Language Identification (LID)

Applications of LID System

- Preprocessing for various speech tasks
  - Language specific speech recognition
  - Language translation systems
- Multilingual voice controlled information retrieval system
- Route the telephone call to the human operator based on the language
- Multilingual applications

Language Specific Features

- Frame level
  - WLPCC, LP residual, Phase of the LP residual
- Syllable level
  - Spectral & Prosodic features
- Multi-syllable level
  - Prosody characteristics
  - Phonotactics
  - Positional features

Language Specific Aspects of Speech

- Acoustic-phonetics
  - Articulatory configuration of vocal tract for different sound units
- Phonotactics
  - Rules governing the way different phonemes are combined to form sub-words are different in different languages
- Prosody
- Vocabulary & lexical structure
  - Word roots and lexicon are different for different languages

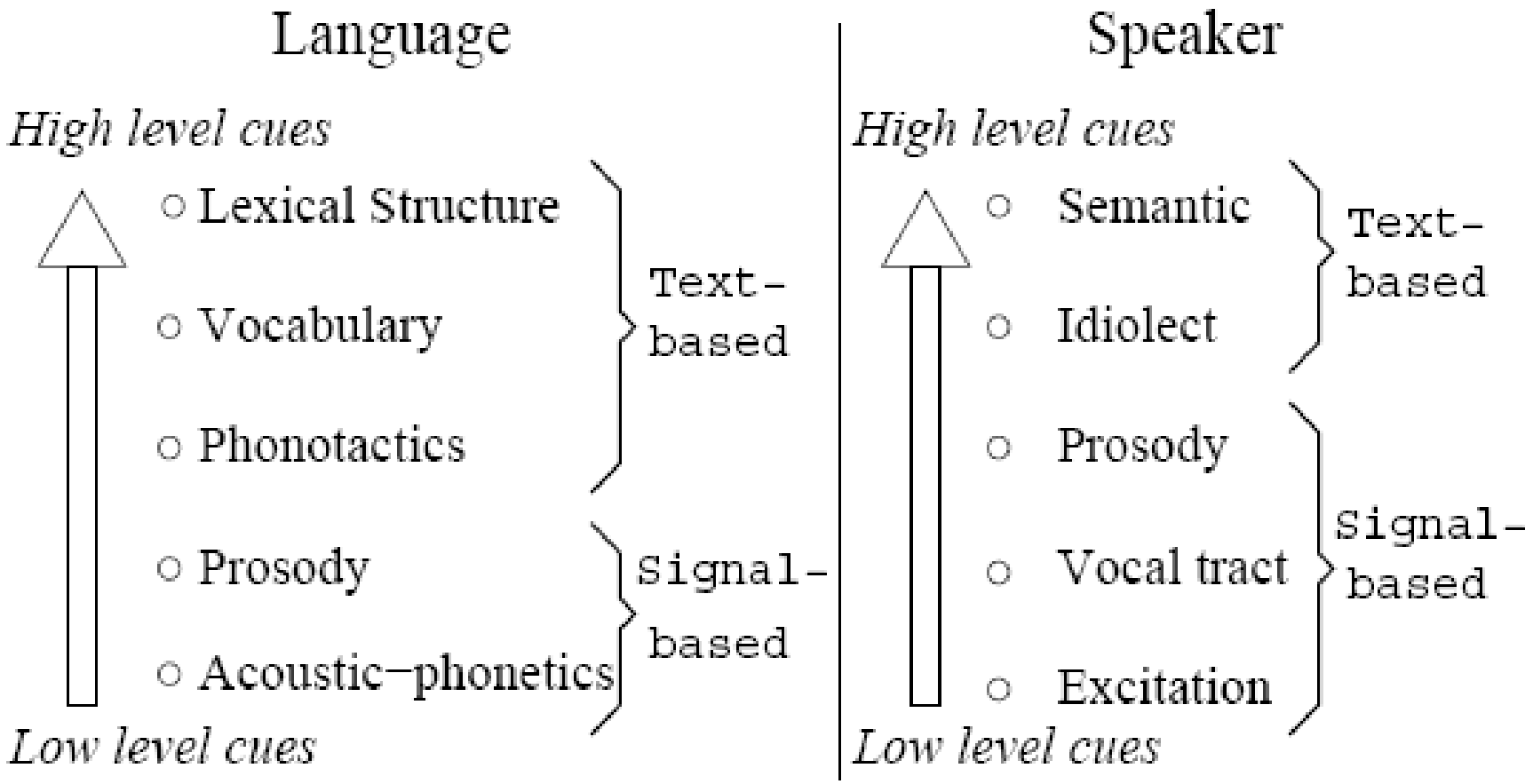
Implicit LID Features

- Frame level
  - Production level constraints specific to language
- Syllable level
  - Acoustic variations in the realization of syllables
- Multi-syllable level
  - Phonotactic constraints and Prosodic constraints

Issues in LID

- Variability in speaker characteristics
- Variability in accents
- Variability in the environment and channel characteristics
- Extraction and representation of language specific prosody
  - Rhythm of a language
  - Melody of a language

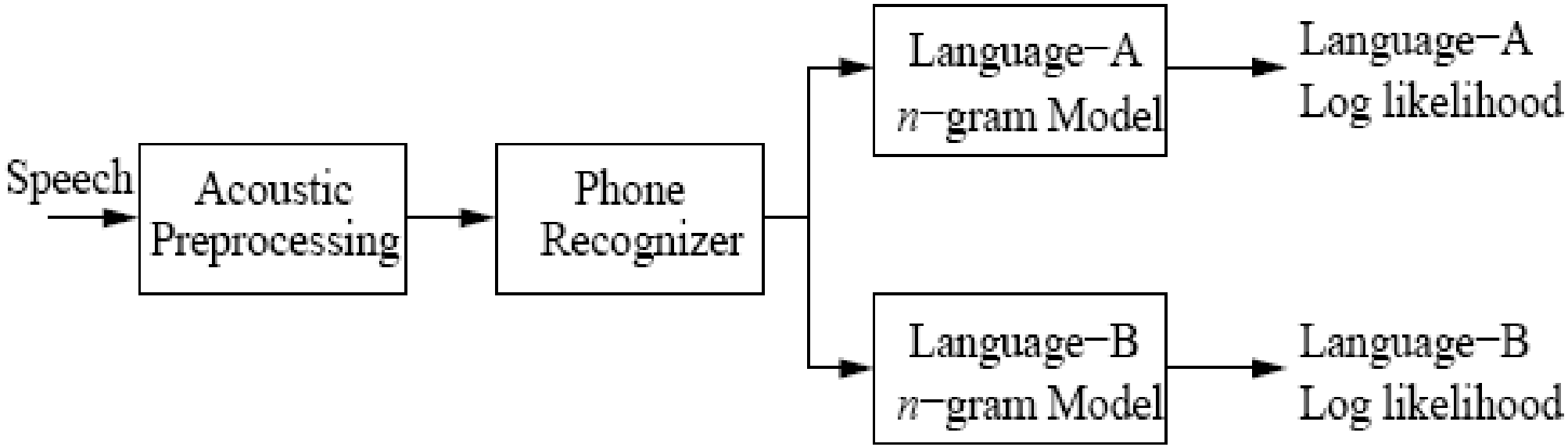
Language and Speaker-Specific Cues and their Levels of Manifestation



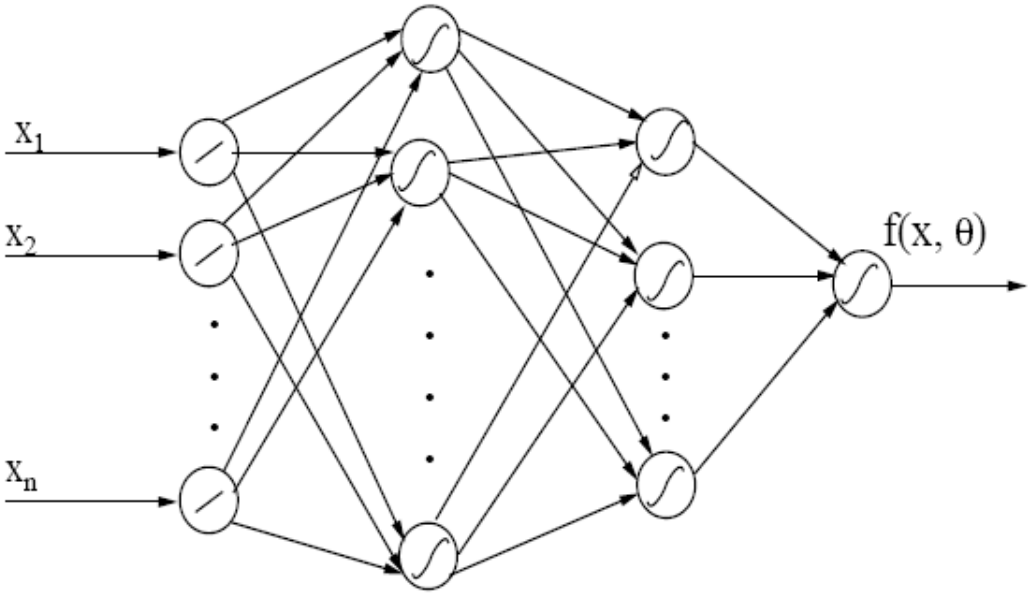
Language Discriminating Cues and their Representation

Cues for LID	Representation
Acoustic-phonetics	Spectral features (MFCC, LPCC)
Prosody	Features from duration, F_0 and amplitude
Phonotactics	Sequences of subword labels
Vocabulary and lexical structure	Sequences of word transcriptions

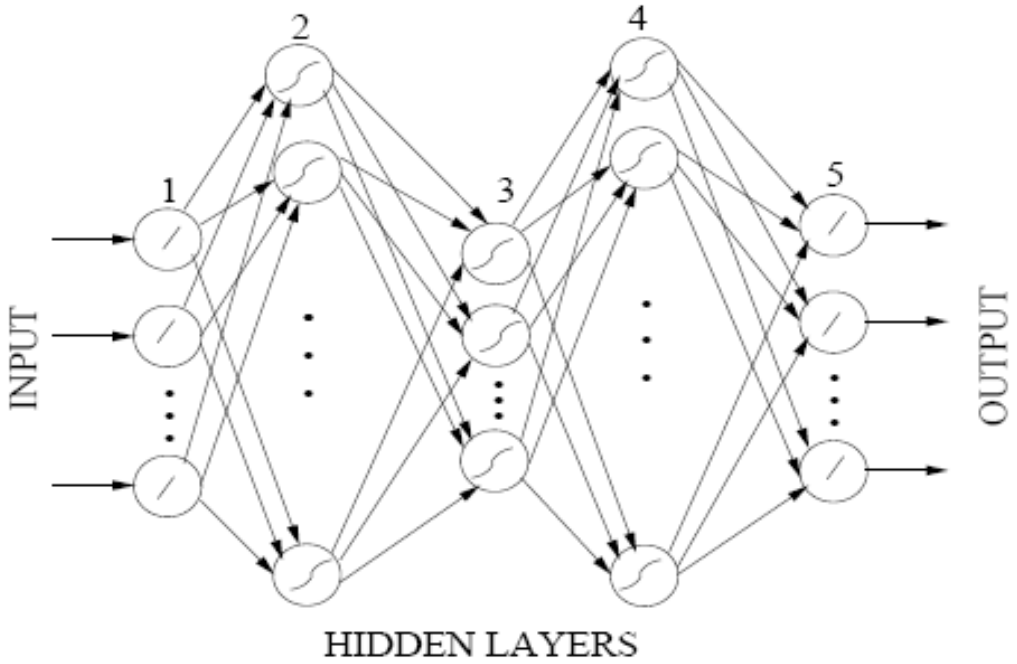
Phone based LID System



NN Models for Developing Language Models



FFNN Model

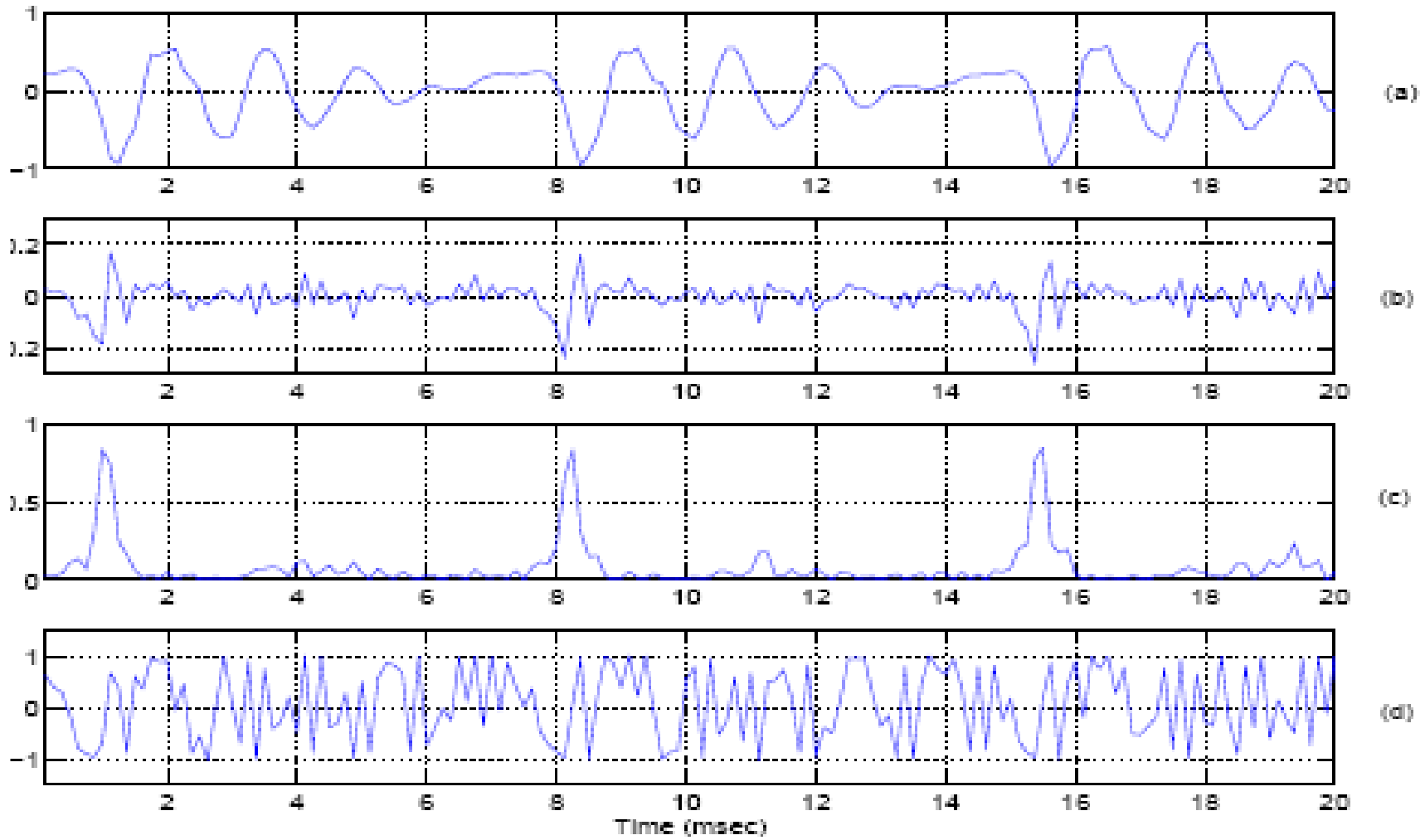


AANN Model

LID system using frame level features

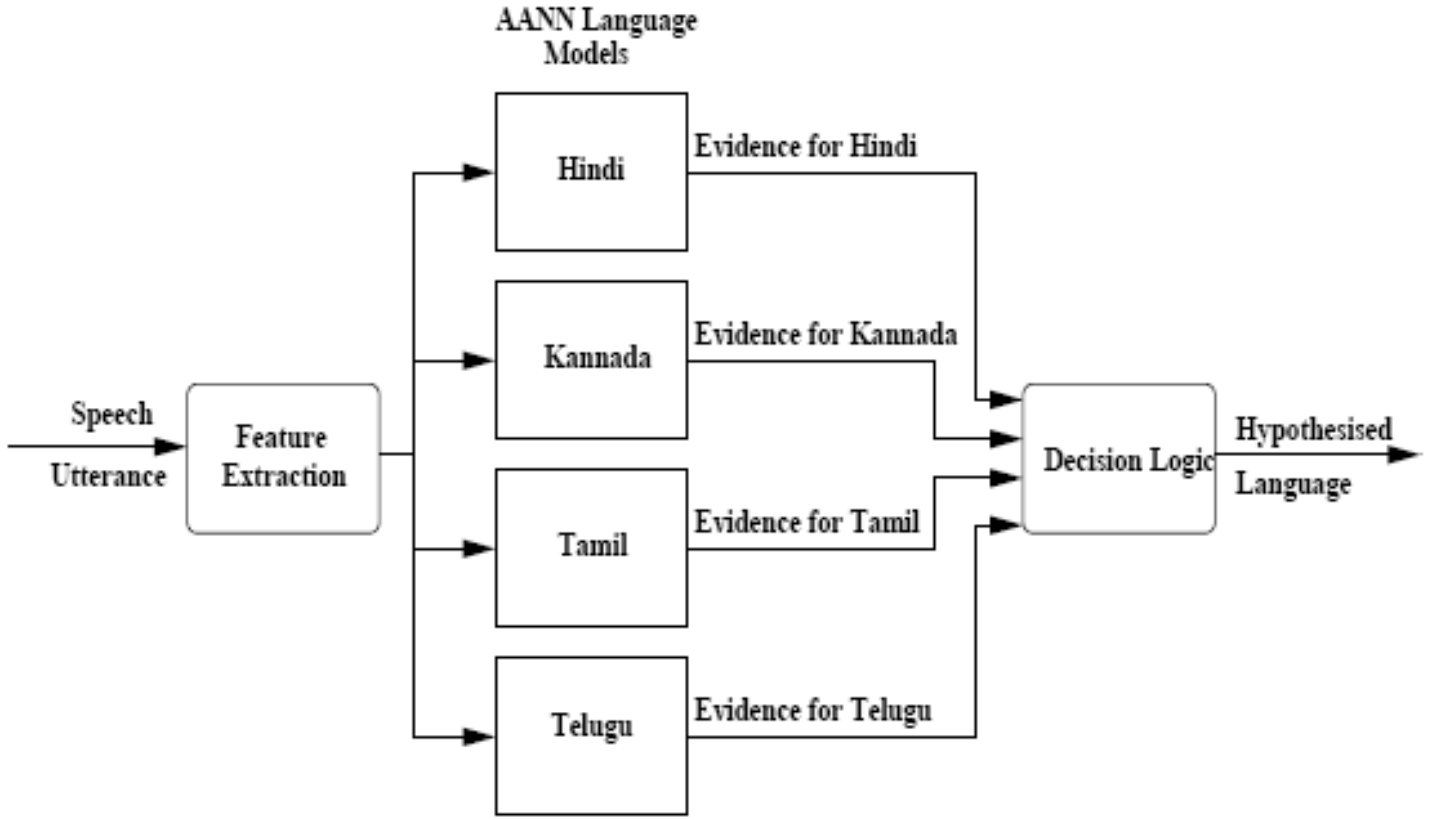
- Spectral features (WLPCC)
- LP residual
- Phase of the LP residual

Excitation source features at frame level

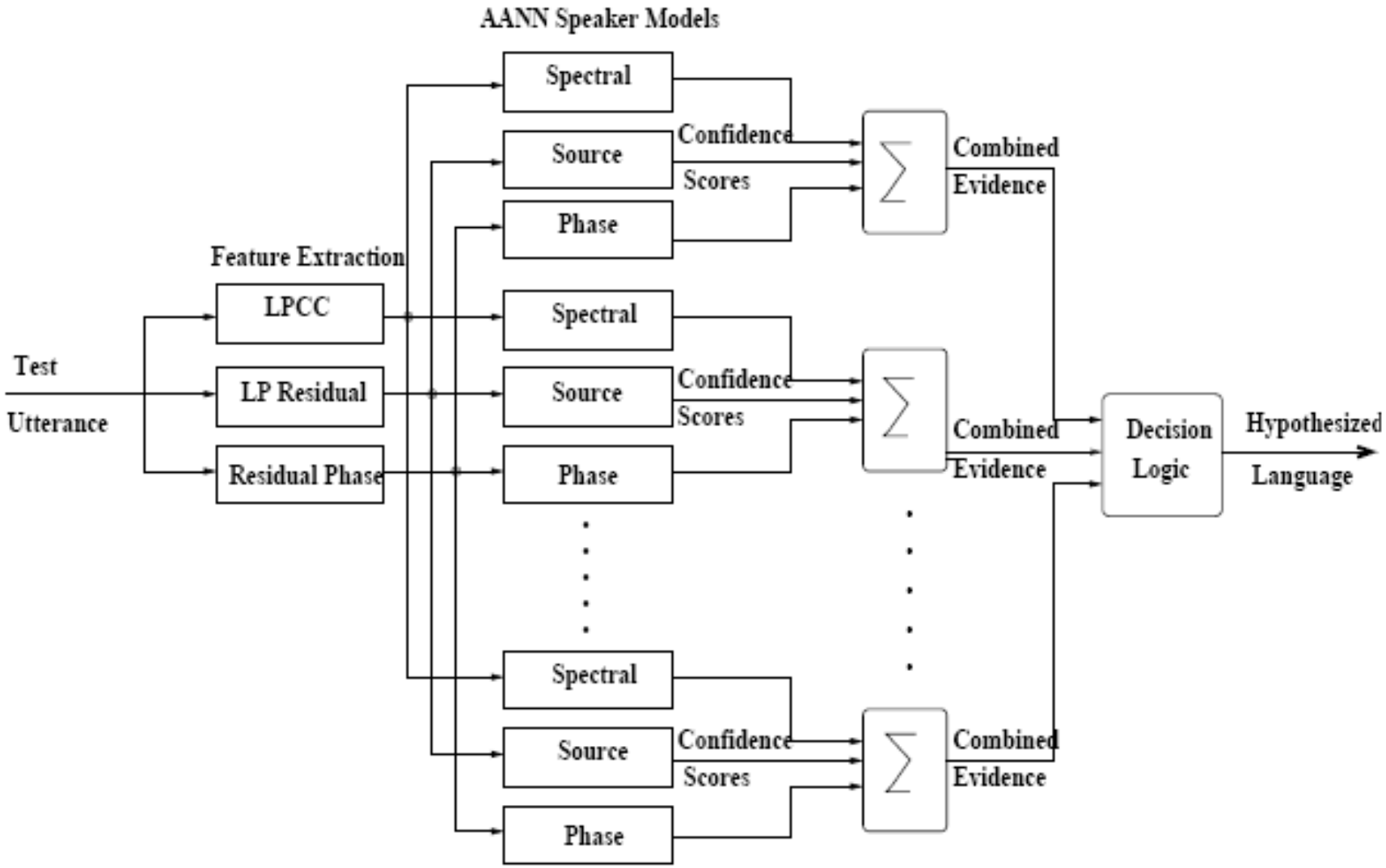


(a) Speech segment, its (b) LP residual, (c) Hilbert envelope of the LP residual and (d) Phase of the LP residual

LID using spectral features

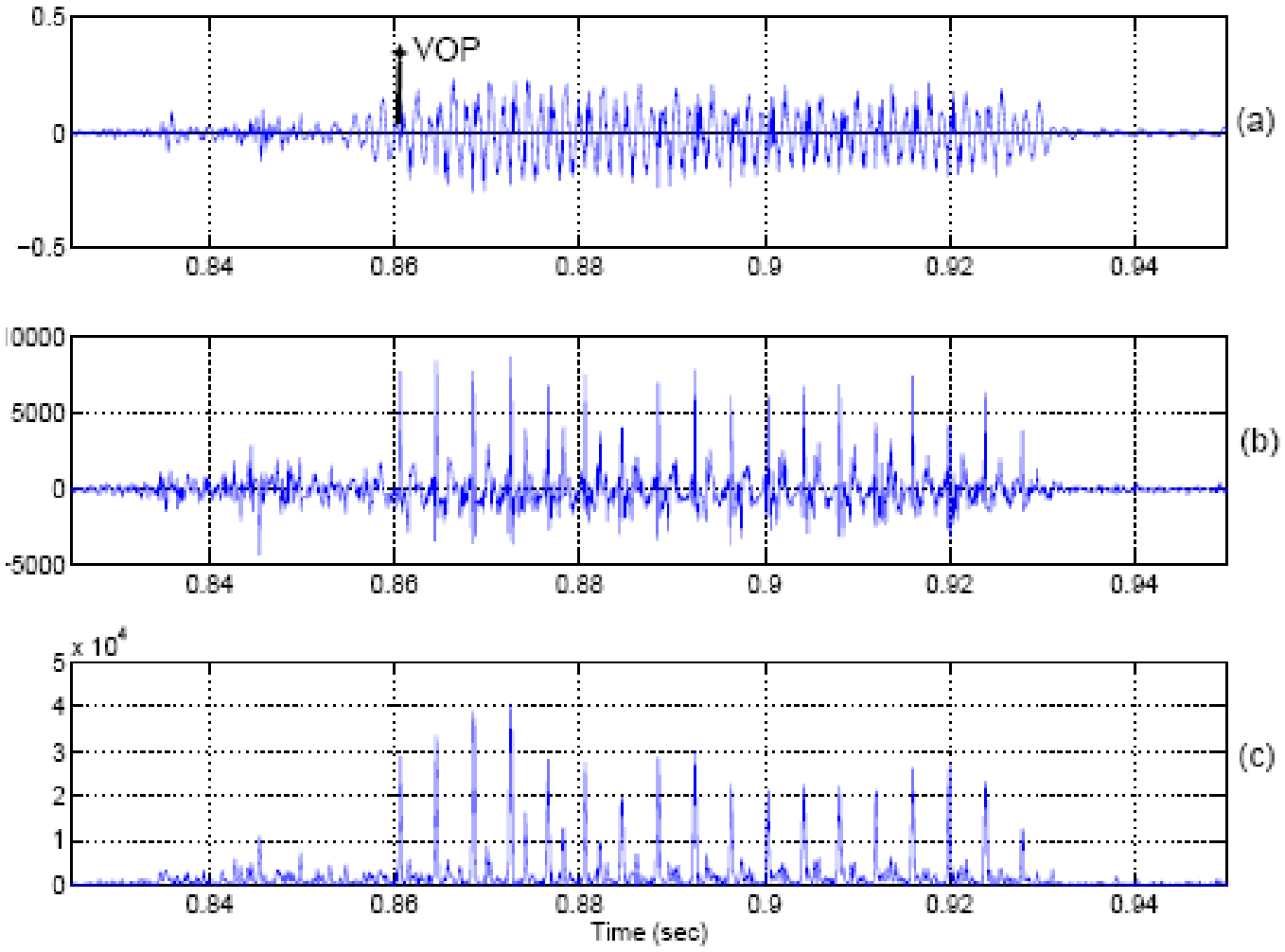


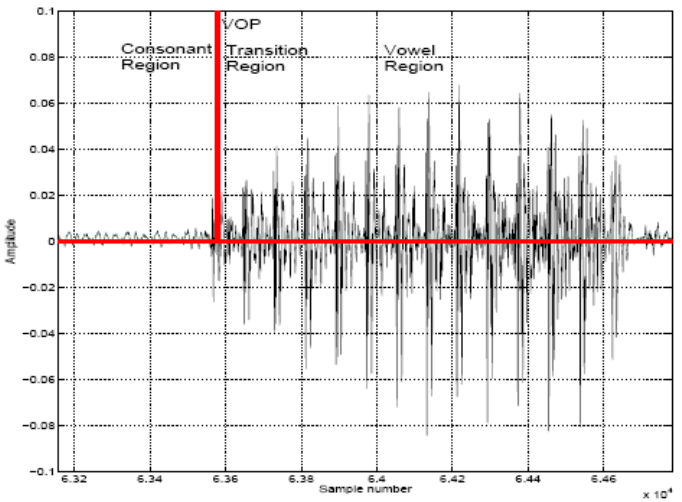
LID system using frame level features (Spectral + Source + Phase)



Syllabic features for Language Identification

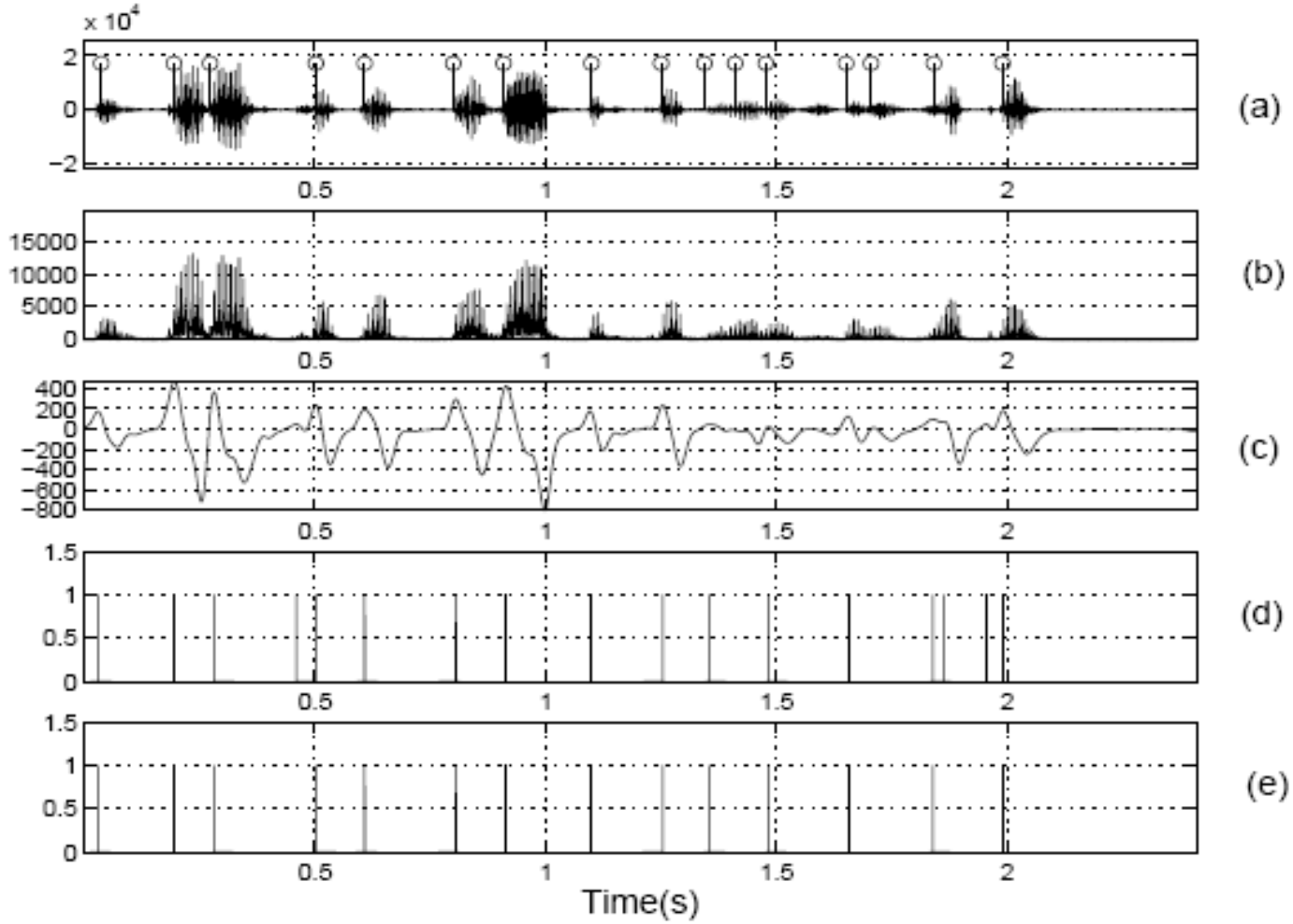
Choice of CV as the Basic Unit





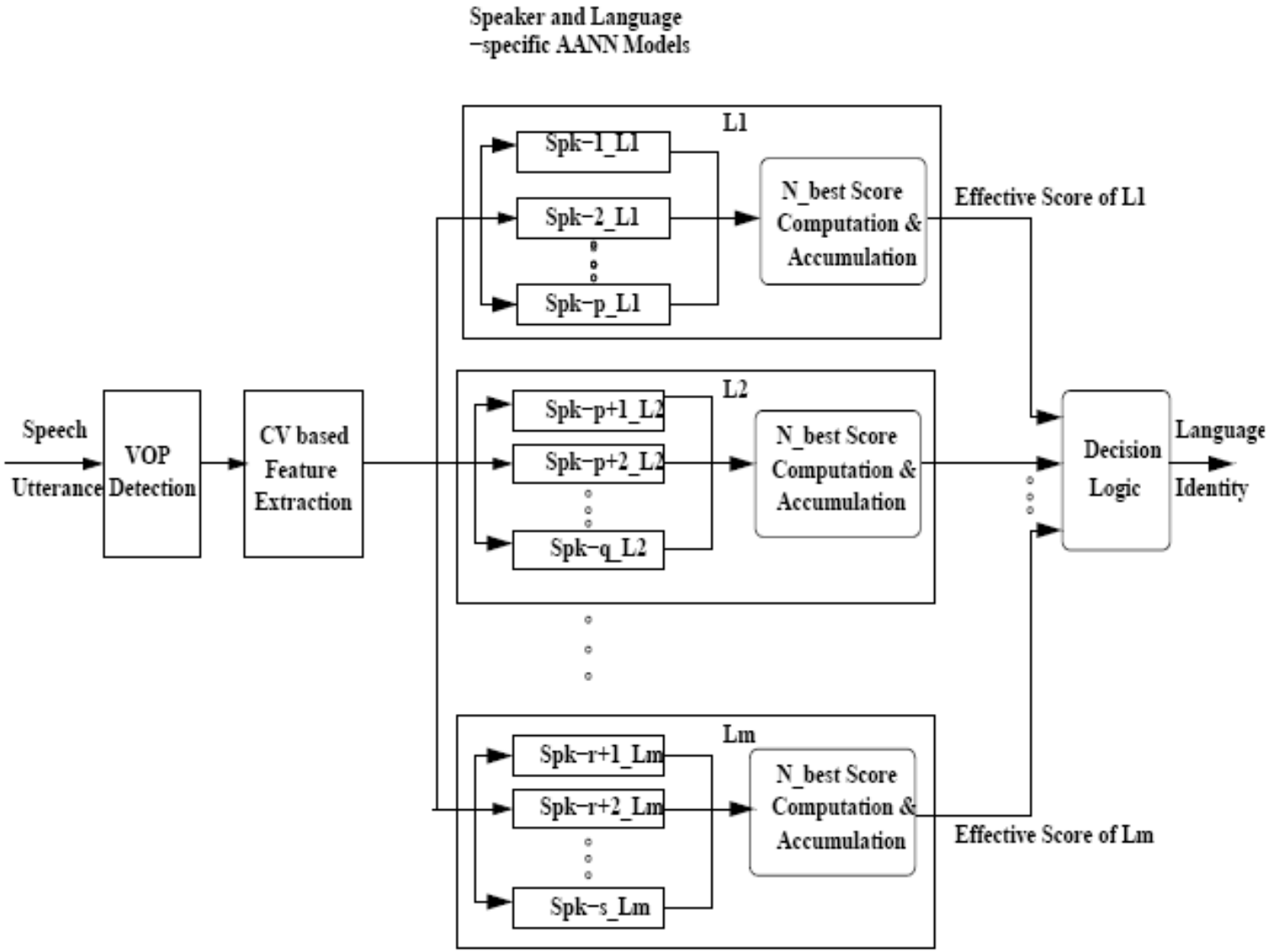
Strength of excitation for locating VOP

Detection of VOPs



(a) Speech signal, its (b) Hilbert envelope of the LP residual, (c) VOP evidence plot (d) Output of the peak picking algorithm and (e) hypothesized VOPs after removing the spurious ones

LID System using Syllable Features



LID using Multisyllabic Features

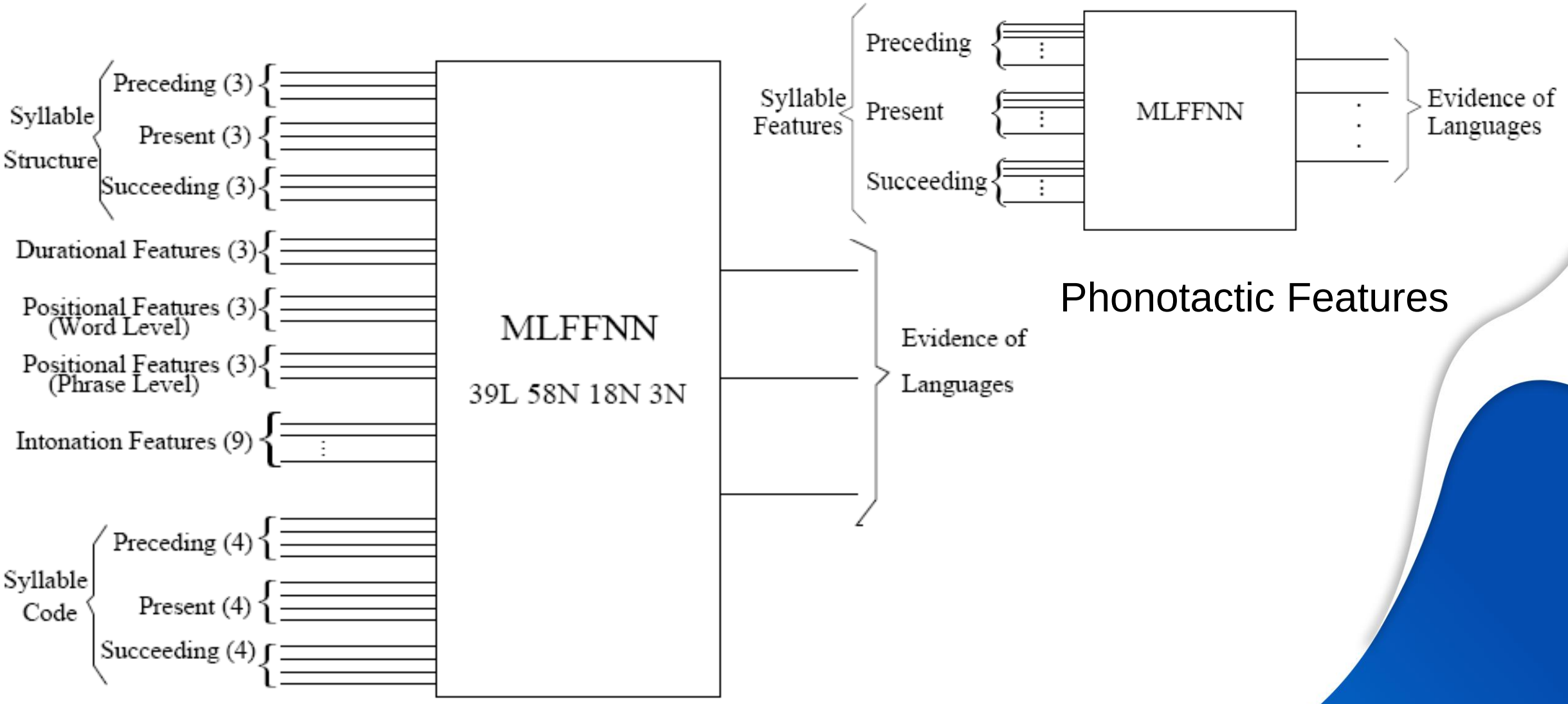
Multisyllabic Features

- Phonotactics
- Broad phonotactics (POA & MOA)
- Prosodic features
- Prosodic + Phonotactic features

Multisyllabic features (cont..)

Characteristics	Variable	# Parameters	Representation	Dimension
Phonotactics	Syllable identity	4	trisyllabic	12
Broad phonotactics	Category of syllable	4	trisyllabic	12
Phonotactics and prosody	Syllable identity	4	trisyllabic	39
	Syllable structure	3	trisyllabic	
	Intonation	3	trisyllabic	
	Duration	1	trisyllabic	
	Positional	6	monosyllabic	
Prosody	Syllable structure	3	trisyllabic	27
	Intonation	3	trisyllabic	
	Duration	1	trisyllabic	
	Positional	6	monosyllabic	

Phonotactic+Prosodic Features



Phonotactic Features

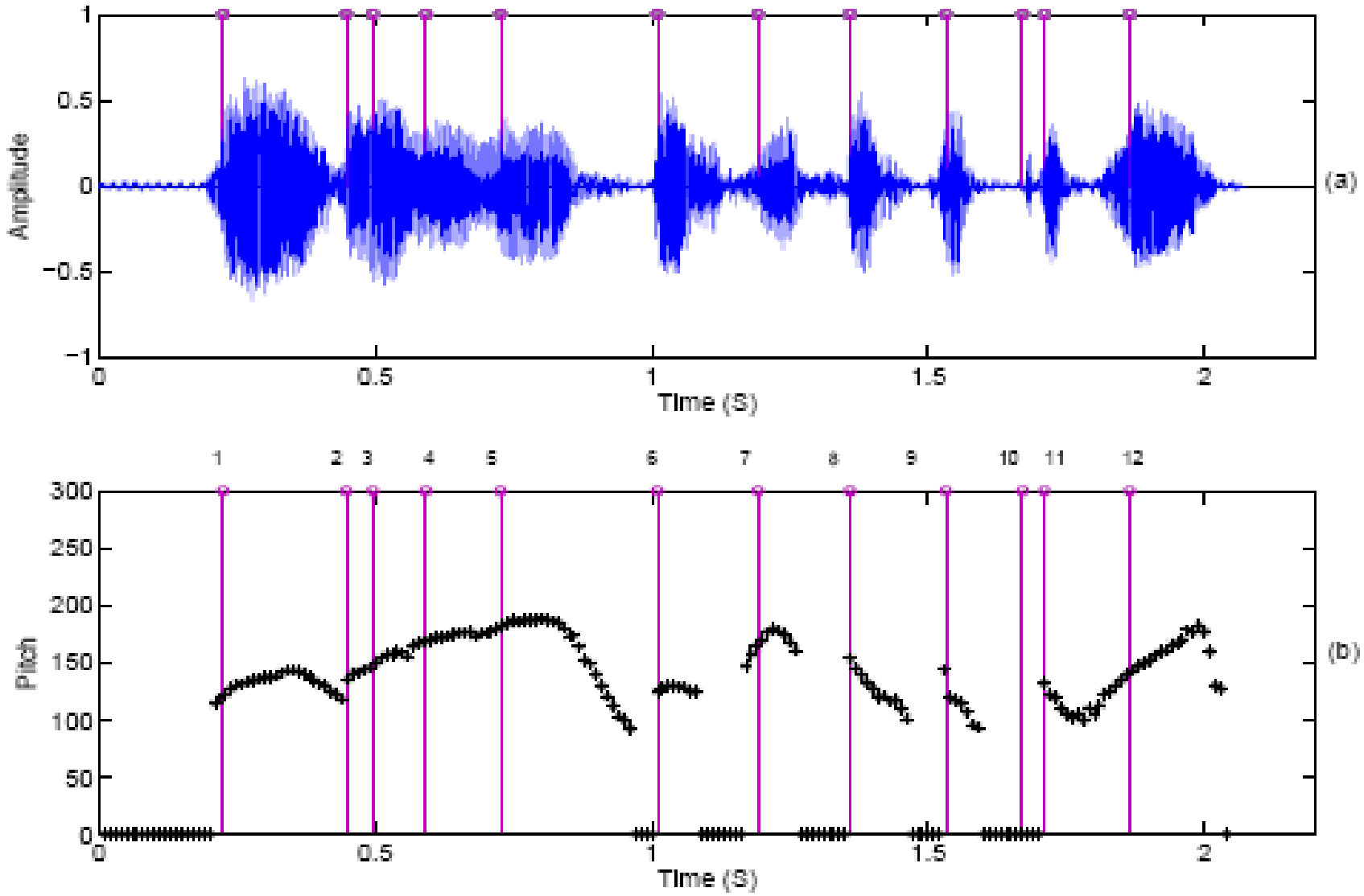
Phonotactic + Prosodic Features

Deriving Prosodic Features from Unlabelled Speech

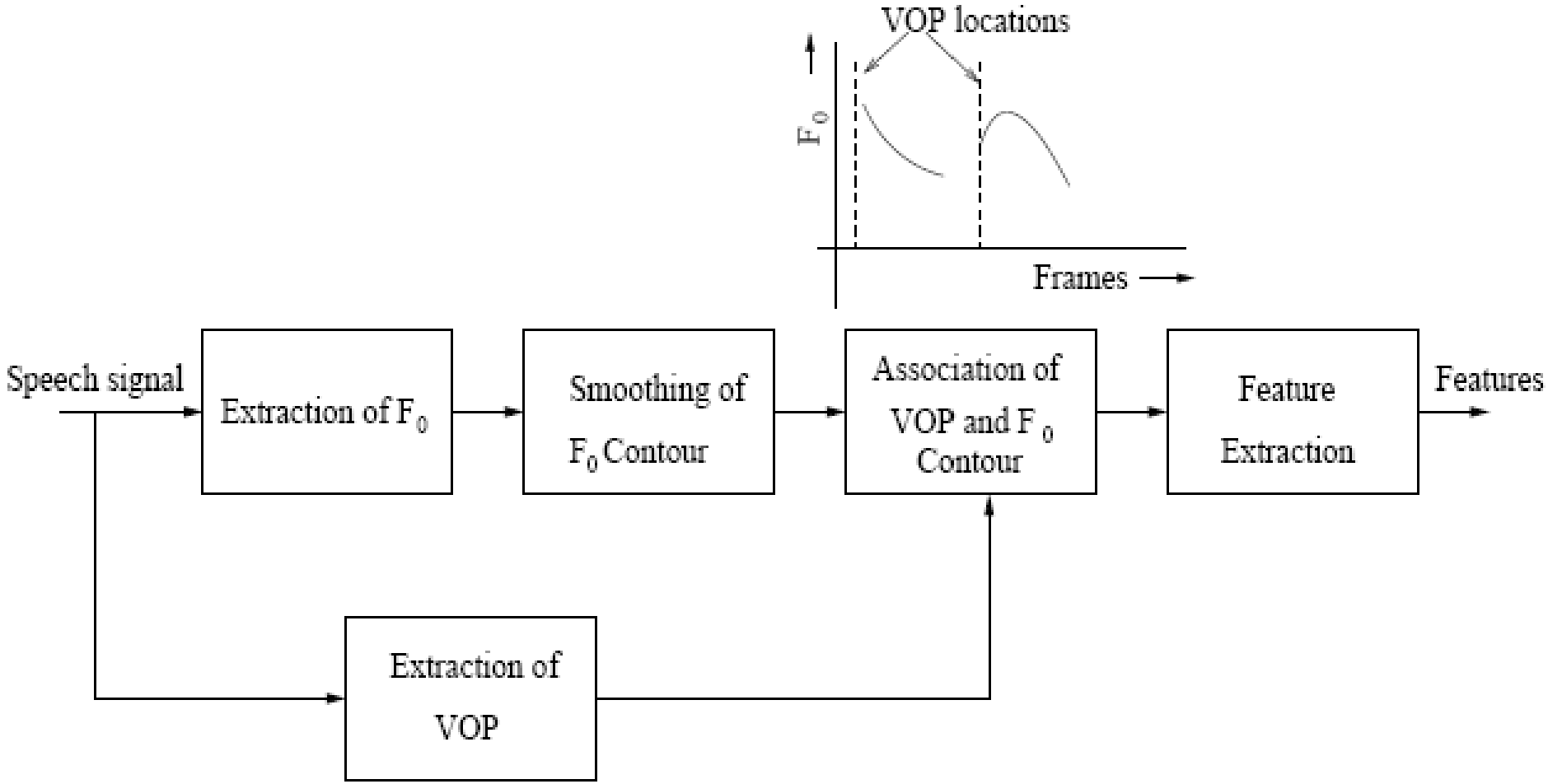
Prosodic Parameters

- Intonation
  - Change in F0
  - Distance of F0 peak from VOP
  - Amplitude tilt
  - Duration tilt
- Rhythm
  - Syllable duration
  - Duration of voiced region
- Stress
  - Change in the log-energy in the voiced region

F0 contours associated with VOPs



Deriving the F0 features using VOP locations



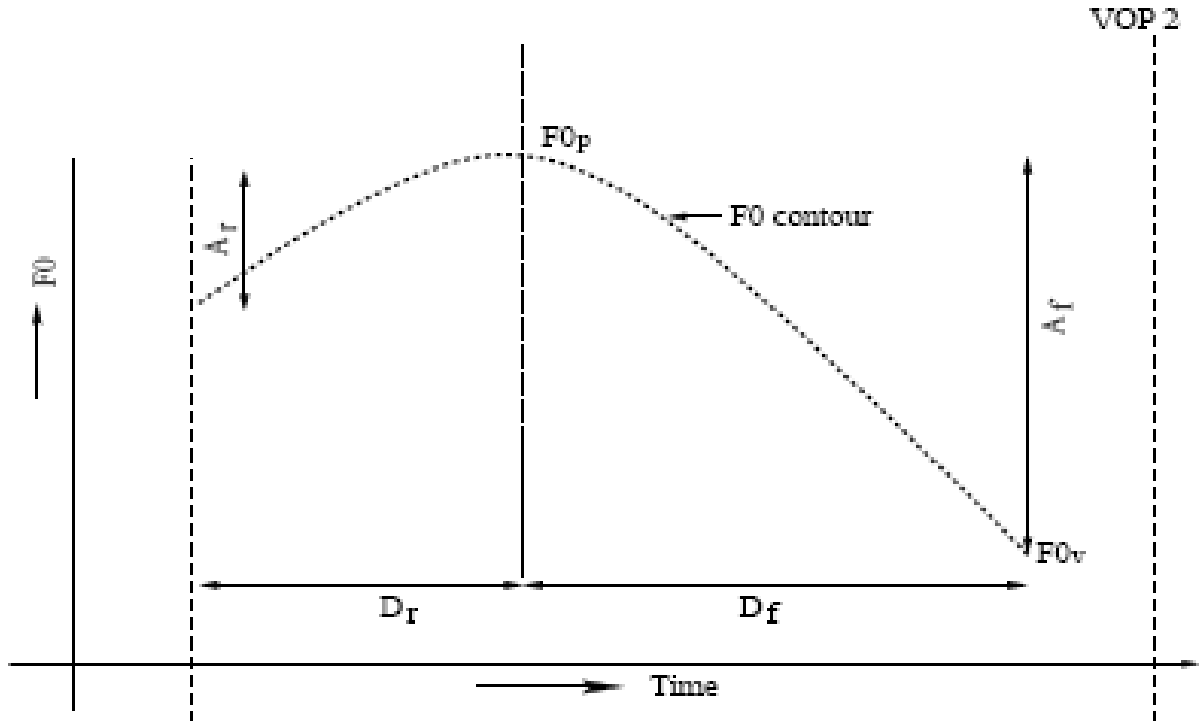
F0 parameter extraction

Change in F_0 (ΔF_0)

Distance of F_0 peak with respect to VOP (D_p)

Amplitude tilt (A_t)

Duration tilt (D_t)



$$A_t = \frac{|A_r| - |A_f|}{|A_r| + |A_f|},$$

$$D_t = \frac{|D_r| - |D_f|}{|D_r| + |D_f|},$$

Illustration of F0 representation using tilt parameters

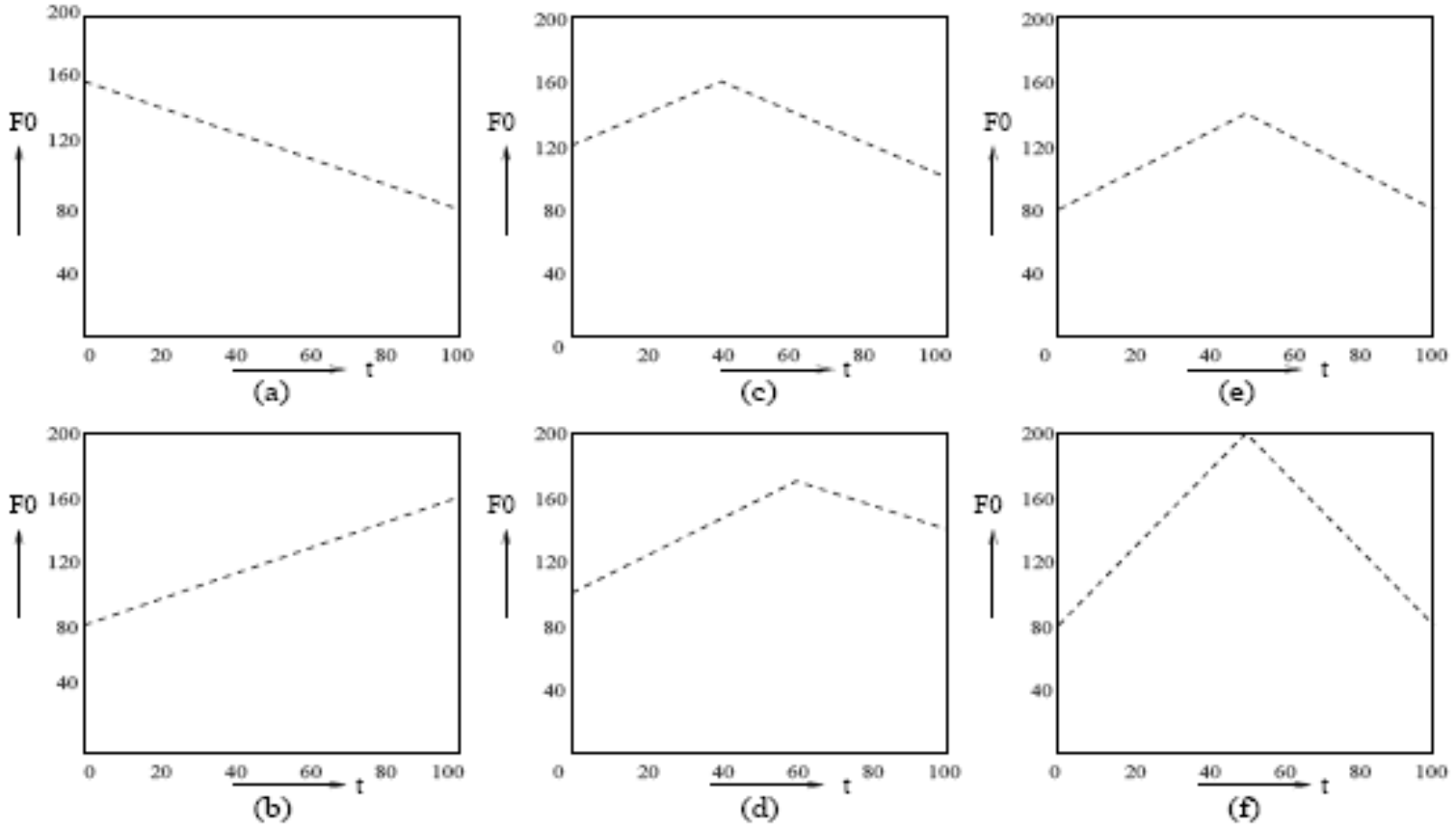
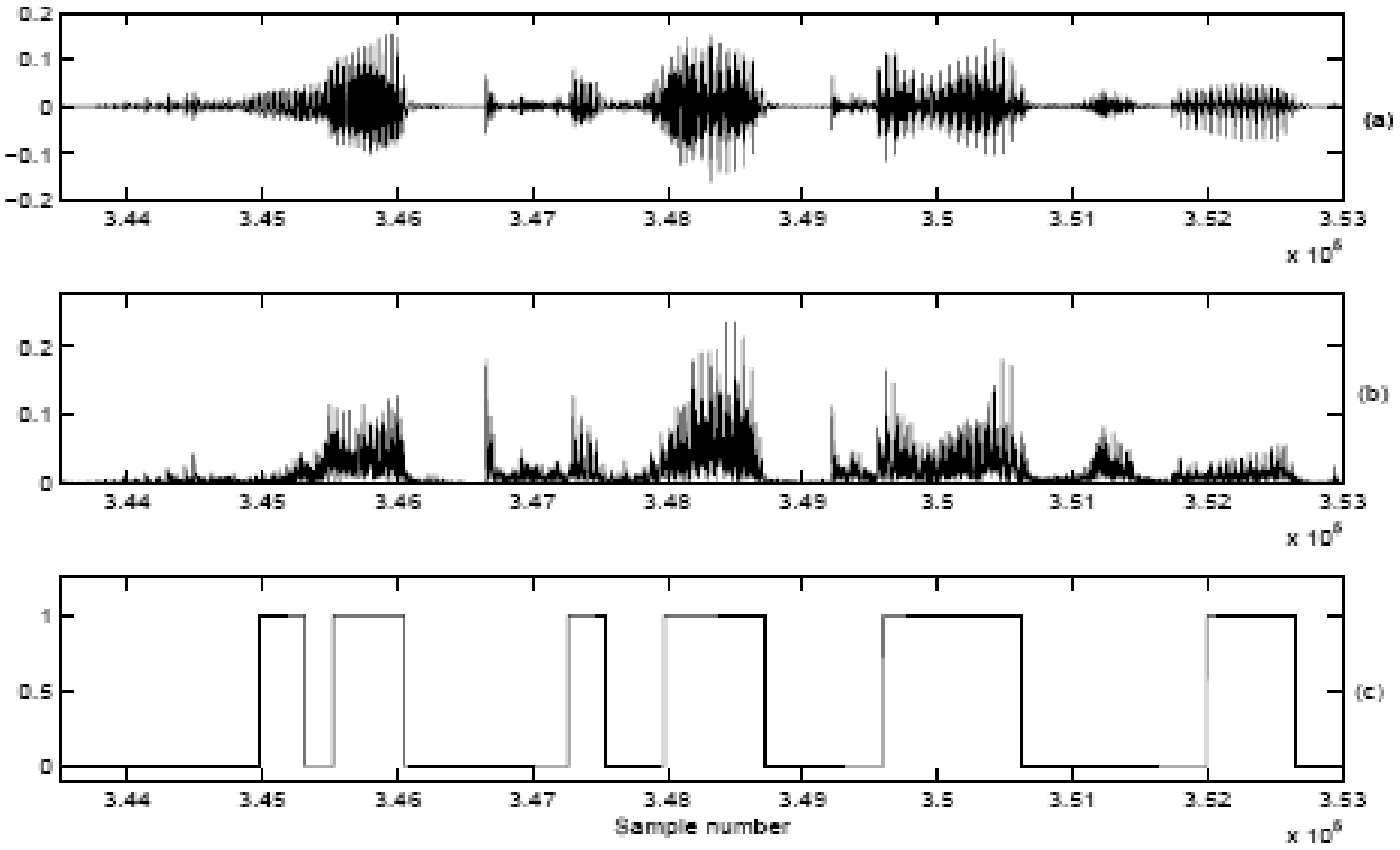
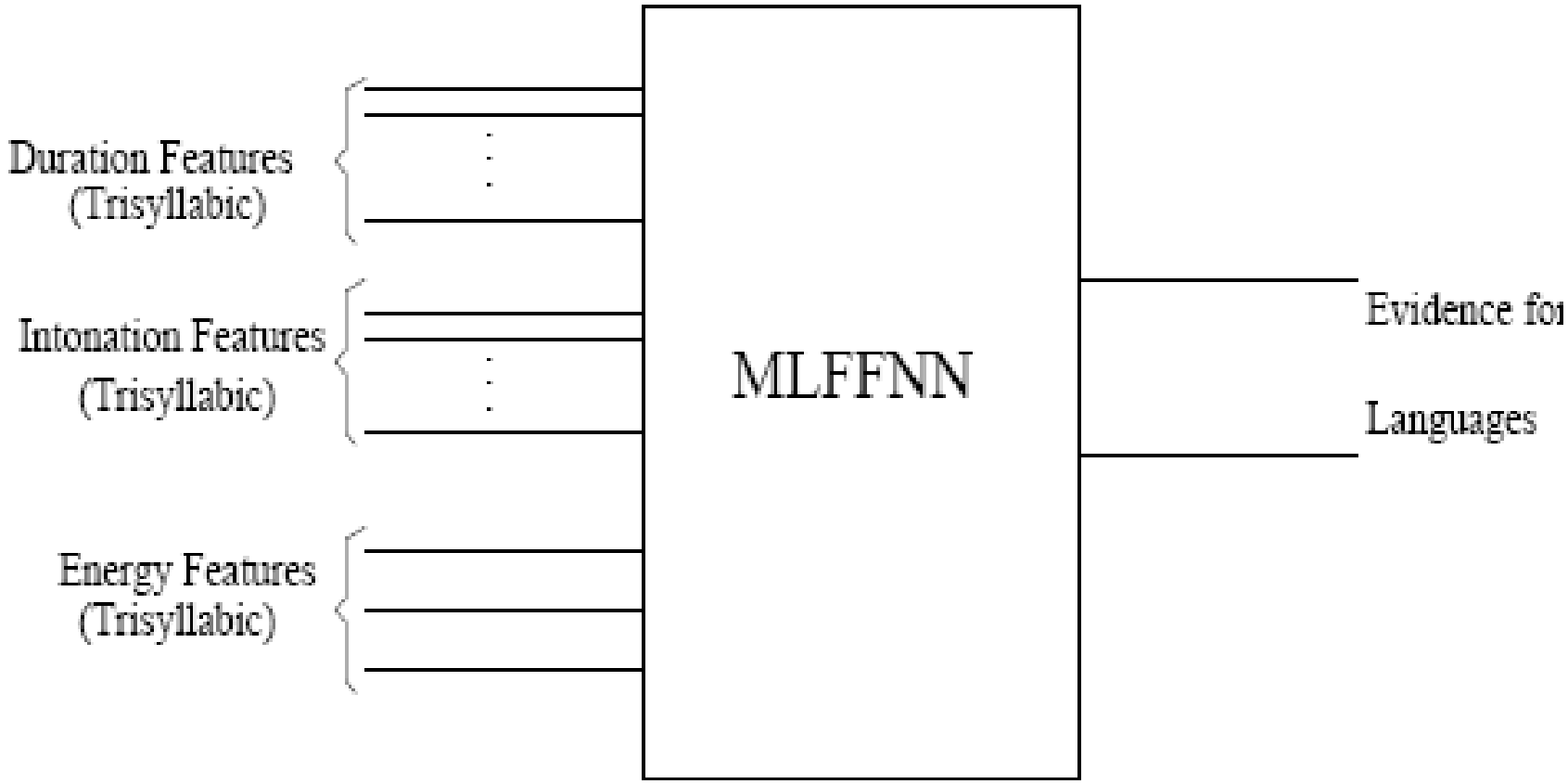


Fig. 6.8: Illustration of F_0 contours with various tilt parameters. (a) A_t =-1, D_t=-1; (b) A_t=1, D_t=1; (c) A_t=-0.2, D_t=-0.2; (d) A_t=0.4, D_t=0.2; (e) A_t=0, D_t=0; (f) A_t=0, D_t=0.

Detection of voiced regions



Prosody based neural network for language identification



Thank You