Voice Coding

Notes on Speech and Audio Processing

Chia-Ping Chen

Department of Computer Science and Engineering
National Sun Yat-Sen University
Kaohsiung, Taiwan ROC

Introduction

- We introduce the vocoders, also called the analysis-synthesis systems. Basically, the components include spectral envelope estimation (analysis) and speech synthesis.
- The primary goal of voice (or speech) coding is to reduce the bit rate of the transmission of voice. It is also used for efficient storage. The theoretical basis is the source coding theory.

Coding Standards

- Standard bandwidth and coding rates for audio signal is shown in Figure 31.1. It is obvious that the higher the audio quality we want, the larger the bit rate we need.
- For the telephone the bit rate is lowered by a factor of 4, as a result of research, while maintaining the quality of the transmitted signal.

Channel Vocoders

- Refer to Figure 31.2. The transmitter analyzes the audio signal to extract and encode the spectral, voicing and pitch estimates. These signals are transmitted over the channel. The receiver decodes and synthesizes the signal.
- Given the overall system block diagram, there are a few crucial design parameters to be considered:
  - transmission rate (bit rate) R
  - number of filters N
  - filter central frequencies and bandwidths
  - update rate

Energy Measurements

- For each bandpass filter, the output is processed to produce an estimate of the spectrum.
- The first step is deriving the spectral magnitude using, e.g., the full-wave or half-wave rectifiers.
- This followed by a low-pass filter which eliminates the high-frequency part of magnitude signal while preserving the spectrum near dc.
- We keep the slow variation in the range of $5 - 15$ Hz because this is similar to the rate of change of the articulator.

Vocoder Design for 2400 bps

- Suppose that the bit rate is 2400 bps and 400 bps is used in excitation coding.
- Case 1: Suppose we use differential pulse code modulation (DPCM) across bands. The first uses 3 bits while the others use 2 bits, for a total of 20 channels. Suppose the frame rate is 50. Then the bit rate is $(3 + 2 * 19) * 50 + 400 = 2450$ bps.
- Case 2: Suppose instead we use 15 channels and use 3 bits for each channel. In order to meet the 2400 bit rate, the frame rate should be 44 Hz, since $(3 * 15) * 44 + 400 = 2380$ bps.

μ-Law Quantization

- The more bits, the less quantization error. However, the bits per channel is constrained by the bit rate.
- The μ-law quantization is defined for input $x > 0$

$$y = X \frac{\log\left[1 + \mu(\frac{x}{X})\right]}{\log(1+\mu)}$$

where μ is a non-negative parameter and X is the maximal input value. Figure 31.7 shows a 2-bit case.

- The non-linear quantization can be generalized by considering probability distribution for the input and construct quantization regions corresponding to equal probability masses.

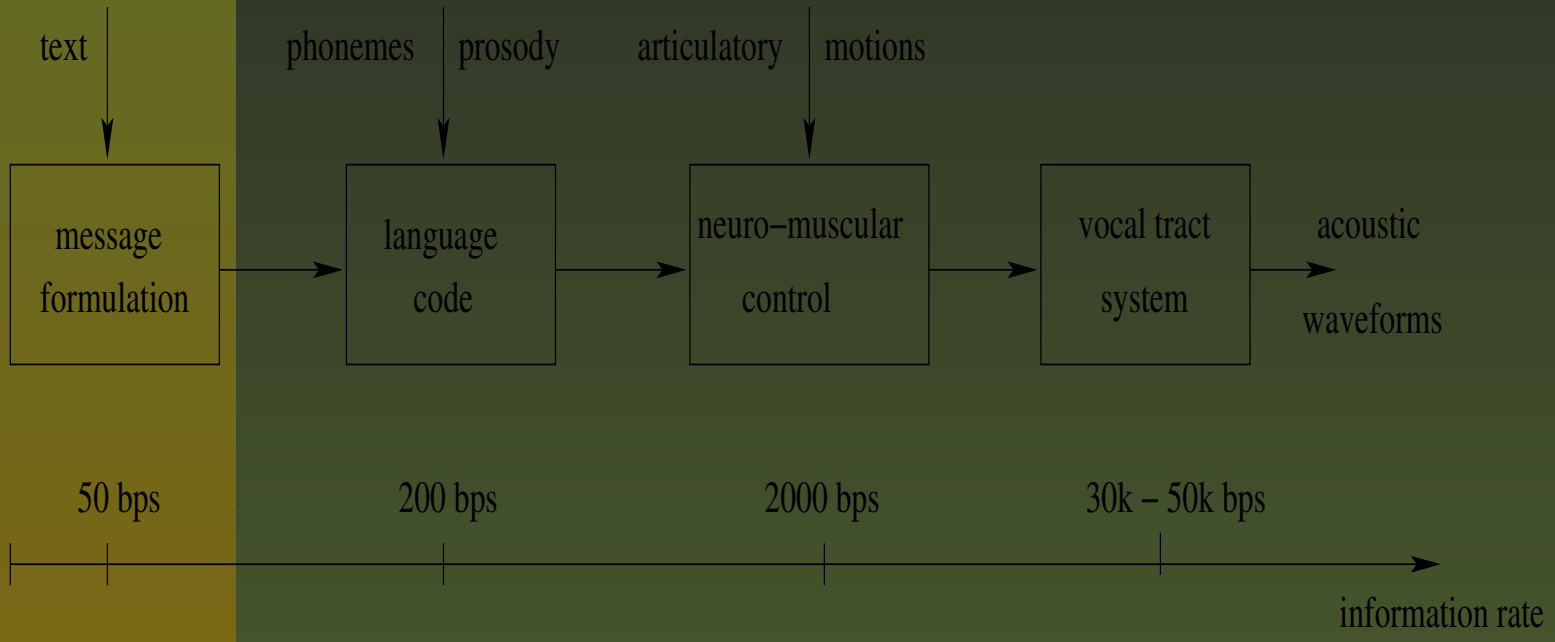
Linear Transformations

- Another way to reduce bit rate is to reduce the number of parameters to be transmitted.
- The magnitude signals are correlated to some degree. It is natural to take advantage of this redundancy.
- There are time-domain correlation and frequency-domain correlation.
- Common methods such as PCA and DCT are used. Note that in PCA one needs to find the principle axes, while DCT is data-independent.

LPC Vocoder

- In contrast to the channel vocoders, an LPC vocoder deal with the error signals.
- The error signal ought to be codable at a low bit rate assuming the spectral information is captured in the LPC spectral analysis.
- The error signals are encoded as well as other parameters.

Information Rates of Speech



Low-Rate Vocoders

- Low-rate: $R \leq 2400$ bps
- Further bit rate reduction
  - interpolation: exploiting correlation in the time and frequency domain
  - vector quantization: transmitting code index
  - reduce the number of parameters to be transmitted
  - recognition/synthesis approach

Frame Fill

- Basic idea: transmit from analyzer to synthesizer every Mth frame.
- $M = 2$ case: For an omitted frame, say frame N, compare it with frame $N - 1$ and $N + 1$ or some weighted combination of both.
- Select the best match and append the I.D. code to the frame that is to be transmitted.
- For LPC vocoder, higher order coefficients are omitted for unvoiced frames for further bit rate reduction.

Vector Quantization

- Basic idea: represent a vector by an index, which is mapped to a reproduction vector.
- Assume that we can use 20 bits to represent a spectrum. How do we decide these 2^{20} patterns?
  - A large amount of data must be collected and analyzed.
  - A distance measure must be defined.

Bit Rates for An Ideal System

- Based on the recognition/synthesis approach
  - recognition $\longrightarrow$ transmission $\longrightarrow$ synthesis.
  - reduce the bit rate to approximately $50 - 100$ bps
  - some information is lost, such as the speaker's identity and style.
- The bit rate for 64 phonemes and 10 phonemes per second is 60 bps. The rate for 1000 allophones (different phones for the same phoneme) and 10 phonemes per second is 100 bps. This is virtually the lower bound for vocoders.

Medium-Rate Voice Coding

- Bit rates below 2400 bps sacrifice sound quality.
- A medium-rate, typically $4800 - 16000$ bps, can deliver more robust and higher quality speech.
- Always involve coding of the excitation in addition to the vocal tract parameters. For example, an error-signal excited LPC vocoder
  - excites the LPC synthesizer with the complete error signal, reproducing the original speech.
  - a compressed version of the error signal can be sufficient for quality purpose.

Difference and Adaptive Coding

- Instead of $s(n)$, we quantize $e(n) = s(n) - s(n-1)$, assuming that this requires less bits.
- Sometimes a fast rising or falling signal would cause the decoded signal to “lag”.
- In this case, we can adaptively change the quantization steps: if error signals are of the same sign for three successive samples, increase the step size; otherwise decrease it.

CELP

- The difference in excitation from frame to frame is not explicitly given by pulses but by an index corresponding to a codeword, representing a sequence.
- The CELP (Code-Excited Linear Predictive) coder chooses the codeword with the least distortion.
- Since the pitch period and excitation sequence change more frequently than the spectrum, the codebook index and pitch estimates are transmitted more often than LP coefficients.

A Standard for 4800 bps Using CELP

- 10 LP parameters using 34 bits every 30 ms. 1133 bps.
- Pitch estimates is sent 4 times faster. Each is 12 bits on average (5 bits for gain, 8/6 bits for pitch period). 1600 bps.
- Excitation codebook indices is also sent every 7.5 ms, using 14 bits (5 bits for gain, 9 bits for code index). 1867 bps.
- 200 bps for synchronization and forward error correction.

Codebook Search

- The critical step in the CELP is the search for the optimal codeword.
- Each codeword is associated with a distortion score by going through the pitch synthesis filter, the LP filter and the perceptual weighting filter.
- A number of solutions have been proposed to reduce the search cost.

Fast Codebook Search

- Preliminary search followed by more refined search.
  - multi-resolution codebook search
  - partial sequence elimination
- Tree-structured delta codebook design
- Adaptive codebook: choosing from previous excitation samples.
- Linear combination of codebooks: K basis vectors spanning 2^K codewords.