Prosodic feature representation and language identification

Dr. Anil Kumar Vuppala, IIIT Hyderabad

Contents

- Introduction to prosody
- Language Identification
- Prosody representation
- Language identification using prosodic features
- References

Introduction

- ***Prosody***

  It refers to certain properties of the speech signal such as pitch, duration and intensity in speech.

  - **Intonation:** The dynamics of pitch or F_0 patterns over time is known as intonation contour.

  - **Duration:** The sequence of length of syllables is known as duration patterns.

  - **Intensity:** The dynamics of intensity patterns over time is known as intensity contour.

Applications of Prosody

1. Text to speech synthesis

2. Language identification

3. Emotion recognition

4. Speech and Speaker recognition

Issues in Prosody

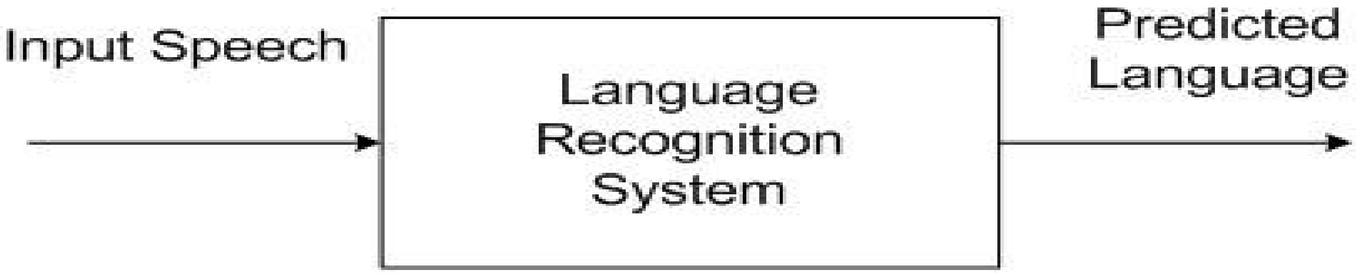
1. Duration modeling

2. Intonation modeling

3. Intensity modeling

We can model prosody using linguistic context and production constraint features. FFNN can be used for modeling.

Language Identification (LI)



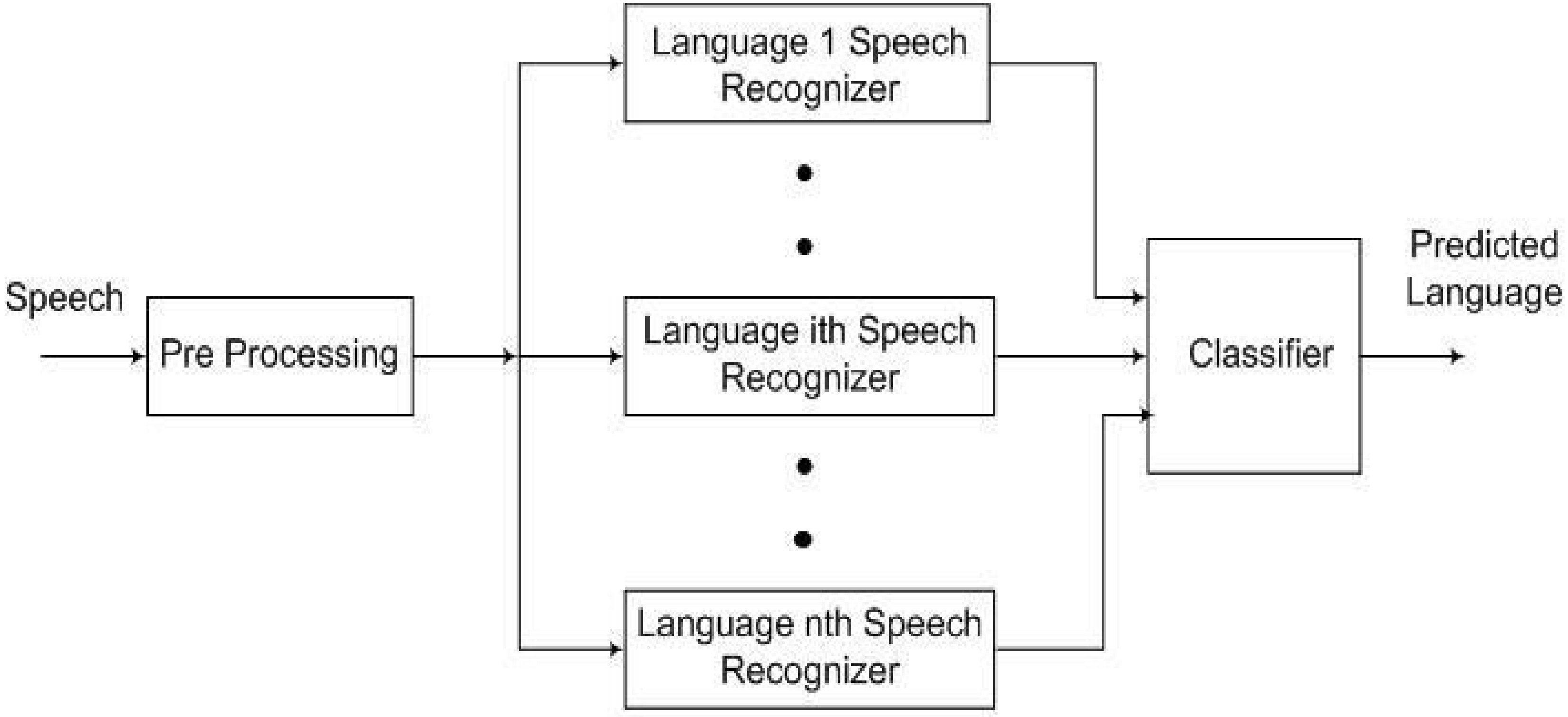
Block Diagram of Language Recognition

Applications of LI

- ❑ Real life applications of language recognition:-
  1. A front end for automatic speech recognition
  2. Speech to speech translation
  3. Assistance for speech activated automated system
  4. Information retrieval from databases

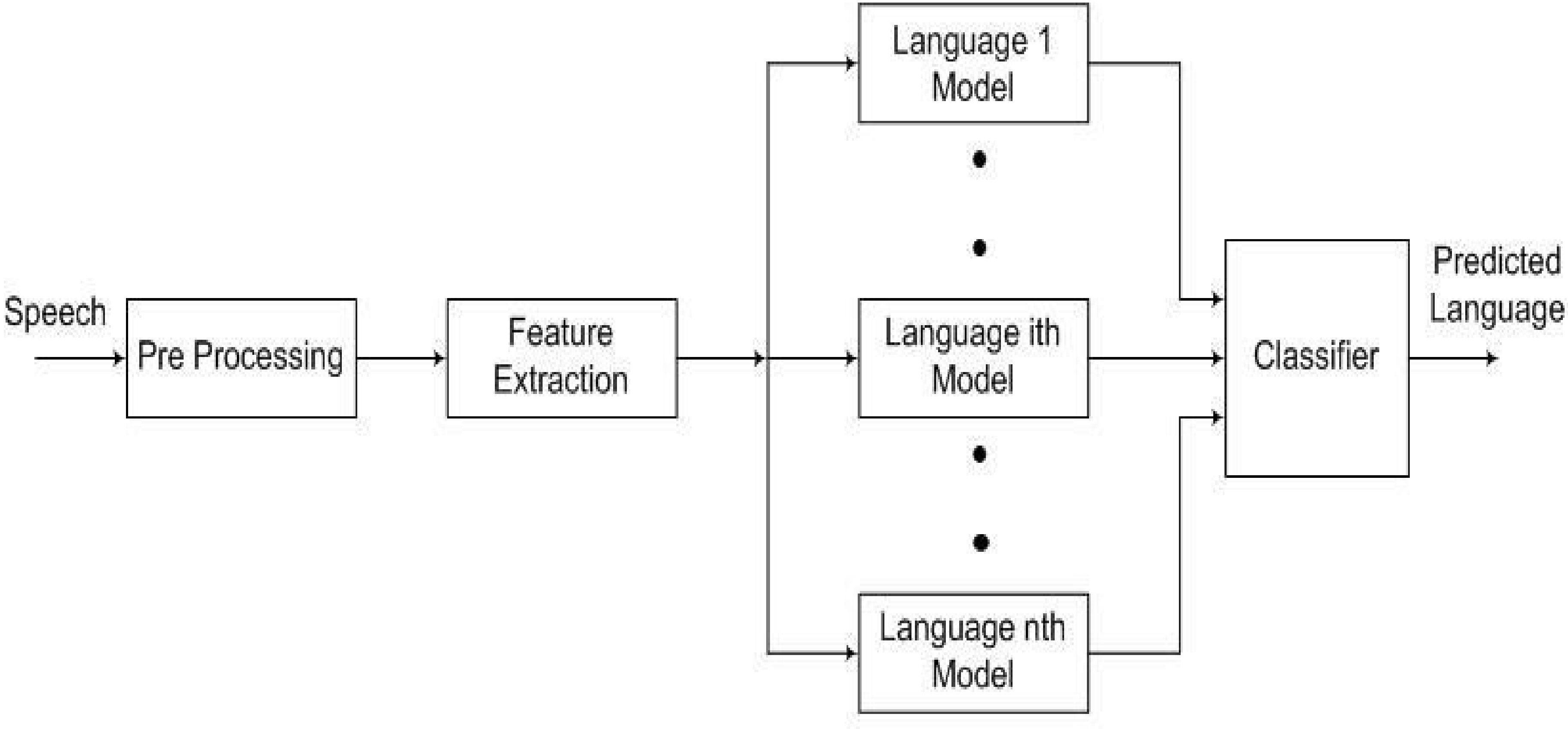
- ❑ Conditions for sophisticated language recognition system:-
  1. System should not be biased to specific speakers
  2. Tolerance for degradation in input speech should be high

Explicit Language Identification System


Speech
Pre Processing
Language 1 Speech Recognizer
Language ith Speech Recognizer
Language nth Speech Recognizer
Classifier
Predicted Language

Explicit Language Recognition System

Implicit Language Identification System



Implicit Language Recognition System

Issues in Language Identification

- ❑ Variation in speaker
- ❑ Variation in channel and background
- ❑ Variation in dialects
- ❑ Similarities in languages

Features used for LI system

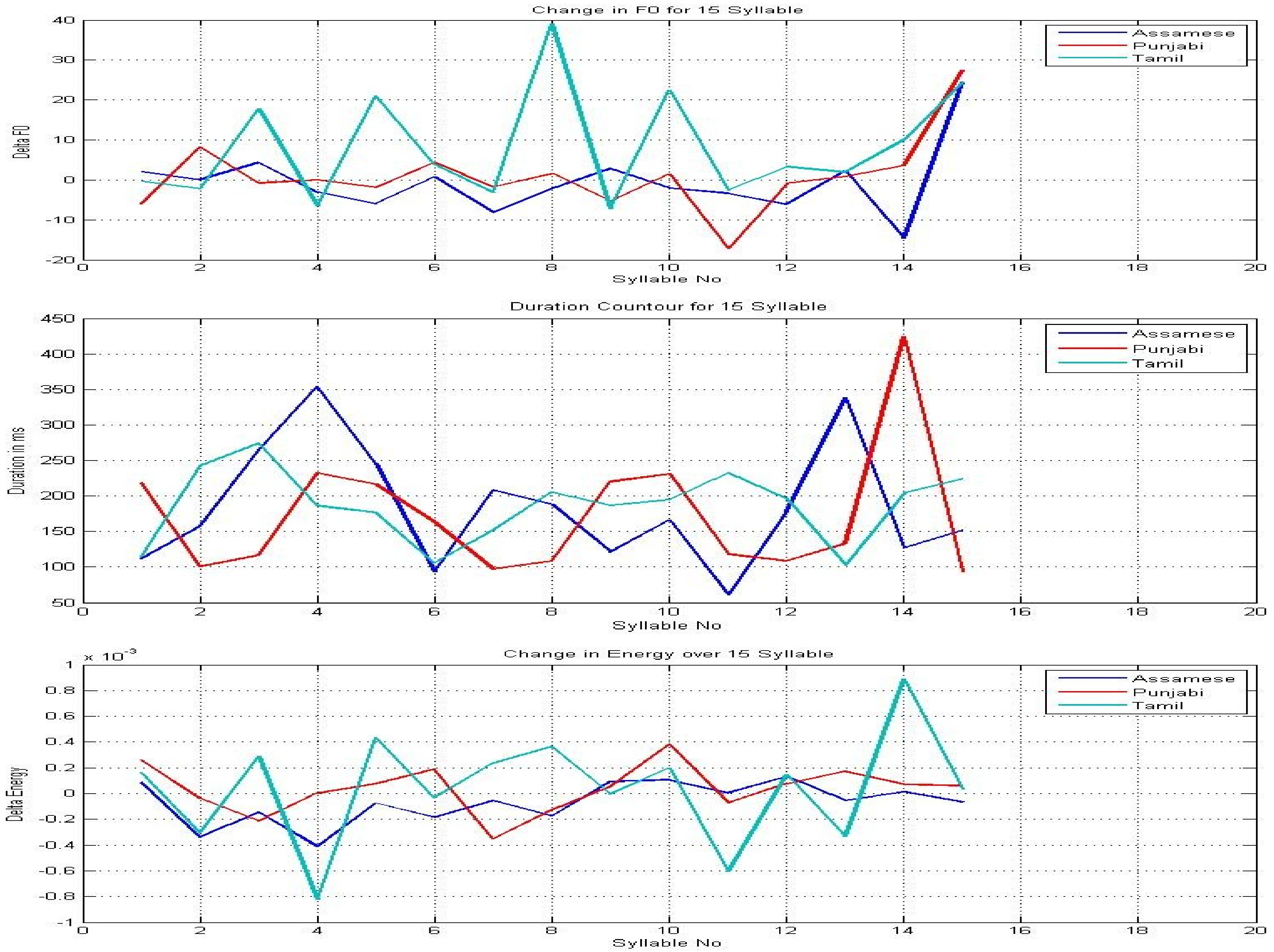
- Spectral
  - ➢ Linear Prediction Cepstral Co-efficient (LPCC)
  - ➢ Mel-frequency Cepstral Co-efficient (MFCC)
- Prosody
  - ➢ Intonation
  - ➢ Rhythm
  - ➢ Stress

Prosodic Features

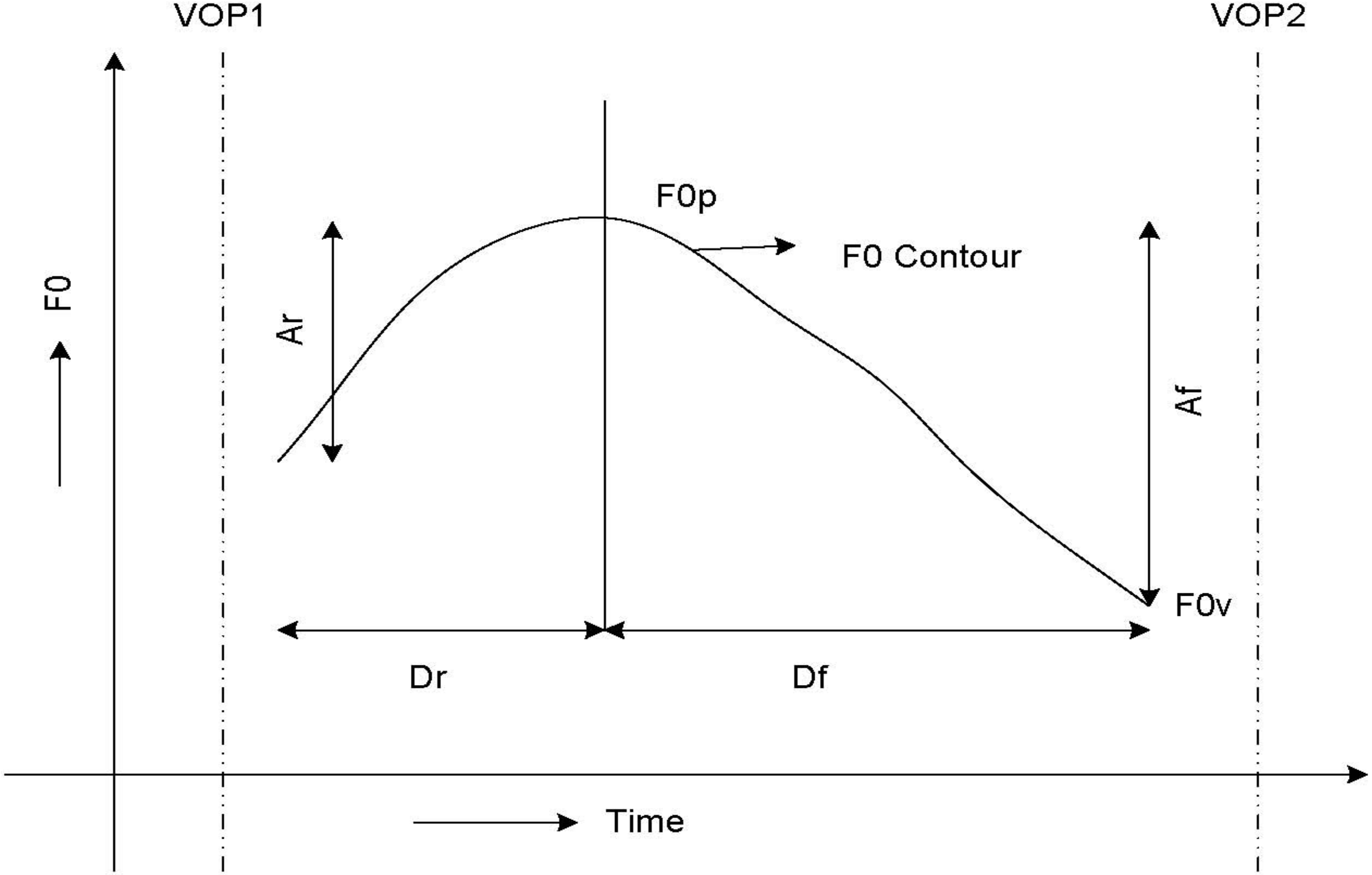
- Intonation
  - ➢ Change in F0
  - ➢ Distance of F0 peak with respect to VOP
  - ➢ Amplitude Tilt
  - ➢ Duration Tilt
- Rhythm
  - ➢ Distance between successive VOP
  - ➢ Duration of voiced region
  - ➢ Change in F0
- Stress
  - ➢ Change in log energy in voiced region
  - ➢ Change in F0
  - ➢ Distance between successive VOP

1. Change in F0
2. Distance of F0 peak with respect to VOP
3. Amplitude Tilt
4. Duration Tilt

5. Distance between successive VOP
6. Duration of voiced region
7. Change in log energy in voiced region


Change in F0 for 15 Syllable
Delta F0
40
30
20
10
0
-10
-20
0
2
4
6
8
10
12
14
16
18
20
Syllable No
Assamese
Punjabi
Tamil
Duration Countour for 15 Syllable
Duration in ms
450
400
350
300
250
200
150
100
50
0
2
4
6
8
10
12
14
16
18
20
Syllable No
Assamese
Punjabi
Tamil
Change in Energy over 15 Syllable
x 10^-3
Delta Energy
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1
0
2
4
6
8
10
12
14
16
18
20
Syllable No
Assamese
Punjabi
Tamil

Tilt Parameter Calculation



Database Collection

- Number of languages : 27
- Source : TV news bulletin, talk shows, live shows and interviews and AIR (All India Radio) news bulletins.
- Number of speakers per language : 5 male and 5 female.
- Amount of speech per speaker : 5 to 6 minutes.
- Sampling Frequency : 16 kHz.
- Audio sample size: 16 bit.
- Channels : Mono.
- Audio format : Pulse-code modulation (PCM).

Description of Indian Language Speech Corpus

Language	Region	Speaking Population (Mil)	Speakers		Duration In Minutes
			F	M	
Arunachali	Arunachal Pradesh	0.41	6	15	72
Assamese	Assam	13.17	6	8	67.33
Bengali	West Bengal	83.37	14	10	69.78
Bhojpuri	Bihar	38.55	5	7	59.82
Chhattisgarhi	Chhattisgarh	11.5	9	11	70
Dogri	Jammu and Kashmir	2.28	8	12	70
Gojri	Jammu and Kashmir	20	3	12	44
Gujrati	Gujarat	46.09	7	6	48.96
Hindi	Uttar Pradesh	422.05	14	24	134.7
Indian English	All over India	125.23	12	13	81.66
Kannada	Karnataka	37.92	4	8	69.33
Kashmiri	Jammu and Kashmir	5.53	2	19	59.64
Konkani	Goa and Karnataka	2.49	5	15	50
Manipuri	Manipur	1.47	11	11	64
Mizo	Mizoram	0.67	3	8	48
Malyalam	Kerala	33.07	7	12	81.09
Marathi	Maharashtra	71.94	7	9	74.33
Nagamese	Nagaland	0.03	11	9	60
Neplai	West Bengal	2.87	7	6	54.19
Oriya	Orissa	33.02	10	4	59.87
Punjabi	Punjab	29.1	7	10	80.91
Rajasthani	Rajasthan	50	10	10	60
Sanskrit	Uttar Pradesh (UP)	0.014	0	20	70
Sindhi	Gujarat and Maharashtra	2.54	14	6	50
Tamil	Tamil Nadu	60.79	7	10	70.96
Telugu	Andhra Pradesh (AP)	74	7	8	73.72
Urdu	UP and AP	51.54	5	16	86.49

LI System using Syllable Level Prosody

- Data : 5 male and 5 female speakers speech data.
- Model : GMM.
- Features: Intonation, Rhythm and Stress (IRS) features.
- Testing : using leave one speaker out each speaker’s speech data with three different utterance duration (5, 10 and 20 sec.) are used.
- Result : 32.00%

LI System using Word Level Prosody

- Data : 5 male and 5 female speakers speech data.
- Model : GMM.
- Features: Intonation, Rhythm and Stress (IRS) features of syllables for previous, present and next syllable (total 21 dimension).
- Testing : 1 male and 1 female speech data with three different utterance duration (5, 10 and 20 sec.) are used.
- Decision Making : Maximum posterior probability.
- Result : 35.22%

LI System using Global Level Prosody

- Data : 5 male and 5 female speakers speech data.
- Model : GMM.
- Features: F0 , energy and duration variation for continuous 15 syllable in a sentence.
- Testing : 1 male and 1 female speech data with three different utterance duration (5, 10 and 20 sec.) are used.
- Decision Making : Maximum posterior probability.
- Result : F0 – 28.50%, Energy – 21.57 % and Duration – 25.18%

Combination of Features

Features	Performance (%)
Global level prosodic features (F0 + Energy + Duration variation)	33.79%
Syllable + word level prosodic features	37.58%
Syllable + word + global level prosodic features	39.46%
Prosody + spectral	62.13%

Summary

- ❑ Prosodic features such as intonation, rhythm and stress related to syllable can be used for Language Identification along with conventional spectral features.
- ❑ Different prosodic and spectral features are combined for further improvement in performance of LI system.

References

[1] E. Ambikairajah, H. Li, L. Wang, B. Yin, and V. Sethu, "Language identification: A tutorial," IEEE Circuits and Systems Magazine, vol. 11, no. 2, pp. 82-108, 2011.

[2] J. Benesty, M. Sondhi, and Y. Huang, "Springer handbook of speech processing," Springer-Verlag, 2007.

[3] T. Nagarajan, "Implicit System for Spoken Language Identification", PhD thesis, Indian Institute of Technology, Madras, January 2004.

[4] L. Mary and B.Yegnanarayana, "Extraction and representation of prosodic features for language and speaker recognition", Speech Communication, vol. 50, pp. 782-796, 2008.

[5] L. Lamel and J. Gauvain, "Cross lingual experiments with phone recognition," in IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. 507-510, April,1993.

[6] L. Lamel and J. Gauvain, "Language indentification using phone-based acoustic likelihoods", in IEEE International Conference on Acoustics, Speech, and Signal Processing,pp. 293-296, April 1994.

[7] K. Berkling, T. Arai, and E. Bernard, "Analysis of phoneme-based features for language identification," in IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. 289-292, April 1994.

[8] R. Tucker, M. Carey, and E. Paris, "Automatic language identification using sub-words models," in IEEE International Conference on Acoustics, Speech, and Signal Processing,pp. 301-304, April 1994.

References

[9] M. Zissman, "Comparison of four approaches to automatic language identification of telephone speech," IEEE Transactions on Audio, Speech and Language Processing, vol. 4,pp. 31-44, january 1996.

[10] Y. Yan and E. Barnard, "A approch to automatic language identification based on language-dependant phone recognition," in IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. 3511-3514, May 1995.

[11] J. Foil, "Language identification using noisy speech," in IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. 861-864, April 1986.

[12] M. Sugiyama, "Automatic language recognition using acoustic features," in IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. 813-816, May 1991.

[13] J. Balleda, H. A. Murthy, and T.Nagarajan, "Language identification from short segments of speech," in International Conference on Spoken Language Processing, pp. 1033-1036,October 2000.

[14] A. Jayaram, V. Ramasubramanian, and T. Sreenivas, "Language identification using parallel sub-word recognition," in IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. 32-35, April 2003.

[15] L. Mary, Multilevel Implicit Features for Language and Speaker Recognition. PhD thesis, Indian Institute of Technology, Madras, June 2006.

[16] B. Ma, H. Li, and R. Tong, "Spoken language recognition using ensemble classifiers,“ IEEE Transactions on Audio, Speech and Language Processing, vol. 15, pp. 2053-2062,September 2007.

References

[17] D. Reynolds, "Speaker Identification and Verification Using Gaussian Mixture Speaker Models," Speech Communication, vol. 17, no. 1-2, pp. 91-108, 1995.

[18] K. S. R. Murty and B. Yegnanarayana, "Epoch extraction from speech signals," IEEE Trans. Audio, Speech, and Language Processing, vol. 16, pp. 1602-1613, Nov. 2008.

[19] A. K. Vuppala, J. Yadav, S. Chakrabarti, and K. S. Rao, \Vowel onset point detection for low bit rate coded speech," *IEEE Trans. Audio, Speech and Language Processing*, 2012. DOI: 10.1109/TASL.2012.2191284.

[20] K. N. Reddy, Speech technology: issues and implications in indian languages," in *26th All India Conference of Dravidian Linguists*, (Trivendrum, India), 1998.

[21] Sudhamay Maity, Anil Kumar Vuppala, K. Sreenivasa Rao and Dipanjan Nandi,"IITKGP-MLILSC Speech Database for Language Identification”, Eighteenth National Conference on Communications(NCC), 2012.

[22] K. S. Rao and B.Yegnanarayana, “Intonation modeling for Indian languages”, in Proc. 8th Int. Conf. on Spoken Language Processing (Interspeech-2004), Jeju Island, Korea, pp. 733-736, Oct. 2004.

[23]K. S. Rao and B. Yegnanarayana, “Modeling durations of syllables using neural networks”, Computer Speech and Language, Vol. 21, pp. 282-295, Apr. 2007.

[24] K. S. Rao and B. Yegnanarayana, “Two-stage duration model for Indian languages using neural networks”, in Lecture Notes in Computer Science, Neural Information Processing, Springer, pp. 1179-1185, 2004.

Thank you