Speech Coding

Notes on Spoken Language Processing

Chia-Ping Chen

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

Introduction

- Speech coding refers to the study of encoding speech signals (digitally), often under bandwidth limitation.
  - Digital representation can be transmitted over data network.
  - Bandwidth limitation often requires the speech signal to be compressed.
- Other audio signals such as music may need to be compressed too, such as MP3. That is referred to as audio coding.

Coder Attributes

- bit rate (bandwidth)
  - telephone speech ($300 - 3400$ Hz, sampled at 8kHz)
  - wideband speech ($50 - 7000$ Hz, sampled at 16kHz)
  - audio coding (44.1kHz)
- fixed-rate vs. variable-rate
- lossy vs. lossless

Quality Measure

- mean opinion score (MOS)
  - excellent
  - good
  - fair
  - poor
  - bad
- signal-to-noise ratio (SNR)

$$\mathrm{SNR} = \frac{\sigma_x^2}{\sigma_e^2} = \frac{E\{x^2[n]\}}{E\{e^2[n]\}}$$

Delays

- algorithmic delay: due to block (frame) encoding
- computational delay: time for frame processing
- multiplexing delay: time for adding error correction
- transmission delay: time to traverse the channel
- decoder delay: time to reconstruct signal
- Delay of more than 150 ms is not acceptable in an interactive conversation.

Graceful Degradation

- Transmission errors may occur, especially in a wireless setting.
  - channel errors
  - missing frames
- Graceful degradation of speech quality under channel errors is desired.
- Likewise, graceful degradation of speech quality with missing frames is also desired.

Pulse Code Modulation (PCM)

- quantize a signal sample to 2^B levels and use B bits for representation
  - linear PCM: the quantization levels are linearly spaced
  - μ-law PCM: the idea is to have same SNR regardless of the signal level; the step size is proportional to the signal
  - adaptive PCM: the step size is proportional to signal standard deviation

Delta Modulation (DM)

- Instead of quantizing the samples, DM quantizes the differences in subsequent samples.

$$d[n] = x[n] - \tilde{x}[n]; \; \hat{d}[n] = d[n] + e[n]; \; \hat{x}[n] = \tilde{x}[n] + \hat{d}[n]$$

  where $\tilde{x}$ is the predicted signal, $\hat{d}$ is the quantized difference.

- The simplest case is the 1-bit DM.

$$\hat{x}[n] = \tilde{x}[n] \pm \Delta$$

- Adaptive DM (ADM) has a step size that increases if subsequent errors have the same sign, in order to "catch up".

Improved DPCM

- Uses linear prediction of past quantized values instead of just the previous one.

$$\tilde{x}[n] = \sum_{k=1}^{p} a_k \hat{x}[n-k]$$

- The coefficients a_k can be adapted.
- ADPCM combines differential quantization with adaptive step size. Used in ITU-T Recommendations.

Frequency-Domain Coder

- advantages of working in frequency domain
  - Frequency domain components are approximately uncorrelated.
  - The masking effect can be more easily implemented in the frequency domain.

Masking

- The masking effect is a phenomenon that human cannot perceive a sound below a certain level in the presence of another sound of a near frequency.
- We don’t need to encode such a sound.
- This is the basic idea of the MPEG-1 Layer I audio encoding standard.
- MP3 stands for MPEG-1 Layer III, which is not far from the main idea introduced here.

Adaptive Spectral Entropy Coding

- ASPEC is used in high-quality music signals.
- The DFT coefficients are grouped into 128 subbands, with 128 scalar quantizers.
- Entropy coding is used to encode the coefficients of that subband.
- Suppose subband j has k_j levels of step size T_j, then

$$k_j = 1 + 2 \times \text{rnd}\,(\frac{P_j}{T_j}),$$

where rnd() is a rounding function (to the nearest integer) and P_j is the quantized magnitude of largest component.

Consumer Audio

- Dolby Digital: multichannel, lossy AC-3 coding, sampling rate 48 kHz, up to 24 bits.
- MPEG: MPEG-1 up to 384 kbps; MPEG-2 16-bit linear PCM at 48 kHz.
  - MPEG-1: up to 384 kbps
  - MPEG-2: 16-bit linear PCM at 48 kHz; used in DVD audio; variable rate from 32 to 912 kbps.
- Digital Theater Systems (DTS): multichannel, 20-bit PCM at 48 kHz; variable rate from 64 to 1536 kbps

Digital Audio Broadcasting

- DAB is a radio with high sound quality, service availability, flexible coverage scenarios, and spectrum efficiency.
- Most widely used system is the Eureka 147 DAB.
  - Each channel has a bandwidth of 1.536 MHz, with a raw data rate of 2304 kbps.
  - The useful data rate is between $600 - 1800$ kbps, used for audio and data programs.
  - Audio programs are compressed by MUSICAM (MPEG-1 Layer II).

LPC Vocoder

- Speech production can be modelled by an excitation source driving a time-varying filter.
  - For voiced speech, the source is a periodic impulse train.
  - For voiceless speech, the source is a random white noise.
- Using linear prediction removes the redundancy in the signal, so a simpler quantizer can be used on the residual signal.
- For example, Federal Standard 1015 is based on a 2.4 kbps LPC vocoder.

Code Excited Linear Prediction

- The residual is defined by

$$e[n] = x[n] - \sum_{i=1}^{p} a_i x[n-i].$$

- The residual signal is quantized using VQ. It is represented by a codeword.
- In CELP, the LP coefficients are quantized and transmitted, as well as the codeword index.
- The prediction using LPC is called short-term prediction. The prediction of the residual based on pitch is called long-term prediction.

Open-loop Estimation

- We first estimate LPC and quantize them.
- We can obtain the transfer function of the LPC filter.
- Then we can estimate the optimal codeword index for excitation vector, as well as the optimal gain, based on the minimization of error signal (analyss by synthesis).
- The quantized LPC, the codeword index, and the optimal gain are transmitted.

CELP Standards

- Various standards for bit-rate/quality constraints.
- VoIP: voice over internet protocol, adopts many audio coding standards.
  - G.728: toll-quality, low-delay CELP at 16 kbps.
  - G.729: toll-quality, 10-ms delay CELP at 8 kbps.
  - G.723.1: 30-ms delay CELP at $5.3/6.3$ kbps.
- They are under the H.323 umbrella standard.

GSM

- General System for Mobile communication
- voice coding: regular pulse excited-linear predicative coder (RPE-LPC) with long-term predictor loop
  - full rate = 13 kbps
  - half rate = 5.6 kbps
- Enhanced full-rate (EFR) standard based on ACELP achieves toll quality.

Low-Bit-Rate Speech Coders

- CELP is an example of waveform-approximating coder, which approximates the speech waveforms.
- In contrast, a low-bit-rate coder does not do that. Rather, the goal in low-bit-rate coding is to compress the signal into a perceptually equivalent one.
- Consequently, low-bit-rate coders have small SNR and is sensitive to noise.