Introduction

We have to distinguish speech coding and speech vocoding.

- Speech coding balances high re-synthesis quality and low transmittion bit rate.
- Speech vocoding focuses on such parameters that are adequate to model underlying structure of speech. Compression (equivalent to transmittion rate) in less important.

Historical speech coders

- Based on analysis of parameters of linear speech model
- Transmit the parameters across the transmission channel
- Re-synthesise a reproduction of the speech signal with a linear model.

Channel vocoder - 1939

Figure: Homer Dudley (1896-1987)

Features

- Excitation: a pulse train or random noise
- System response (vocal tract model): 10 bandpass filters
- Quality: intelligible, but not very high quality: `http://www.youtube.com/watch?v=5hyI_dM5cGo`

Formant vocoder - 1953

Similar to the channel vocoder, but transmits information about formants directly.

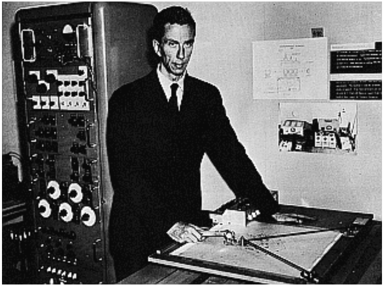


Figure: Gunnar Fant (1919-2009) and his OVE (Orator Verbis Electris) - a cascade formant synthesizer

Features

- Higher quality
- Synthesises speech with a small numbers of dumped resonators or poles (connected in parallel or in cascade)
- Formants are difficult to estimate reliably

LPC vocoder - 1970

LPC vocoders automatically captures formants (if they are dominant), and so it avoids the problem of formant tracking.
Formant and later LPC vocoders aimed to improve one well-known problem – a buzzy quality of vocoded speech.

- multi-pulse excitation
- regular-pulse excitation
- code-excited linear prediction

Current parametric speech coding

There are two main approaches to speech coding:

1. Parametric coding – that aims at reproducing the speech waveform as faithfully as possible. Typically the parameters are specified by a linear speech production model.
2. Waveform coding – that preserves only the spectral properties of speech in the encoded signal. Most of the effort has been done on excitation modelling.

ITU-T standardisation

Standardised waveform and parametric coding techniques are summarised by Tab (next slide). For now, no ITU-T 4 kb/s standard has yet been named. The standardisation effort has begun in 1994, but it has been shown that it is difficult to achieve toll-quality performance in all conditions, roughly represented by:

- Intelligibility,
- Quality,
- Speaker recognizability,
- Communicability,
- Language independence,
- Complexity.

ITU-T standardisation

Table: *ITU-T standards, upper part is waveform coding, below part is parametric coding.*

Standard/coder	Bandwidth	Bit rate	Notes
G.726 (ADPCM, 1986)	8 kHz	32 kbs	standardised in 1984 as G.721
G.728 (LD-CELP, 1992)	8 kHz	16 kbps	Low-Delay CELP
G.729 (CS-ACELP, 1998)	8 kHz	8 kbps	Conjugate-Structure algebraic CELI
– (MELP/CELP, 2002)	8 kHz	4 kbps	not standardised, waiting for you ☺
– (MELP, 1996)	8 kHz	2.4 kbps	parametric coding, US MIL-STD 30
– (MELPe, 2001)	8 kHz	1.2 kbps	US STANAG 4591 standard
– (MELPe, 2006)	8 kHz	600 bps	ext. US STANAG 4591, quality bett

Speech quality

The Figure compares quality depending on the bit rates.

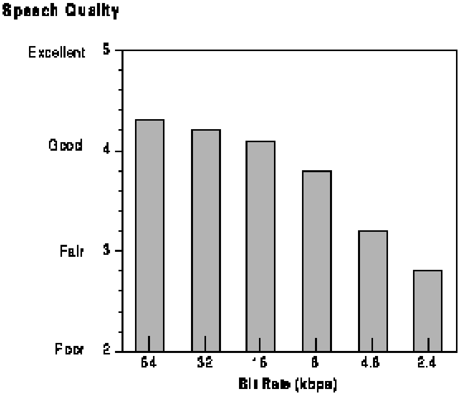


Figure: The speech quality mean opinion score for various bit rates.

Ideal low bit rate speech coder

- Let us define R as a bit rate of speech coding and H an entropy of the source coding.
- Shannon's source coding theorem says that source can be encoded with arbitrary small error probability, if $R > H$.
- However, what is H of a speech signal?

Estimation of speech entropy

- Information entropy of the source H quantifies the number of bits needed to describe the data. Entropy of the source alphabet with N symbols can be defined as $H = log_2(N)$.
- The information content of speech varies along two main dimensions, (i) the intrinsic one (phonetic/articulatory and speaker information) and (ii) the extrinsic one (phonological level represented be prosody information). Then, the H_{speech} can be estimated as:

$$H_{speech} = \frac{H_{phonetic}}{T_{phonetic}} + \frac{H_{speakers}}{T_{speakers}} + \frac{H_{prosodic}}{T_{prosodic}}. \qquad (1)$$

An example: English

Let us suppose that

- English has 38 phonemes with average duration $T_{phonetic} = 0.1$ (s),
- an average listener can distinguish 1000 speakers in average time $T_{speakers} = 1$ (s),
- and prosody can be characterised by roughly 100 symbols (such as 36 different part-of-speech tags, 15 different ToBI tags, 16 different basic emotions, and so far), estimated again by average phoneme duration $T_{prosodic} = T_{phonetic} = 0.1$ (s).

Final theoretical estimate

- Then, $H_{phonetic} = log_2(38)$, $H_{speakers} = log_2(1000)$ and $H_{prosodic} = log_2(100)$.
- Then we have an entropy estimate for the intrinsic speech information content in range of $50 - 60$ bits and extrinsic speech content $60 - 70$ bits.
- From the source coding theorem we can estimate that the minimal achievable bit rate is around $110 - 130$ bits per second.

Reality: rates about 1.000 – 2.000 b/s

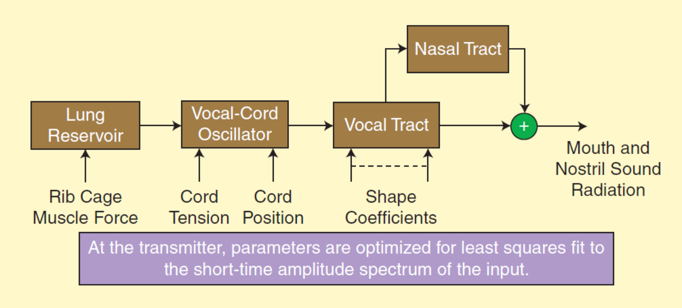


Figure: Components of the "speech mimic" system (Flanagan'2010).

Fourier coefficients

- Fourier coefficients can be used as parameters of the LPC residual signal $r[k]$.

$$r[k] = \frac{1}{N} \sum_{n=0}^{N-1} X[n] \exp(jk\frac{2\pi n}{N}) \tag{2}$$

where N is the pitch period, n is the frequency index, and $X[n]$ is the FT.

- Since $r[k]$ is real, we can write

$$r[k] = \sum_{n=0}^{N/2} A[n] \cos(k\frac{2\pi n}{N} + \phi_n) \tag{3}$$

where $A[n]$ are magnitudes and ϕ_n are phases of the LP residual harmonics.

- Excitation is synthesised as a sum of harmonic sine waves.

An idea of mixed excitation

- low-pass filtered pulses
- high-pass filtered noise
- sometimes a they are combine with multiband algorithm with individual voicing decisions

Mixed-excitation linear prediction (MELP)

- Different mixtures of a number (5) of frequency bands
- Only two filters are needed regardless the number of frequency bands
- The periodicity in each band is determined as normalised auto-correlation $c[t]$

$$c[t] = \frac{< x[k], x[k+t] >}{\sqrt{\sum_{k=0}^{N-1} x^2[k] \sum_{k=0}^{N-1} x^2[k+t]}} \tag{4}$$

Components of the MELP coding

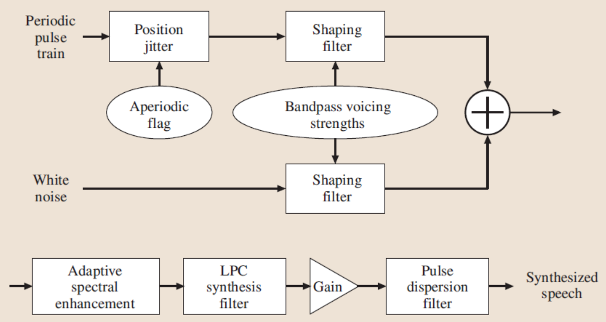


Figure: Mixed-excitation linear prediction analysis and synthesis.

MELP improvements 1

- to *mimic* erratic glottal pulses (typical in communication or a vocal fry), the periodicity of pitch periods is destroyed with jitter distributed up to $\pm 25\%$
- Jittery voicing is detected using peakness p from LP residual:

$$p = \frac{\sqrt{\frac{1}{N}\sum_{k=0}^{N-1} r^2[k]}}{\frac{1}{N}\sum_{k=0}^{N-1} |r[k]|} \tag{5}$$

- Encoder transmits: voiced, unvoiced and jittered flags.

MELP improvements 2

- Adaptive spectral enhancements: formant matching algorithm. In natural speech, resonances typically do not completely decay during one pitch period. This enhancement is to assure the same in the LPC modelled speech.
- Pulse dispersion filter: enhancements of re-synthesised speech in frequency bands that do not contain formants. It introduces additional excitation for longer pitch periods.

An idea of Sinusoidal coding

- Model speech as a sum of sine waves

$$x[k] = \sum_{l=0}^{L} A[l] \cos(k \frac{2\pi l}{L} + \phi_l) \tag{6}$$

- With higher frequency resolution, the model works also for unvoiced speech.

Sinusoidal transform coder (STC)

1. Signal windowing with a duration approximately 2 pitch periods.
2. STFT
3. Find maximums of the sine wave frequencies
4. Estimate magnitude and phase of the located complex spectra
5. Re-synthesis: phase trajectory can be modelled with a cubic polynomial as a function of time.

Properties of the STC

- Use of linear decomposition model
- Voiced speech frequencies are assumed to be harmonics – the encoder does not encode all sine waves. The search of harmonics is based on the pitch value.
- Parametric model for phase as well: voiced excitation is assumed to have zero phase.
- Parametric model of sine wave amplitudes (using LP coefficients in frequency or time domain)

Waveform interpolation

- Excitation signal for a voiced sound is frame-by-frame similar. Therefore one can extract these glottal flow cycles (more specifically LP residuals) at a slower rate, quantize them, and reconstruct missing cycles at a receiver.
- Analysis includes an alignment process in which each extracted cycle to correlation with the previous one. Extracted signal do have similar shapes.
- Harmonic sine wave synthesis of the excitation signal followed by LPC synthesis.
- Extracted signals are decomposed by low-pass and high-pass filter (with cut-off around 20 Hz) for two components: slowly evolved waveform and rapidly evolved waveform.

Very low bit rate (VLBR) speech coding

Very Low Bit Rate (VLBR) speech coding targets bit rates typically about 100 – 150 bps. A VLBR system can be achieved by the integration of symbol recognition (as an encoder) and speech synthesis (as a decoder), where:

- a sequence of *symbols*, such as phonemes, is transmitted instead of a compressed audio signal.
- Additional information such as *pitch*,
- and *duration* of the symbols is required to recover the original prosody.

HMM-based VLBR speech coding

- Within the last two decades, automatic speech recognition (ASR) and text to speech (TTS) technologies have almost completely converged around a single paradigm: the hidden Markov model (HMM).
- The HMM framework is almost completely data-driven. That is, it responds automatically to data with little human interaction required.
- In general the peripheral technologies, such as speech coding, advantageously share the HMMs' data driven capabilities. They allow, for example, tuning to a particular user after a few minutes.

Components of HMM-based VLBR system

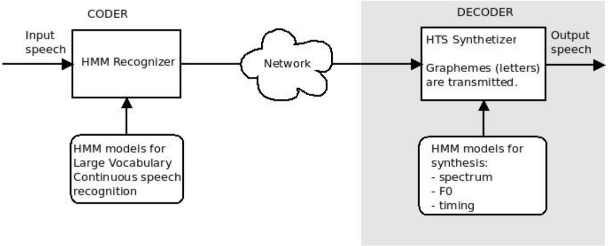


Figure: Hidden Markov Model (HMM) parametric speech coding.

The recognition/synthesis paradigm

- Use phoneme automatic speech recognition (ASR) for symbol encoding.
- Use prosody encoder and prosody reconstruction for
  - pitch
  - duration
- Use HMM-based speech synthesis (HTS system) for re-synthesis.

Speaker adaptation 1

- HTS technique is a new TTS paradigm that has emerged based on ASR technology, and can be thought of as an inversion of an HMM that allows speech to be synthesized as well as recognized.
- Although the HMM and HTS paradigms unify the general theory of ASR and TTS, and there is still a significant practical gap between the two approaches, they can be integrated into an elegant solution of very low bit-rate speech coding.
- Voice adaptation in HTS starts with HMMs trained on many speakers (HTS average) and uses HMM adaptation techniques drawn from speech recognition, to adapt the models to a new speaker (of the same language and with the same accent).

Speaker adaptation 2

1. The Vocal Tract Length Normalisation (VTLN)
2. A Maximum Likelihood Linear Regression (MLLR) based adaptation performs much better, but estimated bit-rates are much higher:

   $$\hat{\mu} = A\mu + b. \tag{7}$$

   The transform matrices A and b needs to be transmitted.
3. An approximation to MLLR-based adaptation might be multi-regression HMMs

   $$\hat{\mu} = \mu + R_0 + R\xi. \tag{8}$$

   The only difference with the is that MLLR applies a transform to the mean vector, whereas multi-regression HMMs applies the transform to the auxiliary vector ξ.

Symbol coding

The basic issue here is to select a suitable symbol set.

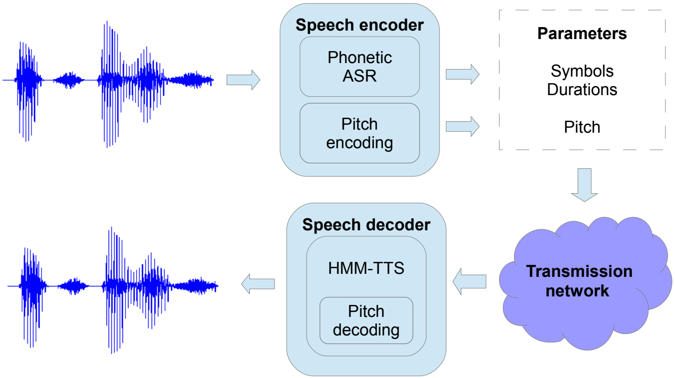
- Data-driven approach, where the symbol set is found automatically using a vector quantization.
- Knowledge-based approach, where the symbol set is a phoneme set of a particular language or shared phonemes set.
- Lossless coding is further applied here – it means no loss of any information during symbol coding. In other words it allows perfect reconstruction. An example – Huffman coding.

Huffman Coding

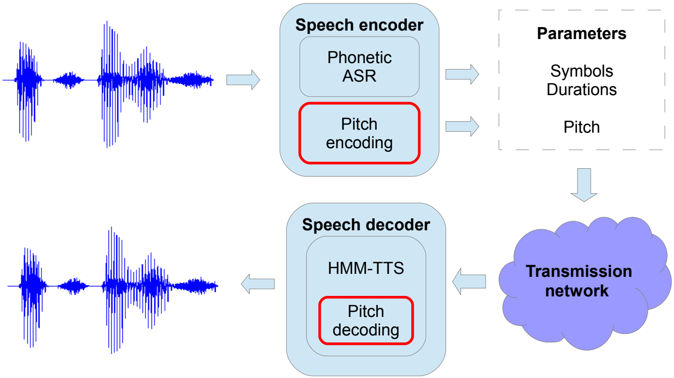
- Assign short codewords to frequent inputs
- Assign long codewords to less frequent inputs
- Similar to the Morse code
- Design:
  1. Merge together two least probable inputs, assign new probability.
  2. Repeat the merging until there is only one input remaining.

Another popular lossless algorithm is Lempel-Ziv coding.

Purpose of pitch encoding

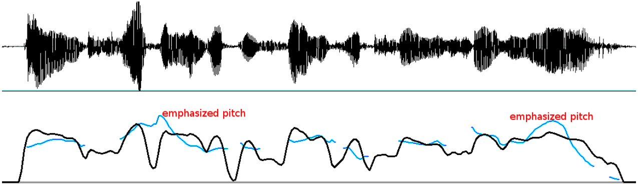


Purpose of pitch encoding



Pitch information

- Current pitch coding techniques work on the segmental level, as pitch quantisation[1], or contour/piecewise linear approximation[2].
- Pitch conveys both segmental (e.g. tone) supra-segmental information (e.g. emphasis)



I am talking about the same picture you showed me!

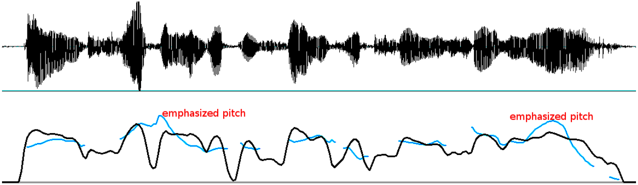
- Can we encode pitch on a higher-than segmental level?

[1] T. Nose and T. Kobayashi, Very low bit rate F0 coding, ICASSP'11

[2] K.S. Lee and R.V. Cox, A very low bit rate coding, IEEE TSAP'01

Pitch information

- Current pitch coding techniques work on the segmental level, as pitch quantisation[1], or contour/piecewise linear approximation[2].
- Pitch conveys both segmental (e.g. tone) supra-segmental information (e.g. emphasis)



I am talking about the same picture you showed me!

- Can we encode pitch on a higher-than segmental level?

[1] T. Nose and T. Kobayashi, Very low bit rate F0 coding, ICASSP'11

[2] K.S. Lee and R.V. Cox, A very low bit rate coding, IEEE TSAP'01

Theorethical minimal pitch coding rate

- Let us define R as a bit rate of pitch coding and H an entropy of the source coding. Shannon's source coding theorem says that source can be encoded with arbitrary small error probability, if $R > H$. However, what is H_{pitch} of a pitch signal?
- The pitch signal can by described by 15 different ToBI tags, theoretically changed with each phoneme (every 100ms) and then H can be roughly estimated as:

$$H = \frac{H_{pitch}}{T_{pitch}} = \frac{log_2(15)}{0.1} = 40bits \tag{9}$$

Theorethical minimal pitch coding rate

- Let us define R as a bit rate of pitch coding and H an entropy of the source coding. Shannon's source coding theorem says that source can be encoded with arbitrary small error probability, if $R > H$. However, what is H_{pitch} of a pitch signal?
- The pitch signal can by described by 15 different ToBI tags, theoretically changed with each phoneme (every 100ms) and then H can be roughly estimated as:

$$H = \frac{H_{pitch}}{T_{pitch}} = \frac{log_2(15)}{0.1} = 40bits \tag{9}$$

Idea of the parametric pitch coding

- Pitch coding is "embedded" in audio coding – it is not parametrized.
- In waveform coding that make assumptions about possible decomposition of the signal with a source-filter model of speech production, it is transmitted frame-by-frame
- In parametric coding the pitch can be directly parametrized, as here we make the assumption that the speech signal contains supra-segmental cues – "syllables".

Method of the parametric speech coding

Syllable-based technique

1. Calculate raw F0.
2. Segment the stream on syllable boundaries.
3. For unvoiced syllable do nothing.
4. Parametrize the longest pitch contour of the voiced syllable, which have more than 3 voiced segments.
5. Transfer the pitch contour parameters along with the timing.

Parameterization using curve fitting technique

- A segment of a pitch contour with the length of $N+1$, $f(i/N)$, is approximated using discrete (Legendre) orthogonal polynomial as

$$\hat{f}\left(\frac{i}{N}\right) = \sum_{j=0}^{J-1} a_j \cdot \phi_j\left(\frac{i}{N}\right), \quad 0 \le i \le N \tag{10}$$

- where the parameters are

$$a_j = \frac{1}{N+1} \sum_{i=0}^{N} f\left(\frac{i}{N}\right) \cdot \phi_j\left(\frac{i}{N}\right). \tag{11}$$

- and J represents the order of approximation[3].

[3] S.H. Chen and Y.R. Wang, Vector quantization of pitch information, IEEE Trans. on Communications 1990.

An example of the coder

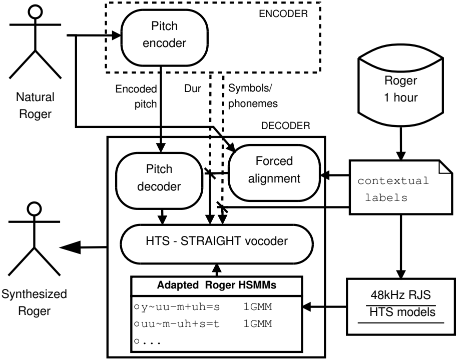


Figure: *VLBR speech coding experimental setup with recognition-synthesis architecture, abstracting the encoder (dotted lines) except for pitch encoding and decoding modules.*

Duration coding

- Duration in recognition/synthesis speech coding system is coded using a vector quantisation method – so called a lossy coding.
- The input is discretisized, and the loss of information is related to the resolution of the discretization. We cannot use a prior knowledge about the duration.

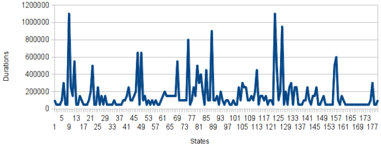


Figure: Duration of HMM states of an example speech.

Linde-Buzo-Gray (LBG) algorithm – initialisation

1. Given training set $T = x_1, x_2, \ldots, x_M$, and error $\epsilon = 0.001$
2. Let the number of codewords $N = 1$ and centroid $c_1^* = \frac{1}{M}\sum_{m=1}^{M} x_m$. Then we calculate average distortion

$$D_{ave}^* = \frac{1}{Mk}\sum_{m=1}^{M} \| x_m - c_1^* \|^2 \tag{12}$$

where k is dimensionality of the training example x_m.

3. For $i = 1..N$:

$$\begin{aligned} c_i^{(0)} &= (1+\epsilon)c_1^* \\ c_{N+i}^{(0)} &= (1-\epsilon)c_1^* \end{aligned} \tag{13}$$

Set $N = 2N$.

Linde-Buzo-Gray (LBG) algorithm – iterations

- For all training examples find index n^* that achieves the minimum of $\| x_m - c_n^{(i)} \|, \forall m \in M, n \in N$, set $Q(x_m) = c_{n^*}^{(i)}$
- Update codevectors as average of training examples in the coding region:

$$c_n^{(i+1)} = \frac{\sum_{Q(x_m)=c_n^{(i)}} x_m}{\sum_{Q(x_m)=c_n^{(i)}} 1}, \forall n \in N \tag{14}$$

- $i = i + 1$
- Distortion error:

$$D_{ave}^{(i)} = \frac{1}{Mk} \sum_{m=1} M \| x_m - Q(x_m) \|^2 \tag{15}$$

- If $(D_{ave}^{(i-1)} - D_{ave}^{(i)})/D_{ave}^{(i-1)} > \epsilon$, make new iteration

Final codewords and distortion: $c_n^* = c_n^{(i)}, D_{ave}^* = D_{ave}^{(i)}$.