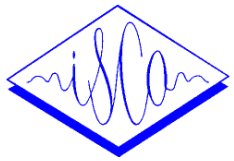


International Symposium on Chinese Spoken Language Processing (ISCSLP 2002)
Taipei, Taiwan
August 23-24, 2002

Concatenative Chinese speech synthesis and quality evaluation

Li Haizhou, PhD

Vice President, InfoTalk

haizhou.li@infotalkcorp.com

Infotalk Corp.

23 Aug, 2002

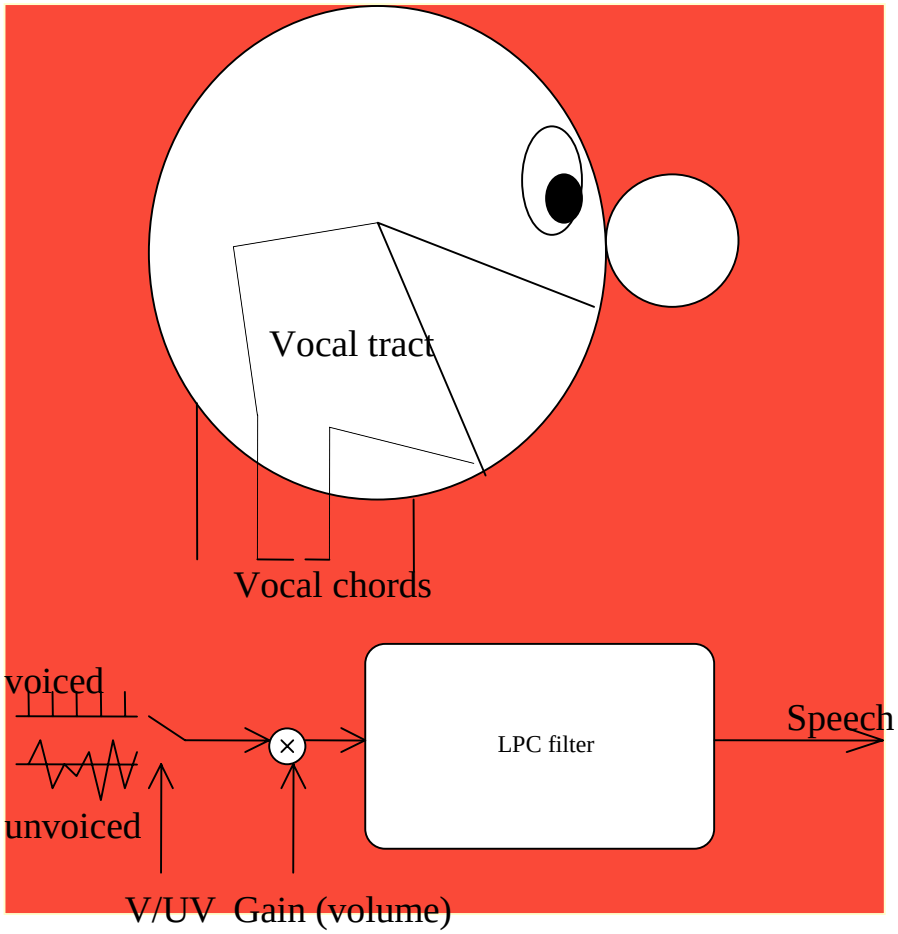
Content

1. Introduction
2. Linguistic processing
3. Prosody generation
4. Speech synthesis
5. Effective objective measure
6. Summary

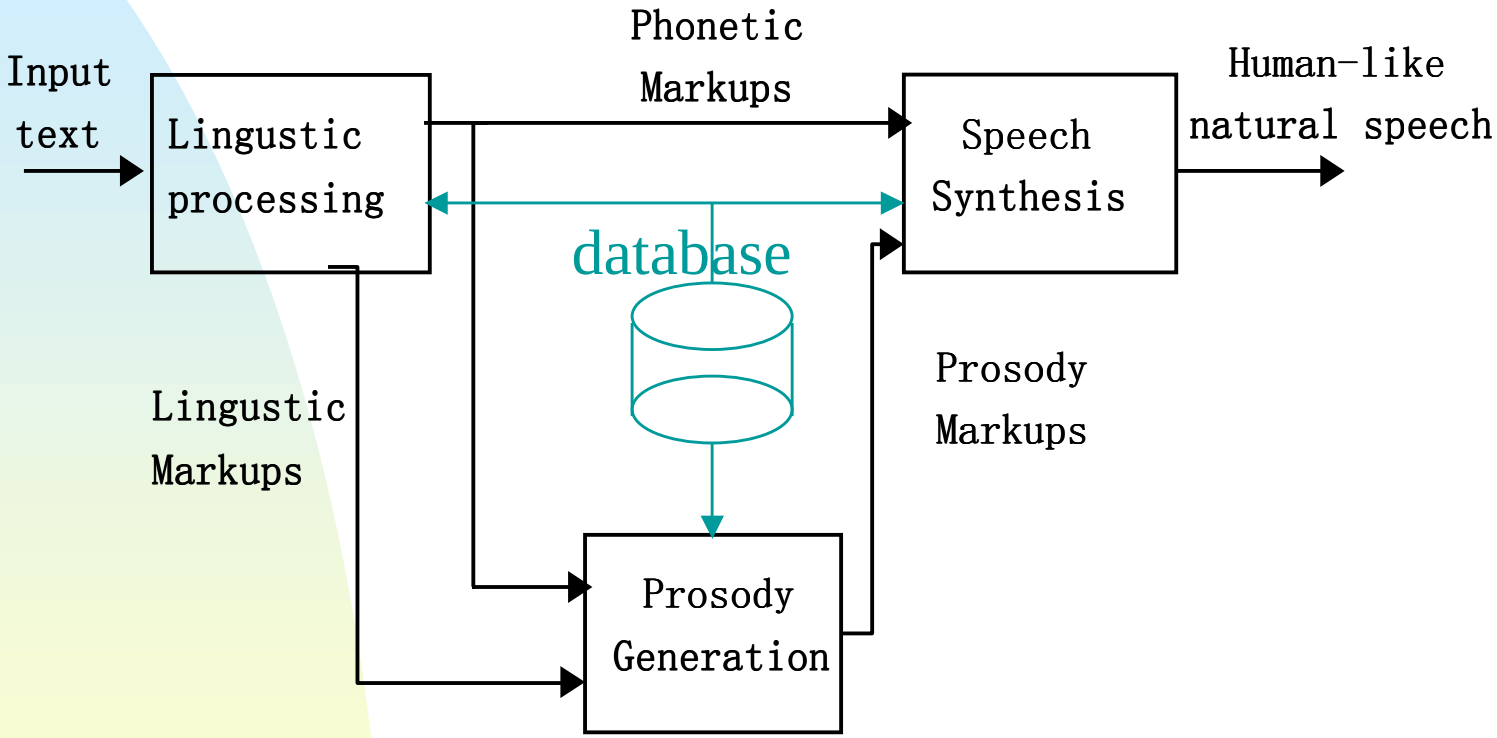
Content

1. Introduction
2. Linguistic processing
3. Prosody generation
4. Speech synthesis
5. Effective objective measure
6. Summary

1.1 The LPC module



1.2 Concatenative module



1.3 Successful factors

- Increasing computing power
- Efficient of ASR search technique
- Automatic labeling of speech corpus
- Effective objective measure

1.4 Establishment of voice library

- Balanced speech corpus
  - Position in phrase, phonetic context and tonal context
- Automatic phonetic labeling
- Glottal voice alignment
- Effective inventory management
- Number of syllables: ~2000
- Voice inventory per syllable: ~150

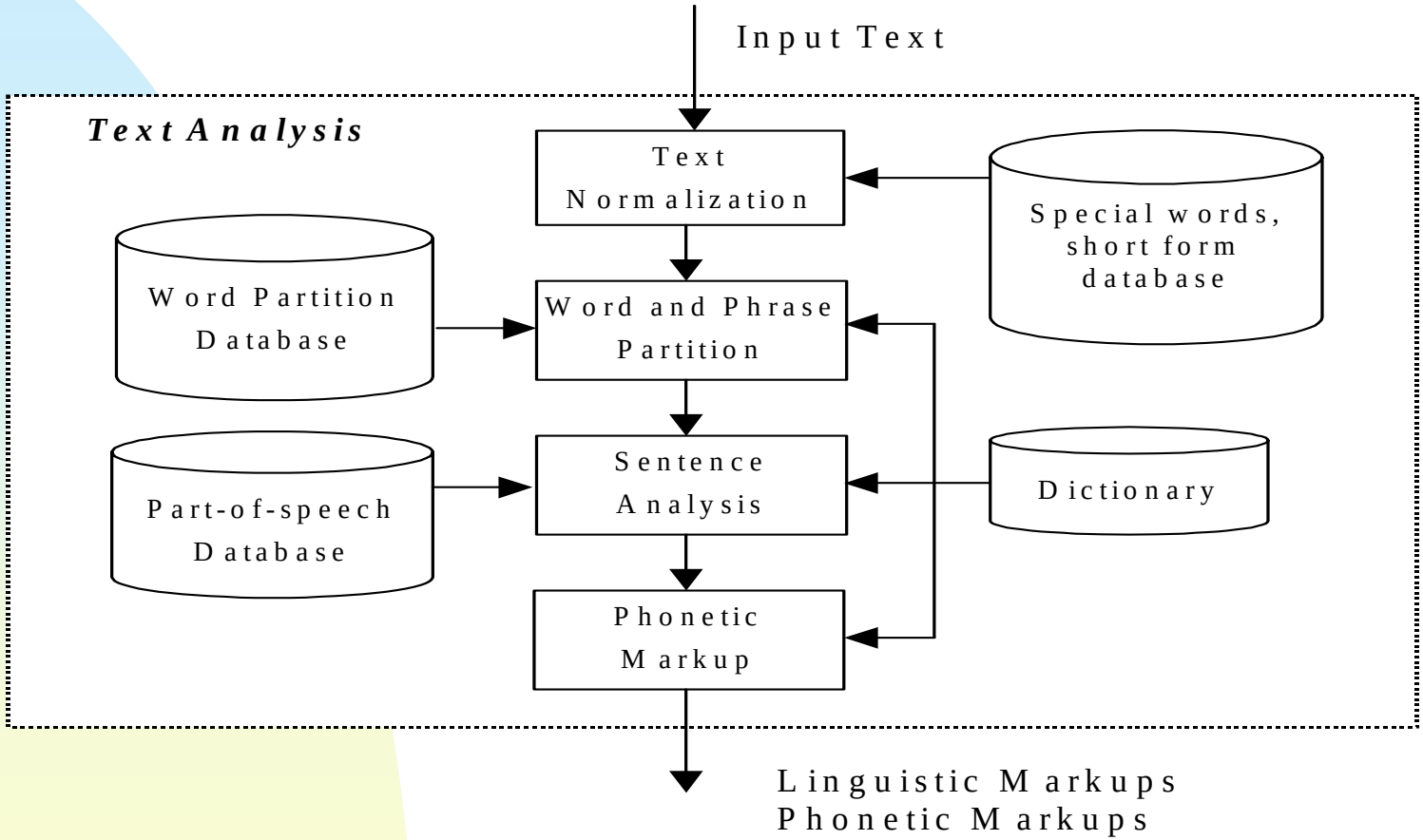
Content

1. Introduction
2. Linguistic processing
3. Prosody generation
4. Speech synthesis
5. Effective objective measure
6. Summary

2.1 Linguistic processing task

- Input: orthographic message (e.g. sentence, word, letter)
- Processing:
  - Text normalization
  - Word and phrase segmentation
  - Sentence analysis
- Output: annotated phonetic message

2.2 Linguistic processing module



2.3 Text-normalization

- Convert a text into a phonetic transcription
  - Text: link with actual pronunciation may be vague or absent
  - Phonetic transcription: unambiguous representation of sounds
- Examples
  - 晚間最低氣溫-10℃。 
  - Apec 
  - 手機13901297788 

2.4 Word and Phrase Segmentation

- No space between words or phrases
- The process of phrase segmentation is to markup the boundary of words and phrases
- Examples
  - 患有嚴重口吃的小明坐在廟口吃小吃。
  - 我這次得到美國方面的鼎力相助。我這次得到美國去出差。

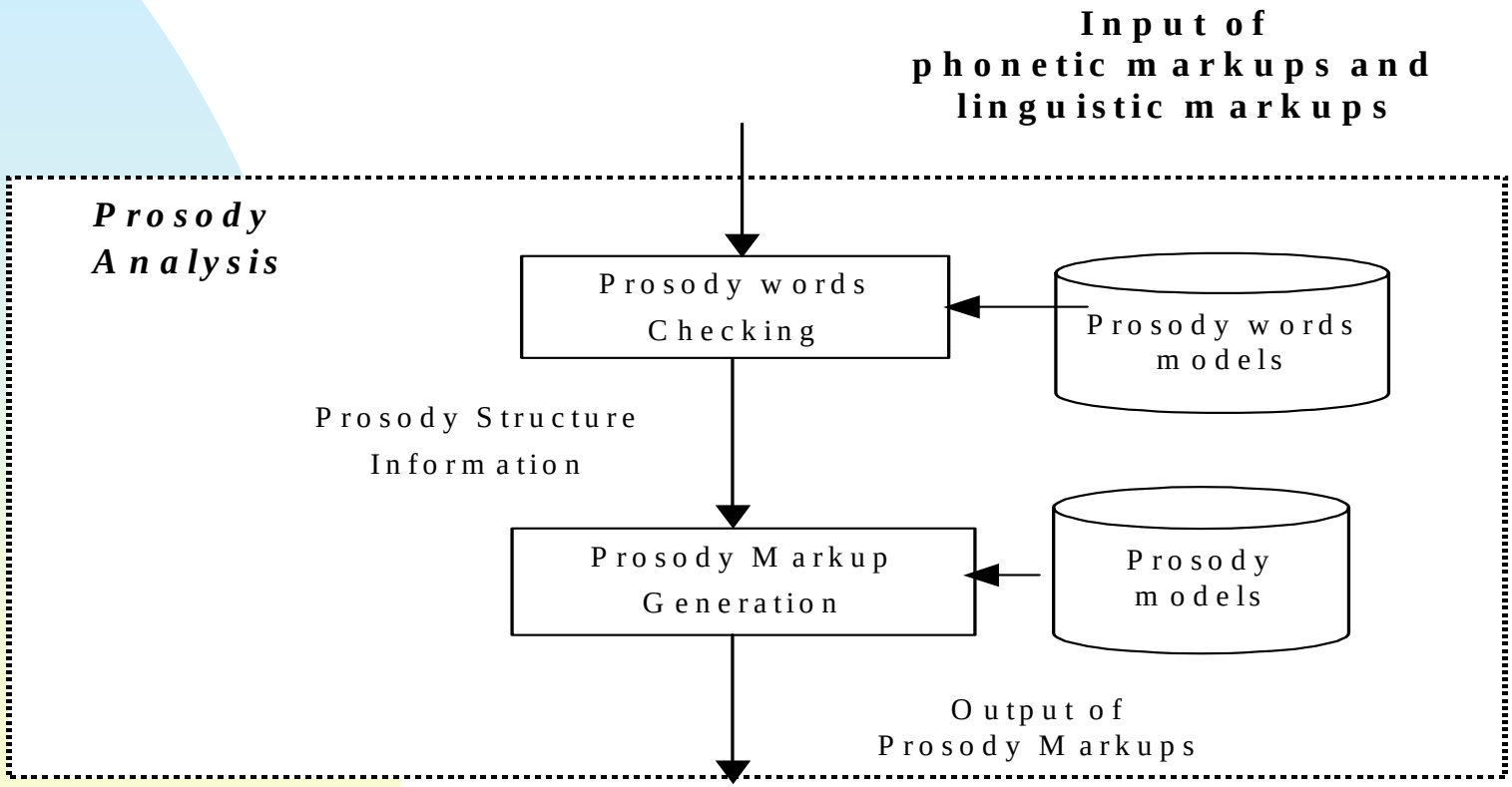
2.5 Sentence Analysis

- Part-of-speech annotation
- Examples
  - 他的競選口號爲 "爲社會除害"。 
  - 曾叔叔曾到本所訪問。 

Content

1. Introduction
2. Linguistic processing
3. Prosody generation
4. Speech synthesis
5. Effective objective measure
6. Summary

3.1 Prosody module



3.2 Prosody words

- Input Text
  - 请你给我一杯水
- Text Partition
  - 请 | 你 | 给 | 我 | 一 | 杯 | 水
- Prosody words
  - 请你 | 给我 | 一杯水

3.3 Prosody modeling

- Data-driven modeling techniques
  - Factors of consideration
    - Pitch
    - Duration
    - Energy
  - Data mining based on annotation of relevant linguistic features
- Modeling ↔ Transplantation
  - flexibility ↔ naturalness

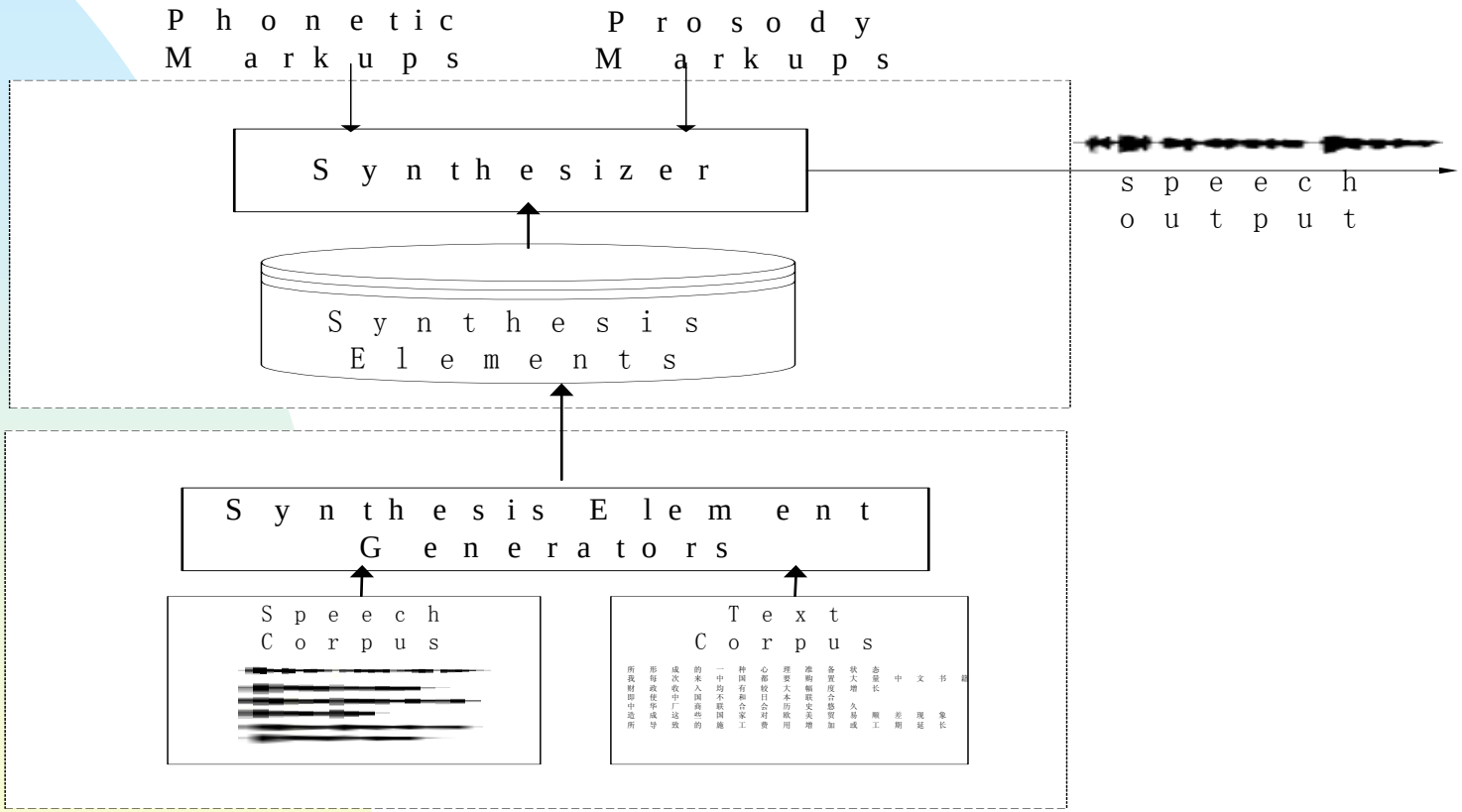
3.4 Prosody generation

- Supra-segmentals structure the speech flow:
  - expression level: encoding of linguistic intentions
  - perception level: stress, intonation, tone, rhythm, speech rate, loudness
  - acoustic level: fundamental frequency, intensity, duration
- Importance of prosody in TTS:
  - Intelligibility
  - Naturalness
- Demonstration:

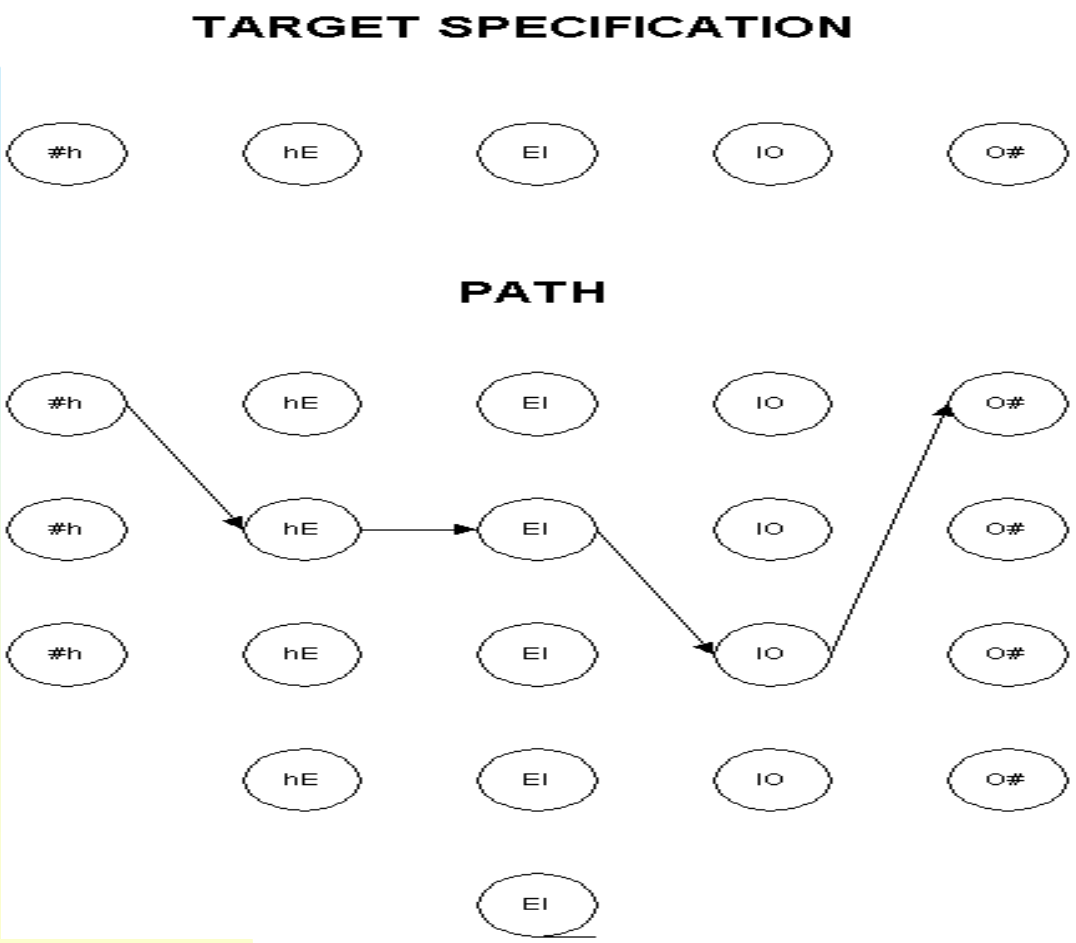
Content

1. Introduction
2. Linguistic processing
3. Prosody generation
4. Speech synthesis
5. Effective objective measure
6. Summary

4.1 Synthesis module



4.2 Unit selection



4.3 Cost Functions

- Candidate to Target matching
  - Node Cost (unigram)
  - Sum of contributions from various symbolic and acoustic features
  - Typically ~ 20 features per unit
- Candidate to Candidate matching
  - Transition Cost (bigram)
  - Numerical match between properties of untis at boundaries
  - Typically ~ 5 features at unit boundary
- Balanced match

Content

1. Introduction
2. Linguistic processing
3. Prosody generation
4. Speech synthesis
5. Effective objective measure
6. Summary

5.1 Quality assessment

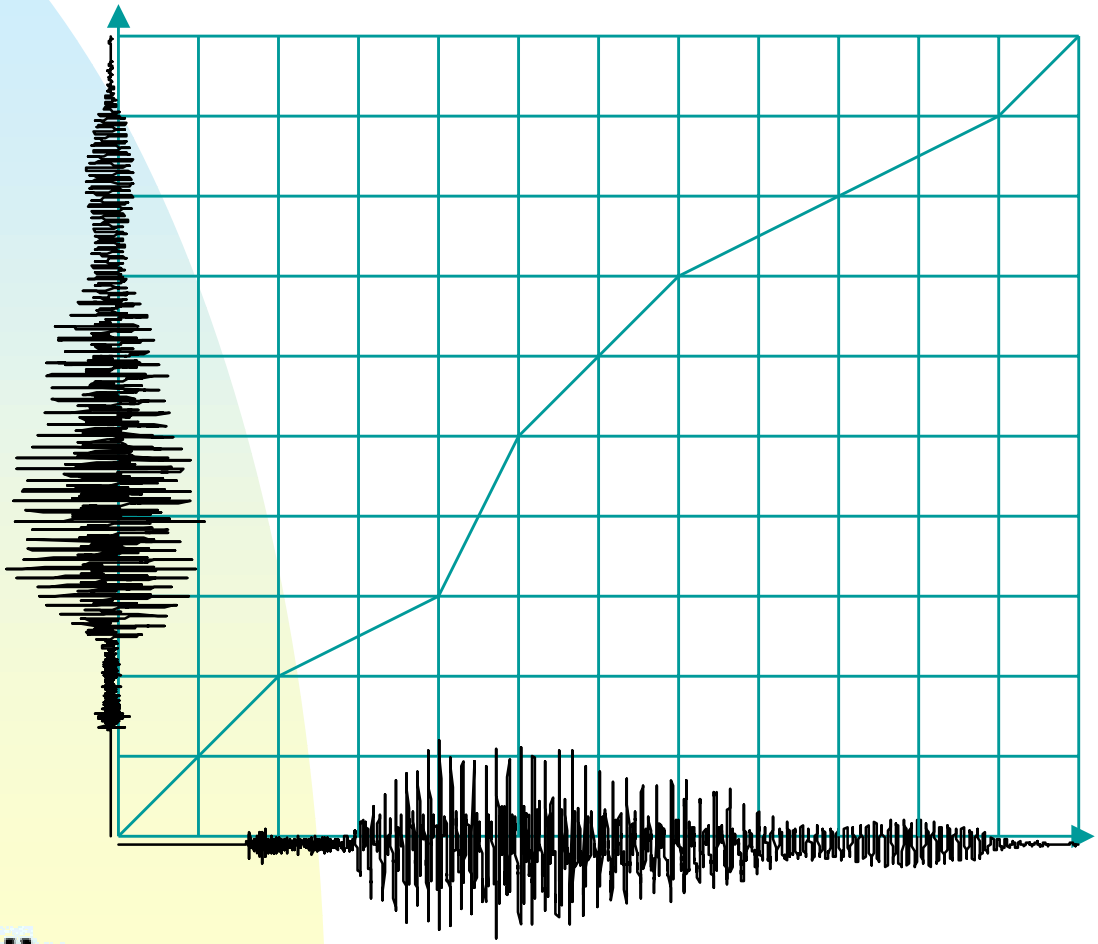
- Subjective vs. objective
- Input-output based vs. output based
- Mean Opinion Score
- Challenges: No input, costly subjective test
- Regression training: slow, unfeasible

5.2 Corpus-based TTS advantages

- Automatic objective test becomes possible
- Objective measures for input-output based test
  - Segmental SNR
  - Log spectral distance
  - Itakura distance
  - LPC
  - Cepstral distance
- Dynamic Time Warping for sentence scoring
- Regression training

5.3 Dynamic warping of test signal

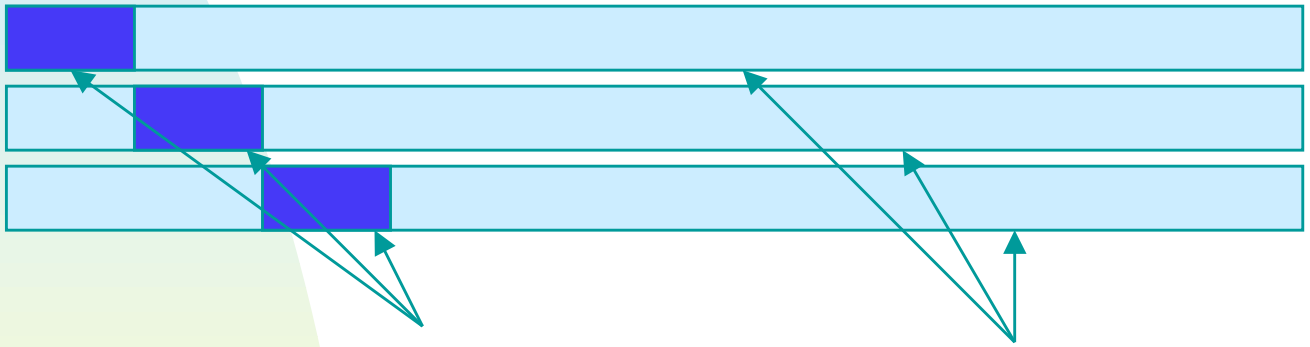
Original



synthesized

5.4 Regression training

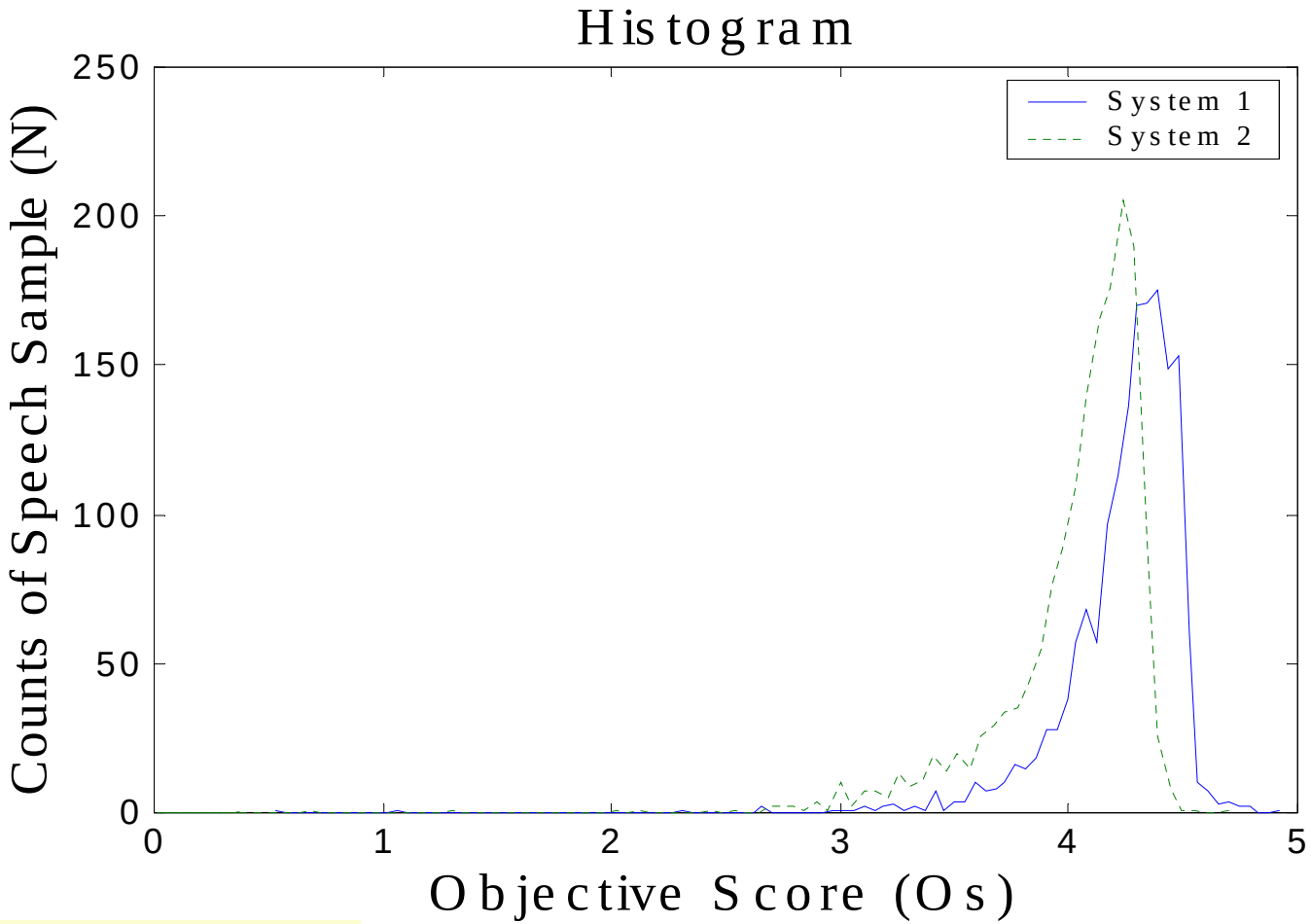
- To beautify the voice by using recorded corpus
- Ten 10% withholding test



Withholding data for testing

Full corpus

5.5 Mean Opinion Score histogram



Content

1. Introduction
2. Linguistic processing
3. Prosody generation
4. Speech Synthesis
5. Effective objective measure
6. Summary

6. Summary

- Concatenative Synthesis provides good quality Chinese TTS
  - Putonghua Engine 
  - Cantonese Engine 
  - Mandarin Engine 
- An effective quality evaluation strategy is available for objective measure of corpus-based TTS system

Thanks