L17: Speech synthesis (front-end)

Text-to-speech synthesis

Text processing

Phonetic analysis

Prosodic analysis

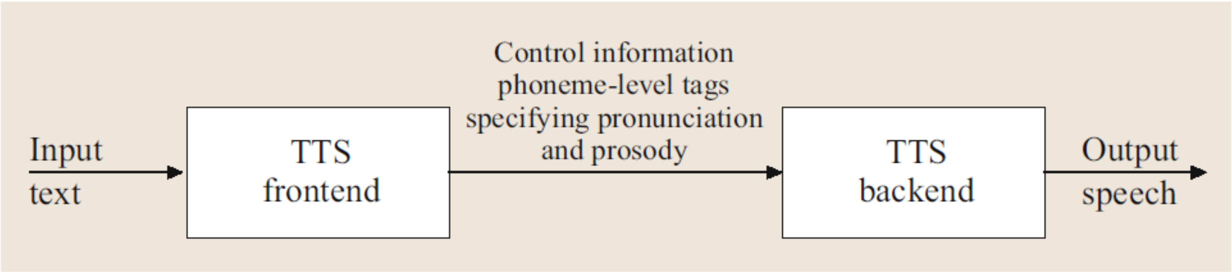
Prosodic modeling

[This lecture is based on Schroeter, 2008, in Benesty et al., (Eds); Holmes, 2001, ch. 7;
van Santen et al., 2008, in Benesty et al., (Eds);]

Text to speech synthesis

Introduction

– The goal of text-to-speech (TTS) synthesis is to convert an arbitrary input text into intelligible and natural sounding speech
  - TTS is not a “cut-and-paste” approach that strings together isolated words
  - Instead, TTS employs linguistic analysis to infer correct pronunciation and prosody (i.e., NLP) and acoustic representations of speech to generate waveforms (i.e., DSP)
  - These two areas delineate the two main components of a TTS system
    – the front-end, the part of the system closer to the text input, and
    – the back-end, the part of the system that is closer to the speech output



[Schroeter, 2008, in Benesty et al., (Eds)]

- **TTS front-end (the NLP component)**
  - Serves two major functions
    - Convert raw text, which may include numbers, abbreviations, etc., into the equivalent of written-out words
    - Assign phonetic transcriptions to each word, and mark the text into prosodic units such as phrases, clauses and sentences
  - Thus, the front-end provides a symbolic linguistic representation of the text in terms of phonetic transcription and prosody information
- **TTS back-end (the DSP component)**
  - Often referred to as the "synthesizer," the back-end converts the symbolic linguistic representation into sounds
  - A number of synthesis techniques exist, including
    - Formant synthesis
    - Articulatory synthesis
    - Concatenative synthesis
    - HMM-based synthesis

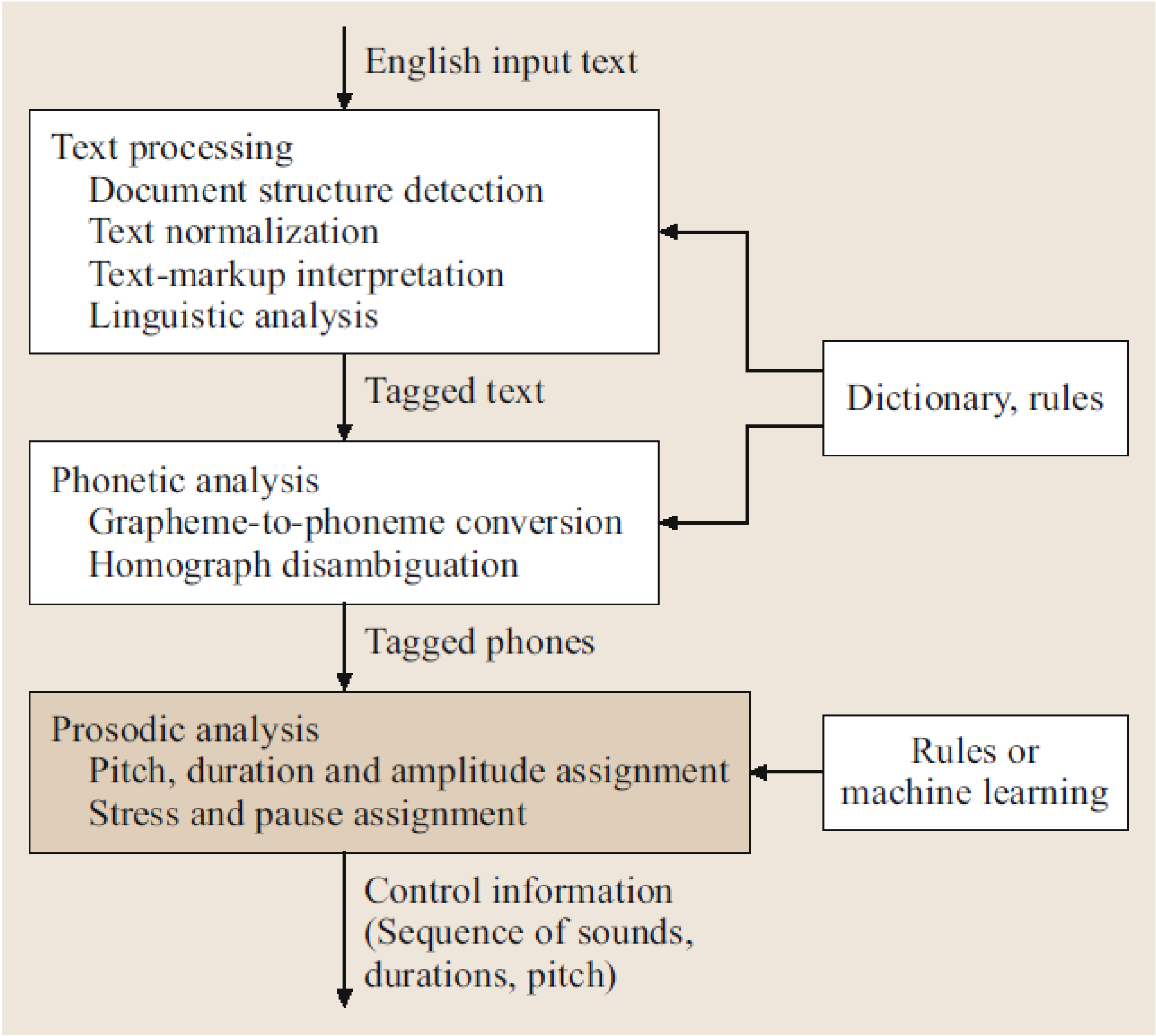
http://en.wikipedia.org/wiki/Speech_synthesis

Components of a front-end

- Text processing
  - Responsible for determining all knowledge about the text that is not specifically phonetic or prosodic
- Phonetic analysis
  - Transcribes lexical orthographic symbols into phonemic representations, maybe also diacritic information such as stress placement
- Prosodic analysis
  - Determines the proper intonation, speaking rate and amplitude for each phoneme in the transcription

- Proper treatment of these topics would require a separate course
  - Here we just provide a brief overview of the different steps involved in transforming text inputs into a representation that is suitable for synthesis

Tasks and processing in a TTS front-end



[Schroeter, 2008, in Benesty et al., (Eds)]

Text processing

Purpose

- Text processing is responsible for determining all knowledge about the text that is not specifically phonetic or prosodic
  - In its simplest form, text processing does little more than converting non-orthographic items (e.g., numbers) into words
  - More ambitious systems attempt to analyze white spaces and punctuations to determine document structure

Tasks

- Document structure detection
  - Depending on the text source, may include filtering out headers (e.g., in email messages)
  - Tasks are simplified if document follows the standard generalized markup language (SGML), an international standard for representing e-text
- Text normalization
  - Handles abbreviations, acronyms, dates, etc. to match how an educated human speaker would read the text
    - Examples: ‘St.’ can be read as ‘street’ or as ‘saint’, ‘Dr.’ as ‘drive’ or ‘doctor’, spelling out ‘IBM’ or ‘MIT’ but not ‘NASDAQ’ or ‘NATO’

- Text markup interpretation
  - Can be used to control how the TTS engine renders its output
    - Examples: using ‘address mode’ for reading a street address, rendering sentences with various emotions (e.g., angry, sad, happy, neutral)
  - Easier if text follows the speech synthesis markup language (SSML)
- Linguistic analysis (a.k.a. syntactic and semantic parsing)
  - May include tasks such as determining parts-of-speech (POS) tags, word sense, emphasis, appropriate speaking style, and speech acts (e.g., greetings, apologies)
    - Example: in order to accentuate the sentence ‘*They can can cans*’ it is essential to know that the first ‘can’ is a function word, whereas the second and third are a verb and a noun, respectively
  - Most TTS systems forego fully parsing the input text in order to reduce computational complexity and also because text input oftentimes consists of isolated sentences or fragments

Phonetic analysis

Purpose

- Phonetic analysis focuses on the phone level within each word, tagging each phone with information about what sound to produce and how to produce it

Tasks

- Morphological analysis
  - Analyzes the component morphemes of a word (e.g., prefixes, suffixes, stem words)
    - Example: the word 'antidisestablishmentarianism' has six morphs
  - Decomposes inflected, derived and compound words into their elementary graphemic units (their morphs)
    - Rules can be devised to correctly decompose the majority of words (about 95% of those in a typical text) into their constituent morphs
  - Why morphological analysis?
    - A high proportion of English words can be combined with prefixes and/or suffixes to form other words, and the pronunciation of the derived words are closely related to that of their roots

- Homograph disambiguation
  - Disambiguates words with different senses to determine pronunciations
    - Examples: ‘object’ (verb/noun), ‘resume’ (verb/noun), ‘contrast’ (verb/noun), ‘read’ (present/past)…
- Grapheme to phoneme (G2P) conversion
  - Generates a phonemic transcription of a word given its spelling
  - Two approaches are commonly used for G2P conversion
    - Letter-to-sound rules (LTS)
    - Lookup dictionaries (Lexicon)
  - LTS rules are best suited for languages with a relatively simple relation between orthography and phonology (e.g., Spanish, Finnish)
  - Languages like English, however, generally require a lexicon to achieve highly accurate pronunciations
    - The lexicon should at least include words whose pronunciation cannot be predicted from general (LTS) rules
    - Words not included in the lexicon are then transcribed through LTS rules
    - LTS rules may be learned by means of classification and regression trees

Prosodic analysis

Purpose

- Prosodic analysis determines the progression of intonation, speaking rate and loudness across an utterance
- This information is ultimately represented at the phoneme level as
  - amplitude
  - duration, and
  - pitch (F0)

Roles of prosody in language

- In the case of tonal languages, pitch is used to distinguish lexical items
- Prosody helps structure an utterance in terms of phrases, and indicates relationships between phrases in utterances
- Prosody helps focus attention on certain words
  - Highlight a contrast (contrastive stress)
  - Emphasize their importance
  - Enhance the intelligibility of words that may be unpredictable from their context

Loudness/intensity

- Mainly determined by phone identity
  - e.g. voiceless fricatives are weak, most vowels are strong
- However, loudness also varies with stress
  - e.g., stressed syllables are normally a little louder
- It is fairly easy to include rules to simulate these effects
- The effect of loudness is not critical in the synthesized speech (when compared to pitch and duration) and most TTS system ignore it

Duration

- The second most important prosodic element, it helps with
  - Stress: phones become longer than normal
  - Phrasing: phones get noticeably larger prior to a phrase break
  - Rhythm
- Properties
  - Intrinsic duration vary considerably between phones, e.g. ‘bit’ vs. ‘beet’
  - Durations is affected by speaking rate, by steady sounds (vowels, fricatives), which vary more than transient sounds (stops)
  - Duration depends on neighboring phones: e.g., vowels before voiced Cs (‘feed’) are longer than before unvoiced Cs (‘feet’)
  - Other rules include
    - If a word is emphasized, its most prominent syllable is normally lengthened
    - At the end of a phrase syllables tend to be longer than in other positions

Pitch

- The most important prosodic element
- As with duration, some general rules are known
  - F_0 contours typically show maxima closed to stress syllables
  - There is generally a globally downward trend of the F_0 contour over the duration of a phrase
  - Trend is reversed for the final syllable in yes/no questions or in non-terminal phrases, but further accelerates downward in terminal phrases
- Pitch is a controversial topic with many different schools of thought
  - British school: evolved from old style prescriptive linguistics, concerned with teaching 'correct' intonation to non-native speakers
  - Autosegmental-metrical school: seeks to provide a theory of intonation that work cross linguistically
  - Fujisaki model: aimed to follow known biological production mechanisms
  - Tilt model: built purely for engineering purposes

Prosodic models

History of prosodic models

- Rule-based approaches
  - Developed during the period of formant synthesizers
  - Models employ a set of rules derived from experiments or the literature
  - Examples
    - Duration: Klatt's model, used for the MITTalk system
    - Intonation: Pierrehumbert's model, which is the basis for ToBI
- Statistical approaches
  - Developed during the period of diphone synthesizers
  - Examples:
    - Duration: sums-of-products model of van Santen
    - Intonation: tilt model of Taylor
- Use as-is approaches
  - Developed with unit-selection systems
  - Approach is to use a large corpora of natural speech to train prosodic models and serve as a source of units for synthesis
    - Instead of having one token per diphone, corpus contains several tokens with different phonetic and prosodic context characteristics

Klatt's duration model

– The model assumes that
  - Each phonetic segment has an inherent duration
  - Each rules tries to effect a % increase or decrease in the phone's duration
  - Segments cannot be compressed beyond a certain minimum

$$Dur = MinDur \frac{(InhDur - MinDur)Perc}{100}$$

  – where $Perc$ is determined according to 10 different rules that take into consideration the phonetic environment, emphasis, stress level, etc.

  – Each rule produces a separate $Perc$, which are then combined multiplicatively

  - However, the model does not account for interactions between rules

Other duration models

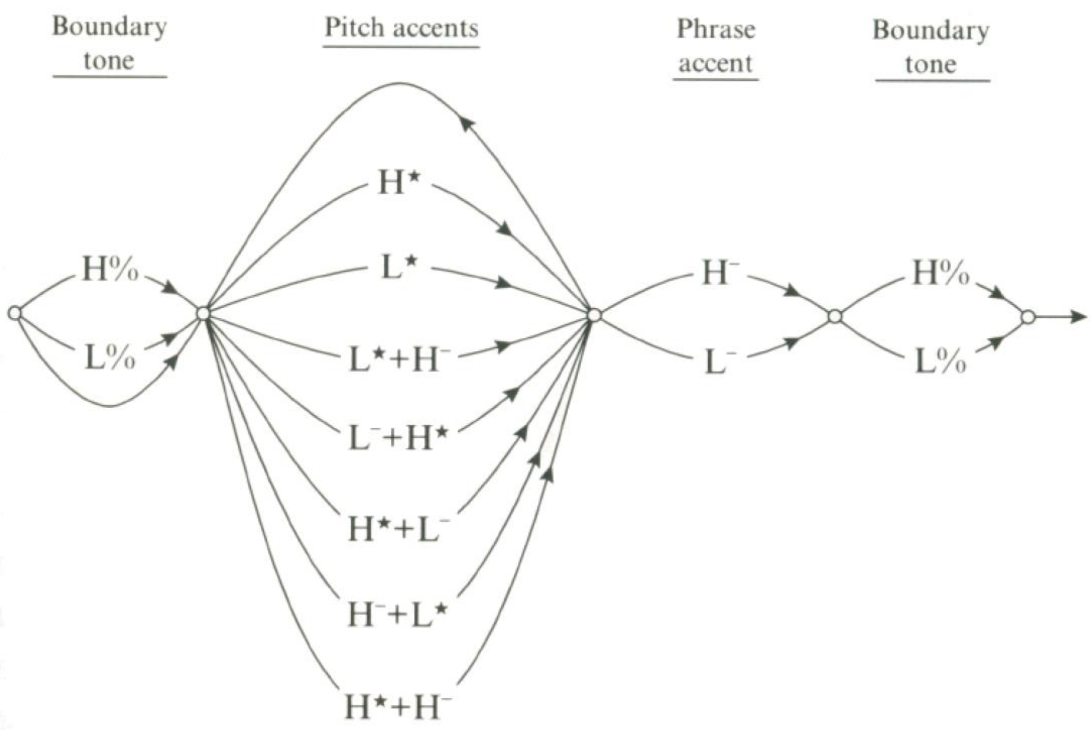
– CART-based models (used in Festival)

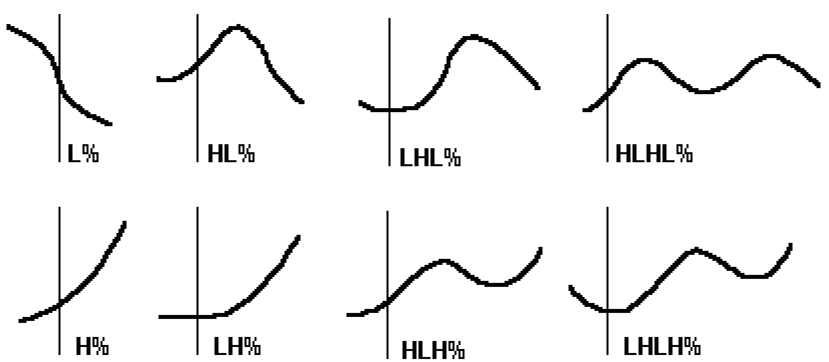
– Neural-network-based models (Campbell)

– Sums-of-products (van Santen)

Pierrehumbert's intonation model

- Considers intonation to be a sequence of high (H) and low (L) tones
- The H and L tones are the building blocks for three larger tone units
  - Pitch accents, used to mark prominence
    - Can be single tones (H*,L*) or pairs of tones (L+H*, L*+H,H*+L,H+L*), where the asterisk (*) denotes alignment with the stressed syllable
  - Phrase accents, link the last pitch accent to the phrase boundary
    - Denoted by (L-,H-)
  - Boundary tones, determine the boundary of intonational phrases
    - These are represented by (%H,%L,H%, L%), where the % denotes the alignment of the boundary tone with the onset or offset of the intonation
- Pierrehumbert's theory of intonation led to the ToBI (tones and break indices) prosody annotation standard
  - ToBI is just a labeling system, but does not provide F_0 contours
  - Several methods have been developed to convert ToBI labels into actual F_0 contours



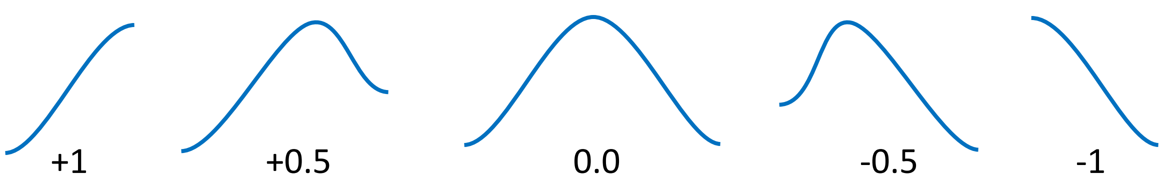


http://www.linguistics.ucla.edu/people/jun/ktobi/K-tobi.html

Tilt model

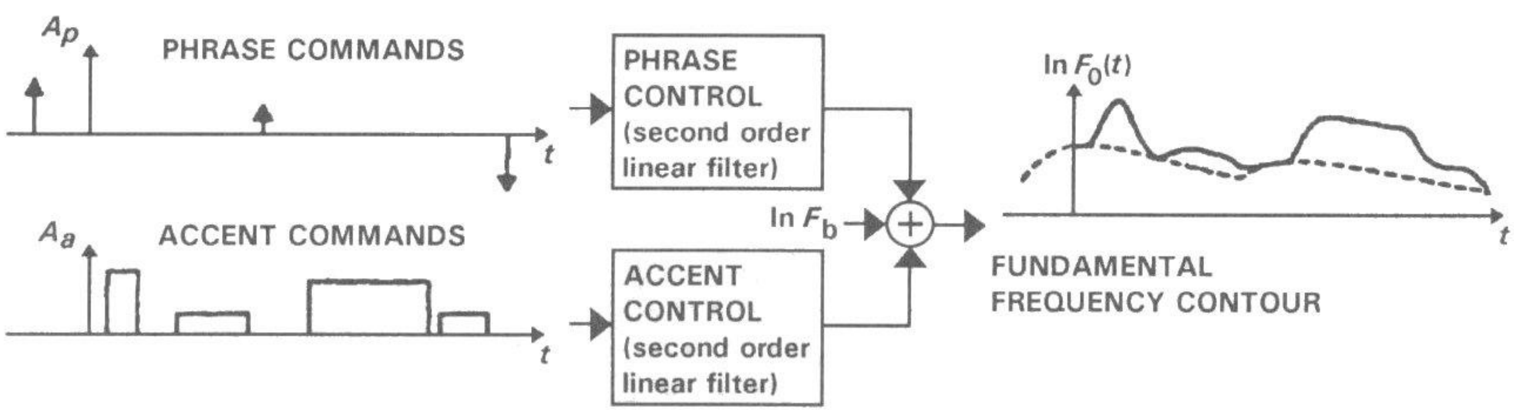
- Developed explicitly as a practical engineering model of intonation
- Considers intonation to be a sequence of four types of events
  - Pitch accents, boundary tones, connections, and silences
- Pitch accents and boundary tones are modeled by piece-wise combinations of parameterized quadratic functions (rising or falling)
  - Connections are modeled by straight-line interpolations
- Amplitude and duration of these functions are defined by three parameters

$$tilt_{amp} = \frac{|A_{rise}| - |A_{fall}|}{|A_{rise}| + |A_{fall}|}; \quad tilt_{dur} = \frac{|D_{rise}| - |D_{fall}|}{|D_{rise}| + |D_{fall}|}; \quad tilt = \frac{tilt_{amp} + tilt_{dur}}{2}$$



Fujisaki's intonation model

- Considers the $\log F_0$ contour to be the addition of two components
  - A phrase command
    - Characterizes the overall trend of the intonation
    - Modeled by pulses, placed at intonational phrase boundaries
  - An accent command
    - Highlights extreme excursions (e.g. for stressed syllables)
    - Modeled by step functions, placed around accent groups



[Holmes, 2001]