Speech and Audio Coding - a Brief Overview

U. Heute
Inst. for Circuit and System Theory
Faculty of Engineering, Christian-Albrecht University
D-24143 Kiel, Germany

Abstract

The historic „coding gap“ between (narrow- and wide-band) high-rate and (narrow-band) low-rate coding of speech has been filled more and more during the past 15 years. The GSM coder of 1990 was a very important step right into the gap, and it caused more research towards better quality and higher compression, which, together with other activities world-wide, closed the gap. The concepts behind this are explained, with a stress on the basis of the final break-through to good quality at medium-to-low and even wide-band speech at medium rates. The same concepts followed for speech were also applied to music, with some strong differences in the results: While time-domain approaches prevail for speech, frequency-domain coding was successful for audio, and it was accompanied by much more exploitation of psycho-acoustic effects.

Speech Coding and the “Coding Gap”

In the “classsical” speech-coding literature, “high-rate” and “low-rate” codings were discriminated: While digital speech transmission or storage was possible with “telephone quality”, i.e., a bandwidth of *300...3400 Hz*, a sampling rate f_s*= 8 kHz*, and a signal-to-noise ratio (SNR) of *35...38 dB*, via various schemes between log. PCM at a bit rate f_B *= 64 kbit/s* (ITU G.711) and ADPCM at f_B *= 32 kbit/s* (ITU G.726), really low rates like f_B *= 2.4 kbit/s* were possible, for the same narrow-band speech, only with “synthetic” quality via VOCODER techniques like LPC-10, where no waveform reproduction was aimed at and, therefore, no SNR description could be used appropriately. Early attempts to code wide-band speech (*50...7000 Hz,* f_s *= 16 kHz*) somehow “naturally” worked at “high rates”, like a split-band ADPCM at *64 kbit/s* (ITU G.722). Between these groups, a “gap” was seen – topic of “lab’ research” mainly in the 70’s and beginning 80’s, but then invading realisation in the 90’s [1].

The last-named system (standardised in 1986) was also reckoned to be applicable for music. But in the later 80’s, research began to strive for a much better quality, namely that of the CD, at, in a first step, much higher rates: Audio coding became a field of its own.

Speech Coding: Origin of the Gap

The rate reduction from that of a simple linear A/D converter with some *w = 11(...12) bits* and $f_B = w*f_s$ *= 88 kbit/s* over *64 kbit/s* down to *32 kbit/s* at roughly the same quality uses three essential steps: Non-linear quantisation, adaptivity, and, last but not least, the concept of linear prediction (LP): A non-recursive filter with a transfer function $H_o(z) = \sum_{i=1}^{n} a_i z^{-i}$ computes a linear combination of past signal values as an estimation of the next sample according to $\widehat{x} = \sum_{i=1}^{n} a_i x_{k-i}$, which is subtracted from the actual sample x_k, yielding the difference or residual signal $d_k = x_k - \widehat{x}_k$ from a prediction-error filter *H(z) = 1- $H_o(z)$*. Excitation of *1/H(z)* by d_k in the receiver regenerates x_k.

The minimisation of the residual variance leads to the optimal predictor-coefficient vector $\underline{a} = \underline{R}^{-1}\underline{r}_o$, found from the correlation matrix $\underline{R}$ and vector $\underline{r}_o$. Due to this minimisation, d_k is “smaller” than x_k and can be quantised with less bits, even considerably less if both prediction and quantiser are adapted with time – but, if the *SNR quality* is to be kept, not less than *w=4*. Together with f_s*= 8 kHz*, this limits the rate reduction to the above-mentioned value of *32 kbit/s*. On the other hand, large parts of the signal information reside in the predictor and quantiser parameters, either adapted *backwards* (ADPCM) in both transmitter and receiver or transmitted as *side-information* with f_B*=...2...kbit/s* (APC). Exactly this is the data rate of the (LPC-) VOCODER, synthesising speech from parameters – *with synthetic quality*.

Speech Coding: Approaches to Close the Gap

RELP Coding: If “parameters-only” representations reduce quality and residual rates cannot be reduced by smaller values of *w,* f_s should be diminished. This is possible not for speech, but for the residual, which is whitened by the (necessarily) decorrelating filter *H(z)*: After band-limiting d_k to $(f_s/2)/r$, the

remaining (more or less constant) base band needs only a sampling rate f_s/r; in the receiver, a spectrally widened signal can be applied to *1/H(z)* in a base-band residual-excited LP system (BB-RELP). However, some unnatural, metallic distortions are unavoidable. Two reasons cause these artifacts: An unflexible *down-sampling*, which does not take care of spectral *fine structures* in d_k, namely its lines at harmonics of the fundamental frequency f_p (often termed "pitch").

Further Redundancy Reduction - LTP: Spectral lines are due to periodicity, they carry redundancy like the spectral shape (more or less) removed by *H(z)*. So, they can be removed in a similar way by a second, "pitch" predictor, which uses correlation not at distances of n/f_s ($n \approx 10$ in ADPCM/APC), but of $m_p/f_s \approx 1/f_p$ ($m_p \in (32,160)$) for normal voices). This long time range gives rise to the notation of a "long-term predictor" (see Fig. 1).

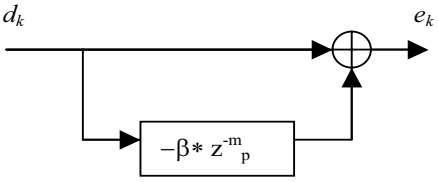

Fig. 1: Long-term predictor

Flexibilities in Down-Sampling - RPE and MPE: Instead of taking every *r-th* sample of the residual (d_k or e_k, if an LTP is used) in a naive manner, one may think of dissolving the mis-fit between sampling and "pitch" frequencies by, e.g., a decimation switching between various factors *r*, with a reasonable mean reduction. However, this is much less successful than a simple switch between the possible *r phases* of the down-sampling grid, according to some optimality criterion (see Fig. 2). What remains is a decimated residual (with or without LTP), applied in the receiver as a regular-pulse excitation (RPE) of *1/H(z)*.

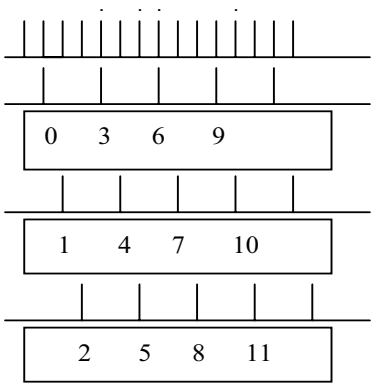

Fig. 2: RPE with *r* phases (e.g., *r=3*)

Beyond, *optimised* amplitudes may be applied instead of just samples, yielding a generalised RPE. Furthermore, the N/r remaining values after down-sampling a block of N data could as well be placed on any optimal positions rather than a grid, leading to the concept of multi-pulse excitation (MPE).Both RPE and MPE can, on the other hand, be simplified by allowing only simple amplitudes like *{-1,0,1},* with a common gain factor per (sub-) block, thus also separating shape and gain of the excitation.

Vector Quantisation (VQ): After quantisation, the reduced number of samples allows for a reduced set of possible excitation blocks. Another direct way to such excitations applies a table ("code-book") with, say, $L = 2^w$ length-N (non-decimated) vectors from which the best one is chosen – and coded by w bits per N samples; for, e.g., $L = 1024$ and $N = 40$, this means that $w = ¼$ *bits/sample* are used in such a "VQ" scheme. The same principle can the also be applied to parameters like the predictor-coefficient vector $\underline{a}$ (or any equivalent, transformed coefficient set like the so-called reflection-coefficient vector $\underline{k}$, the log-area-ratio vector $\underline{lar}$, or the line-spectral-frequencies' vector $\underline{lsf}$, to be calculated from a polynomial transformation of *H(z)*.

Speech Coding: Closing the Gap

The first important realisation step right *into* the gap was the GSM full-rate (FR) codec (see Fig. 3), standardised at the end of the 80's by ETSI and applied world-wide since 1992 [2]. It uses those of the above ideas which were realisable with then available hardware and yielded the best quality achievable thereby. Quality, however, could still not be measured by means of SNR: New instrumental measures of perceived quality, based on psycho-acoustics, became another field of research, at that time.

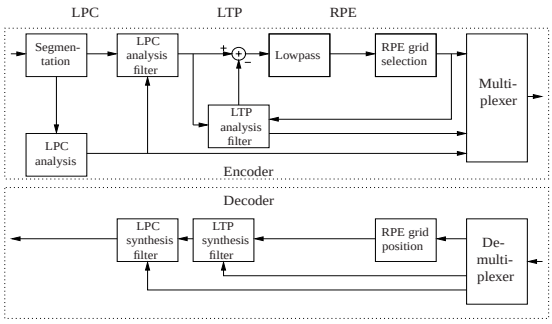

Fig. 3: GSM-FR system (RPE-LTP-RELP)

The limited quality of GSM-FR coding as well as the tremendous growth of user numbers caused a wave of research towards improvements. They relied on an enormous increase in computational power, on some of the above-named tools, plus two decisive ideas.

Speech Coding: The Break-Through Basis

The *VQ* principle turned out to be a very powerful concept. It is applied to both the receiver-excitation signal and the parameters. Especially, LSF-VQ with very good quality requires only some *20...24 bits* per *20 ms* frame, i.e. a rate of *1...1.2 kbit/s*.

Improved code-book designs via efficient training enhanced quality. By means of simple entries (like in simplified RPE/MPE) realisability was achieved. The simple (shape) vectors could be combined with separate gains and also be efficiently transformed into complex varieties by means of linear combinations ("vector sums") or matrix operations ("algebraic code-books").

Above all, however, a better optimisation criterion was decisive: Instead of minimising the MSE of the residual approximation, the distortion of the receiver-output signal – synthesised in the transmitter! – was reduced to a minimum. Furthermore, this "*analysis-by-synthesis*" technique included a perceptual weighting (as a crude perception model) before the MSE evaluation, and it finally also enclosed the LTP viewed as a second, (pitch-) adaptive code-book into its *closed loop*.

Speech Coding: CELP Coders

Systems with a code-book excitation and analysis-by-synthesis are termed CELP systems, with variants to be understood from the above considerations. Fig. 4 shows a typical coder (ITU G.729).

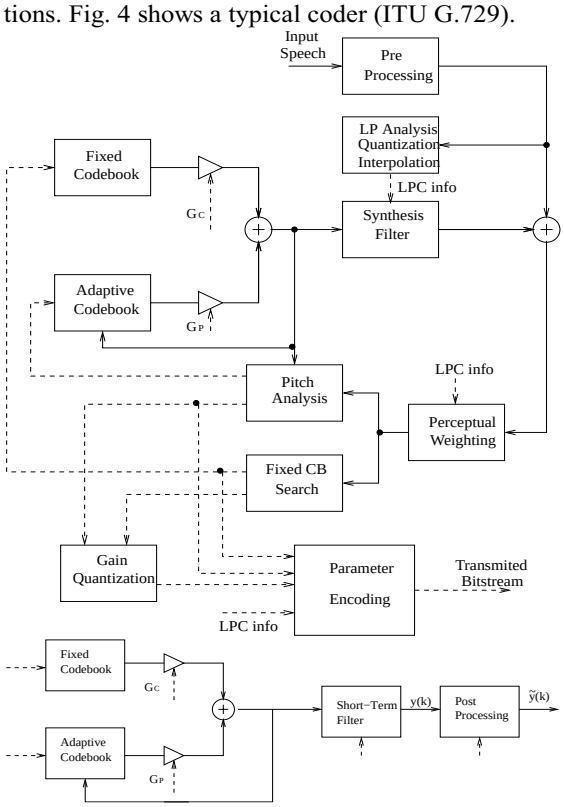

Fig. 4: Typical CELP (G.729 ACELP)

Many such systems are in use, nowadays [3-6]:

- ETSI Half-Rate (GSM-HR): VSELP; f_B= *5.6 kbit/s*
- ETSI Enhanced Full-Rate (GSM-EFR): ACELP; f_B= *12.2 kbit/s*
- ITU G.729: ACELP; f_B= *8 kbit/s*
- ITU proposal: "eXtended" CELP; f_B= *4 kbit/s*
- ITU Low-Delay Codec G.728: LD-CELP (n=50!); f_B= *16 kbit/s*
- MPEG-4 Scalable-Rate Codec: ACELP; f_B= *4...24 kbit/s*
- ETSI Adaptive Multi-Rate System (GSM-AMR): ACELP; f_B= *4.75...12.2 kbit/s*
- ETSI Wide-Band AMR System (GSM WB-AMR): ACELP; f_B= *6.6...23.85 kbit/s*

(both adapting to changing transmission channels, with 8 and 9 rates, respectively; the latter to be applied also in EDGE and UMTS).

Speech Coding: Frequency-Domain Coding

In the beginning of the GSM definition phase, the majority of the proposals was of a completely different kind [1]: They aimed at "medium rates" by means of "continuous" sub-band or "block-oriented" transform coders (SBC, TC). In fact, both can be viewed as systems with analysis-synthesis filter-banks, whether explicitly built as such or realized by filter-bank equivalent DFTs or DCTs; the true differences were their numbers M of frequency channels / components: M = ...8...16... in SBC, M =...128... in TC. The corresponding down-sampling of the narrow-band signals by r=M caused the misleading classification. Fig. 5 displays a typical TC system with adaptive quantisation and bit-allocation following the spectral envelope (ATC). This was realised for f_B= *14 kbit/s* with a quality rather close to GSM-FR and a similar quality. Nevertheless, frequency-domain coders turned out to be much less successful than the predictive systems discussed above.

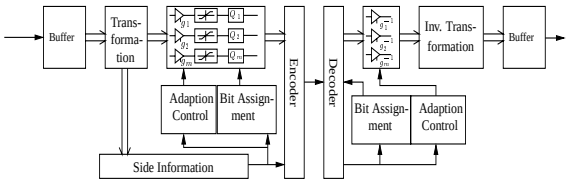

Fig. 5: Adaptive Transform Codec (ATC)

Speech Coding: Other Ways to Close the Gap

Describing speech segments as a finite sum of sinusoids with suitable amplitudes (coded directly or via an envelope by means of LPC coefficients), frequencies, and phases was intensively investigated in the late 80's. A trade-off between quality, computational load, and bit-rate was possible by choosing more or less refined descriptions especially of the frequencies and phases. A simplified version of this "sinusoidal

modelling“ (SM) applies a (pitch) frequency grid, and this “harmonic-coding“ (HC) scheme can be augmented by allowing narrow-band noise in unvoiced sections. From here, the way is not far to hybrid time-frequency codecs with mixed or multi-band excitation (MELP, MBE, IMBE, EMBE, aiming at rates near 4 kbit/s and below [6]).

Audio Coding: An even Shorter Overview

This part can be short because the important principles have all been named above. Nevertheless, there are important differences between today's successful speech and audio codecs: For music, *frequency-domain* coding prevails, and the coarse perception model of the CELP (or RPE/MPE) weighting filter is replaced by much more detailed *models of the hearing process*. These models include outer and middle-ear frequency responses, the critical-band (or Bark-scale) separation on the basilar membrane, non-linearities of the loudness perception, and time as well as frequency masking abilities of the ear – though not in all details in every coder. Fig. 6 shows the block diagram of the MPEG-1/Layers-I+II system. It's similarity with the ATC (Fig. 5) is obvious, with the exception of the inherent psycho-acoustic model. It's most popular variant, the Layer-3 system (“MP3”), adds a refinement of the sub-bands by means of a modified cosine transform (MDCT, with M=18), HUFFMAN coding, window-length switching (against “pre-echos”) etc., and it achieves CD quality at f_s= *44.1 kHz* / f_B= *192 kbit/s*.

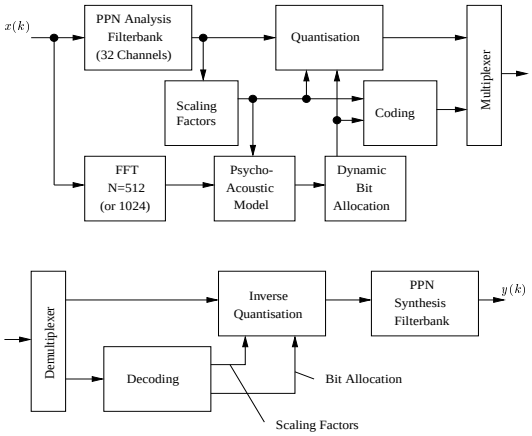

Fig. 6: MPEG-1 (Layers I+II) codec

The way to “MP3” in 1992 went from an ATC with a DCT (as that for speech), later an MDCT plus simultaneous-masking ideas in the OCF systems of 1988 in Germany or a similar (DFT-based) system termed PXFM in the US via a combination of both with additional octave-band QMF banks, inclusion of time-masking effects, and Bark-scale approximating QMF trees combined with transformations, again in Germany (MUSICAM, 1990)) and the US (ASPEC, 1991) as well as in France (1989) to the 3-layer standard including all successful steps.

Audio Coding: Further Work and Results

All the other ideas named in the speech-coding parts can of course be also tried for audio signals: VQ, “pitch” filters, linear prediction (in sub-bands), MPE, or SM are applicable. Especially, new time-frequency representations are of interest (wavelets). Finer spectral resolution, window-form switching, and more sophisticated psycho-acoustic models can be use advantageously, as in the advanced audio codec (AAC) chosen for MPEG-2 and MPEG-4. The latter standards aim at, respectively, enhanced stereo-surround representations (5 full channels with f_B= *320 kbit/s*!) and scalable rates and qualities. As a final goal of research, (still!) *transparent* quality is envisaged at f_B= *16 kbit/s*.
For more details on audio coding, the presentation [7] at this conference, and for a very broad overview, the recent report [8] are recommended.

Speech Coding: Further Results and Work

Good telephone quality will soon be achieved at f_B= *4 kbit/s*, while good wide-band quality is available at f_B= *...12... kbit/s*. Good telephone speech with f_B= *...2.4... kbit/s* is worked on – e.g., by means of waveform-interpolation concepts [9], models of the hearing process used in the LP excitation [10], or based on phoneme classification [11].
For more details, the reader is referred to [6] and to an updated list of literature in an extended version of this contribution [12].

References

[1] Heute, U.: Medium-Rate Speech Coding. Speech Comm., vol. 7 (1988), pp. 125-149.
[2] ETSI Rec. GSM 06.10:”GSM-FR Transcoding, 1988.
[3] Heute, U.: Speech Coding: Approaches, Trends, Standards (in German). Proc. ITG Conf. Source & Channel Cod., Munich, 1994, pp.437-448.
[4] Noll, P.: Wideband Speech and Audio Coding. IEEE Commun. Mag., Nov. 1993, pp. 34-44.
[5] Noll, P.: Digital Audio Coding for Visual Comm's. Proc. IEEE, vol. 83 (1995), pp. 925-943.
[6] Vary, P., Heute, U., Hess, W.: Digitale Sprachsignalverarbeitung. Teubner, Stuttgart, 1998.
[7] Mourjopoulos, J.: The Evolution of Digital Audio Technology. DAGA'02, Bochum.
[8] Painter, T., Spanias, A.: Perceptual Coding of Digital Audio. Proc. IEEE, vol. 88 (2000), pp. 451-513.
[9] Kleijn, B., Haagen, J.: A Speech Coder Based on Decomposition of Characteristic Waveforms. Proc. IEEE ICASSP'95, pp. 508-511.
[10] Ambikairajah, E., Epps, J., Lin, L.: Wideband Speech and Audio Coding Using Gammatone Filterbanks. Proc. IEEE ICASSP'01, Speech L-7.4.
[11] Ehnert, W.: Sprachcod. variabler Bitrate mit phonembasierten Ansätzen. Diss. LNS/TF/CAU, Kiel, 2000.
[12] Heute, U.: Speech and Audio Coding. To appear in 2003.