TTS−System “Papageno”

Overview

- Scaleable speech synthesis engine
- Unified kernel for embedded and server solutions
- Covers applications from 5 MB to 100 MB footprint
- Multi−lingual system due to separation of engine and speaker/language dependent knowledge bases
- 3 main parts:
  - Preprocessing: tokeniser, tagger, grapheme−to−phoneme
  - Prosody: symbolic and acoustic part
  - Acoustic: unit selection and concatenation

Knowledge bases – Tokeniser

- Language dependent
- Different types of characters:
  - Separators: space, tabulator, new line …
  - Numbers: digits, floating point (comma), ordinal point (th, nd, rd, th), dates (colon, point), relations (slash, colon) …
  - Words: graphemes, dashes, points (abbreviations)
  - Punctuation marks: brackets, points, comma, slashes …
- Training material for handling of numbers (date, time, ordinal – cardinal, relations)
- Handling of abbreviations, acronyms

Knowledge bases – Tagger

- Language dependent
- Part−of−speech (POS) tagging
- Additional features (gender, number, case, definiteness), esp. for German ordinal numbers “der 1. (erste) Versuch" – "ein 1. (erster) Versuch"
- Requirements: lexicon and tagged text
- Features as detailed as possible
- Training of neural network or n−gram

Knowledge bases – G2P

- Language dependent
- Pronunciation of words, abbreviations, acronyms and numbers
- Requirements: pronunciation dictionary
- Features:
  - Huge dictionary
  - Special handling of foreign words (different phoneme sets, mapping between the sets)
  - List of proper names (or special label in the dictionary)
  - Distinguishing features for homographs (POS, semantic information)
- List of syllable cores and stressable phonemes
- G2P of words by neural networks (first : phoneme sequence incl. syllable boundaries; second: word stress)
- Pronunciation of numbers by graphs trained on examples
- Abbreviations and acronyms with dictionaries

Knowledge bases – symbolic prosody

- Speaker dependent (but may be used for other speakers too)
- Position of phrase accents and breaks
- Requirements: labelled text
- Features: one break (B3) and one accent level (PA)
- Labels manually by ONE expert, given the acoustic representation of the text
- inter− and intra−labeller variance too big
- Training of neural networks (first: position of breaks; second: position of accent within one phrase)

Knowledge bases – text

- Should contain every type of sentences (declaration, question, exclamation)
- Long and short sentences
- Consistency between written and spoken text:
  - Abbreviations
  - Cardinal numbers (100 can be “hundert” or “einhundert”)
  - ordinal numbers in graphemic form (“erstens” instead of “1.”)
  - For homographs different entries in the dictionary (“modern_adj” for modern and “modern_verb” for molder)
- Consistent file names and extensions