Inclusion of a Prosodic Module inSpoken Language Translation Systems

Robert Eklund & Bertil Lyberg
robert.eklund@haninge.trab.se
bertil.lyberg@haninge.trab.se

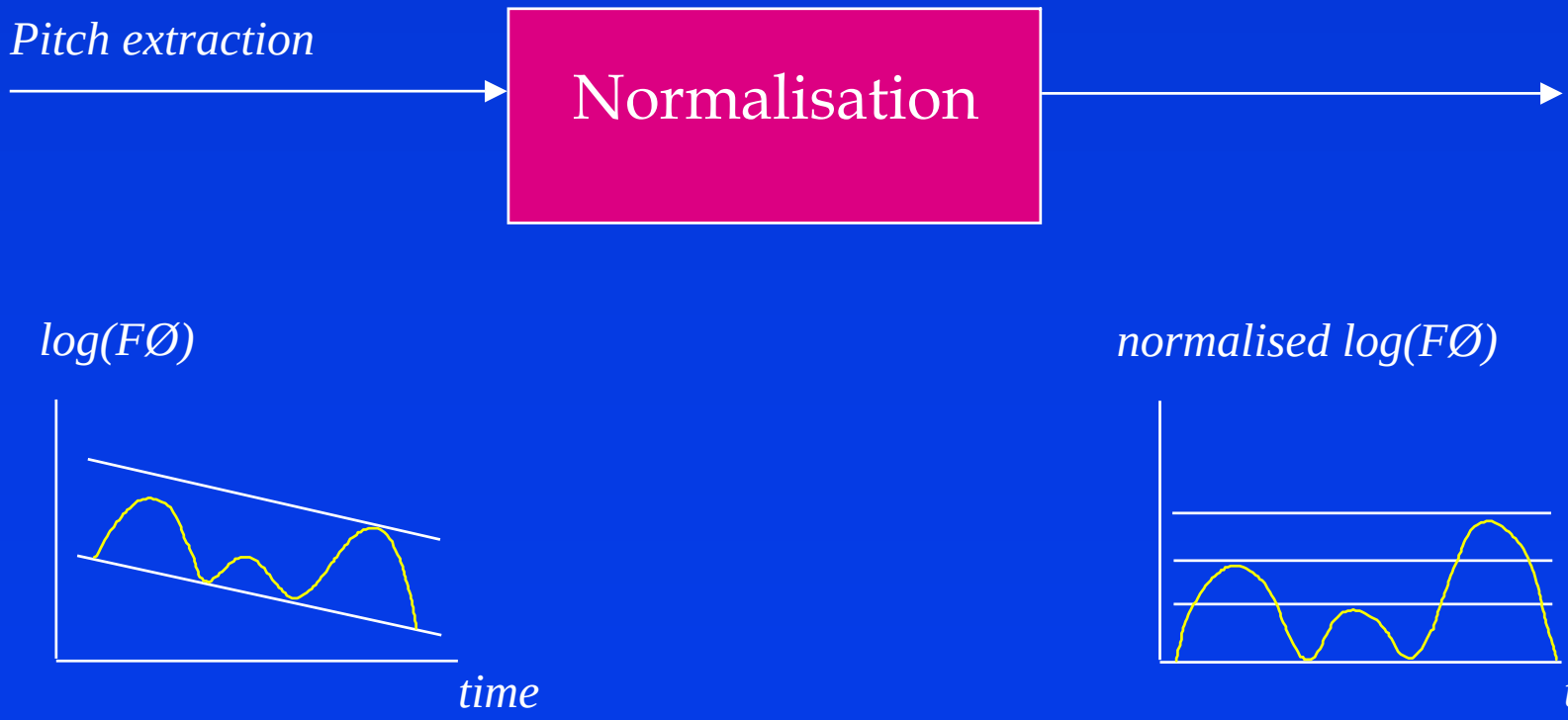
Telia Research AB
Sweden

Abstract

Current speech recognition systems mainly work on statistical bases and make no use of information signalled by prosody, i.e. the segment duration and fundamental frequency contour of the speech signal. In more advanced applications for speech recognition, such as speech-to-speech translation systems, it is necessary to include the linguistic information conveyed by prosody. Earlier research has shown that prosody conveys information at syntactic, semantic and pragmatic levels. The degree of linguistic information conveyed by prosody varies between languages, from languages such as English, with a relatively low degree of prosodic disambiguation, via tone-accent languages such as Swedish, to pure tone languages. The inclusion of a prosodic module in speech translation systems is not only vital in order to link the source language to the target language, but could also be used to enhance speech recognition proper. Besides syntactic and semantic information, properties such as dialect, sociolect, gender and attitude etc is signalled by prosody. Speech-to-speech recognition systems that will not transfer this type of information will be of limited value for person-to-person communication. A tentative architecture for the inclusion of a prosodic module in a speech-to-speech translation system is presented.

Fundamental Frequency Normalisation

Fundamental frequency, FØ, is extracted from the speech signal. The estimated FØ declination is subtracted from the actual FØ to give a normalised representation of FØ variation. The output signal is given in musical intervals.

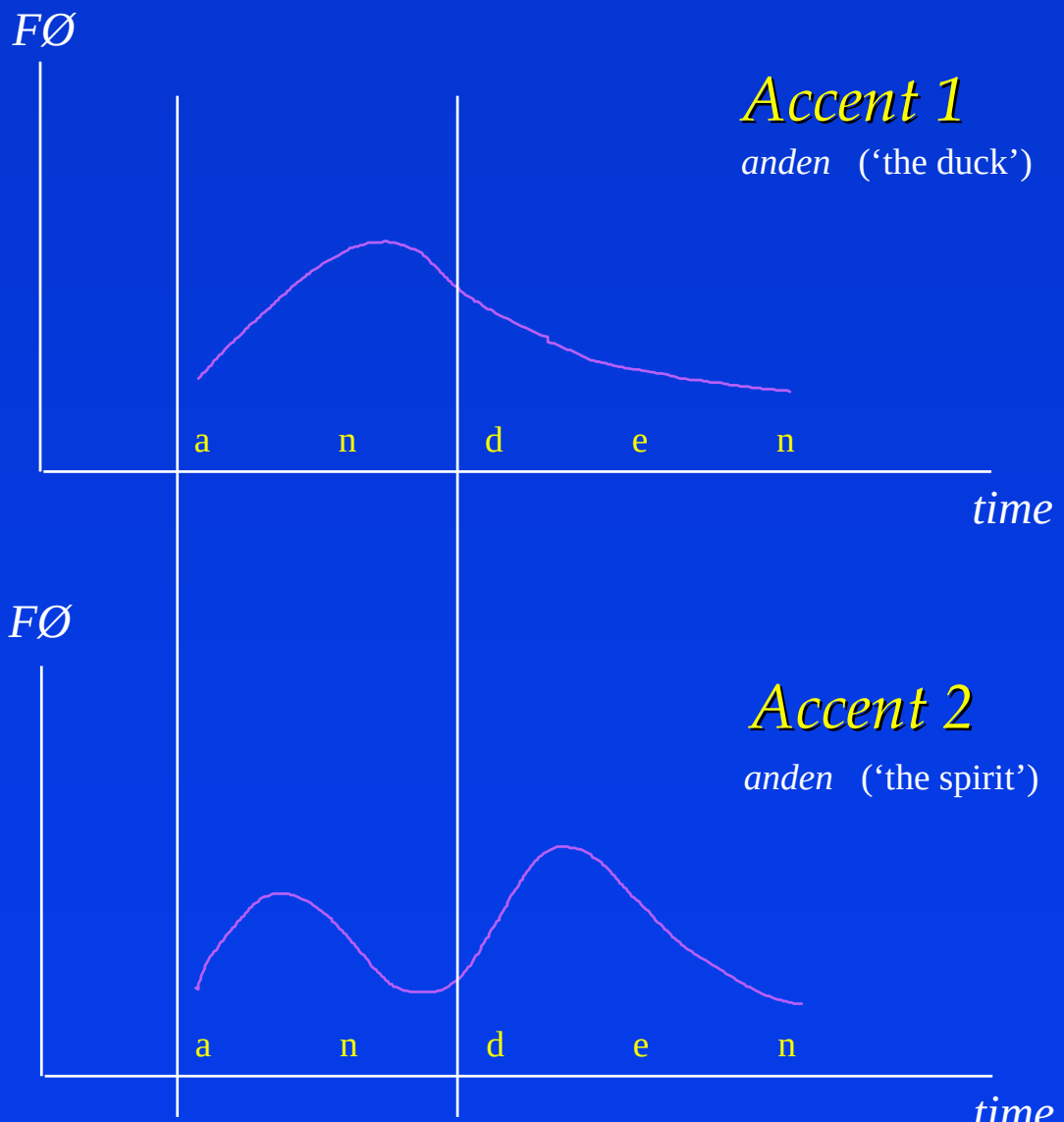


Swedish Tonal Accents

In tone accent languages such as Swedish, tone alone differentiates between lexical items. Tone accents are not stable the way tones are in tone languages, and may change or disappear through processes like compounding or stress.

In Swedish, accent is predictable from morphology. Sentence Accent is characterised by a rise in the main stressed syllable in Accent 1 words and a rise in the secondary stressed syllable in Accent 2 words. The main stressed syllable in Accent 2 words is characterised by a fall. The Accent 2 pattern is also typical of compound words.

Although fundamental frequency is the most important parameter in distinguishing between Accents 1 and 2, they also differ in intensity and duration.



Prosody: Its Features and Acoustic Correlates

Linguistic information

- *Segmental*
- *Suprasegmental*
  - *Statement / question*
  - *Accent 1 / Accent 2*
  - *Stressed / unstressed syllable*
  - *Focus*
  - *Quantity*
  - *Juncture*

Acoustic correlates

- *Fundamental frequency*
- *Intensity*
- *Duration*

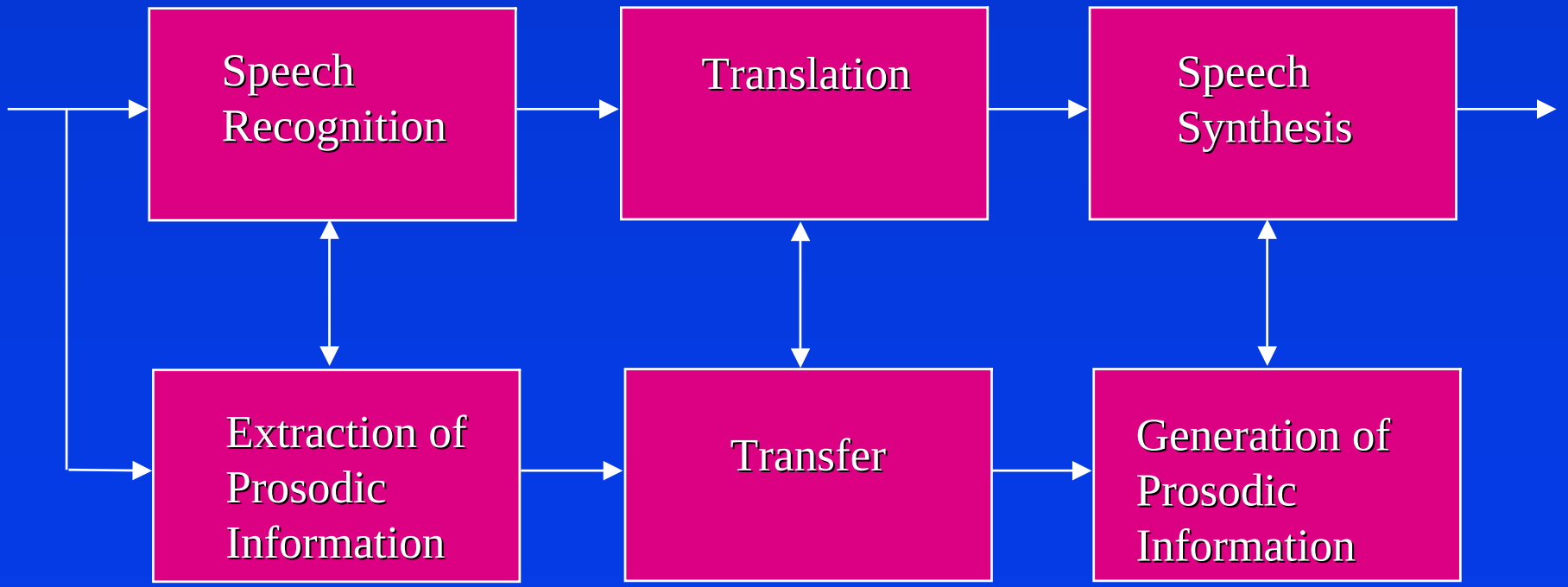
Extra-linguistic information

- *Gender*
- *Attitude*
- *Speech rate*

The term prosody normally refers to the features pitch, stress and quantity, whose acoustic correlates are fundamental frequency, intensity and duration.

To be noted is that there is no one-to-one relationship between the prosodic parameters and their acoustic manifestation.

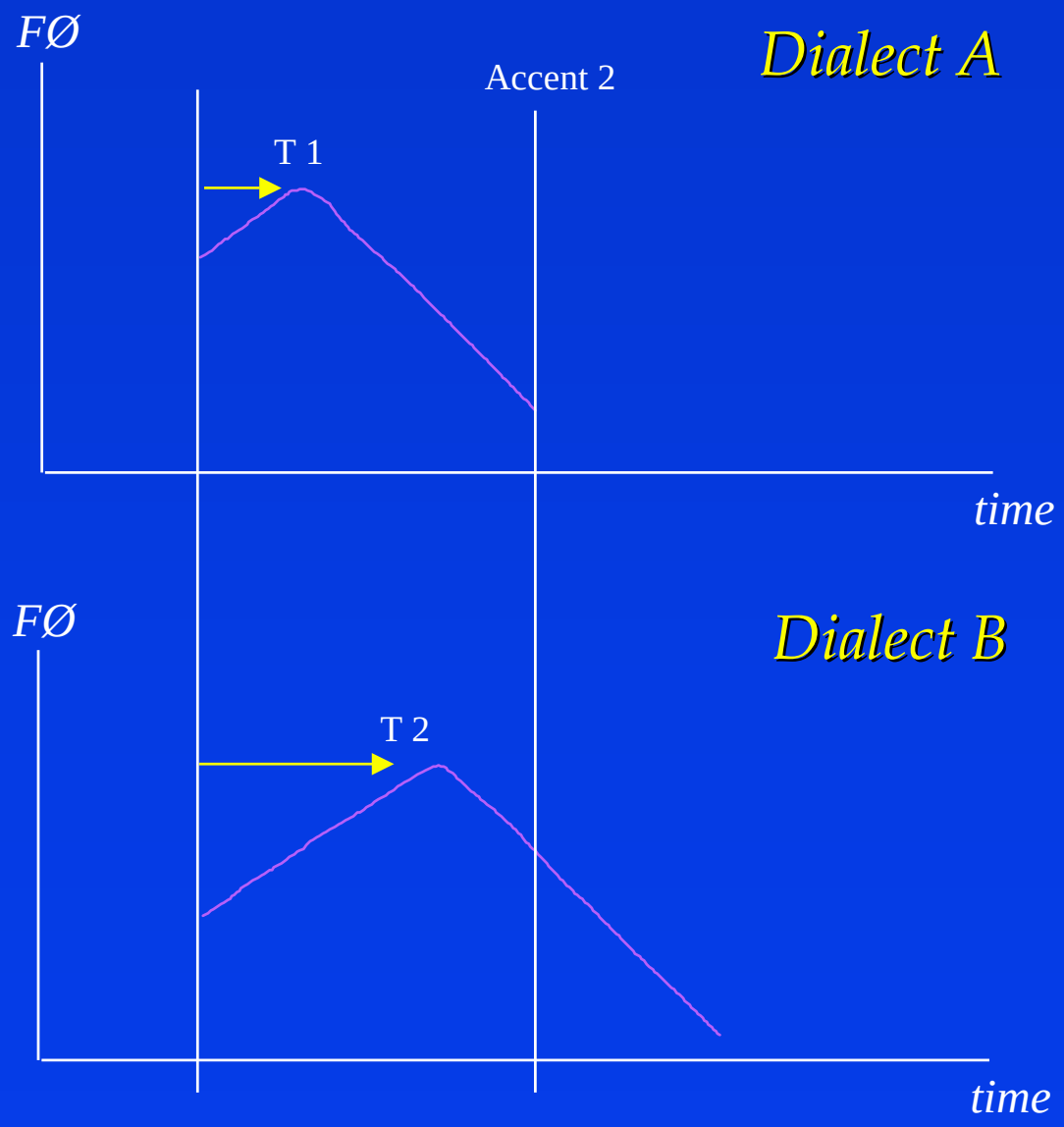
Spoken Language Translation with Transfer of Prosodic Information



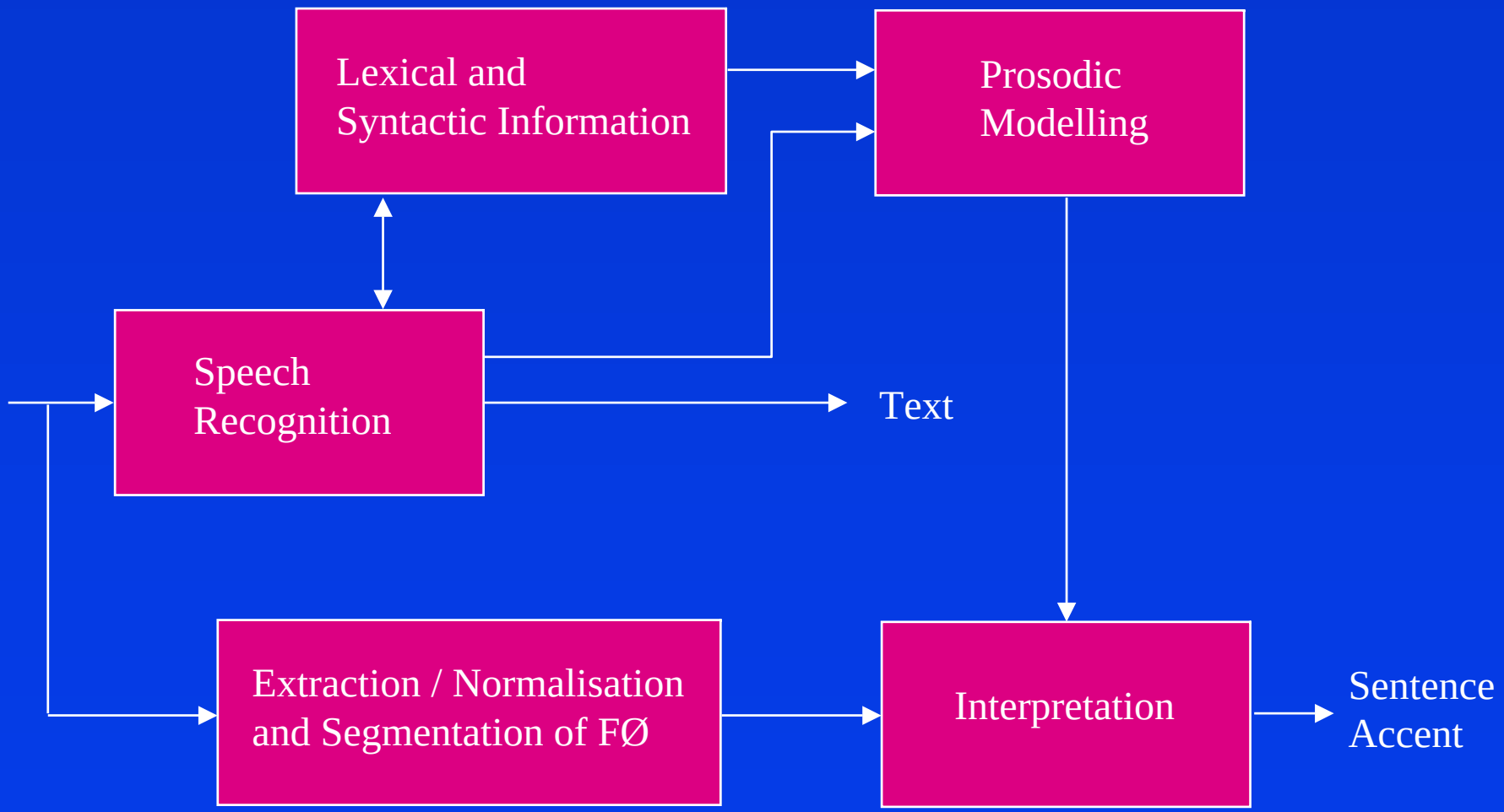
Dialect Recognition & Generation

Dialects of a language differ in their prosodic realisation, allophonic variation, vocabulary and syntax.

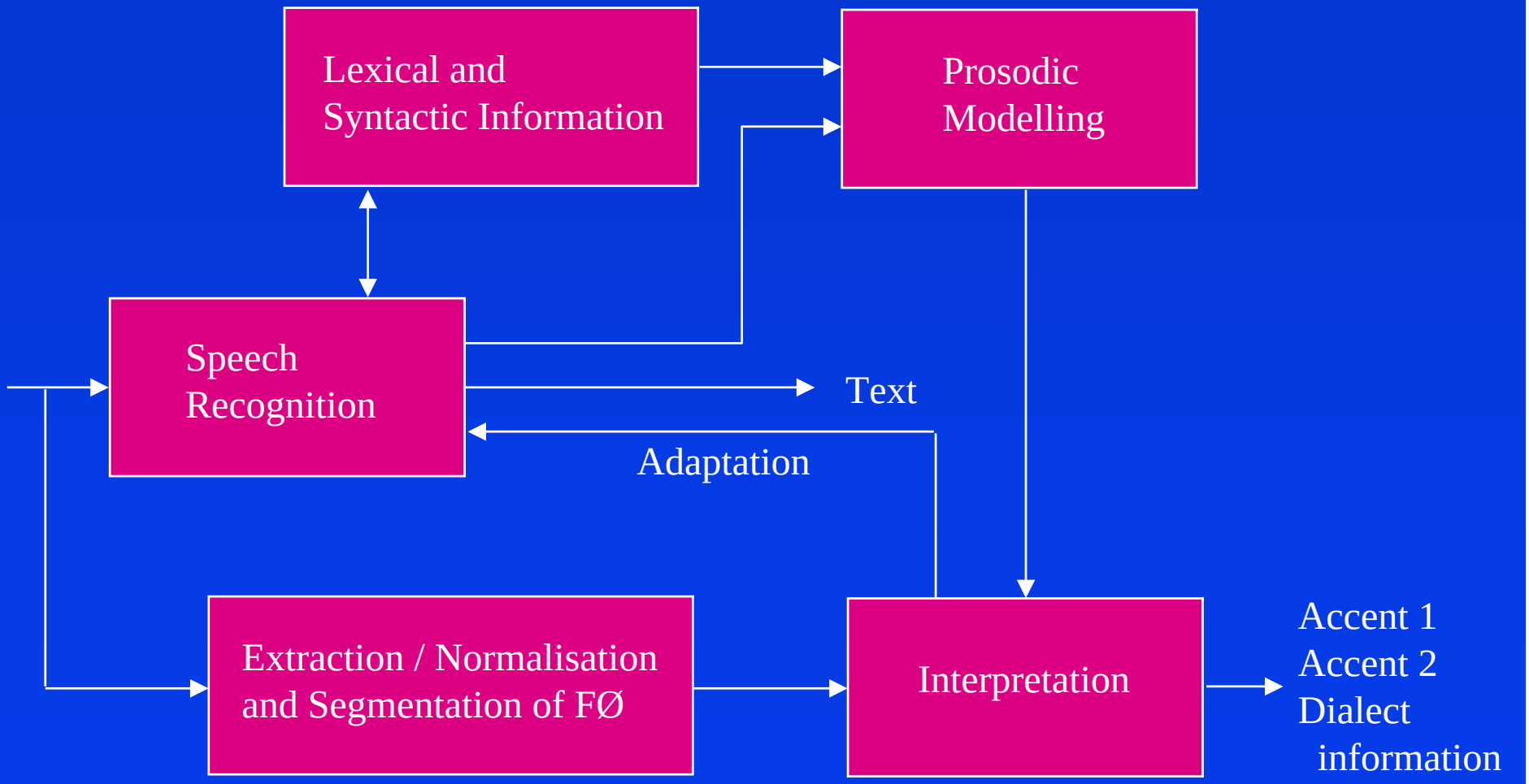
Accent 1, Accent 2 and Sentence Accent are realised in different ways in Swedish dialects. By comparing the segment string and the fundamental frequency contour, one may determine what dialect is being spoken.



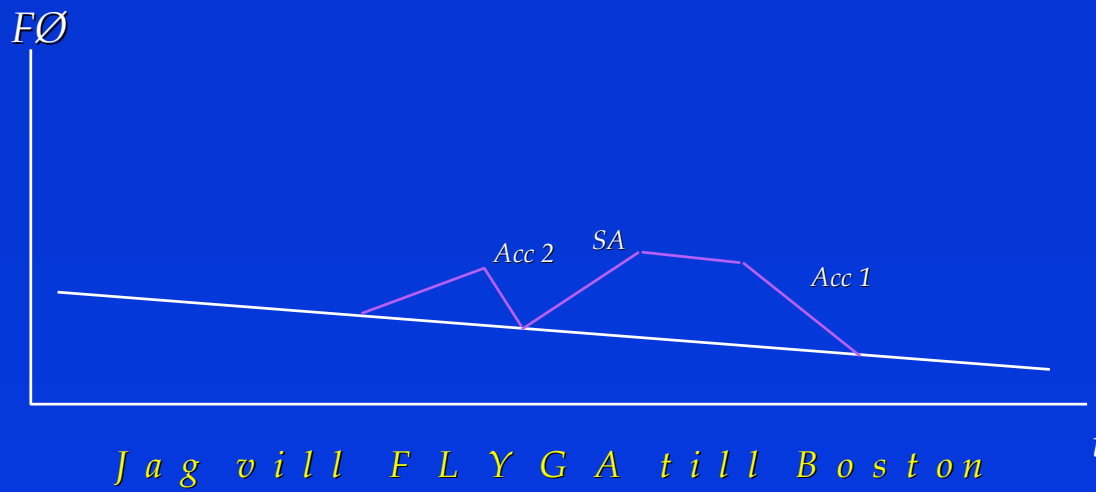
Extraction of Sentence Accent



Extraction of Tone Accent & Dialect Information in Swedish

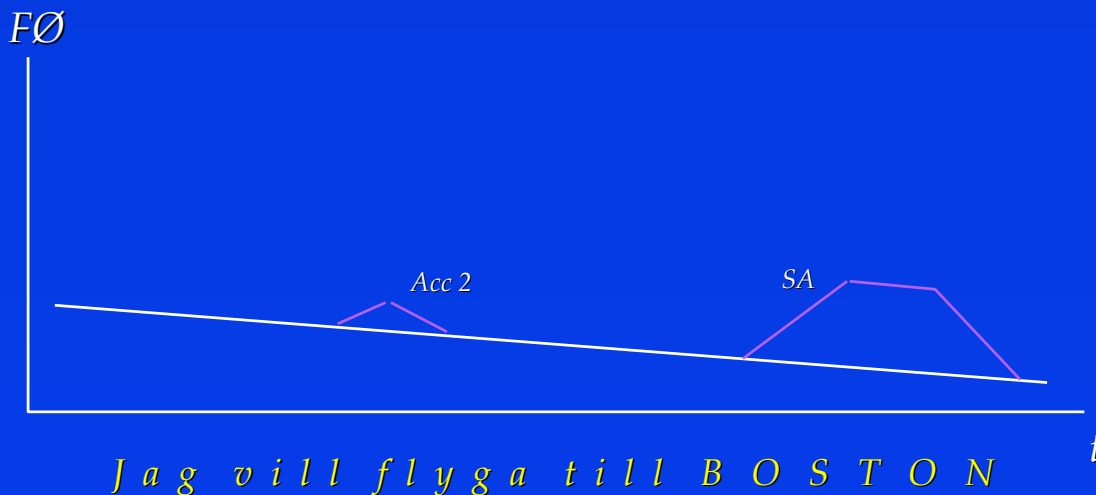


Sentence Accent Generation



Generation of the utterance

Jag vill FLYGA till Boston.
('I want to FLY to Boston')



Generation of the utterance

Jag vill flyga till BOSTON.
('I want to fly to BOSTON')