Speech coding and compression

Corso di
Networked Multimedia Systems

Master Universitario di Primo Livello in
Progettazione e Gestione di Sistemi di Rete

Carlo Drioli

Università degli Studi di Verona
Facoltà di Scienze Matematiche,
Fisiche e Naturali

Dipartimento di Informatica

Speech coding and compression: OUTLINE

Introduction to voice coding/compression

PCM coding

Speech vocoders based on speech production

LPC based vocoders

Speech over packet networks

Introduction

Approaches to voice coding/compression

- Waveform coders (PCM)
- Voice coders (vocoders)

Quality assessment

- Intelligibility
- Naturalness (involves speaker identity preservation, emotion)
- Subjective assessment: Listening test, Mean Opinion Score (MOS), Diagnostic acceptability measure (DAM), Diagnostic Rhyme Test (DRT)
- Objective assessment: Signal to Noise Ratio (SNR), spectral distance measures, acoustic cues comparison

PCM coding of speech signals

Simplest PCM coding of speech signals involves:

- Anti-aliasing filtering
- Sampling
- Companding
- Quantization

Example: ITU-T G.711 PCM

Bit rate	64 Kbit/s
Sampling freq.	8 KHz
Quant. bits	8 bits
Companding	logarithmic a-law and μ-law

PCM coding of speech signals

DPCM/ADPCM

- Code difference of current sample from previous sample
- May use prediction coding: difference is from estimate of current sample based on past samples
- Difference is quantized through scalar quantization
- Quantization step-size and predictor coefficients can be adaptive (ADPCM)

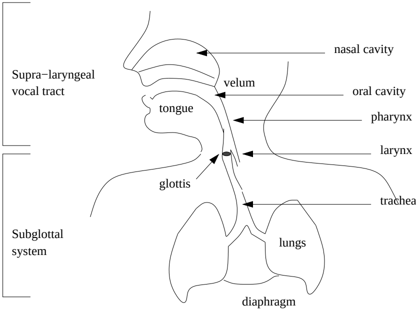
Example: ITU-T G.726 PCM

Bit rate	32 Kbit/s
Sampling freq.	8 KHz
Bits per sample	4 bits

Speech vocoders based on speech production

Physiology of the voice production

- Diaphragm, lungs, and trachea
- Larynx, vocal folds, and glottis
- Pharynx
- Oral and nasal cavities



Speech vocoders based on speech production

(Rough) classification of phonation sounds

- Voiced: vocal folds oscillation produces a periodic excitation source which excites the vocal tract resonances; different VT configurations correspond to different vowels
- Unvoiced: airflow is forced through a vocal tract constriction which can occur in several places between glottis and mouth, producing a noise sound with no harmonic structure

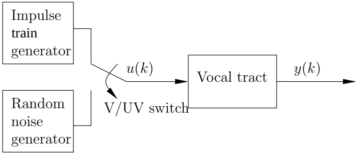
Principal acoustic parameters of voiced sounds

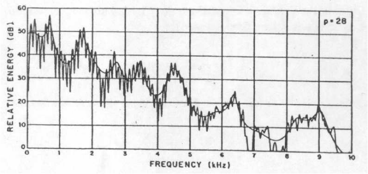
- Pitch: fundamental frequency of excitation source
- Formant frequencies: location of energy emphasys/deemphasys in the spectrum due to vocal tract resonances

LPC based vocoder

LPC Model-based scheme

- Speech production is modelled as a linear filter excited by a switched noise pulse train





LPC based vocoder

LPC analysis

- The signal $y(n)$ can be predicted from its past samples, with a certain error $e(n)$ (prediction error):
  $y(n) = \sum_{i=1}^{N} a_i y(n-i) + e(n)$
- LPC analysis finds the best predictor coefficients a_i
- Prediction error $e(n)$ is interpreted as the voice excitation (i.e., $e(n) = G \cdot u(n)$). Different models can be used (pulse trains, excitation codebooks, etc.)
- Predictor coefficients are encoded through scalar quantization
- Alternative representations of LPC coefficients are preferred (e.g., reflection coefficients or LSPs) since more robust against quantization and interpolation

LPC based vocoder

Example: LPC-10 standard

- Bitrate: 2.4kbps
- The 10 LPC coefficients $\{a_k\}_{k=1}^{10}$ are represented as LSP parameters $\{\omega_k\}_{k=1}^{10}$
- If frame size is 20 msec (50 frames/sec) $\rightarrow$ each frame: 48 bits. A possible allocation:

Parameter Name	Notation	bits per frame
LSP	$\{\omega_k\}_{k=1}^{10}$	34
Gain	G	7
Voiced/Unvoiced and Period	V/UV, T	7

- Gain and LSP coefficients are encoded through scalar quantization

LPC based vocoder

Problems of LPC-based coding schemes

- V/UV classification: gross simplification
- Simplified pulse train excitation: perfectly periodic spectra, innatural speech synthesis

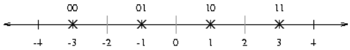
Refined LPC-based coding schemes

- Multi-Pulse Excited (MPE) and Regular-Pulse Excited (RPE) codecs: excitation $u(n)$ is given by a fixed number of non-zero pulses for every speech frame
- MELP: Mixed-excitation linear prediction
- CELP: Code[book]-excited linear prediction uses Vector Quantization (VQ, catalogs of excitation waveforms)

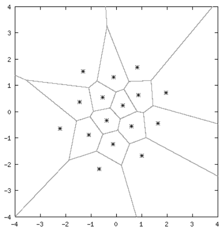
LPC based vocoder

Scalar and Vector Quantization

- Scalar quantization



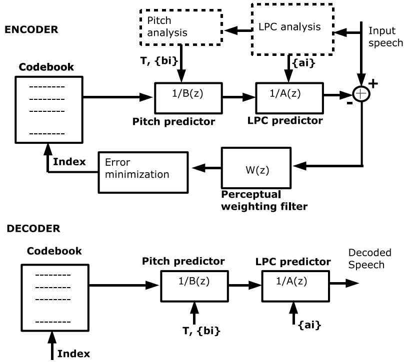
- Vector quantization
  Example of 2D Vector quantization



LPC based vocoder

CELP vocoder

- Excitation is encoded through Vector Quantization technique
- Analysis-by-synthesis scheme to select code words
- A pitch prediction filter is included
- Uses perceptually weighted distortion measure for optimal code word selection



LPC based vocoder

Example: FS-1016 CELP vocoder

- Bitrate: 4.8 kbps
- The 10 LPC coefficients are represented as LSPs
- Frame size is 30 msec (33.3 frames/sec) $\rightarrow$ each frame: 144 bits.
- Frames are further divided into four 7.5 ms sub-frames
- An adaptive codebook is used to model excitation of the LPC filter
- A fixed codebook containing pseudo-random codes is also searched

Parametric vocoders

Other vocoder schemes used in current Standards

- Sub-band coders (SBC): based on filter bank front-end processing. Subbands are encoded separately.
- Sinusoidal analysis/synthesis coders: speech is modeled as sum of time-varying sinusoids
- Multi-Band Excitation (MBE) codecs: adopts an excitation/vocal tract separation model. The excitation spectrum is divided in voiced and unvoiced bands.
- Prototype Waveform Interpolation (PWI): information is sent about a single pitch cycle every 20-30 ms, and interpolation is used to reproduce a smoothly varying quasi-periodic waveform for voiced speech segments.

Speex: an open-source state-of-art speech codec

Speex characteristics

- Free software/Open-source speech codec
- Targeted at VoIP and file-based compression
- Supports narrowband (8 kHz), wideband (16 kHz), and ultrawideband (32-48 kHz) mode.
- Wide range of bit-rates available (2-44 kbps)
- Dynamic bit-rate switching and Variable Bit-Rate (VBR)
- Voice activity Detection (VAD)
- Variable complexity
- Encoding based on CELP scheme

Speech over packet networks

Overview

- Voice over Internet Protocol (VoIP) aims at providing real-time voice communication over packet-switched networks with the quality of circuit-switched network
- Problems due to packet-switched network unreliability: delays, packet losses, out-of-order packets, jitter, echo.

Speech over packet networks

Components of VoIP

- Signaling
  Create and manage connections between endpoints (SS7, SIP, H.323)
- Encoding
  Conversion of voice in digital format
- Transport
  Transportation of voice packets across available network media (UDP, RTP)
- Gateway control
  Interoperation with different IP-based schemes or with PSTN

Speech over packet networks

Voice coders

- Transform, compress and organize voice into packets
- Usually there is a trade-off between voice quality and bandwidth used
- Available vocoders have bitrates ranging form 1.2 to 64 kbps
- Available vocoders have processing delay ranging form 5 msec to 20 msec
- In VoIP, total delay is made up of various components delays in the network

Speech over packet networks

A voice coding standard suited for VoIP: ITU G.729

- Based on CELP and CELP variants
- It offers toll quality speech at a low bit rate of 8Kbps.
- Operates on 10ms frames with short algorithm delays.
- Short-term synthesis filter is a 10th order LP filter.
- Long-term, or pitch synthesis, filter relies on adaptive-code book approach.
- G.729 Annex B provides description of Voice Activity Detection (VAD), Discontinuous Transmission (DTX), and Comfort Noise Generator (CNG) algorithms for transmission rate reduction during silence periods

References

A. S. Spanias, "Speech Coding: A Tutorial Review."*Proceedings of the IEEE*, 82(10):1541–1581, 1994.