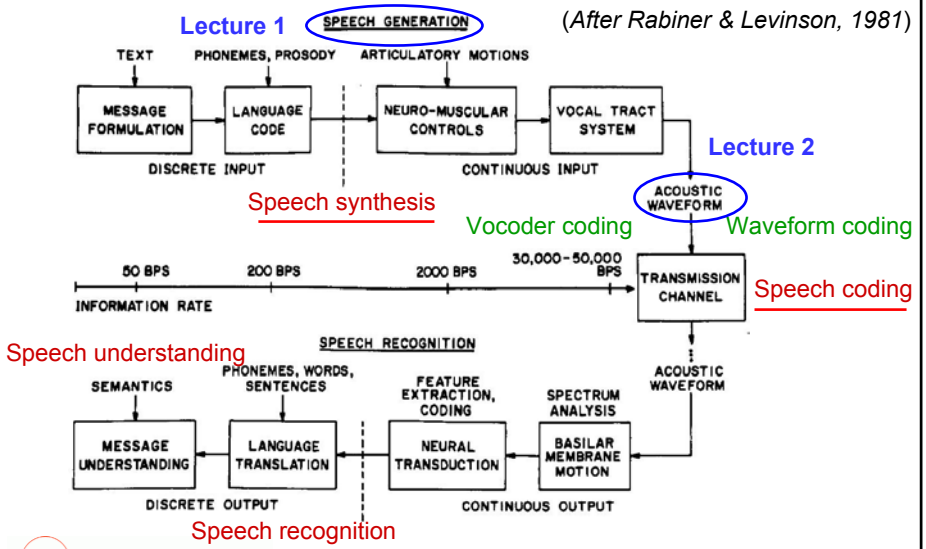
Speech Communication, Spring 2006

Lecture 3: Speech Coding and Synthesis

Zheng-Hua Tan

Department of Communication Technology
Aalborg University, Denmark
zt@kom.aau.dk

Human speech communication process



Part I: Speech coding

- Speech coding
  - Waveform coding
  - Parametric coding (vocoder)
  - Analysis-by-synthesis
- Speech synthesis
  - Articulatory synthesis
  - Formant synthesis
  - Concatenative synthesis

Speech coding

- Definition: analogue waveform → digital form
- Objectives: (for transmission and storage)
  - High compression - reduction in bit rate
  - Low distortion - high quality of reconstructed speech
  - But, the lower the bit rate, the lower the quality.
- Theoretical foundation
  - Redundancies in the speech signals
  - Properties of speech production and perception
- Applications
  - VoIP
  - Digital cellular telephony
  - audio conferencing
  - voice mail

Speech coders

- Waveform coders
  - Directly encode waveforms by exploiting the characteristics of speech signals, mostly (scalar coders) sample-by-sample.
  - High bit rates and high quality
  - Examples: 64kb/s PCM (G.711), 32 kb/s ADPCM (G.726)
- Parametric (voice coder i.e., vocoder) coders
  - Represent speech signal by a set of parameters of models
  - Estimate and encode the parameters from frames of speech
  - Low bit rates, good quality
  - Examples: 2.4 kb/s LPC, 2.4 kb/s MELP
- Analysis-by-synthesis coders
  - Combination of waveform and parametric coders
  - Medium bit rates
  - Examples: 16 kb/s CELP (G.728), 8 kb/s CELP (G.729)

Time domain waveform coding

- Waveform coders directly encode waveforms by exploiting the temporal (time domain) or spectral (frequency domain) characteristics of speech signals.
  - Treats speech signals as normal signal waveforms.
  - It aims at obtain the most similar reconstructed (decoded) signal to the original one.
  - So SNR is always a useful performance measure.
- In the time domain:
  - Pulse code modulation (PCM)
    - Linear PCM, μ-law PCM, A-law PCM
    - Adaptive PCM (APCM)
  - Differential PCM (DPCM)
    - Adaptive DPCM (ADPCM)

Linear PCM

- Analog-to-digital converters perform both sampling and quantization simultaneously.
- Here we analyse the effects of quantization: each sample → a fixed number of bits, *B*.
- Linear PCM
  - *B* bits represent 2^B separate quantization levels
  - Assumption: bounded input discrete signal

$$|x[n]| \le X_{\max}$$

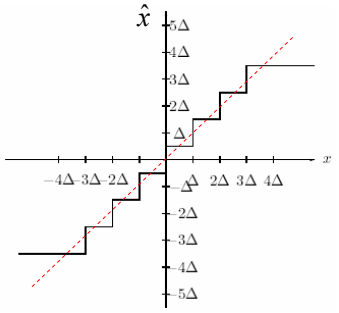
  - Uniform quantization: with a constant quantization step size Δ for all levels x_i

$$x_i - x_{i-1} = \Delta$$

Linear PCM (cont'd)

- Two common uniform quantization characteristics:
  - mid-riser quantizer
  - mid-tread quantizer
- Two parameters for a uniform quantizer:
  - the number of levels $N=2^B$
  - the step size Δ.

$$2 \cdot X_{\max} = \Delta \cdot 2^B$$



Three-bit (N=8) mid-riser quantizer

Quantization noise and SNR

- Quantization noise: $e[n] = x[n] - \hat{x}[n]$

  if $2 \cdot X_{\max} = \Delta \cdot 2^B$, $-\frac{\Delta}{2} \le e[n] \le \frac{\Delta}{2}$

- Variance of e[*n*] which is uniformly distributed.

$$\sigma_e^2 = E[(e[n]-\mu)^2] = E[e^2[n]] = \int_{-\frac{\Delta}{2}}^{\frac{\Delta}{2}} e^2[n] \cdot \frac{1}{\Delta} \cdot de[n] = \frac{\Delta^2}{12} = \frac{X_{\max}^2}{3 \times 2^{2B}}$$

- SNR of the quantization

$$SNR(dB) = 10\log_{10}(\frac{\sigma_x^2}{\sigma_e^2}) = (20\log_{10} 2) \cdot B + 10\log_{10} 3 - 20\log_{10}(\frac{X_{\max}}{\sigma_x})$$

indicating each bit contributes to 6 dB of SNR

11~12-bit PCM achieves 35 dB since signal energy can vary 40 dB

Applications of PCM

16-bit linear PCM

- Digital audio stored in computers: Windows WAV, Apple AIF, Sun AU

- Compact Disc – Digital Audio
  - A CD can store up to 74 minutes of music

  Total amount of data =

  44,100 samples/(channel*second) * 2 bytes/sample * 2 channels * 60 seconds/minute * 74 minutes

  = 783,216,000 bytes

μ-law and A-law PCM

Human perception is affected by SNR →
constant SNR for all quantization levels →
the step size being proportional to the signal value rather than being uniform →
a logarithmic compander

$$y[n] = \ln |x[n]|$$

+ a uniform quantizer on y[n] so that

$$\hat{y}[n] = y[n] + \varepsilon[n]$$

$$\hat{x}[n] = x[n]\exp\{\varepsilon[n]\} \cong x[n](1+\varepsilon[n]) = x[n] + x[n]\varepsilon[n]$$

thus SNR is constant for all levels

$$SNR = \frac{1}{\sigma_\varepsilon^2}$$

μ-law and A-law PCM (cont'd)

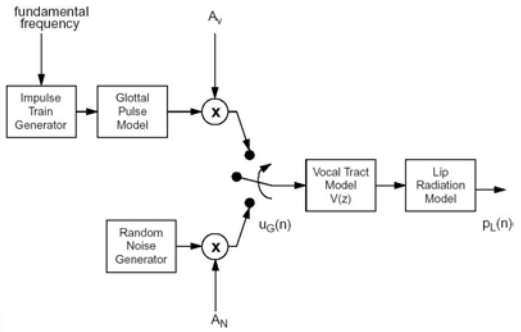
- μ-law approximation

$$y[n] = X_{\max} \frac{\log[1+\mu \frac{|x[n]|}{X_{\max}}]}{\log[1+\mu]} sign\{x[n]\}$$

- A-law approximation
- G.711 standardized telephone speech coding
  - 64 kbps = 8 kHz sampling rate * 8 bits per sample
  - Approximate 35 dB SNR ←→ 12 bits uniform quantizer
  - Whose quality is considered toll and an MOS of about 4.3, a widely used baseline.

Parametric coding (vocoder)

- Are based on the all-pole model of the vocal system
- Estimate the model parameters from frames of speech (speech analysis) and encode the parameters on a frame-by-frame basis
- Reconstruct the speech signal from the model (speech synthesis)



Parametric coding (vocoder) (cont'd)

- Does not require/guarantee similarity in the waveform
- Lower bit rate, but the quality of the synthesized speech is not as good both in clearness and naturalness
- Example – LPC vocoder
  - The source-filter model & LPC vocoder

Source → Filter → Output

Vocal tract

linear predictive coding → an *LPC vocoder*

Analysis-by-synthesis - CELP

- CELP (code excited linear prediction): a family of tech. that quantize the LPC residual using VQ, thus the term *code excited*, in addition to encoding the LPC parameters.

- CELP based standards

	kbps	MOS	Delay
G.728	16	4.0	low
G.729	8	3.9	10ms
G.723.1	5.3/6.3	3.9	30ms
EFR GSM	12.2	4.5	

Speech coders attributes

- Factors: bandwidth (sampling rate), bit rate, quality of reconstructed speech, noise robustness, computational complexity, delay, channel-error sensitivity.
- In practice, coding strategies are the trade-off among them.
- Telephone speech: bandwidth 300~3400Hz, sampled at 8kHz
- Wideband speech is used for a bandwidth of 50-7000Hz and a sampling rate of 16kHz
- Audio coding is used to dealing with high-fidelity audio signals with a sampling rate of 44.1kHz

Mean Opinion Score (MOS)

- The most widely used measure of quality is the Mean Opinion Score (MOS), which is the result of averaging opinion scores for a set of subjects.
- MOS is a numeric value computed as an average for a number of subjects, where each number maps to a subjective quality

excellent	good	fair	poor	bad
5	4	3	2	1

Organisations and standards

- The International Telecommunications Union (ITU)

Standard	Method	Bit rete (kb/s)	MOS	Complexity (MIPS)	Release Time
ITU G.711	Mu/A-law PCM	64	4.3	0.01	1972
ITU G.729	CS-ACELP	8	4.0	20	1996

- The European Telecommunications Standards Institutes (ETSI)

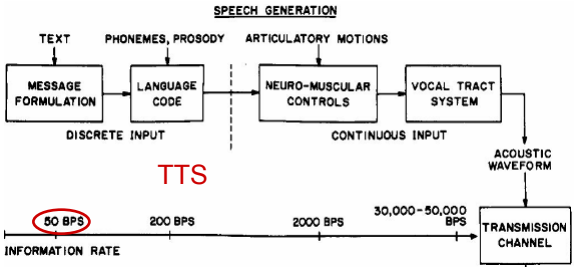
Standard	Method	Bit rete (kb/s)	MOS	Complexity (MIPS)	Release Time
GSM FR	RPE-LTP	13			1987
GSM AMR	ACELP	4.75-12.2			1998

Part II: Speech synthesis

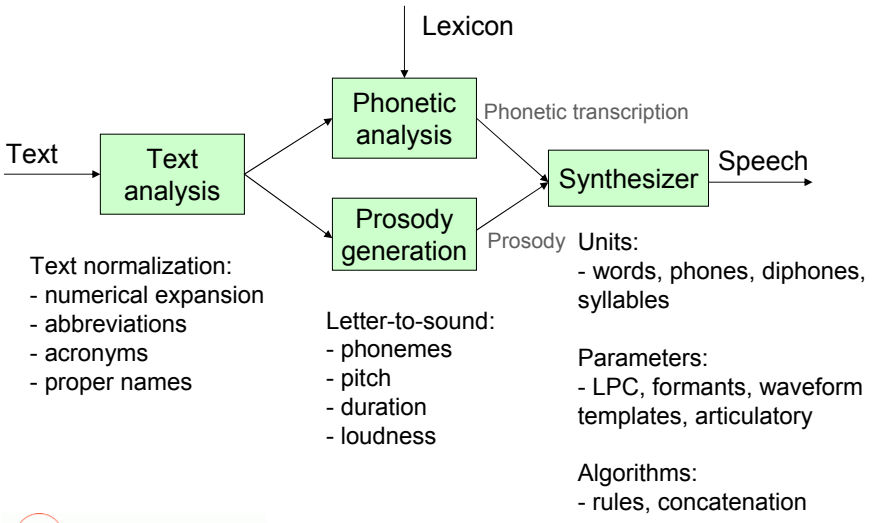
- Speech coding
  - Waveform coding
  - Parametric coding (vocoder)
  - Analysis-by-synthesis
- Speech synthesis
  - Articulatory synthesis
  - Formant synthesis
  - Concatenative synthesis

Text-to-speech (TTS)

- TTS converts arbitrary text to intelligible and natural sounding speech.
- TTS is viewed as a speech coding system with an extremely high compression ratio.
- The text file that is input to a speech synthesizer is a form of coded speech. What is the bit rate?



Overview of TTS



Text analysis

- document structure detection
  - to provide context for later processes, e.g. sentence breaking and paragraph segmentation affect prosody.
  - e.g. email needs special care. This is easy :-) ZT
- text normalization
  - to convert symbols, numbers into an orthographic transcription suitable for phonetic conversion.
  - Dr., 9 am, 10:25, 16/02/2006 (Europe), DK, OPEC
- linguistic analysis
  - to recover the syntactic and semantic features of words, phrases and sentences for both pronunciation and prosodic choices.
  - word type (name or verb), word sense (river or money bank)

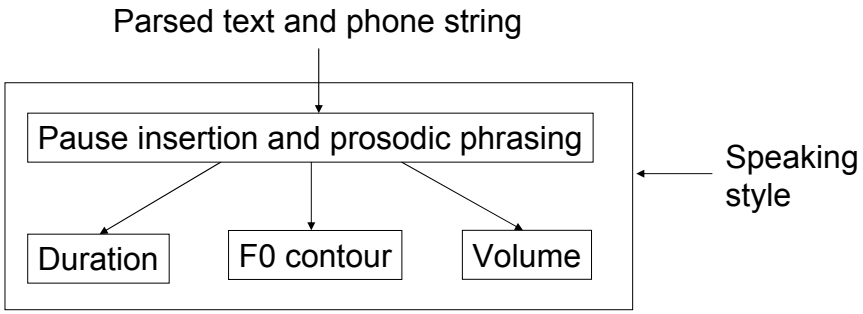
Letter-to-sound

- LTS conversion provides phonetic pronunciation for any sequence of letters.
- Approaches
  - Dictionary lookup
  - If lookup fails, use rules.
    - knight: k -> /sil/ % _n
    - Kitten: k -> /k/
    - Classification and regression trees (CART) is commonly used which includes a set of yes-no questions and a procedure to select the best question at each node to grow the tree from the root.

Prosody

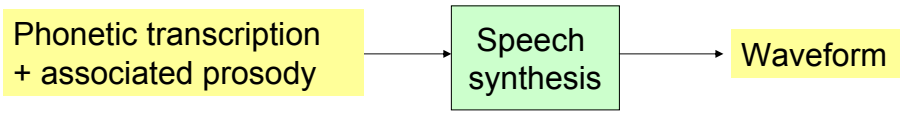
- Pause: indicating phrases and having break
- Pitch: accent, tone, intonation
- Duration
- Loudness

Block diagram of a prosody generation system



Speech synthesis

A module of a TTS system that generates the waveform.



Approaches:

- Limited-domain waveform concatenation, e.g. IVR
- Concatenative systems with no waveform modification, from arbitrary text
- Concatenative systems with waveform modification, for prosody consideration
- Rule-based systems – as opposed to the above data-driven synthesis. For example, formant synthesizer normally uses synthesis by rule.

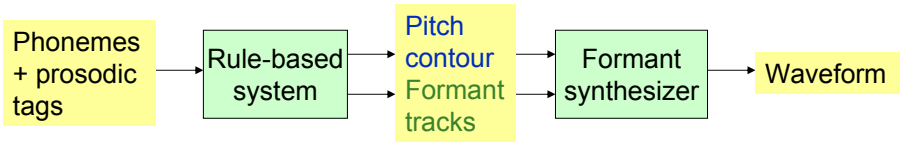
Types according to the model

- Articulatory synthesis
  - uses a physical model of speech production including all the articulators
- Formant synthesis
  - uses a source-filter model, in which the filter is determined by slowly varying formant frequencies
- Concatenative synthesis
  - concatenates speech segments, where prosody modification plays a key role.

Formant speech synthesis

- A type of synthesis-by-rule where a set of rules are applied to decide how to modify the pitch, formant frequencies, and other parameters from one sound to another

- Block diagram



Concatenative speech synthesis

- Synthesis-by-rule generates unnatural speech
- Concatenative synthesis
  - A speech segment is generated by playing back waveform with matching phoneme string.
    - cut and paste, no rules required
    - completely natural segments
  - An utterance is synthesized by concatenating several speech segments. Discontinuities exist:
    - spectral discontinuities due to formant mismatch at the concatenation point
    - prosodic discontinuities due to pitch mismatch at the concatenation point

Key issues in concatenative synthesis

- Choice of unit
  - Speech segment: phoneme, diphone, word, sentence?
- Design of the set of speech segments
  - Set of speech segments: which and how many?
- Choice of speech segments
  - How to select the best string of speech segments from a given library of segments, given a phonetic string and its prosody?
- Modification of the prosody of a speech segment
  - To best match the desired output prosody

Choice of unit

- Unit types in English (After Huang et al., 2001)

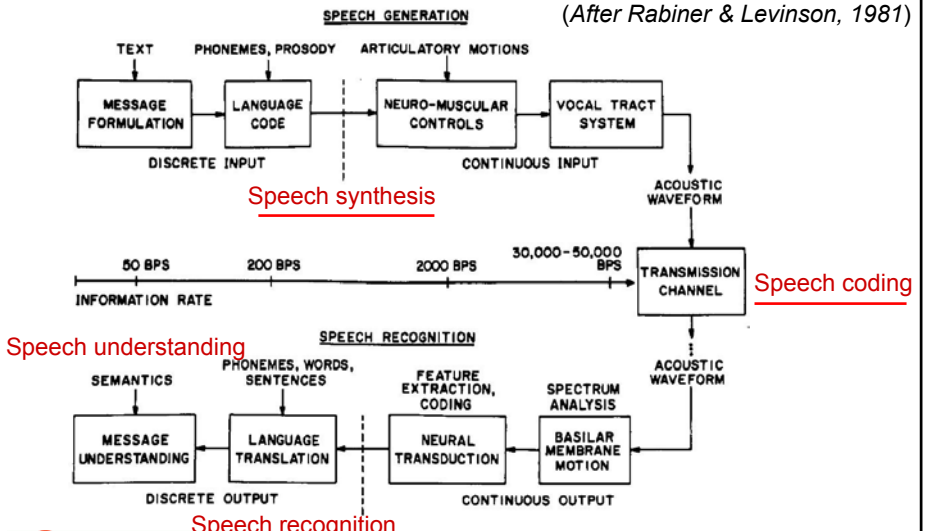
Unit length	Unit type	# units	Quality
Short ↓ Long	Phoneme	42	Low ↓ High
	Diphone	~1500	
	Triphone	~30K	
	Semisyllable	~2000	
	Syllable	~15K	
	Word	100K-1.5M	
	Phrase	∞	
	Sentence	∞	

Attributes of speech synthesis system

- Delay
  - For interactive applications, < 200ms
- Momory resources
  - Rule-based, < 200 KB; Concatenative systems, 100 MB
- CPU resources
  - For concatenative systems, searching may be a problem
- Variable speed
  - e.g., fast speech; difficult for concatenative system
- Pitch control
  - e.g., a specific pitch requirement; difficult for concatenative
- Voice characteristics
  - e.g., specific voices like robot; difficult for concatenative

Difference between synthesis and coding

(*After Rabiner & Levinson, 1981*)



Summary

- Speech coding
- Speech synthesis

- Next lectures: Speech Recognition