RESENT RESEARCH PROGRESS ON CHINESE TEXT-TO-SPEECH

AT USTC iFly SPEECH LAB

Ren-Hua Wang

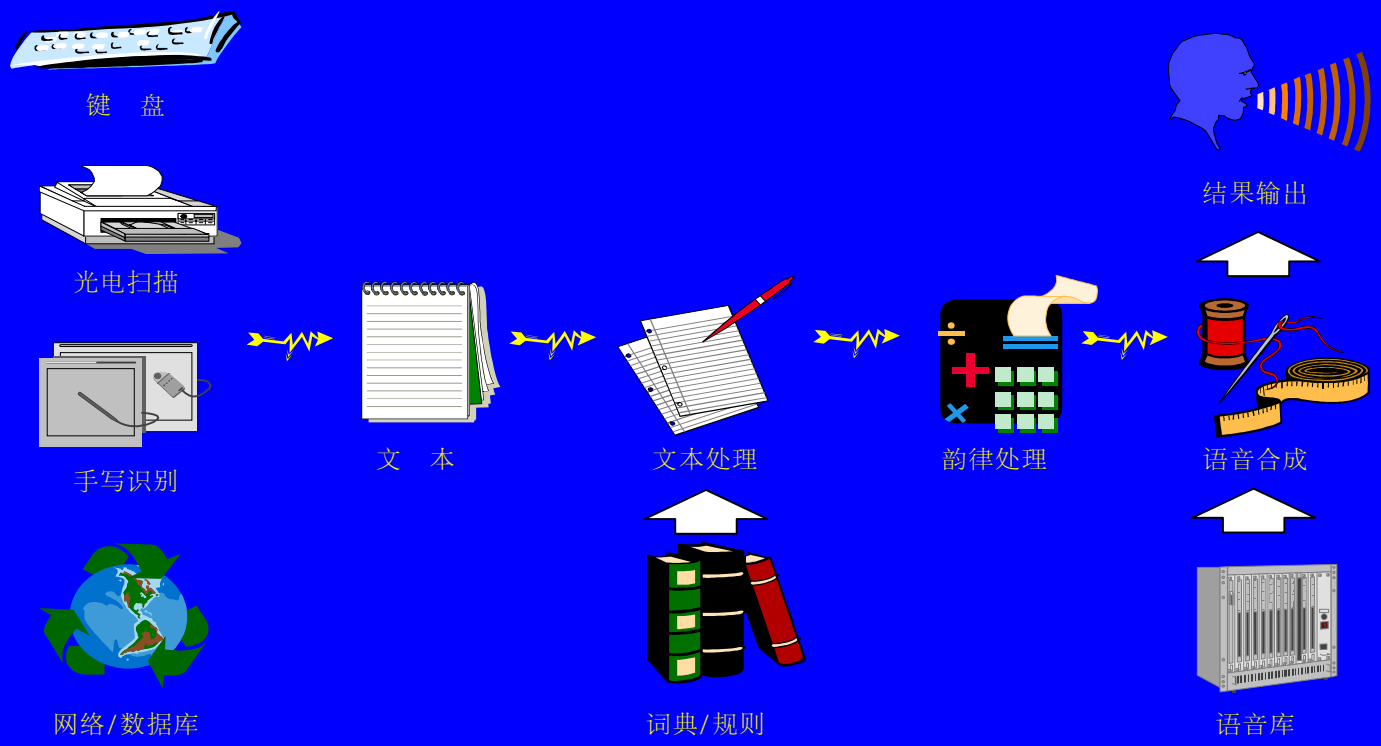
王 仁 华

University of Science & Technology of China
P.O.Box 4, Hefei, 230027 P.R.CHINA
Email: rhw@ustc.edu.cn

1. INTRODUCTION

- Chinese text-to-speech synthesis has made a rapid progress in the past ten years. With the method of waveform concatenating the quality of output speech of the synthesis system has been greatly improved.
- Chinese TTS systems are going to enter the market gradually in a large scale.
- However, higher naturalness, richer expressive ability, and more flexibility are still expecting.

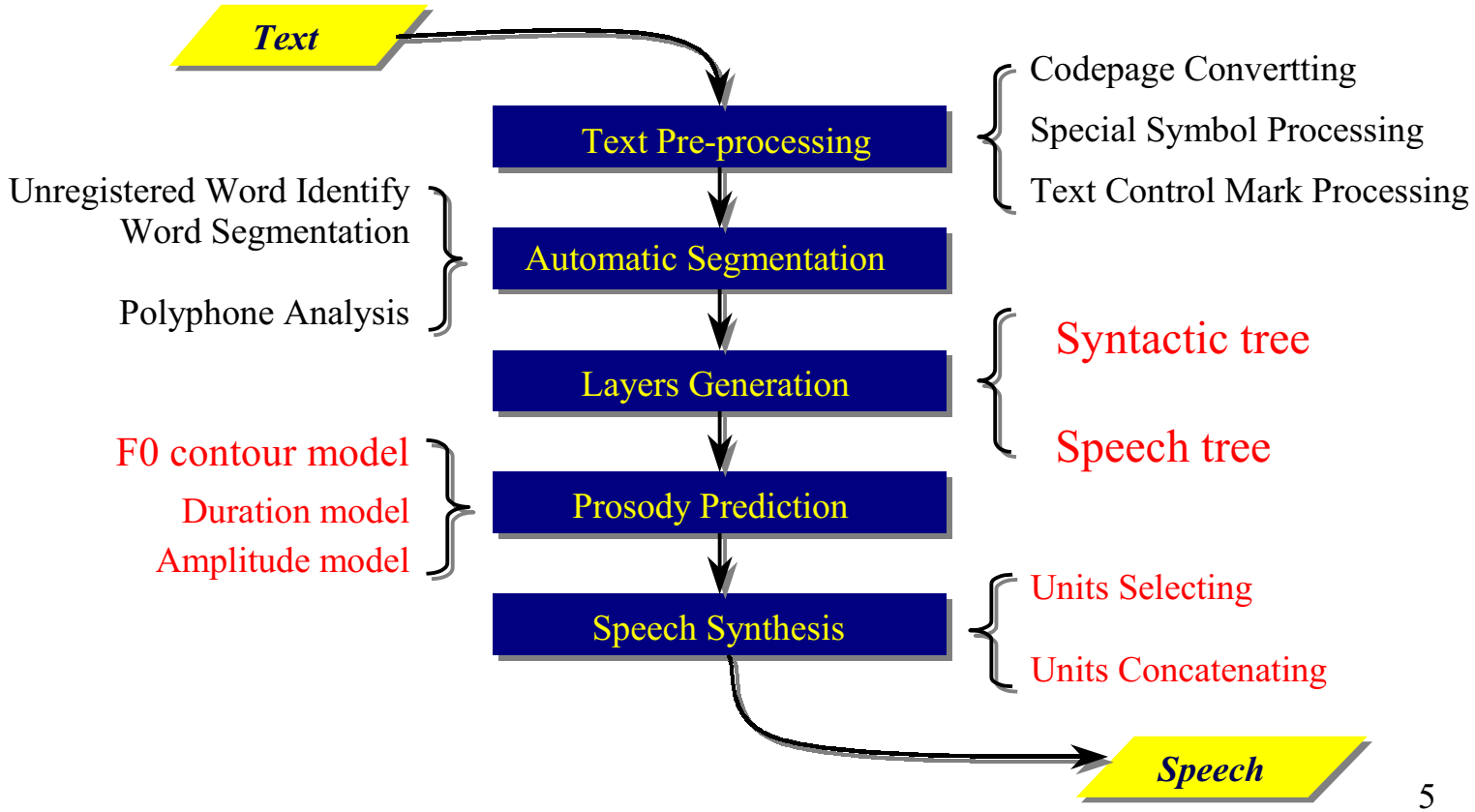
文语转换系统 (TTS)



Key Points

- The text processing ability, which should be able to supply with enough linguistic cues.
  - The correct pronunciation (characters,symbols..)
  - Prosodic specifications (tone and duration)
  - Pause and stress
- The advanced speech synthesizer, which can generate the utterances well-matched with the required specifications
  - Formant synthesizer
  - Waveform concatenating with little modification

TTS Process Flow



CONTENTS

- This paper presents the new progress we made in recent two years on Chinese text-to-speech.
- The main results can be summarized as follows:
  - Hierarchy process idea in Chinese text analysis.
  - Prosody modeling based on the decision tree method.
  - A corpus-based Chinese speech synthesis system

2. HIERARCHICAL TEXT ANALYSIS

- Hierarchical processing idea
- Define different layers for different speech units
- Provide the prosody layer structure in a form of Speech Tree
- Generate speech tree based on the syntactic tree

Hierarchy labeling system

Definition

- L0: Syllable layer
- L1: Speech foot layer
- L2: Prosody words layer
- L3: Master-phrase layer
- L4: Sentence layer

L0: 我|们|的|最|终|目|标|是|得|到|高|自|然|的|语|音

L1: 我们|的|最终|目标|是|得到|高|自然|的|语音

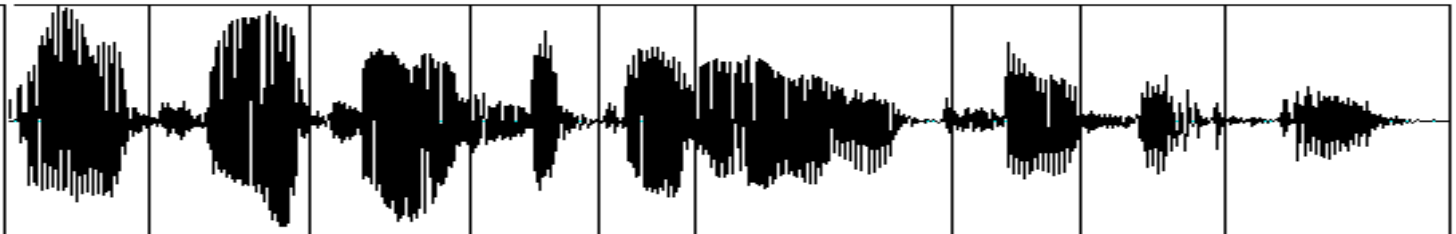
L2: 我们的|最终目标|是|得到|高自然的|语音

L3: 我们的最终目标|是得到高自然的语音

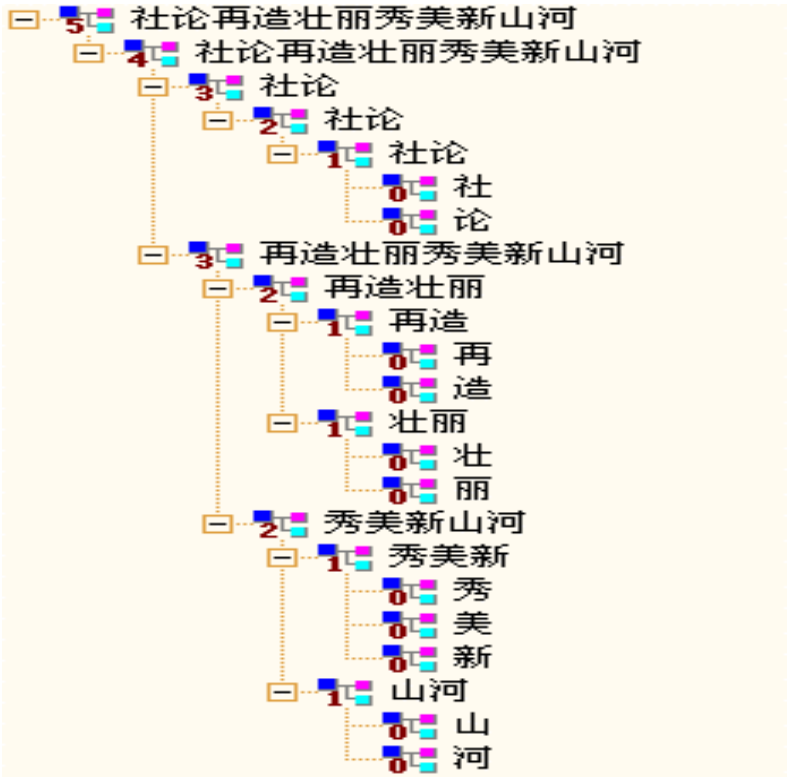
L5: 我们的最终目标是得到高自然的语音

A Labeling Example

(歹徒见势不妙逃走了)

Prosodic Label System									
	dai3	tu2	jian4	shi4	bu2	miao4	tao2	zou3	le0
Foot	dai3tu2		jian4shi4		bu2miao4		tao2zou3le0		
Prosodic Phrase	dai3tu2		jian4shi4bu2miao4				tao2zou3le0		
Intonational Phrase	dai3tu2jian4shi4bu2miao4						tao2zou3le0		
Sentence	dai3tu2jian4shi4bu2miao4tao2zou3le0								

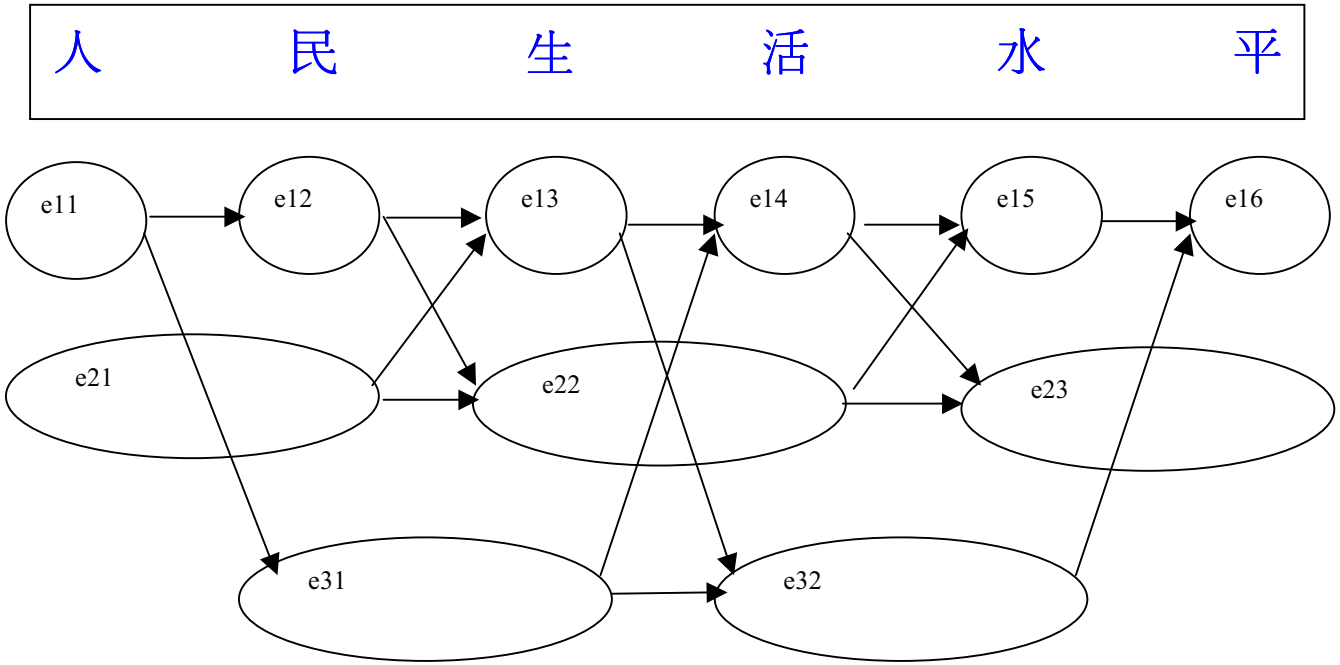
Speech Tree Examples



“社论：再造壮丽秀美新山河。“的语音树

Word segmentation(1)

Full-Expended-Words-Segmentation-Net



Word segmentation (2)

Find optimized path W

$$F(W) = \sum_{i=1}^{n} \ln \frac{P_{cut}(W_i)}{Total_{cut}} + \ln \frac{P_{form}(W_i)}{Total_{form}}$$

Where:

$W = w_1, w_2, \ldots, w_n,$

$P_{cut}(W_I)$ is word's Segmented Word-Frequency (SWF);

$P_{form}(W_i)$ is word's Form Word-Frequency(FWF).

Layers Generation

- L2 is mainly determined by philological knowledge. So we adopt describing rules and corresponding costs to generate the optimizing path for L2
- The decision of L3 depends on the language understanding, which is a difficult subject. We can avoid the problem at certain degree by adopting the statistical method and the information of L2

From syntactic tree to speech tree

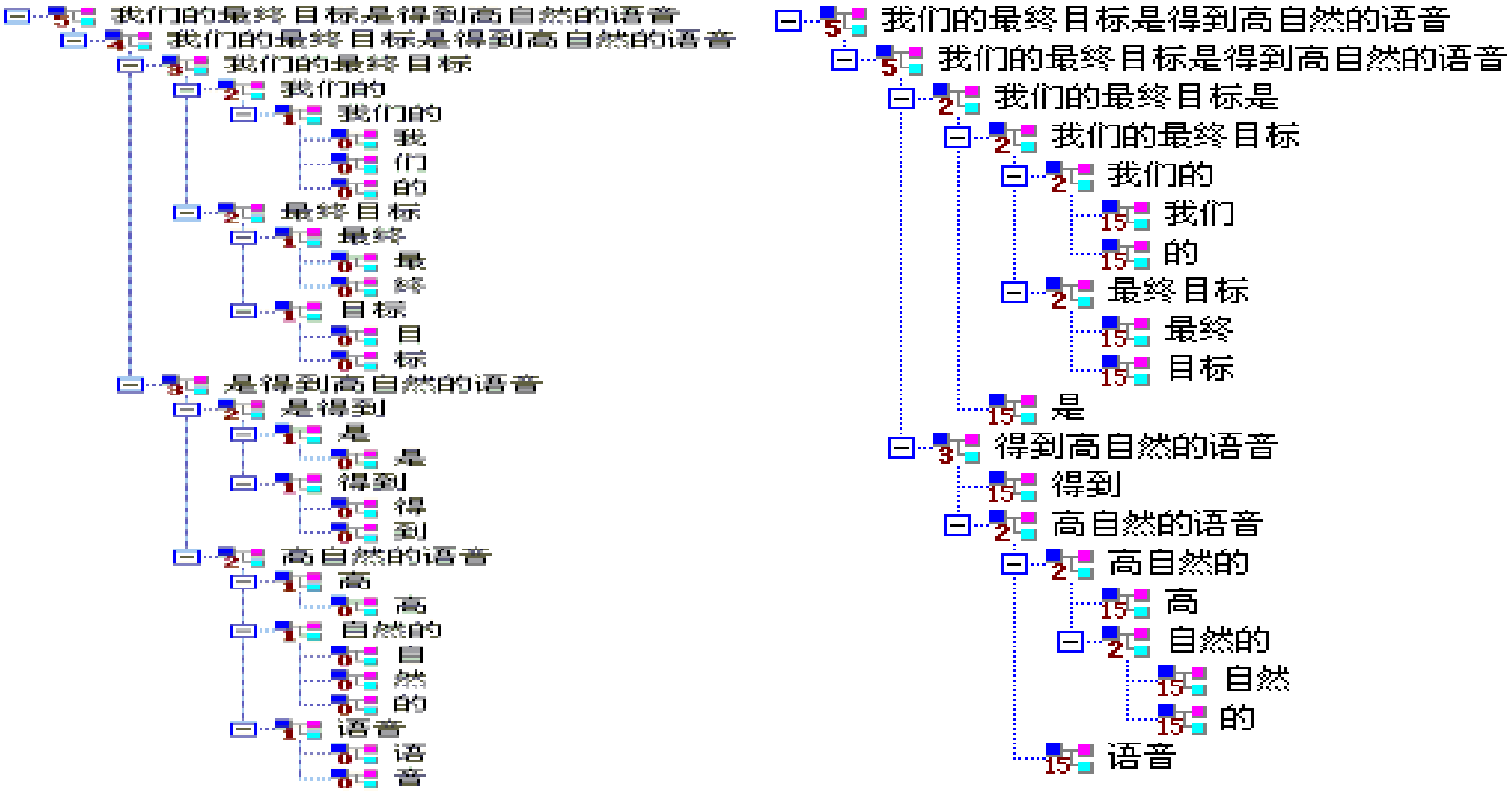


图1：“我们的最终目标是得到高自然的语音”的语音树（左）和语法树

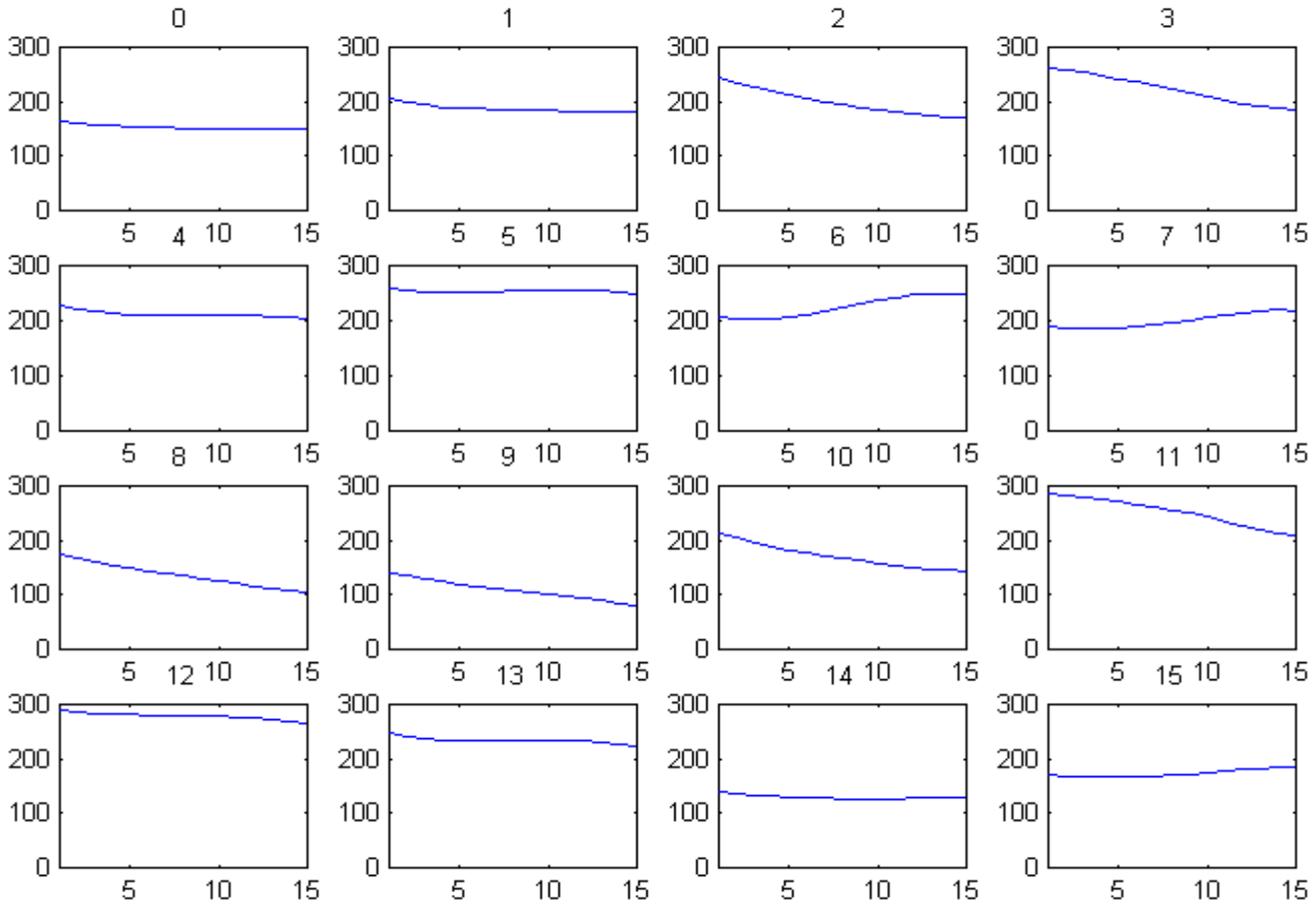
What is more

- OOV problem
  - Insufficient vocabulary
  - Insufficient sense information of word class
- Language model problem
  - Incomplete rules for syntactic analysis
  - Be short of efficient bi-gram tie-in probability
  - Lack association and reasoning
- Stress problem
  - No way for semantic stress

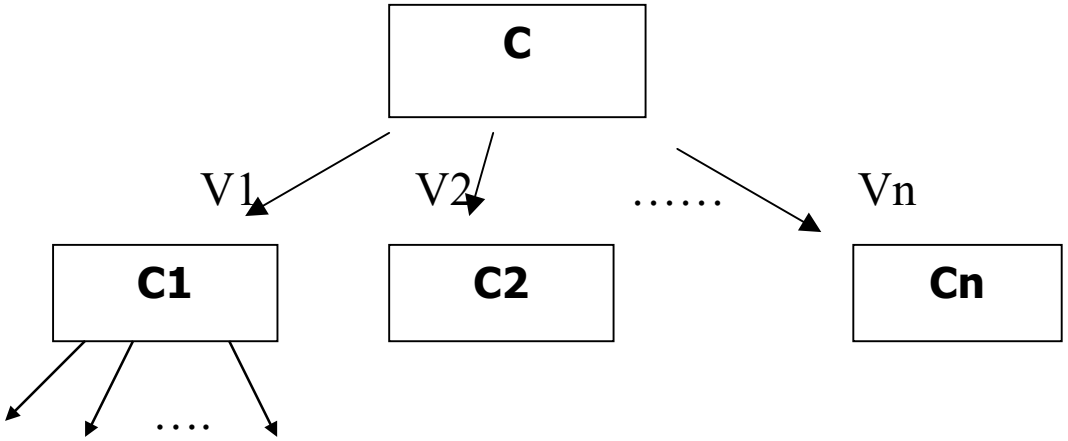
3. PROSODY MODELING

- Based on the statistical decision tree method the modeling for pitch contour and duration achieved important progress.
- These models open out the mapping relations from the prosody context information to prosody feature parameters of units,
- The decision tree method is also used for data learning and prediction in the juncture of pitch contours,

Normalized pitch contours



Generate Decision Tree



Decision Attributes

Attribute Class	Decision Attribute
Self-Attribute	Final class and Initial class
	Tone class
	Word features
Pre-Next-Attribute	Next syllable, tone and initial class
	Pre syllable, tone and final class
Location-Attribute	The relative location of lower layer in the built-in layer
	The absolute location of lower layer in the built-in layer
	The number of the syllables in the word
	The number of the syllables in the sentence

Duration modeling

- Duration normalization

$$T_s^{'} = \frac{T_s - \overline{T_s}}{\sigma^2}$$

- Discrete $T_s^{'}$
- Generate Decision tree

3. CORPUS-BASED CHINESE SPEECH SYNTHESIS

- The concatenating synthesis based on corpus now has become the mainstream because of its higher performance.
- Synthesized speech is generated by catching the optimal speech segments from the corpus and concatenating them together
- most of systems use the Pitch_Synchronous Overlap Add (PSOLA) to perform the required pitch and duration modifications directly on continuous waveforms

Two Problems

- First, what should the corpus include?
  - Corpus design
- Second, how to select the required synthesis units in the corpus for a target sentence to be synthesized?
  - Unit selection
  - Link cost

Corpus Design

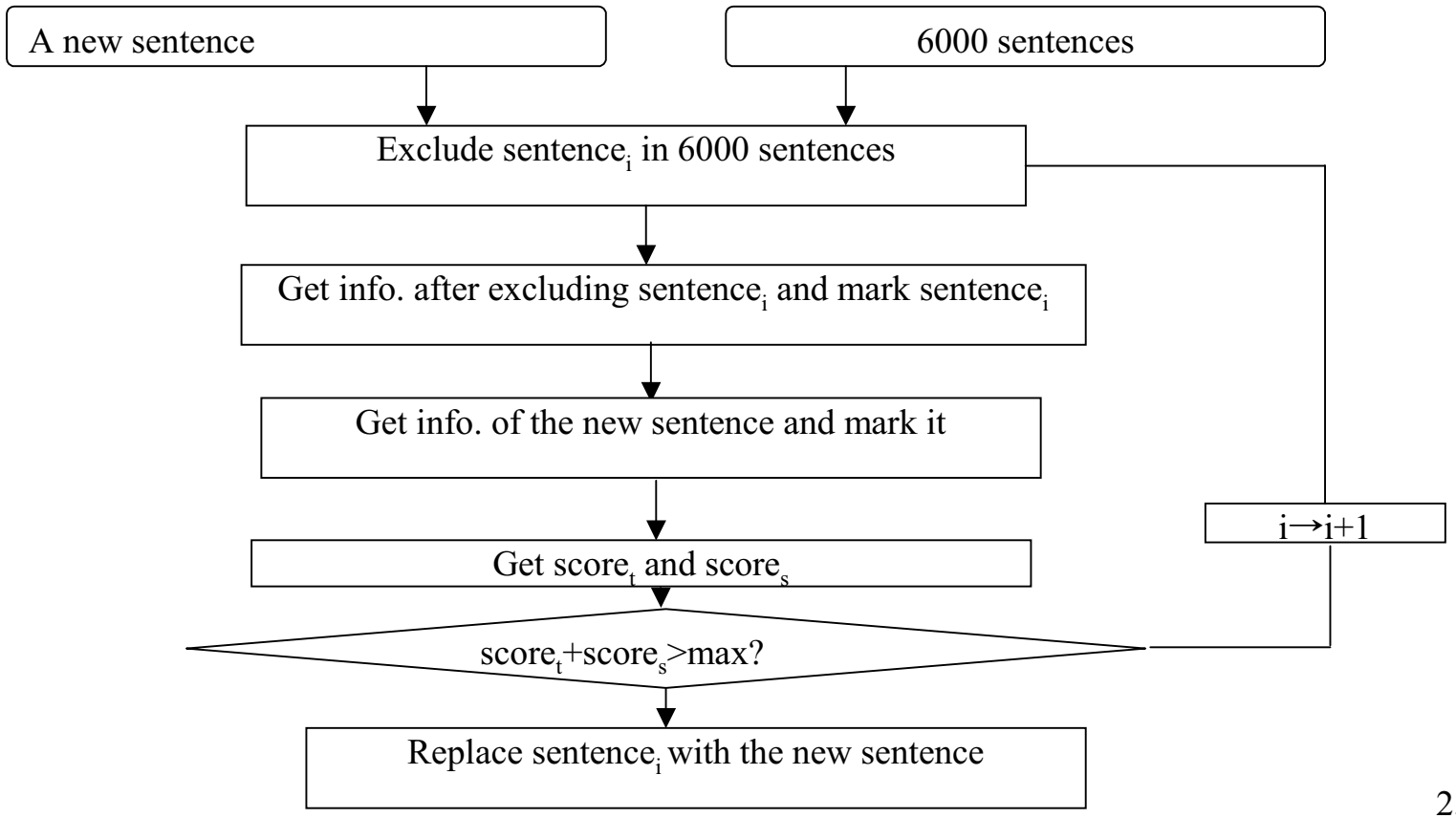
A common principle in corpus design is that natural language phenomena be included as much as possible while text size as small as possible.

- which language phenomena must be considered
- how to solve the contradiction between physical capacity and that of information.

Information For Every Syllable

Information type	a certain syllable
1	Previous and Next Tone Type
2	Sentence Head with Next Tone Type
3	Sentence End with Previous Tone Type
4	Previous Final Type
5	Next Initial Type
6	Position in a Word

Greedy Algorithm Flow Chart



Labeling of Corpus

- The labeling of corpus includes segmentation and prosody labeling
  - The position of each syllable in speech file is decided automatically by the HMM based segmentation tools
  - The prosody structure is generated manually based on the hierarchy labeling system.
- Other preprocessing after the segmentation and labeling
  - Extract the F0 contour parameter
  - Normalize duration and amplitude

Index

- For each syllable, an index is built to get its information by combining its prosody structure and segmentation position with F0 parameter.
- The content of the index is described as following:
  - Position information (beginning, ending and INITIAL-FINAL boundary position in speech file);
  - Prosody parameters information (F0 value of beginning, ending, maximum and minimum);
  - Prosodic structure information (the name of current, preceding and next syllable with extended tone, the position of current syllable in PW, the position of the PW in the MP, the position of MP in Sen.)

Unit Selection (1)

- Target sentence generation
  - Text analysis → speech tree
  - Prosody model → unit prosody parameters
- Unit selection
  - Segment cost
    - Position of syllable in the word
    - Contextual information, such as the final class of pre syllable and initial class of next syllable
  - Prosody cost
    - Pitch contour
    - Duration

Unit Selection (2)

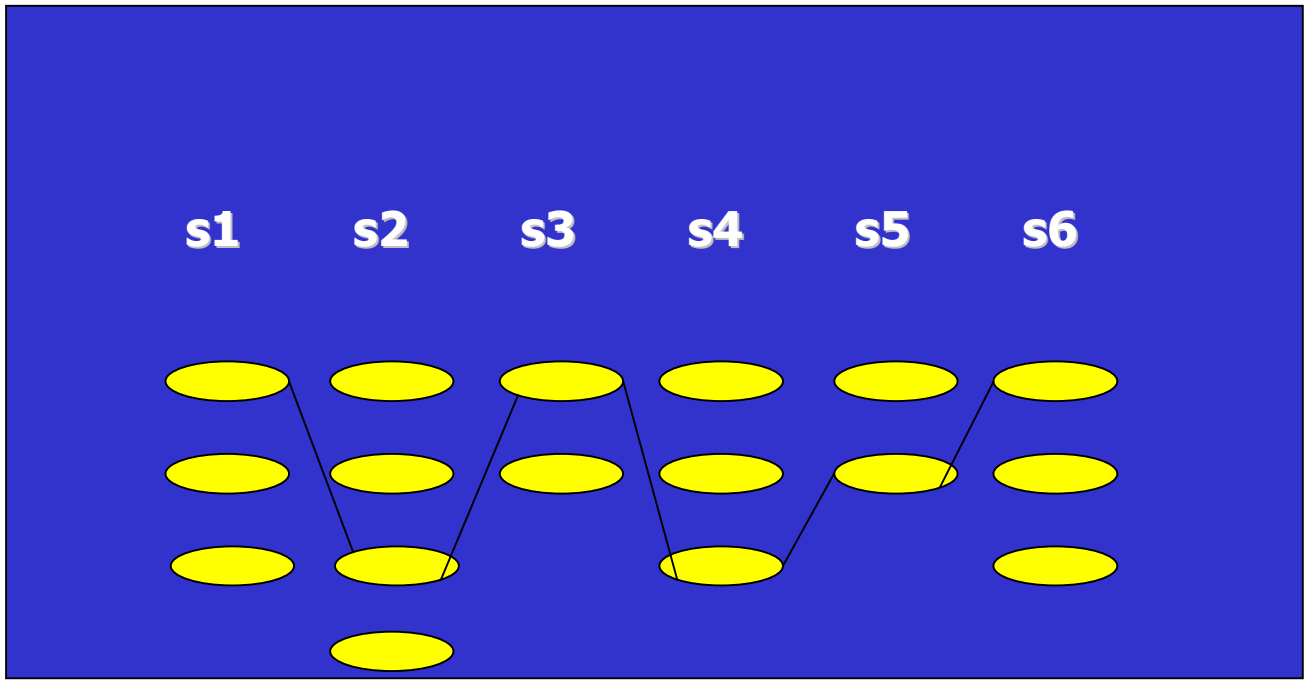
- An instance is selected out as a candidate unit if it has the minimum weighting cost relative to the target.
- For each target syllable, at most ten units are selected out as candidates

Concatenation Cost

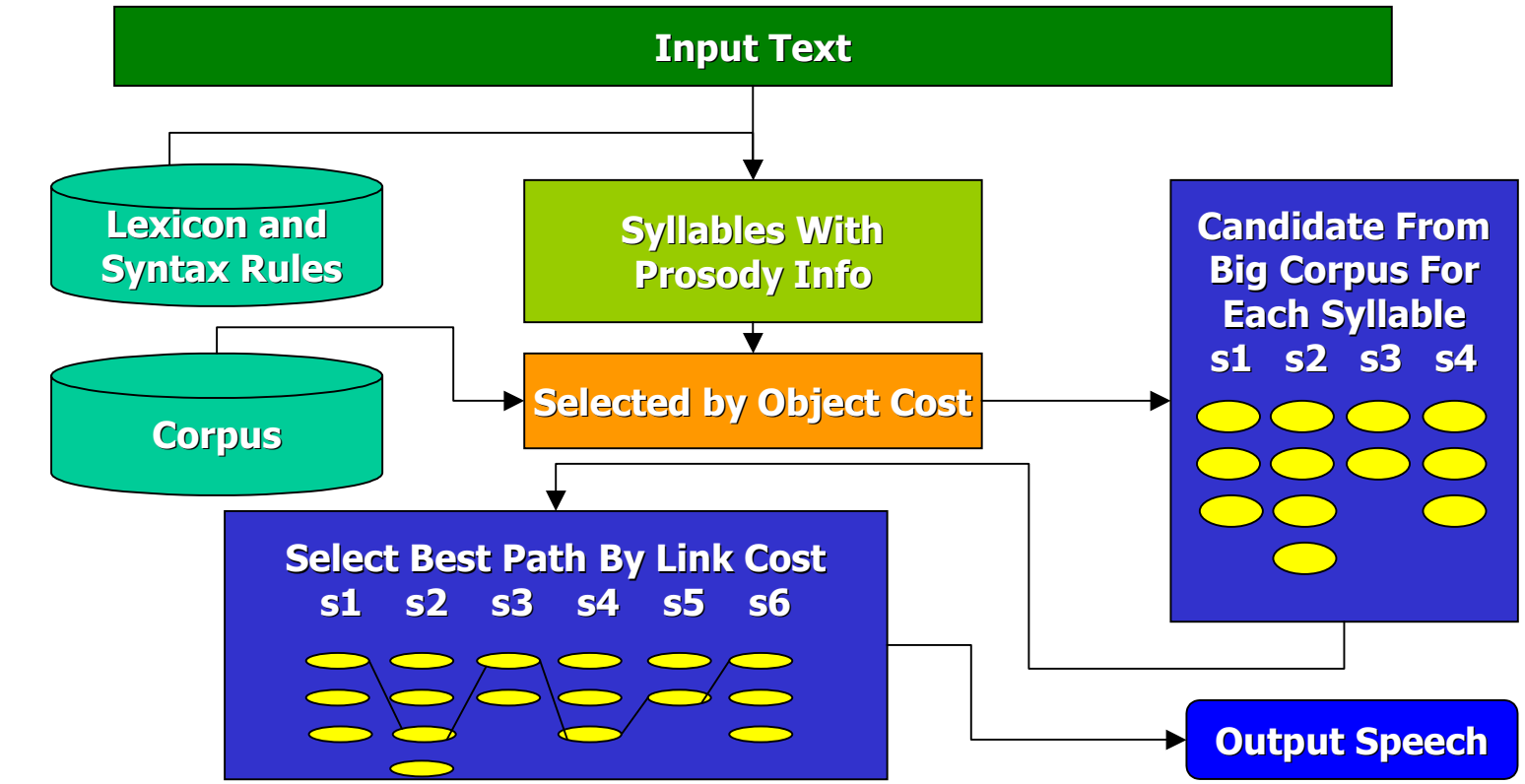
- Juncture cost in for-and-aft pitch contours
  - Pitch difference modeling
- Continuity cost of formants

$$D_f = \sqrt{(F_1^p - F_1^n)^2 + (F_2^p - F_2^n)^2}$$

Select Best Path By Link Cost (By Viterbi search algorithm)



Block diagram of the corpus-based speech synthesizer



4. FURTHER RESEARCHES

- Improve the naturalness of synthesized speech
- Multilingual TTS system
- Enhance the practicability of systems
  - Distributed speech synthesis
  - Embedded TTS
- Enrich the expressional ability of systems
  - Visual TTS
  - Voice conversion

Demonstrations

- Corpus-Based TTS System
- KS-2000 TTS system

Thanks