Multimedia Communications

Speech Compression
Introduction

• No matter what language is being spoken, speech is generated using machinery that is not very different from person to person
• This machinery has to obey physical laws that substantially limit the behavior of the outputs
• Speech can be analyzed in terms of a model and model can be extracted and transmitted to the receiver
• At the receiver the speech is synthesized using the model
Introduction

- Speech is produced by forcing air first through an elastic opening, the vocal cords, laryngeal, oral, nasal and pharynx passages and finally through mouth and nasal cavity.
- First sound is generated and is modulated into speech as it traverses through the vocal tract.
- In order to generate a fragment of speech we have to generate a sequence of excitation signals and the corresponding sequence of vocal tract approximation.

Excitation source  \(\rightarrow\)  Vocal tract filter
Introduction

• Many speech compression schemes:
  – Channel vocoder
  – Linear predictive coder (LPC)
  – Code excited linear prediction (CELP)
  – Sinusoidal coders
  – Mixed excitation linear prediction (MELP)
Vocoder

- Each segment of input speech is analyzed using a bank of band-pass filters called the analysis filters.
- Energy at the output of each filter is estimated at fixed time intervals and transmitted to the receiver.
- A decision is made as to whether the speech in that segment is voiced or unvoiced.
- Voiced sound tend to have a pseudo-periodic structure.
- The period of the fundamental harmonic is called the pitch period.
- Transmitter also forms an estimate of the pitch period which is transmitted to the receiver.
Vocoder

- Unvoiced sounds tend to have a noise like structure
- At the receiver, the vocal tract filter is implemented by a bank of band-pass filters (identical to the filters at transmitter)
- The input to the filters is noise source (for unvoiced segments) or periodic pulse (for voiced)
Vocoder

• Variations of vocoder: formant vocoder, voice excited vocoder
• Vocal tract is a tube of non-uniform cross section that resonates at a number of different frequencies known as formants
• Formant vocoder transmits an estimate of the formant values (usually 4) and an estimate of the bandwidth of each formant
• At the receiver the excitation signal is passed through tunable filters tuned to the formant frequency and bandwidth
Vocoder

• In voice excited channel vocoder, the voice is first filtered using a narrow-band low-pass filter
• Output of the filter is sampled and transmitted to the receiver
• At the receiver, the low-pass signal is passed through a nonlinearity to generate higher order harmonics that together with the low pass signal are used as the excitation signal
• Voice excitation removes the problem of pitch extraction and declaring every segment voiced or unvoiced
Linear Predictive Coder (LPC-10)

• Instead of the vocal tract being modeled by a bank of filters, in LPC, it is modeled as a single linear filter:

\[ y[n] = \sum_{i=1}^{M} b_i y[n - i] + Ge[n] \]

• The input to the vocal tract filter is either the output of a random noise generator or a periodic pulse generator

• At the transmitter a segment of speech is analyzed to make a decision on voiced/unvoiced, the pitch period and the parameters of the vocal tract filter

• In LPC-10 input speech is sampled at 8000 samples per second which is broken into 180 sample segments

• Rate: 2.4 kbps
LPC-10

Decoder

Pitch Period
Pulse Train

V/U
Signal Power

G
H(z)

Random Noise

Synthesized Speech
Multi-pulse linear predictive coding (MP-LPC)

- One of the most important factors in generating natural sounding speech is the excitation signal
- Human ear is especially sensitive to pitch errors
- Using a single pulse per pitch period leads to a buzzy twang
- Multi-pulse linear predictive coding (MP-LPC): several pulses were used during each segment
- Spacing of these pulses is determined by evaluating a number of patterns from a codebook of patterns
MP-LPC

• Each entry in codebook is an excitation sequence that consists of a few nonzero values separated by zeros.
• Codebook entry generating minimum average weighted error is declared the best match.
• Index of the best match is sent to the receiver.
Regular pulse excitation (RPE) coding

• RPE is a modification of MP-LPC
• Instead of using excitation vectors in which the nonzero values are separated by an arbitrary number of zero values, the nonzero values occur at regularly spaced intervals.
• A variation of RPE called regular pulse excitation with long term prediction (RPE-LTP) was adopted as a standard for digital cellular phones by GSM.
Code excited linear prediction (CELP)

- In CELP instead of having a codebook of pulse patterns we allow a variety of excitation signals.
- Given a segment, encoder obtains the vocal tract filter.
- Encoder then excites the vocal tract filter with the entries of the codebook.
- Difference between original speech segment and the synthesized speech is fed to a perceptual weighting filter.
- Codebook entry generating minimum average weighted error is declared to the best match.
- Two examples will be reviewed: Federal Standard 1016 (FS 1016), and G.728.
• Vocal tract filter: \[ y[n] = \sum_{i=1}^{10} b_i y[n-i] + \beta y[n-P] + G e[n] \]

• Input speech is sampled at 8000 samples per second and divided into 30 ms frames containing 240 samples.
• Each frame is divided into four subframes of length 7.5 ms.
• The coefficients \( b_i \) are obtained using the auto-correlation method
• Pitch period is calculated once every subframe
• FS 1016 uses two codebooks: stochastic and adaptive
• An excitation sequence is generated for each subframe by adding one scaled element from the stochastic codebook and one scaled element from the adaptive codebook.
FS 1016

- Scale factors and indices are selected to minimize the perceptual error between the input and synthesized speech
- Stochastic codebook contains 512 entries which are generated using a Gaussian random number generator
- Adaptive codebook consists of the excitation vectors from the previous frame
- FS 1016 provides excellent reproduction in both quiet and noisy environment at rates of 4.8 kbps and above
G.728

- Coding delay: the time between when a speech sample is encoded to when it is decoded if there was no transmission delay.
- Coding delay consists of buffering delay (to store a segment), processing delay.
- Long delays are not acceptable in some applications.
- G728 is a CELP coder with a coder delay of 2 ms operating at 16 kbps (2 bits per sample).
- Reduce delay: reduce the segment size.
- Segment size: 5 samples.
G.728

• The algorithm obtains vocal tract filter parameters in a backward adaptive manner: vocal tract filter coefficients to be used to synthesize the current segment are obtained by analyzing the previous decoded segment.
• G.728 algorithm does not use pitch filter instead it uses a 50th order vocal tract filter.
• Since the vocal tract filter is completely determined in a backward adaptive manner, we have all 10 bits available to encode the excitation sequence.
• 10 bits => 1024 excitation sequences => too many to analyze in 0.625 ms.
G.728

- G.728 uses a product codebook: each excitation is represented by the product of a normalized sequence and a gain term.
- 3 bits used to encode the gain using predictive encoding and 7 bits form the index to a codebook containing 127 sequences.
Figure 14.9 Block diagram of the G.728 encoder.
Sinusoidal Coders

• Main problem with the LPC coder: excitation signal
• CELP coder solves this problem using a codebook of excitation signals
• Sinusoidal coders solve this problem by using an excitation signal that is the sum of sine waves of arbitrary amplitudes, frequencies and phases.
• Since the vocal tract is a linear system, the synthesized speech (based on a sinusoidal excitation) will also be sinusoidal
• Sinusoidal coders directly estimate the parameters required to synthesize the speech at the decoder
Sinusoidal Coders

• Sinusoidal coders divide the input speech into frames and obtain the parameters of the speech separately for each frame.

• Sinusoidal coder uses interpolation algorithms to smooth discontinuities at frame boundaries.
Sinusoidal coder

Figure 8.1 General sinusoidal analysis and synthesis
Mixed excitation linear prediction (MELP)

- MELP is the new federal standard for speech coding at 2.4 kbps
- MELP uses LPC filter to model the vocal tract and a much more complex approach to the generation of the excitation signal
- Excitation signal is a multiband mixed excitation
- Mixed excitation contains both a filtered signal from a noise generator as well as a contribution depending on the input signal
- First step in constructing the excitation signal is pitch extraction
MELP

• Input is also subjected to a multiband voicing analysis using five filters with passband 0-500, 500-1000, 1000-2000, 2000-3000 and 3000-4000 Hz.
• The goal of the analysis is to obtain the voicing strength for each band used in the shaping filters
Figure 17.1 The MELP model of speech production.