Speaker Recognition

Person Identity Characteristics

- Facial
- Voice
- Finger prints
- Iris patterns
- DNA structure
- Applications
  - ✓ Security, Surveillance and Forensic

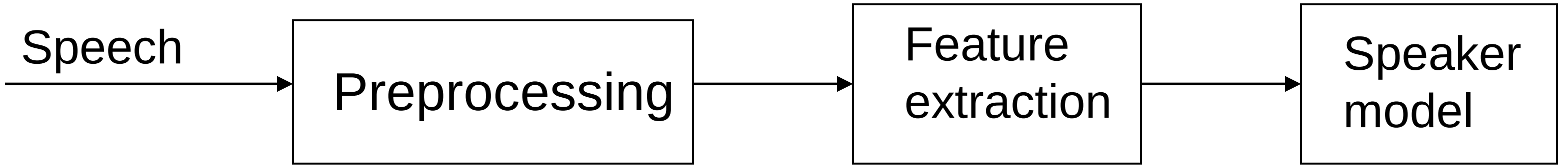
Voice as a Biometric

- Advantages
  - Capture non-intrusively and conveniently with simple transducers
  - Useful for remote access transactions over telephone networks

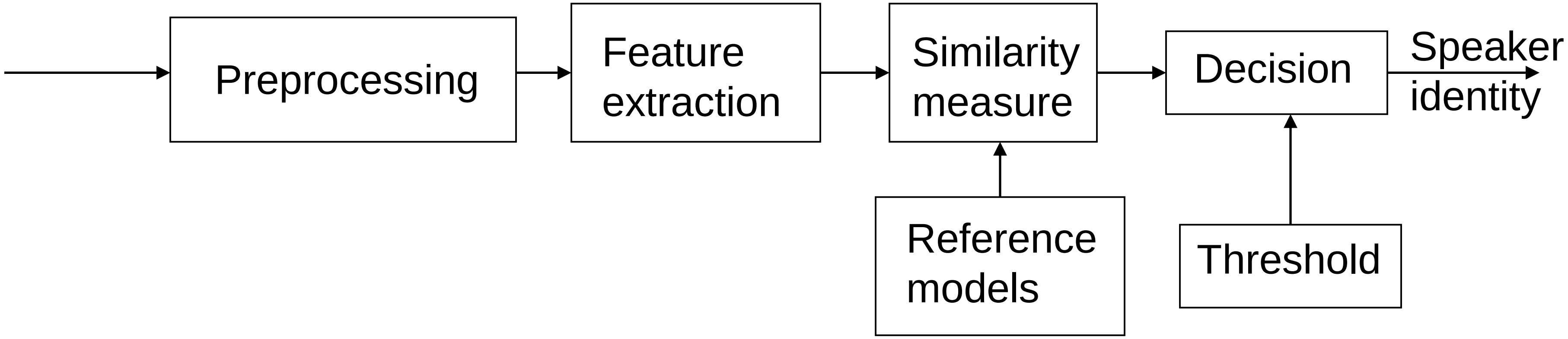
- Drawbacks
  - Subject to many sources of variability
    - Involuntary: inability to repeat the utterance same way
    - Vountary: speaker attempt to disguise their voice
  - Back-ground noise, transmission channels, recording

Basic Approach of Speaker Recognition

Training phase



Testing Phase



Probabilistic Approach for Speaker Recognition

$$S^* = \arg\max_i P(S_i \mid O)$$

$$P(S_i \mid O) = P(S_i \mid R, F, X)$$

$$P(S_i \mid R, F, X) = \frac{P(R, F, X, S_i)}{P(R, F, X)}$$

$$P(S_i \mid R, F, X) = \frac{P(R \mid S_i) P(F \mid S_i) P(X \mid S_i) P(S_i)}{P(R) P(F) P(X)}$$

R : Residual (Excitation)
F : Vocal-tract filter (Frame-level segmental)
X : Supra-segmental

Probabilistic Approach for Speaker Recognition

$$S^* = \arg\max_i P(R \mid S_i) P(F \mid S_i) P(X \mid S_i)$$

$$P(R \mid S_i) = \prod_{j=1}^{k} p(r_j \mid S_i)$$

$$P(F \mid S_i) = \prod_{j=1}^{l} p(f_j \mid S_i)$$

$$P(X \mid S_i) = \prod_{j=1}^{m} p(x_j \mid S_i)$$

$$S^* = \arg\max_i \left[\sum_{j=1}^{k} \log P(r_j \mid S_i) + \sum_{j=1}^{l} \log P(f_j \mid S_i) + \sum_{j=1}^{m} \log P(x_j \mid S_i) \right]$$

Speaker Recognition : Definitions

- Speaker identification
  - Closed-set identification (speaker always present in the set)
  - Open-set identification (speaker may not present in the set
- Speaker verification
  - Verifying the identity claim
  - Open-set identification with #speakers = 1
- Speaker detection
  - Speaker tracking, speaker segmentation, speaker indexing and speaker diarization (multi-speaker speech)
- Text dependent & Text independent speaker recognition

Basis for Speaker Recognition

- Physiological
  - Shape of the vocal tract, moment of the articulators, spectral envelope, formants and their band widths, mass of the glottis (pitch)
- Segmental vs Suprasegmental
  - Individual vs sequence of sounds
- Higher level speaking behavior
  - Choices of words, use of syntactic units, variation of F0, rhythm, breathiness and strength of vocal effort

Extraction of Speaker Characteristics

- Low level features
  - Associated with the periphery in the brain’s perception of speech
  - Spectral correlates: formant locations, bandwidths, pitch periodicity, segmental timing
  - Easier to extract and model (filter bank analysis, LPC)
- High level features
  - Associated to more central locations in the brain’s perception of speech
  - Perception of words, their meanings, syntax, prosody, dialect and idiolect
  - Difficult to extract and model

Applications

- Security
  - Control access to privileged transactions (text dependent)
- Forensic
  - Text dependent or independent
  - Open-set identification or verification
- Surveillance
  - Text independent
  - Open-set identification or verification
- Indexing multimedia data
  - Speaker indexing, segmentation and detection
  - Annotating the audio w.r.t speaker
- Improving the speech recognition using speaker specific models
  - Recognition of multi-speaker speech

Speaker Features : Measurements

- Dominance of machines in speaker recognition
- Acoustic measurements (low level)
  - Short term spectrum, MFCC analysis, CMS, feature warping, Gaussianization, delta, delta-delta coefficients and pitch
  - Application: traditional speaker authentication (passwd)
- Linguistic measurements (high level)
  - Word usage (vocabulary choices, functional word frequency, part-of-speech frequency)
  - Phone sequences and lattices (pronunciation of words and prosodic statistics)
  - Application: indexing broadcast news and passive surveillance

Constructing Speaker Models

- Based on application constraints
  - Fixed passwd test utterances (temporal characteristics of speech specific to speaker)
  - Passive surveillance
    - Less detailed model and model the overall acoustic space

Speaker Models

- Non-parametric approaches
  - Template matching (DTW)
  - Nearest neighbor modeling
- Parametric approaches
  - Vector quantization
  - GMM
    - EM algorithm
    - Background models
    - Speaker adapted background models
    - Text independent speaker recognition studies
  - HMM
    - Text dependent speaker recognition studies
    - Maximum likelihood training
  - SVM
    - Discriminative training
  - Other approaches
    - Hybrid models
    - Eigen voice modeling (confined to low dimensional linear space)
      - Effective when enrollment data is limited
    - ANN (data driven approach)

Adaptation

- Mitigate the effect of mismatch
  - Enrollment and test conditions
  - Different channels & recording devices
  - Different background noises
  - Different linguistic content
- Unsupervised model adaptation
- Threshold adaptation

Decision and Performance

- Close-set identification

$$S^* = \arg\max_i S(y/\lambda_i)$$

- Close-set verification

$$S(y/\lambda_i) >= \theta$$

- Open-set identification
  - Close-set identification + verification

Decision and Performance (Cont..)

- Threshold setting $$\theta^* = \frac{C_{fa}}{C_{fr}} \frac{P_{imp}}{1-P_{imp}}$$

  C_{fa}, C_{fr} Desired false acceptance and rejection rates

  P_{imp} Prior probability of an imposter

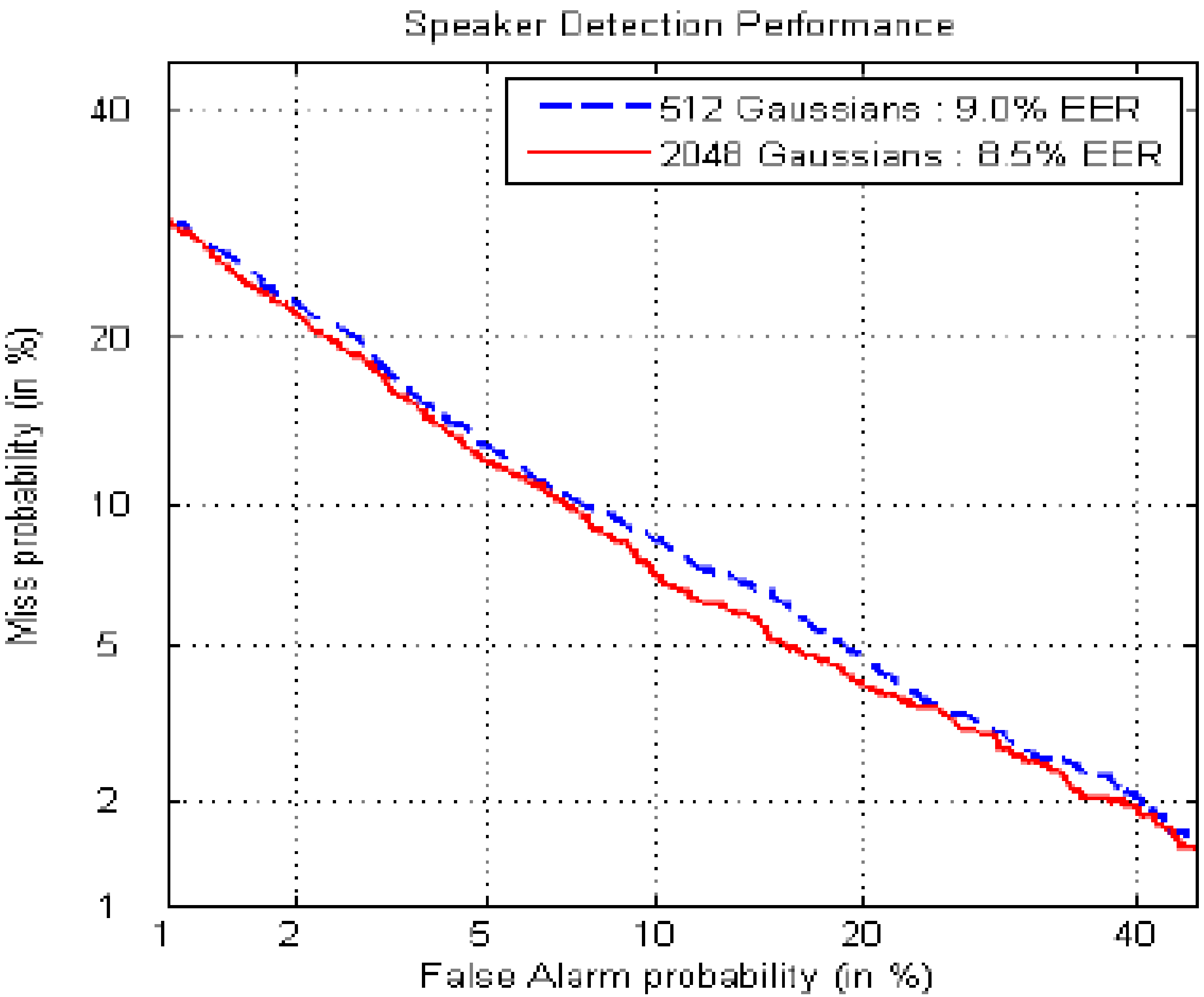
- Score normalization
  - Makes independent across speakers, acoustic conditions, linguistic variations
  - Z-norm, H-norm, T-norm, …

Decision and Performance (Cont..)

- Probability of detection $1 - P_{fr}$
- Receiver operating characteristic (ROC)
- Detection error trade-off (DET)
- Equal error rate (EER)
- Detection cost function

$$C = P_{imp} C_{fa} P_{fa} + (1 - P_{imp}) C_{fr} P_{fr}$$

Illustration of DET Plot



Selected Applications of Automatic Speaker Recognition

- Indexing multi-speaker speech data
  - Initial segmentation using acoustic change detection
  - These segments are clustered using agglomerative clustering algorithm
  - Develop proto-type speaker models using this clustered data
  - Enhance the segmentation using proto-type speaker models
  - Steps 2-4 are iterated

Selected Applications of Automatic Speaker Recognition

- Forensic application
  - Non-expert speaker recognition by lay listeners
  - Expert speaker recognition (linguistic analysis, spectrogram, pitch, timbre, diction, style, idiolect, …
  - Semi-automatic methods
  - Automatic methods
- Customization
  - Email reading to a particular user
  - Caller identification (open-set identification and verification)

Features

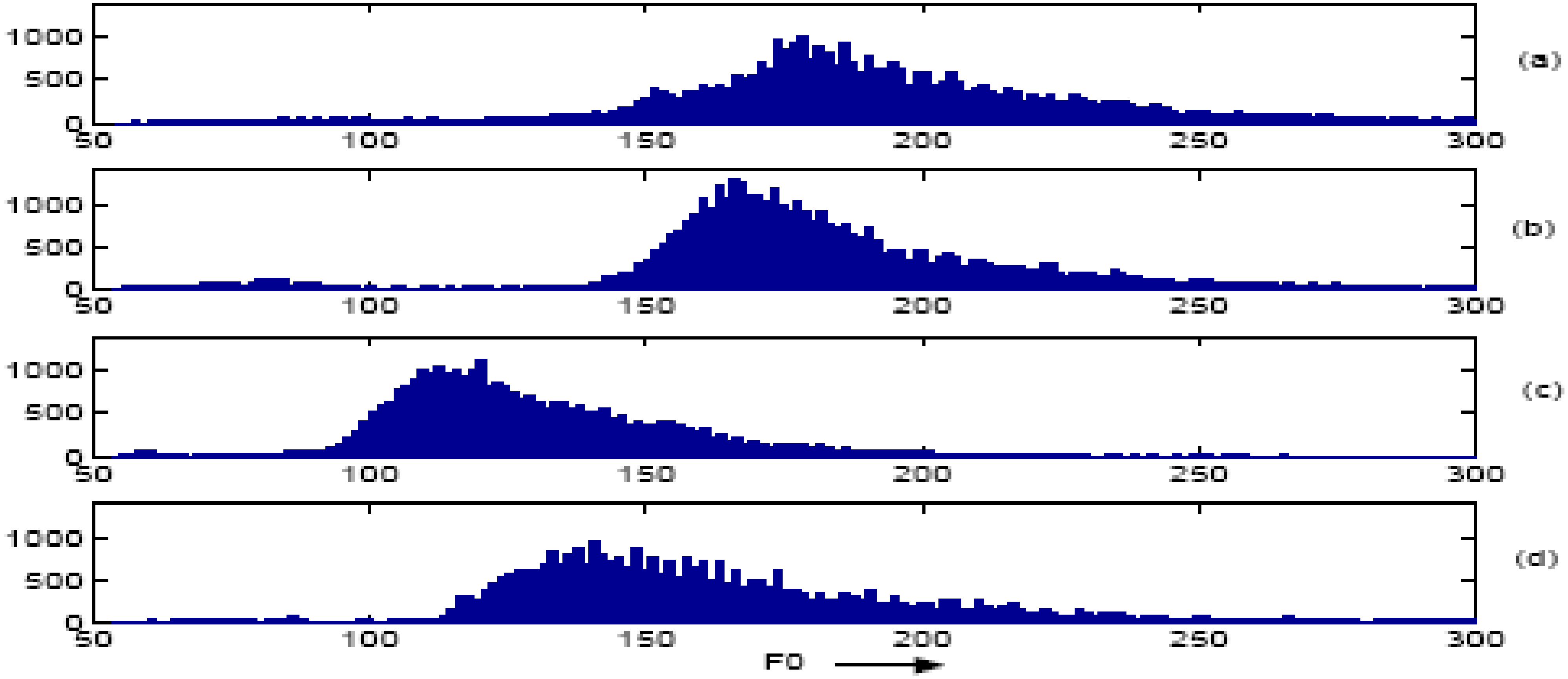
- Vocal tract size and shape
  - Spectral features
- Excitation source
  - LP residual (glottal vibration, glottal pulse shape, glottal open and close phases, …)
- Prosody
  - Intonation and duration patterns, loudness, and stress
- Idiolect
  - Habitual characteristics (usage of certain words and phrases)

Physiological vs. habitual characteristics

Features from Different Levels

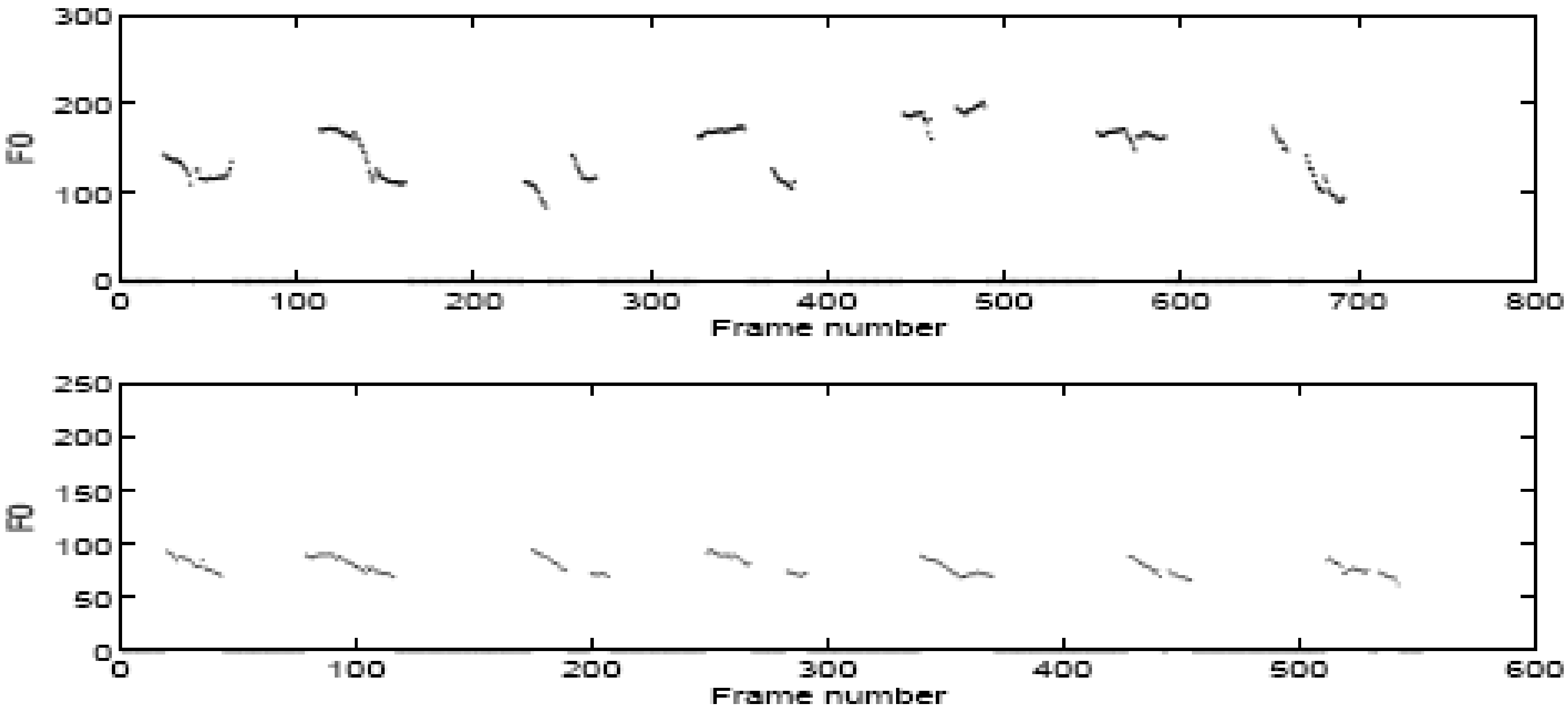
- Sub-segmental features (1-5 ms)
  - LP residual (1-5 ms around glottal closure)
- Segmental features (10-30 ms)
  - Spectral features (WLPCCs)
- Supra-segmental features (>100 ms)
  - Prosodic features
- Idiolect features
  - Features derived from the transcription

Speaker-Specific Aspects of Prosody



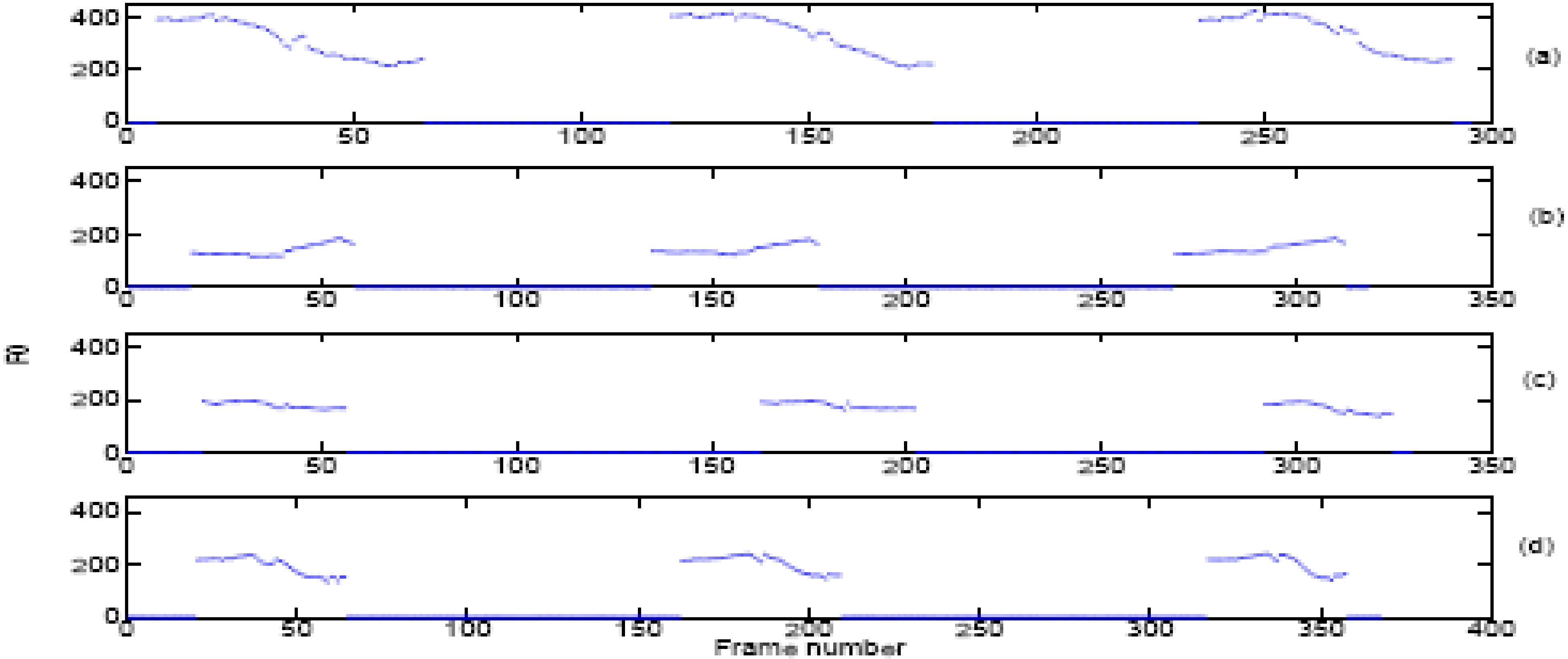
(a) and (b) are the F0 distributions of the two female speakers
(c) and (d) are the F0 distributions of the two male speakers

Speaker-Specific Aspects of Prosody (cont..)



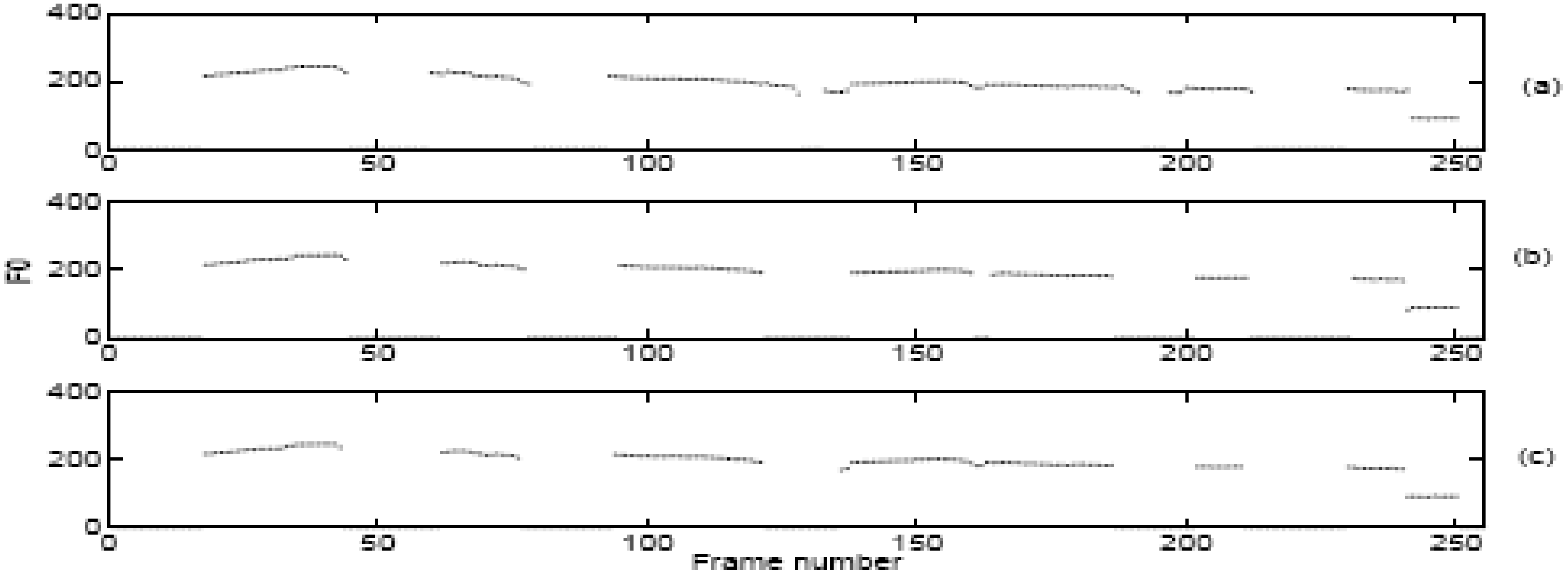
Variations on F0 dynamics for the fixed text by two male speakers
Text: Monday, Tuesday, ….. Sunday (names of the week days)

Speaker-Specific Aspects of Prosody (cont..)



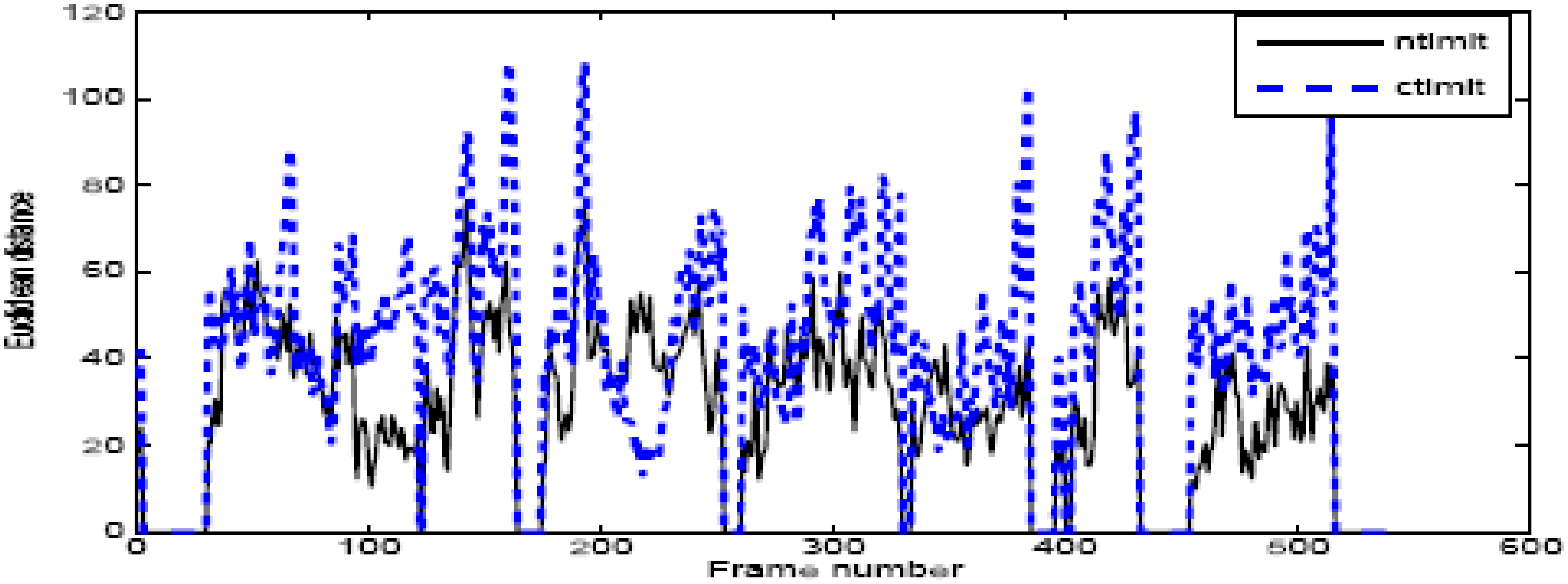
Variations in F0 contour for the fixed text (a) Child, (b) and (c) are two different male speakers, and (d) female speaker

Robustness of Prosodic Features



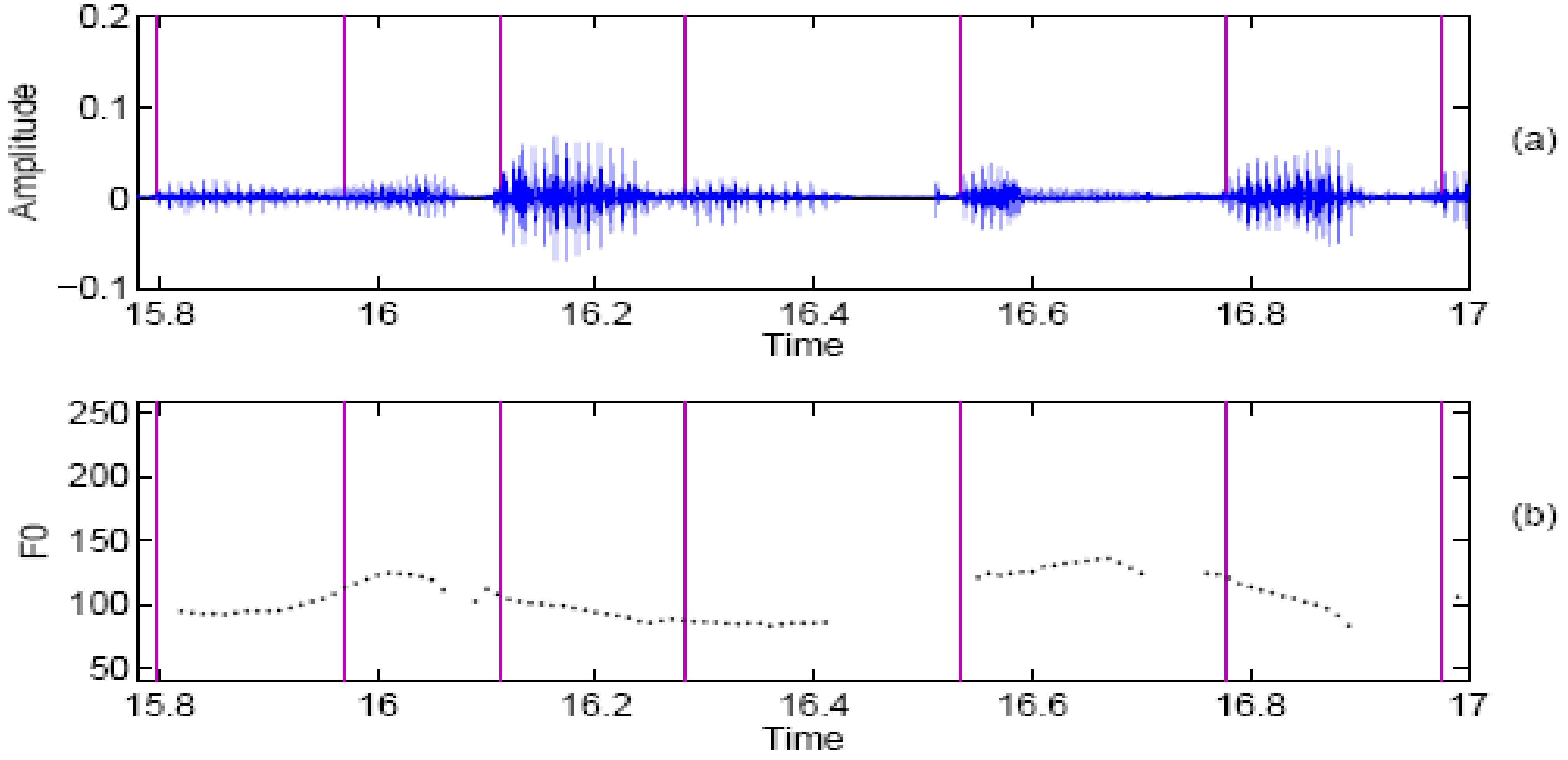
F0 contours of the TIMIT, NTIMIT and CTIMIT data for the utterance “*Don’t carry an oily rag like that*” by the same speaker

Variations in Spectral Features due to Channel Variations



Black: distance of NTIMIT data with TIMIT data for the same utterance
Blue: distance of CTIMIT data with TIMIT data for the same utterance

Extraction and Representation of Speaker-Specific Prosody



(a) Speech signal marked with VOP events
(b) Pitch contour marked with VOP events

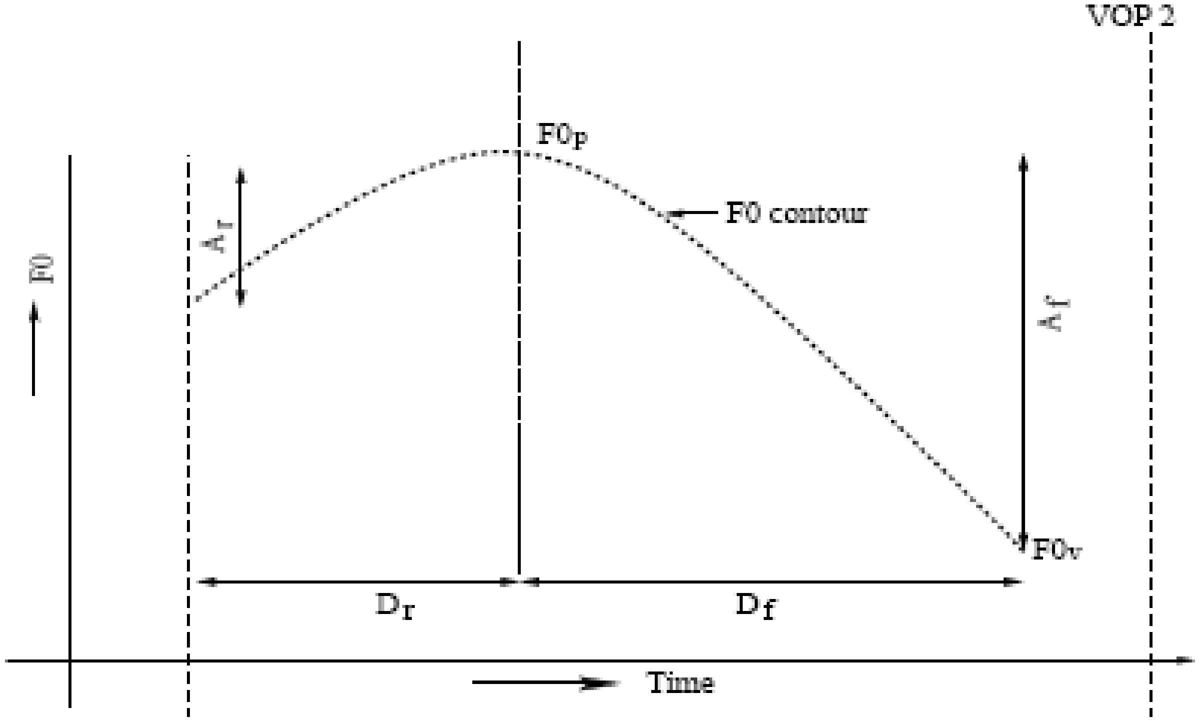
F0 Parameter Extraction

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

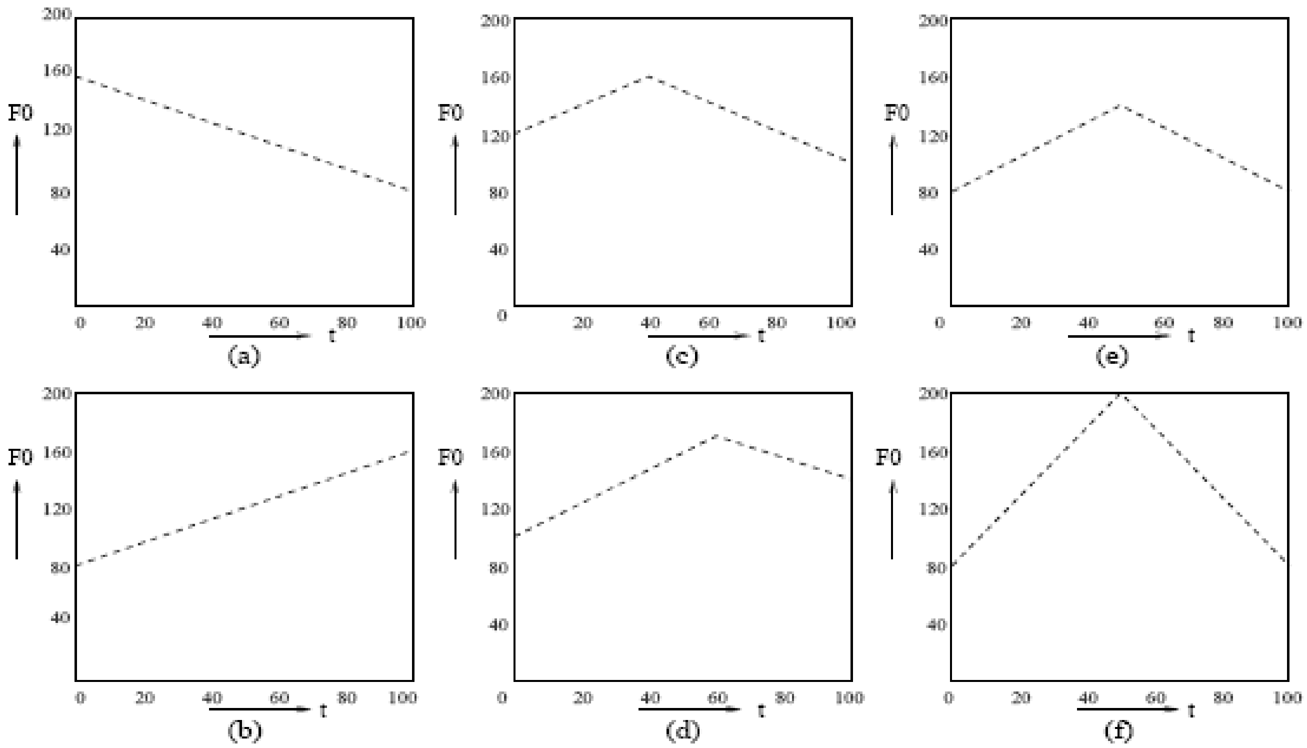


Fig. 6.8: Illustration of F_0 contours with various tilt parameters. (a) A_t =-1, D_t=-1; (b) A_t=1, D_t=1; (c) A_t=-0.2, D_t=-0.2; (d) A_t=0.4, D_t=0.2; (e) A_t=0, D_t=0; (f) A_t=0, D_t=0.

Representation of Speaker-Specific Prosody

- Mean value of F0
- Peak value (maximum) of F0
- Change in F0
- Distance of F0 peak w.r.t VOP
- Amplitude tilt
- Duration tilt
- Change in log energy

Autoassociative Neural Network

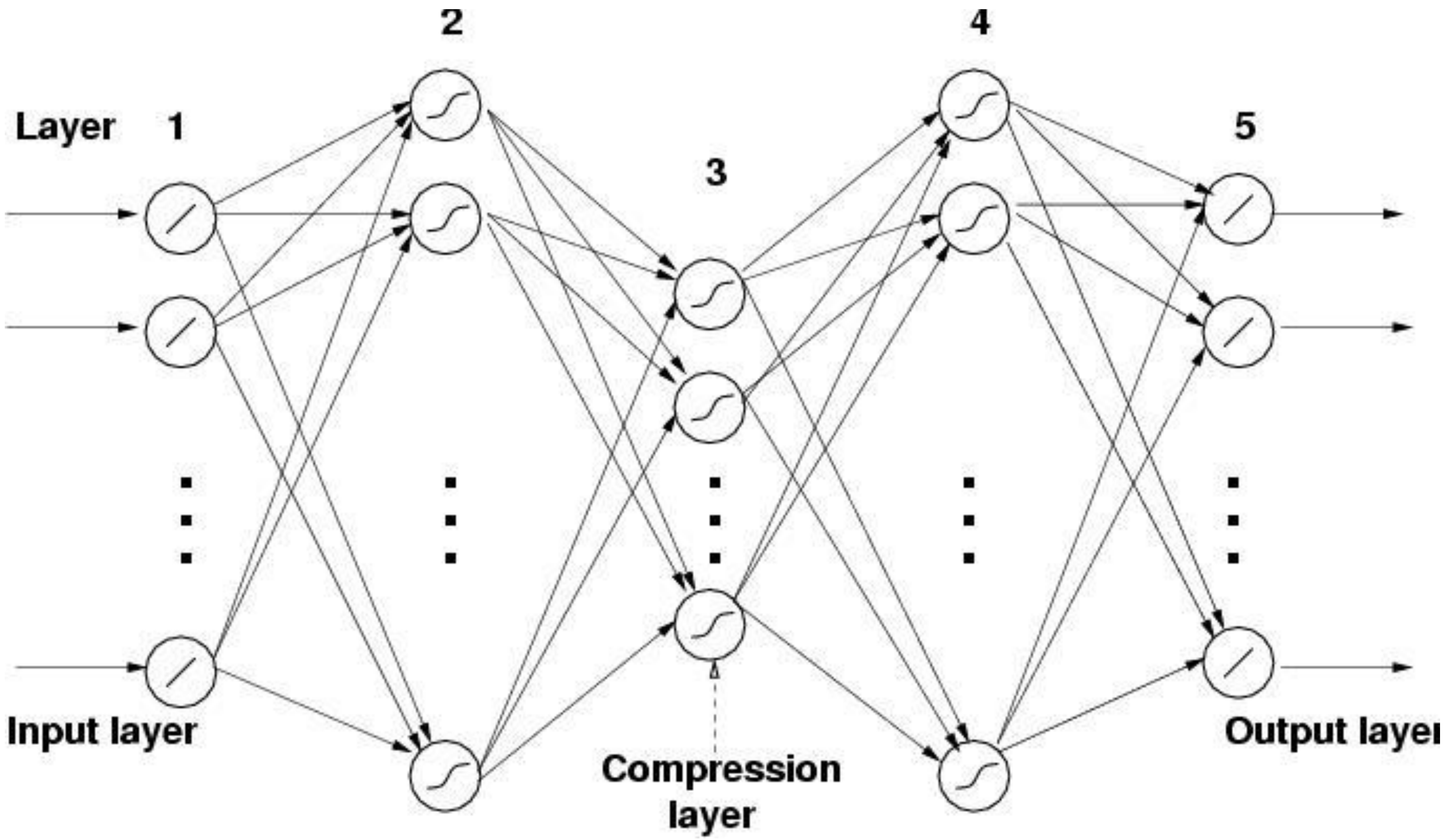
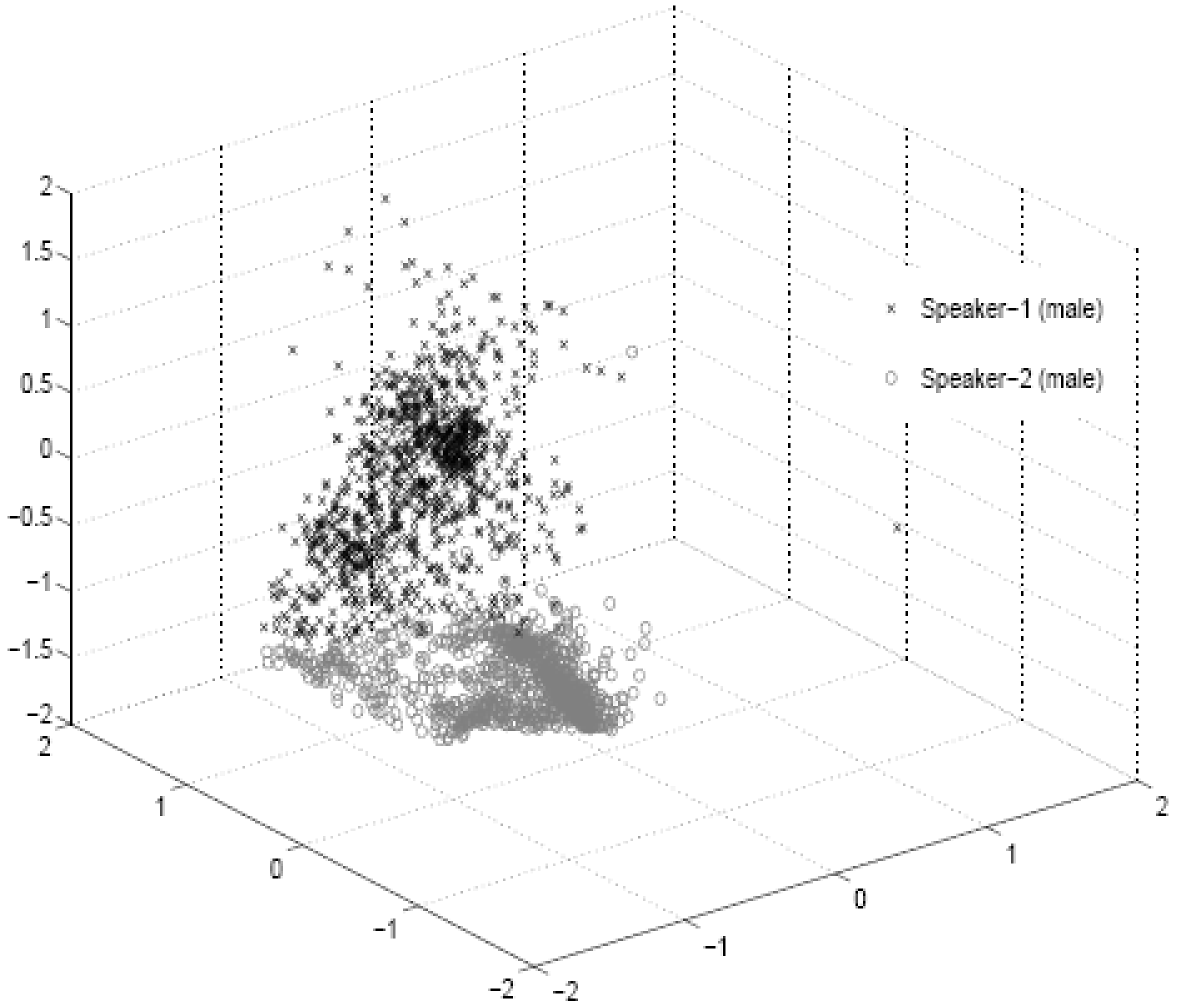


Illustration of Discrimination Property of Speaker-Specific Prosody



AANN model with 7L 14N 3N 14N 7L for deriving the compressed features

Thank You