“From Text to speech: the MITalk system”

Jonathan Allen, M. Sharon Hunnicutt and Dennis Klatt

with Robert C. Armstrong and David B. Pisoni
Overview

- TTL and software
  - Implementation: 5 pages
    - comprehension
  -  
    - phoneme recognition, words recognition
    - intelligibility, comprehension
  - Evaluation: 16 pages

- Klett formant synthesizer
  - (durations, F0, phonetics
  - overview, phonology, prosody
  - Synthesizer: 79 pages

- Letter-to-sound and lexical stress
  - phrase-level parser, morphophonemics
  - Text pre-processing, morphological analysis
  - Analysis: 59 pages

- History of early TTS systems
  - Overview: 12 pages
Analysis

- morphophonemics (post-lexical rules)
- stress assignment
- morphological decomposition
- exception list
- context-dependent LTS rules
- There was no large lexicon:
  - abbreviations etc.
  - finite list of all text processing issues
- Smooth between them
- Generate formant values
- Post prosody so can care about timing

Phonetic component:

- Sentence, clause, word contours with modification
- F0 contour generator (O'Shaughnessy)
- What we not train with CART

Prosodic component:

- Post lexical rules, co-articulation effects
- Phones, stress, metrical boundaries

Phonological component:

Synthesis
bypass path amplitude in dB
sixth parallel formant amplitude in Hz
fifth parallel formant amplitude in Hz
fourth parallel formant amplitude in Hz
third parallel formant amplitude in Hz
second parallel formant amplitude in Hz

third formant bandwidth in Hz
second formant bandwidth in Hz
first formant bandwidth in Hz

nasal zero frequency in Hz
tenth formant frequency in Hz
fourth formant frequency in Hz
third formant frequency in Hz
second formant frequency in Hz
first formant frequency in Hz

amplitude of sinusoidal vocing in dB
amplitude of aspiration in dB
amplitude of aspiration in dB
amplitude of friction in dB
amplitude of voicing in dB
The Kalt Formant Synthesizer

- Fujimura, Bell Labs
  - Lots of early work in synthesis
- Honda, Bateman etc
  - Articulatory/Speech Production
  - Tokuda: HMM Generation
  - Hoste: prosody synthesis
  - Periodic/White voices (voiced/unvoiced)
  - Sounds sources
  - Cascade vs parallel generation
Each requires

- Others □
- Nasal parameters □
- Glottal resonator □
- Formants (Hz, amplitude, bandwidth) □
- Vocalic/Phonation/Auspilation □

The Klatt Formant Synthesizer Parameters
How do you find the parameters?

- Can we derive this for a large data set?
- Are these the right parameters?
- Try all until you find them
- Analyses/Syntheses?
- By expert training?
Intelligibility and Comprehension