TEXT TO SPEECH SYNTHESIS

KESHU

INTRODUCTION

- Language is the ability to express one's thoughts by means of a set of signs, whether graphical, gestual, acoustic or even musical.
- It is a distinctive feature of human beings who use such structured system

Speech

- Speech is major component of a language
- Oldest means of communication
- Levels of speech:
  1. Acoustic
  2. Phonetic
  3. Phonological
  4. Morphological
  5. Syntactic
  6. Semantic
  7. Pragmatic

Perfect TTS Synthesizer

- Human beings
- The reading process involves:
  Seeing, Thinking, Saying, Hearing
- These are most complex processes
- Cannot be imitated

TTS Synthesizer System

- A text to speech synthesizer is a computer based system that should be able to read any text whether it was directly introduced into the computer or through character recognition system (OCR). And speech should be intelligible and natural.

Feature and Multilevel Data Structures

- Plays an important role in contemporary TTS systems for Natural Language Processing

1985)).[6]

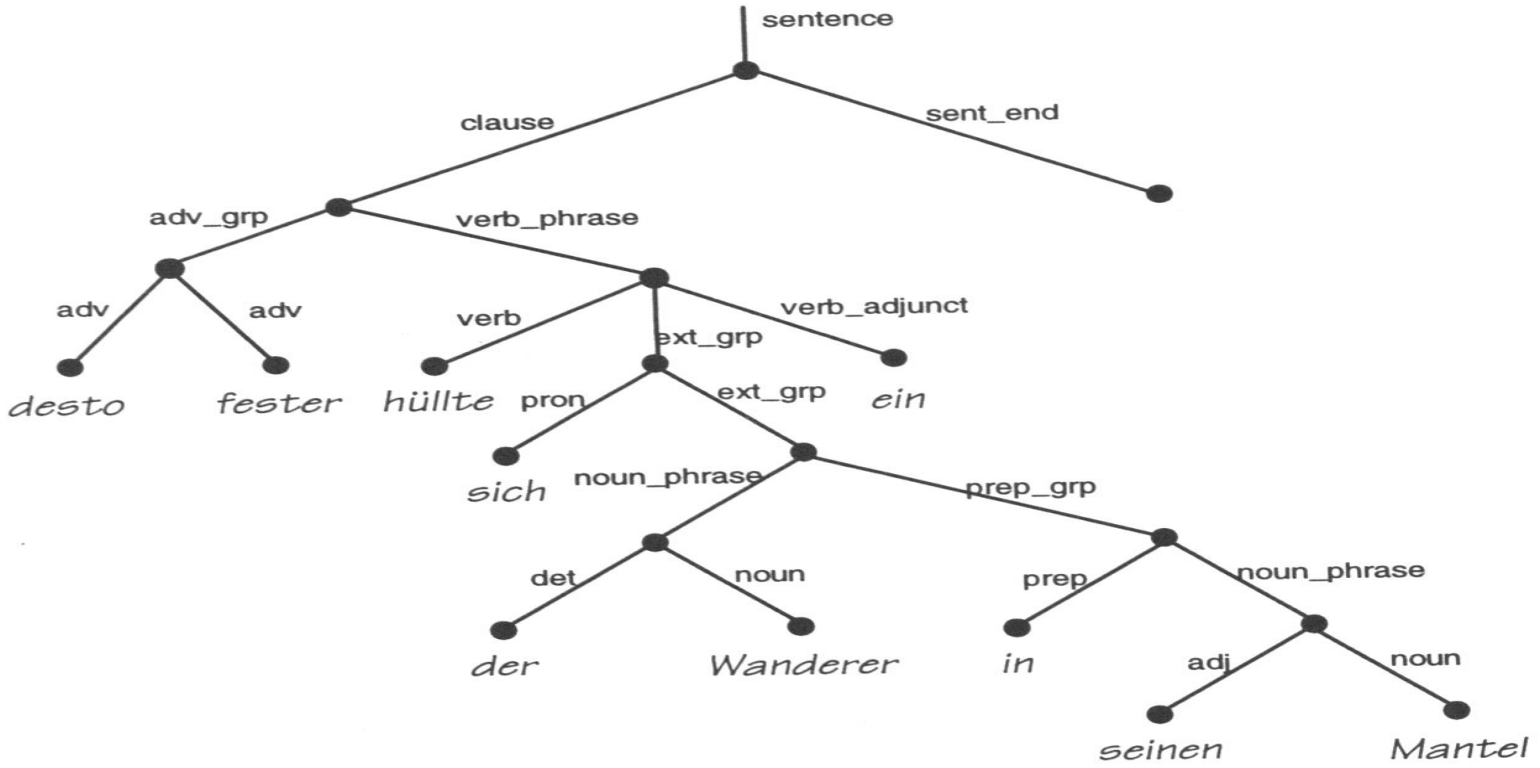

Fig 3.3. An example of a feature structure (depicted as a tree in this case—that is, a DAG with no symbolic sharing of values) when the syntax analysis module has operated on the German sentence fragment *desto fester hüllte sich der Wanderer in seinen Mantel ein* (adapted from Böhm, 1992): "*the more firmly the walker muffles himself up in his coat*").

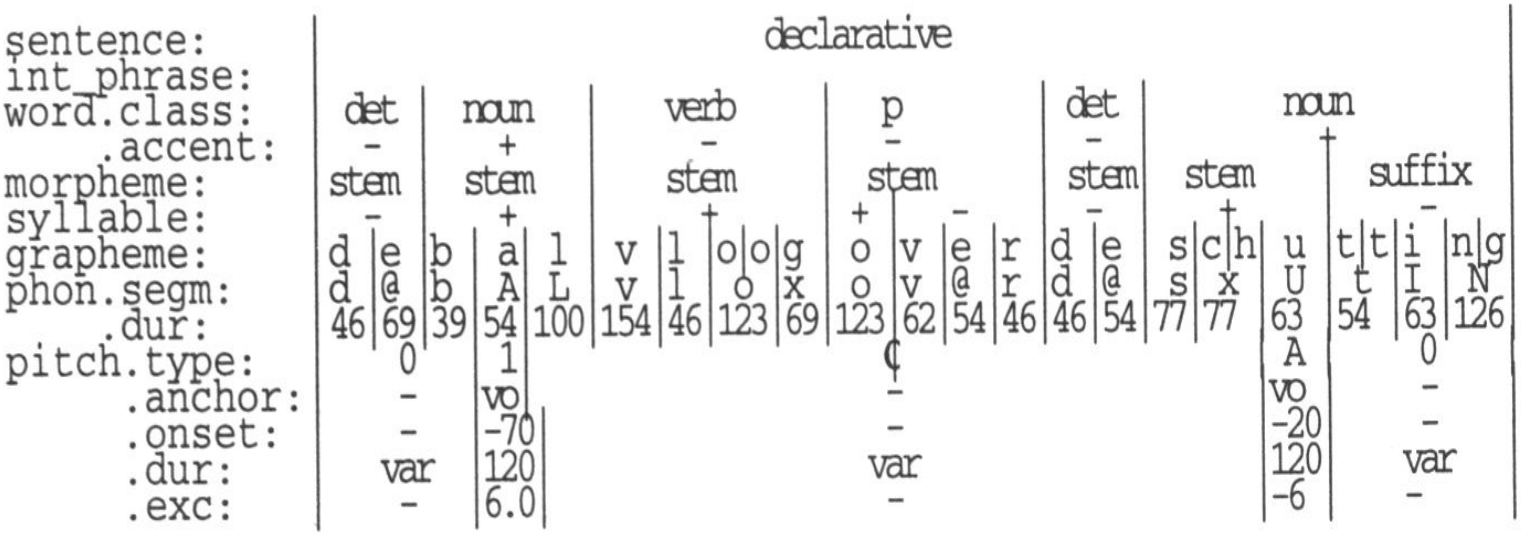

Fig 3.4. An example of a multi-level data structure (MLDS) when all the analysis modules have operated on the Dutch sentence *de bal vloog over de schutting* (*the ball flew over the fence*, after Van Leeuwen and te Lindert, 1993).

Typical TTS Components

- Two components
- Natural Language Processing Module (NLP)
- Digital Signal Processing Module (DSP)

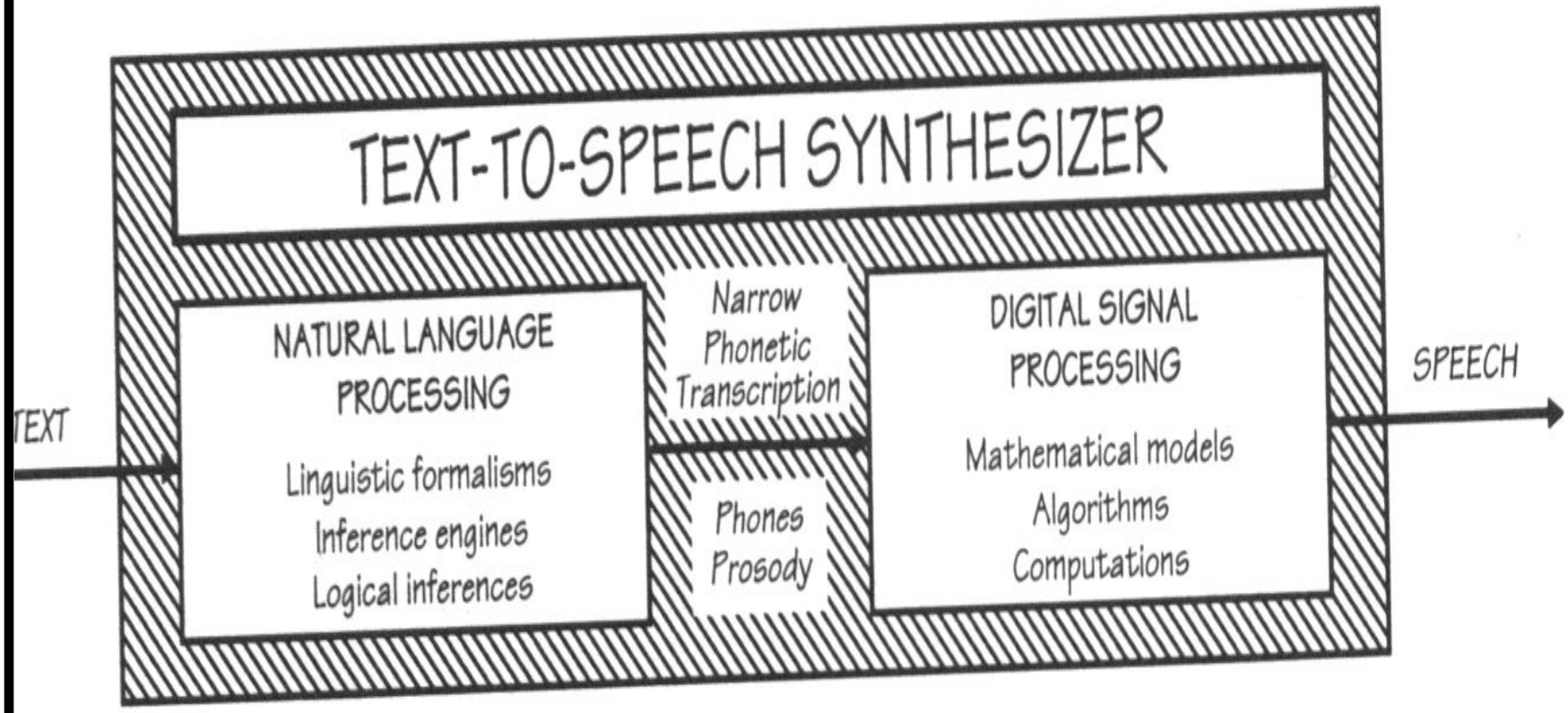

Figure 1.7. A simple but general functional diagram of a TTS system.

NLP and DSP Modules

- The NLP module is capable of producing a phonetic transcription of the text to be read, together with the desired intonation and rhythm. It takes in the text as input and give narrow phonetic transcription as output which is further forwarded to the DSP module. And the DSP module which transforms the symbolic information it receives into natural sounding speech. “Narrow phonetic transcription” which is taken as intermediate varies from synthesizer system to another.

NLP Module of typical TTS system

- Text Analyzer (Morpho Syntactic Analysis)
- Pre-processor
- Morphological Analyzer
- Contextual Analyzer
- Syntactic-Prosodic parser
- Letter to Sound Module

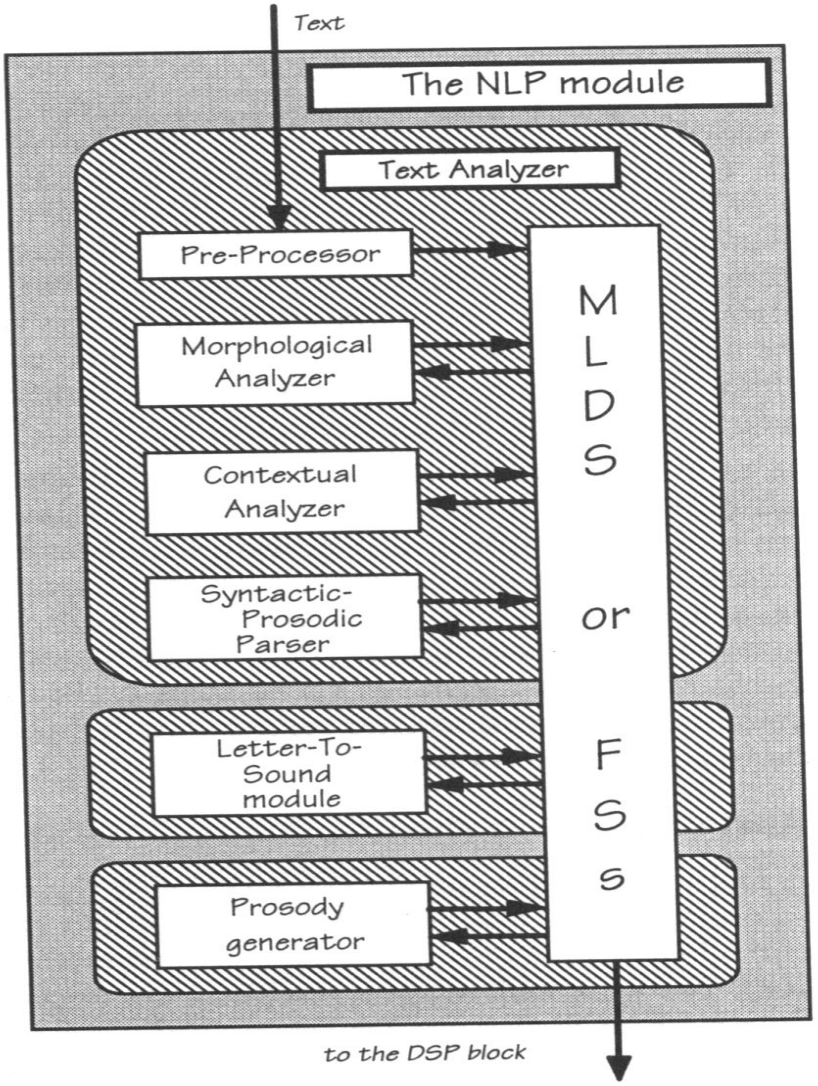

Fig 3.6. The NLP module of a typical text-to-speech conversion system.

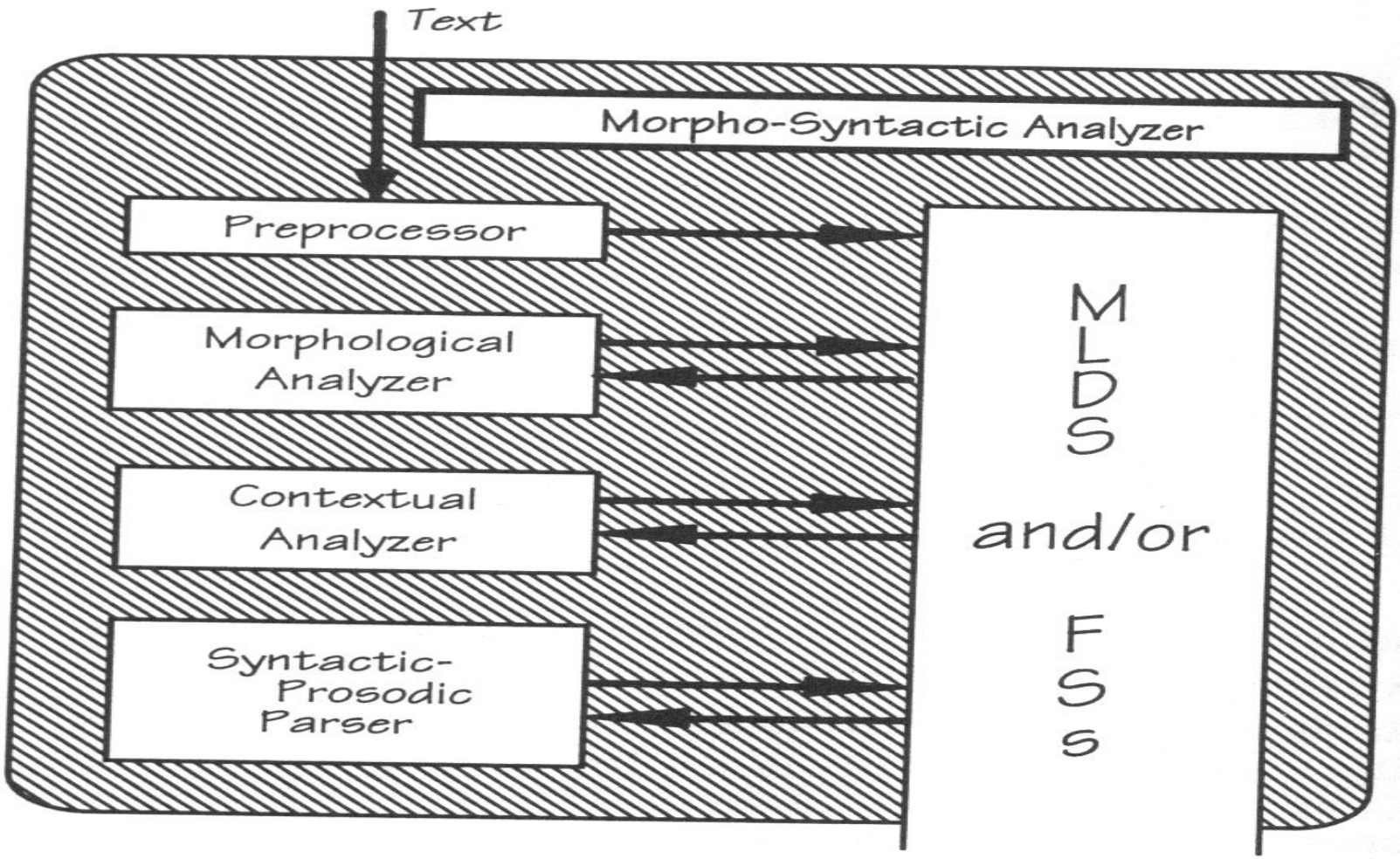

Fig. 4.1. A typical morpho-syntactic analyzer.

Preprocessor

- Takes in texts as strings of ASCII characters
- Transforms text into Broad Segmentation Units (BSU's) following the set:
- A sequence of characters
- A sequence of digits
- A single punctuation mark or another special character
- A sequence of white space characters
- Eg: (I)()(know)()(1)(,)(000)()(words)(,)()(Dr)(.)()(Jones)(.)
- Rewrites the BSU's into a list of word-like units and of syntax bearing punctuation marks called Final Segmentation Units are produced (FSU's).

Preprocessor

- Sentence end detection (semicolon, period – ratio, time and decimal point, sentence ending respectively)
- Abbreviations (e.g. – for instance)

  Changed to their full form with the help of lexicons
- Acronyms (I.B.M – these can be read as a sequence of characters, or NASA which can be read following the default way)
- Numbers (Once detected, first interpreted as rational, time of the day, dates and ordinal depending on their context)
- Idioms (eg. “In spite of”, “as a matter of fact”– these are combined into single FSU using a special lexicon)

Morphological Analysis

- Task is to propose all possible parts of speech categories to each word taken individually on the basis of their spelling
- Words – Function and Content words

Function Words

Function words (determiners, pronouns, prepositions, conjunctions..).

- Can be stored in a lexicon to get their parts of speech categories because of its size.
- Word he:
- <spel> = he
- <syn cat> = pronoun
- <syn num> =
- <syn gen> = masc
- <phon> = /hI/

Content Words

Content words- infinite in number

- Needs Morphology – part of linguistics that describes word forms as a function of reduced set of abstract semantically bearing units called morphemes.
- Inflectional, derivational and compound words (content words) are decomposed into their elementary graphemic units (morphemes)
- Uses regular grammars exploiting lexicons of stems and affixes which is the only way because of its infinite size

Contextual Analysis

- Considers words in their context
- Reduces the list of their parts of speech categories to a very restricted number of highly probable hypotheses, given the corresponding possible parts of speech of neighboring words.
- Achieved by N-grams, multi-layer perceptrons (Neural networks), local stochastic grammars (provided by expert linguistics) etc

Letter to Sound Module

- LTS module is responsible for the automatic determination of the phonetic transciption of the incoming text
- Cannot just look up in a pronunciation dictionary
- Do not follow the rule “one character = one phoneme”
- Examples

  Single character correspond to two phonemes -- x as /ks/

  Several characters producing one phoneme—

  gh in thought

  Single character pronounced in different ways

  c in ancestor, ancient, epic

  Single phoneme resulting in several spellings –

  sh in dish, t in action, c in ancient

Letter to Sound Module

- Some of the cases to consider
- Consonants may be reduced or deleted in clusters (eg. t in softness)
- Assimilation which originates in articulatory constraints and leads to a change of some phonological features of a given phoneme (eg. obstacle)
- Heterophonic homographs which are pronounced differently even though when they have same spelling (eg. record, contrast)
- Phonetic liaisons which affect final consonants of French words immediately followed by a vocalic sound which results in pronunciation of characters that otherwise disappear or in a change of pronunciation
- Schwas (transformation of unstressed vowels into short central phonetic elements is done or simply deletes them – like in thoughtful and interesting
- Vowel lengthening, new words, proper nouns which are really dependent on the language of origin to know the correct pronunciation.

Two Basic Strategies

- Dictionary based and Rule-based

Dictionary Based

- Dictionary based consist of storing a maximum of phonological knowledge into a lexicon and entries are generally restricted to morphemes and pronunciation of surface forms is accounted by inflectional, derivational and compounding morphophonic rules which describe how the phonetic transcriptions of their morphemic constituents are modified when they are combined into words. For those words that are not in the lexicon are transcribed by rule.

Rule Based

- Rule based strategy which transfers most of the phonological competence of dictionaries into a set of letter to sound (grapheme to phoneme) rules. And those words which are pronounced in a such a particular way that they constitute a rule on their own are stored in exceptions directory.

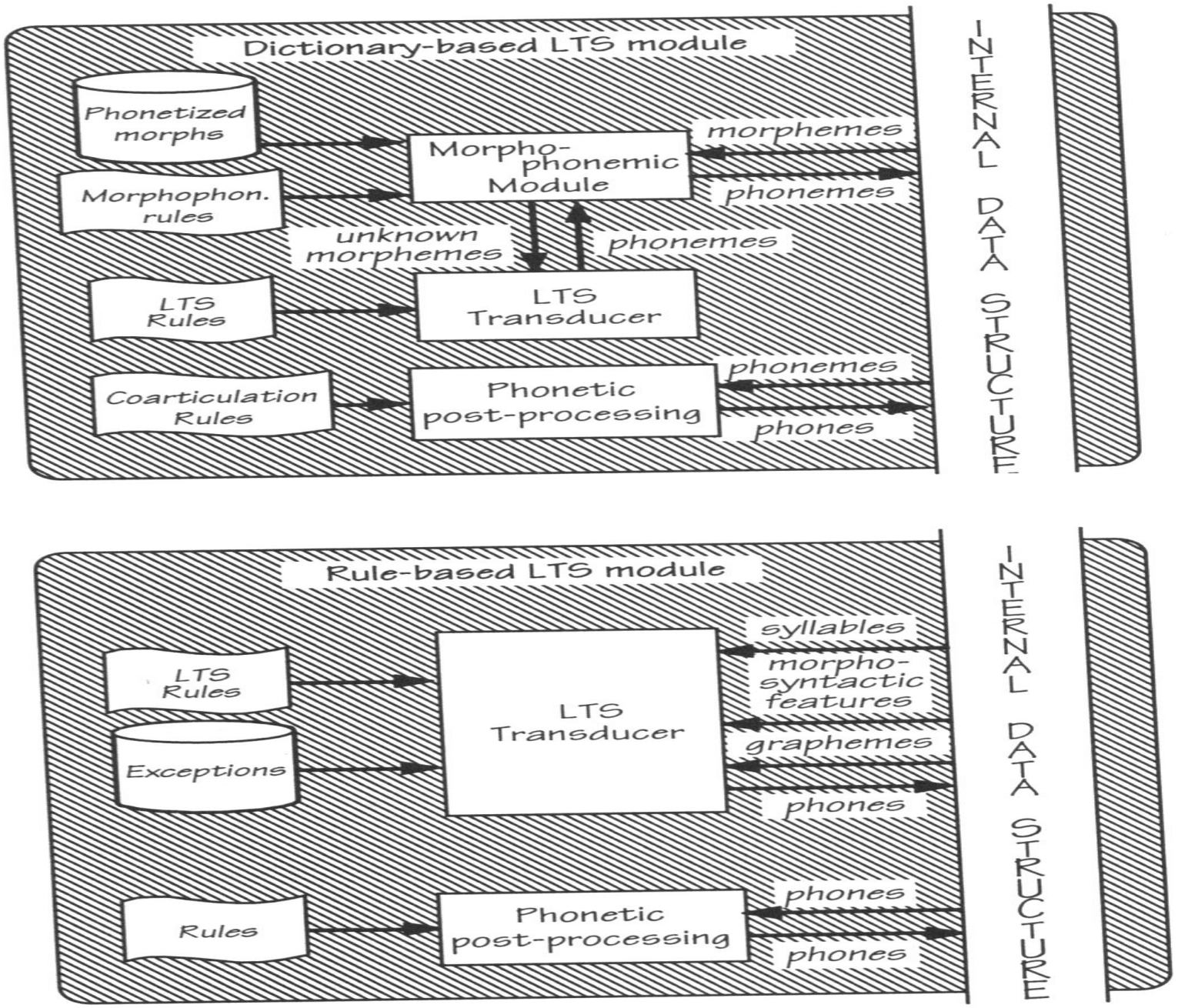

Fig. 5.1. Dictionary-based (top) versus rule-based (bottom) phonetization.

Morpho-Phonemic Module in Dictionary based

- Morphophonology is concerned with phonological changes in the pronunciation of morphemes occurring in the process of word formation.

Morpho-Phonemic Module in Dictionary based

- This module deals with the phonological changes and one distinguishes the following in this module
- Rules for changing phonological features (eg. ion and ure in completion and exposure)
- Rules for deleting or inserting phonemes (eg. buses or landed)
- Rules that account stress shift in languages such as English or German (eg. adApt + ation = adaptation or which doesn't change like in abOrt + ion = abOrtion).
- These are achieved by using rewrite rules and by using Two-level rules[Koskenniemi,1983].

LTS Transducer

- This is the key component that transforms graphemes to phones in the rule based strategy. This is achieved by following Expert rule based systems or trained rule based systems or by neural networks.

Phonetic Post Processing

- In order to increase the intelligibility and the naturalness of synthetic speech, some kind of phonetic post processing is required. After first phonemic transcription of each word has been obtained, this is applied so as to account for coarticulatory smoothing. This smoothing results in high quality speech.

Syntactic Prosodic Parser

- Prosody refers to certain properties of the speech signal which are related to audible changes in pitch, loudness, syllable length. This is also referred as intonation. The features of this are focus, relationships between words, finality. These have specific functions in speech communication.

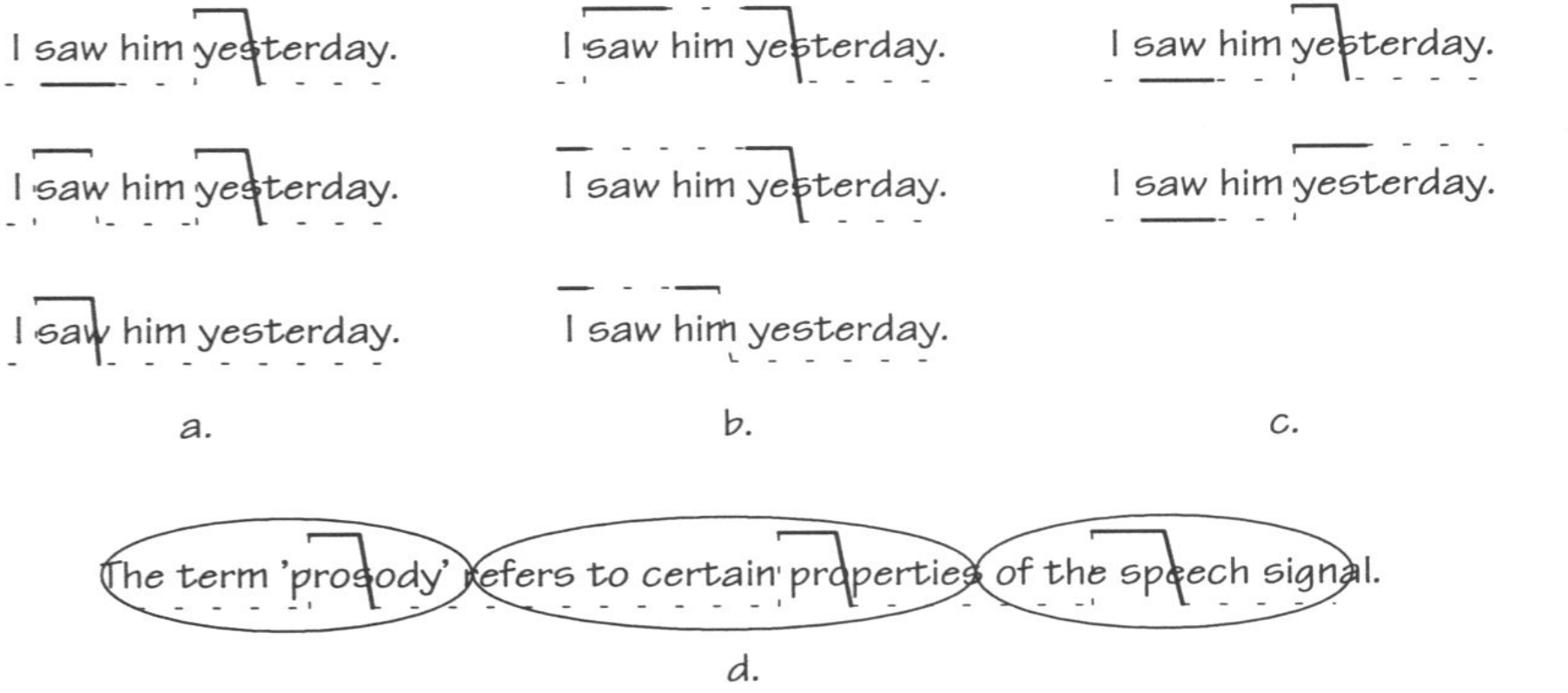

Fig. 6.1. Different kinds of information provided by intonation (lines indicate pitch movements; solid lines indicate stress).

a. Focus or given/new information;
b. Relationships between words (saw-yesterday; I-yesterday; I-him)
c. Finality (top) or continuation (bottom), as it appears on the last syllable;
d. Segmentation of the sentence into groups of syllables.

Syntactic Prosodic parser

- Getting a speech with all those features is impossible.
- Focuses on obtaining an acceptable segmentation and translates it into the continuation or finality but ignores the relationships or contrastive meaning

Syntactic Prosodic Parser

- These prosodic groups are achieved by a recent very crude algorithm termed as chinks ‘n chunks by Liberman and Church [1992] in which prosodic phrases are accounted for by the simple regular rule
- A (minor) prosodic phrase = a sequence of chinks followed by a sequence of chunks

DSP Module

- Takes in the narrow phonetic transcription and gives out speech as output

Why we need TTS system

- There are several advantages of a high quality text to speech synthesis system
- Great use in Telecommunications, relay service, Language Education, aid to handicapped persons, talking books and toys, vocal monitoring, multimedia, man-machine communication etc

Conclusion

- There is longggg waaaay to reach to have a system similar to HAL (Space Odyssey)
- Development in technology and gaining interest in NLP makes everyone think optimistic about reaching the goal soon.