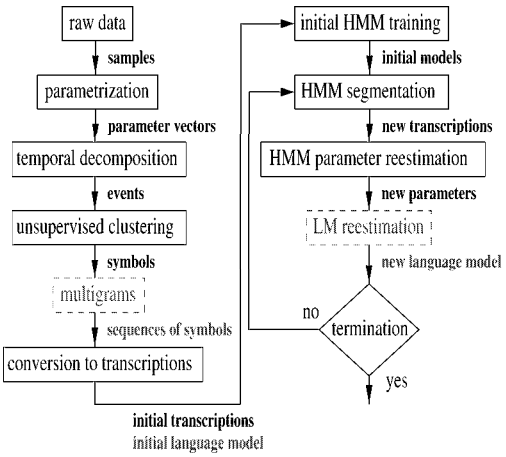

Figure 1: Data-driven derivation of speech units set

The stochastic models of the units are used by the coder to segment the input speech and to label each segment.

For each recognized segment the coder determines a synthesis unit. For this task, the 8 longest representatives per coding unit are selected in the training corpus. The input segment is compared to these representatives by DTW and the information about the choosen representative is transmitted. The transmitted parameters are the index of the coding unit, pitch and energy contours, timing and index of synthesis representative.

The synthesis is for the moment quite primitive. It is done using the selected segments of the training corpus and a LPC synthesizer. Smoothing techniques and better synthesis (PSOLA, HNM) should improve the quality of the coder. Another issue is the encoding of prosody which is not yet done in our work, and should be done on a segmental basis.

Experimental results are so far available for single speaker coding of Swiss French (PolyVar corpus), American English (BU Boston university radio Speech corpus) and Czech (Martin Ruźěk corpus) languages.

At mean rate for coding of units of ~120 bps, for the 3 corpus, we obtained mostly intelligible speech proving the potential of those automatically derived units in speech transcription and coding. Speech files can be downloaded from the web page http://www.fee.vutbr.cz/ cernocky/Icassp98.html.

5 Conclusions

Low bit rate coders from 4 Kbps to 800 bps are already in the stage of standardization and commercial applications. Very low bit rate coders are still in the research stage. Applications such as unified text and speech messagery can be forseen in a brief delay. But works are to be done for speaker and language adaptation, for reducing complexity and delay.

Studies on unsupervised data driven acoustic units determination have much in common with phonetician studies and learning theory.

References

[1] J. Černocký, G. Baudoin, G. Chollet. Segmental vocoder - going beyond the phonetic approach *Proc. ICASSP98*, pp. 605–608, Seattle, 1998.

[2] Spanias. Speech coding: A Tutorial Revew.

[3] C. Jaskie, B. fette. A survey of low bit rate vocoders. *DSP & Multimedia Technology*, pp 26–40, 94.

[4] Picone, G.R. Doddington. A phonetic Vocoder. *Proc. ICASSP-89*, pp. 580–583, 1989.

[5] T.E. Tremain. the government standardLinear Predictive Coding Algorithm: LPC10. *Speech Technology, Vol.1, No2*, pp. 40-49, Apr. 1982.

[6] D. Griffin, J. Lim. Multiband Excitation Vocoder. *IEEE trans. ASSP-36, No 8* pp. 1223, Aug. 1988.

[7] R. McAulay, T.Quatieri. Multirate Sinusoïdal Transform Coding at Rates from 2.4 kbps to 8kbit/s. *Proc. ICASSP-87*, Dallas, 1987.

[8] C. Laflamme & al. Harmonic-Stochastic Excitation (HSX) speech coding below 4 kbit/s. *Proc. ICASSP-96*, pp. 204–207, 1996.

[9] W. Kleijn & al. W.B. Kleijn and K.K. Paliwal Editors, Elsevier *Speech Coding and Synthesis*, 1995.

[10] L.M. supplee, & al. The new federal standard at 2400 bits/s. *Proc. ICASSP-97*pp. 1591-1594, Munich, 1997.

[11] C. Ribeiro, M. Trancoso. Phonetic vocoding with speaker adaptation. *Proc. Eurospeech-97* pp. 1291–1294, Rhodes, 1997.

[12] S. Roucos & al. A segment vocoder at 150 b/s. *Proc. ICASSP-83*, pp.61-64, 1983.

[13] G. Baudoin, J. Černocký, G. Chollet. Quantization of spectral sequences using variable length spectral segments for speech coding at very low bit rate. *Proc. Eurospeech-97*, pp.1295–1298, Rhodes, 1997.