Grenze International Journal of Computer Theory and Engineering , Feb 2016

Extraction and Representation of Prosodic Features for Automatic Speaker Recognition Technology

Nilu Singh[1] and R. A. Khan[2]
[1-2]SIST-DIT, Babasaheb Bhimrao Ambedkar University (A Central University), Lucknow, UP, India
Email: nilu.chouhan@hotmail.com, khanraees@yahoo.com

***Abstract*—To recognize emotion from a speech signal, the feature extraction technique used is known as Prosodic features extraction technique. It is most common methodology to emotion recognition depend on utterance level. Current studies shows that segmental spectral features of a speech signal rely on utterance level measurements also encloses rich information about articulateness and emotion.Automatic Speaker Recognition technology can be defined as it is a task by which recognize speakers from their speech/voice. Speaker Recognition has covered several speaker specific tasks, the task can be categories as text dependent and text independent. Speaker recognition can be dividing in some specific task such as speaker verification, speaker identification, speaker clustering, speaker segmentation, speaker diarization and speaker detection etc.In this paper discussed about the prosodic features for feature extraction from a speech signal. In term of speaker recognition prosodic compute many features as duration, pitch, intensity, speech rate, tone, stress etc.**

***Index Terms*— Prosodic, ASR, Speaker Identification, Speaker verification, Speech signal, speaker modeling, feature extraction technique.**

I. INTRODUCTION

Speaker Recognition is voice based biometric technique which is a field of information security [1] [2]. Since biometric security systems are constructed by human unique features for example voice, iris, retina, thumb impression, face recognition and palm biometric etc. Such type of biometric security systems are used to avoid unauthorized accessibility and help to protect data through user's particular physical and behavioral characteristic. Now day's speaker recognition is become more popular due to by voice/speech secured information retrieved for any user. Today's need wide security because number of users work together with computers and telephone and in the same period they use leading as well as ever-present handheld devices e.g. mobile phone, smart phone, handheld computers etc. which may be hold numerous of personal information. Hence speaker voice characteristics can be used as safe and sound access for remote computers, financial such as credit card accessibility, bank account accessibility etc. [3]

Speech signal hold numerous mixed information such as words, language, emotion, speaking style, tone etc. these factors are sufficient to recognizing the speaker or for a human being. From the speech signal extracted required information by using different voice speech feature extraction techniques [2].The studies show [4] that text-dependent speaker verification systems are susceptible while text-independent speaker

Grenze ID: 01.GIJCTE.2.1.11

verification system robust to this problem. For all speaker recognition systems assessments Equal error rate is a performance metric. Figure (a) provides an overview of an Automatic Speaker Recognition System, the figure shows enrollment and authentication phases. Figure 1 shows the method of Automatic Speaker Recognition Technology system. These are basic required steps of any speaker recognition (SR) system.

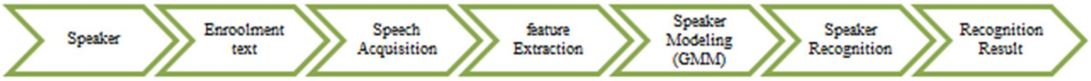


Figure 1: Steps in text-independent Automatic Speaker Recognition System

In modern speaker recognition systems, speaker recognition with affective speech be there used due to the rapid development in affective computing. In real it is not possible that speaker speaks in neutral/normal attitude so automatic speaker recognition process become difficult. It happens many times that speaker enroll with normal attitude, in this case these systems recognize accurately as compared to other. While if speaker enroll with some emotional conditions may be angry, happy then ASR systems undergo emotional state mismatch among training and testing speech data and the performance of systems weakens. It is not possible to make speaker models, to all probable emotional states aimed to improving the systems performance. Hence speaker recognition with affective speech is important since difficulty in acquiring huge data of affective vocalizations from the speakers. System performance can be improved by using speaker recognition with affective speech in communications and telephone based speech recognition [3].

II. Related Work in This Area

In this literature there are limited studies about the speaker recognition with prosodic features. It was' Kersta' who developed the first computerized Speaker verification system in 1960s at Bell Labs, based on spectrographic voice verification [5]. Speaker recognition is interrelated physiological and behavioral features of every speaker's speech production system. These features of speech signal belong to spectral i.e. vocal tract characteristic and supra segmental i.e. voice source characteristics. It is impossible to distinguish such type of features and various speech features are difficult to quantify explicitly, also many speech features are quantify implicitly through different signal measurement. Signal measurements are two types' short-term and long-term spectra. Fundamental frequency is also use to recognize the speaker [6].

Nigel G. Ward and et al. [7] says we make better prediction if we construct the model with cognitive and communicative processes. In his research they prove that make a model using combination of such features it improve 8.4% perplexity advantage and reduces word error rate up to 1.0% in different corpus. Je Hun Jeon and Yang Liu tell in their research paper [8] achieve semi supervised knowledge for detection of prosodic features such as pitch, intonation, phrase boundaries etc. through the co-training algorithm. In his research describe about co-training conditions such as compatible and uncorrelated but in actual voice data do not fulfill these conditions. Here author proposed a framework for co-training data for prosodic detection task.

In 2011 author Ulrike Schild et al [9]. stated spoken category i.e. stressed and unstressed. These spoken categories modulate event related potentials differently. In this paper author describe about how we test that, is the event related potentials phonemes free and involves the lexicon or not. Using this technique result shows phoneme free prosodic feature processing at the lexical level and prosodic features at the pre-lexical as well as lexical level are enlisted by neurobiological speech/spoken word recognition.

In [10] author Lalita Narupiyakul et. al. address that recognition rate can be improved and ambiguity reduce by emphasizing prosody of a sentence by focus part when speaker's produce utterance. Their objective is to analyzing the conception of foci in speech signal, prosody and speakers intention. Their study shows about the understanding and modeling, that how utterance is influenced through the speakers intentions. Since utterance convey different information's to the hearer and such ambiguities can be fixed through different positions of accents. Author also tells that focus information of utterance is required to acquire correct spot for accent.

In 2013,Utpal Bhattacharjee and Kshirod Sarmah shows result in his research paper [27] that when individual acoustic feature i.e. Mel Frequency cepstral Coefficient is used for Speaker recognition and prosodic features for speaker recognition, it is found that the performance of speaker recognition system is better when prosodic features integrated with acoustic features. The result shows that performance of the speaker recognition system improved by 5% as compared to individual MFCC is used while 20% improvement found in case of individual Prosodic considered.

Marc D. Pell in 2007, say that it is noticed that to process vocal expression of human sentiment prosodic features are more helpful. In this study it is shown that for empirical testing prosody is initiated because it helpful to understanding the dissimilar interpersonal perspective of spoken language [28].

III. Basic Terminology Of Speaker Recognition System

A. Speech Signal

Speech signal is a medium to communication and it is also acquire the speaker specific information. Human voice production system is depend on vocal cords, it is also called sound/voice originator. Human vocal cords contain flexible muscle in the throat which are vibrate by breath and expelled by the lungs. This procedure initiating sound vibrations in the air i.e. called sound waves or speech signal or simply known human voice [11].Human voice characteristics are basically determined by three features these are sound generation, emission/production and propagation in the vocal tract. The shape of vocal tract controls prosody while vocal cords control the pitch of voice. To conclude voice quality in reference of voice characteristics two things are responsible one is state of vocal cords and second is state of the vocal tract [12].

Automatic Speaker Recognition

ASR is a research area of signal processing and research doing in this area occurred last five decades [13]. ASR is a method to recognizing a speaker by their voice or speech waves on the basis of specific information contained by speech signal. It is one of the most widely held biometric techniques used in security. ASR used in many where such as access control in organizations, banks, institutes, investigation etc. this technology uses to make available larger security in various areas [14]. Usually speaker recognition systems based on the spectral features for recognizing the human by their voice. Although higher level features also implemented to recognizing the human/speaker [15].

Feature extraction techniques or enrolment

It is most significant element for Automatic speaker recognition (ASR) system. Feature extraction is a process of representing the speech signal in terms of speech characteristic and these characteristics are used for speaker recognition [11]. Also it is source for effect on the system performance. As known that acoustic signal contains significant information about the speakers by which speakers express their message. And feature extracted from speech signal are anticipated to transfer their information, through the system able to distinguish each and every speaker [16]. Feature extraction used to extracts significant small data from speaker's speech and these features used to characterize that identical speaker [14].

Speaker modelling

Speaker modeling is done after feature extraction in speaker recognition. After feature extraction, features are modeled by some modeling techniques or machine learning algorithm. Gaussian Mixture Models (GMM) is the most common speaker modeling technique used for speaker recognition. It gives maximum accurate results in the case of text-dependent speaker recognition but in case of text-independent speaker recognition this gives no improvement to the performance. For every speaker there is a GMM creates [11] [14].

Text-dependent & text-independent

In case of text-independent systems, a random speech/word is articulated to recognize the speaker. While In case of text-dependent systems, a fixed word/sentence is articulated to recognize the speaker [3].

Speaker Verification

ASR also known as speaker authentication gives results only true or false i.e. either accepted or rejected. It is the process where speaker claims their identity and system decides whether identity is true or not. To produce an ID card for verification is an example of verification that means either accepted or rejected (1 to 1).Speaker verification is a procedure to approve the claimed identity, and recognized the speaker is true or imposter. Specific use of speaker verification system is in the user specified pass codes (access control) for secure user logins in security system [1] [11] [17]. In figure 2 descriptions of different techniques of speaker recognition technology.

Speaker Identification

Automatic speaker identification (ASI) is one to many (1: n) process. In this case claimed identity match to many if the speaker exist in the database then return identity, if speaker does not exist in the database i.e. imposter yield no identity. Here claimed identity need to compare all the speakers whom voice present in the

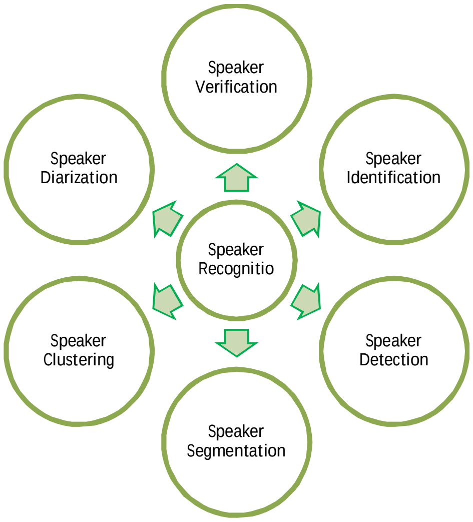


Figure 2: categorization of Speaker Recognition

database. Speaker identification is a method to determine which one are the best matches the input speech sample from a voice database of speakers' voices. Specific application of speaker identification is in the area of forensics and investigation, i.e. if some where is a requirement to determine the identifier of a human being [1] [11].

Imposter
It is just a case if someone try to match an Identity card with some people but actually that person does not exist. In other words it defined as, imposter case used when the speaker voice sample is not stored in the database i.e. speaker has not enrolled but trying to identify them [11].

Robustness
Robustness of a system will not failures and produce bad results when noisy data presented to the system. The term robust is use to define a system which works acceptable even data is noisy [11].

Inter-Speaker Variables And Intra-Speaker Variables
Generally security based Automatic speaker recognition systems falls due to Inter-speaker variables such as channel mismatch, background noise, handsets etc. and intra-speaker variables such as health conditions, emotional state, speakers living area etc. these cause an unacceptably high error rate and also affect the commercial feasibility of the automatic speaker recognition systems [3].

IV. Prosodic Features

The common Prosodic features of a speech signal are pitch, intensity, speaking rate [18].Automatically recognizing the expressive state of a speaker from their voice is known as emotion recognition. Attributes of voice are variable due to different moods of speaker and the affected attributes are pitch, speaking rate, intonation [19].Prosodic features of a speech signal changes continuously throughout the utterance. Hence for analysis short time speech signal frequently used, generally the length of speech signal is 10-30 ms. Integration of spectrum/cepstrum features with prosodic features produce better performance and negligible the chances of the speech feature being counterfeit [13].Prosodic features reflect the static features of the speech signal, the features of speech signal can be categorize as[20]–

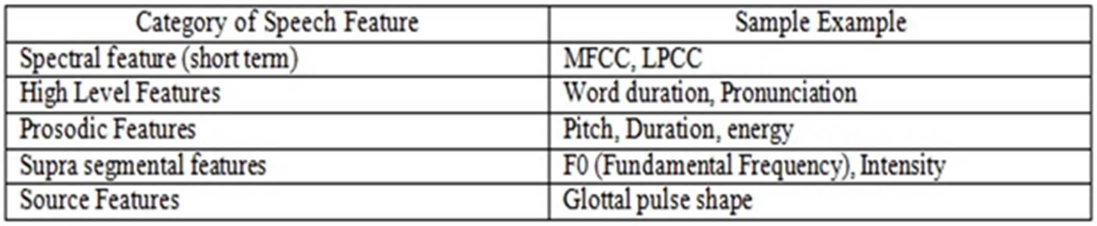
Speech Signal (Deception Induced Physiological Changes)
In this case speech affected by the following factors such as-
TIME: length of communication, response length, rate of speaking, response latency, rate change etc.
VOICEQUALITY: Vocal tension, articulation, satisfaction level (happiness)
FREQUENCY: pitch, pitch range, pitch variations etc.
INTENSITY: amplitude or loudness, loudness diversity etc.

FLUENCY: pauses (silent and filled), speech errors, interruptions etc.

Speech Signal (Deception Induced Cognitive Changes)
In this case speech affected by the following factors such as-
MEMORY: insensible, unpredictable, memory fails
INTELLIGENT (THINKING): delay, quick idea (on emergency) etc.
SENSATION: pressure, panic, tell something untruth etc.
PERCEPTIVENESS: unnecessary positive attitude, exciting etc.
ATTENTION: distraction, lack of concentration
Accuracy rate of emotion recognition depends on length/duration of utterance. Prosodic features can be characterized as different stages or we can say that its shows the speakers mind set. We can correlate acoustic prosody features suggestive of given emotions. For example if fundamental frequency (F0) increases that means speaker is in happy mood and also means that voice intensity and greater variability of F0. And in case of tediousness decreased mean F0 and the first formant frequency mean F1 increased etc. [21] As discussed in [18] it is infeasible to extract all the prosodic features from a speech signal but here listed a few prosodic features extracted from speech signal. To calculate pitch for voiced speech using the pitch tracking algorithm and intensity is measured for autoregressive modelling used. Below discussed some prosodic features-
Pitch, intensity, delta-intensity, delta-pitch
Skewness, Mean and standard deviation, median, maximum, minimum, shimmer, kurtosis, quartiles, range, regression error, linear & quadratic regression coefficients, variances between quartiles etc.
Speaking Rate: Mean and standard deviation of syllable durations, ratio between voiced and unvoiced speech duration

Zero-crossing rate etc.
As discussed in [22] most of the work on prosody to till date done for language modelling where prosodic makes known syntactic structure. It is due to that sometimes it is easy and appropriate prosodic information for a speech into language model as compared to acoustic model. Main factors of speech in which prosodic depend are speaking rate, volume, Pitch height and pitch range. Prosodic features are also termed supra-segmental features. The main phenomena of prosodic that it correlates with fundamental frequency and intensity these are easily voluntarily control through speaker [23].Research result shows that prosodic features are more effective for short speaker utterances whereas phono-tactic features more effective for longer utterances [24]. Table 1 gives the description of different types of speech features and their example with suitable feature extraction techniques.

TABLE I. CATEGORIES OF SPEECH FEATURES AND THEIR AND EXAMPLES



Category of Speech Feature	Sample Example
Spectral feature (short term)	MFCC, LPCC
High Level Features	Word duration, Pronunciation
Prosodic Features	Pitch, Duration, energy
Supra segmental features	F0 (Fundamental Frequency), Intensity
Source Features	Glottal pulse shape

Variation in voice or prosodic variation got attention between speech researcher & scientists; it is one of the achievable solutions to automatic recognition problems. Prosodic variation is often paralinguistic. One of the major problems in Speaker Recognition (SR) belongs to prosody, linguists and phoneticians. Linguists distinguish between linguistic and paralinguistic whereas phoneticians differentiate between vowels and consonants [25].Prosodic also known as Paralinguistic but somewhere they are distinct features to each other such as prosodic cues describe as linguistic function e.g. dialogue, discourse, focus (speech signal) and also provide information regarding to speakers gender, age, physical condition, emotion etc. hence prosodic features are distinguished by variation in duration and silence, pitch, loudness whereas paralinguistic features are vocal and not depend on duration, pitch, loudness.

V. CONCLUSION AND FUTURE WORK

Speaker recognition technology have been various recent improvements and achievement but still many problems occurs which is need to be solved. The problems exist due to speaker variability, channel variability, noise etc. so it is need that to found speech features parameters that are static for long period. And hence to overcome from variation in speaking and robust against voice variation due to cold etc. there is also a need to develop a technique which is able to manage noise and distortion problems occurred by telephone, channel and some background noise. To improve speaker recognition accuracy related to speech features extraction techniques.

The contribution of this study is that, to propose a method of extracting features from speech signals to recognizing a speaker. Prosodic features used to quantify pitch, energy, duration or pause of speech signal these features normally reflected the psychological conditions of a speaker. And hence generally use to recognizing psychological conditions of a speaker. The result shows that performance of the speaker recognition system improved by 5% as compared to individual MFCC is used while 20% improvement found in case of individual Prosodic considered.

ACKNOWLEDGEMENT

This work is sponsored by the CST-UP, Lucknow, India, under CST/D-413.

REFERENCES

[1] Dongdong Li, Yingchun Yang, and Weihui Dai, “Cost-Sensitive Learning for Emotion Robust Speaker Recognition” Hindawi, Publishing Corporation The Scientific World Journal, Volume 2014 (2014), Article ID 628516, 9 pages, http://dx.doi.org/10.1155/2014/628516

[2] Ubai SANDOUK, “Speaker Recognition Speaker Diarization and Identification” The University of Manchester School of Computer Science, year 2012, pp 14-101

[3] Jachym Kolar, Yang Liu, Elizabeth Shriberg, “Speaker adaptation of language and prosodic models for automatic dialog act segmentation of speech” Elsevier, Speech Communication, vol-52, Year (2010), pp. 236–245, www.sciencedirect.com

[4] AANCHAN K MOHAN, “COMBINING SPEECH RECOGNITION AND SPEAKER VERIFICATION” New Brunswick, New Jersey October, 2008, pp.1-108

[5] Greg Kochanski, Chilin Shih and Hongyan Jing , “Quantitative measurement of prosodic strength in Mandarin” Elsevier, Speech Communication vol-41, year (2003) pp. 625–645

[6] Sadaoki furui, “Recent Advances in Speaker Recognition” Elsevier, Pattern Recognition Letters, vol-18, year 1997, pp. 859-872.

[7] Nigel G. Ward, Alejandro Vega and Timo Baumann, “Prosodic and temporal features for language modeling for dialog”, Elsevier Speech Communication vol-54, year (2012) pp. 161–174

[8] Je Hun Jeon and Yang Liu, “Automatic prosodic event detection using a novel labeling and selection method in co-training” Elsevier, Speech Communication vol-54, year (2012), pp. 445–458, www.sciencedirect.com

[9] Ulrike Schild, Angelika B.C. Becker and Claudia K. Friedrich, “Phoneme-free prosodic representations are involved in pre-lexical and lexical neurobiological mechanisms underlying spoken word processing” Elsevier Brain & Language, vol- 136, year 2014, pp. 31–43

[10] L. Narupiyakul et al. “Focus to emphasize tone analysis for prosodic generation” Elsevier, Computers and Mathematics with Applications vol- 55, year 2008 pp. 1735–1753

[11] Mehrdad Ghassemian & Kasper Strange, “Speaker Identification-Features, Models and Robustness” Kongens Lyngby 2009 IMM-M.Sc.-2009-14, pp.1-118.

[12] Yuji Sato, “Voice quality conversion using interactive evolution of prosodic control” Elsevier Applied Soft Computing vol-5, year 2005, pp. 181–192

[13] Md Afzal Hossan, “Automatic Speaker Recognition Dynamic Feature Identification and Classification using Distributed Discrete Cosine Transform Based Mel Frequency Cepstral Coefficients and Fuzzy Vector Quantization”, RMIT University March 2011, pp.16-114

[14] ARUN RAJSEKHAR. G, “REAL TIME SPEAKER RECOGNITION USING MFCC AND VQ” National Institute of Technology Rourkela, year 2008, pp 13-71

[15] Doris Baum, “Recognizing speakers from the topics they talk about” Elsevier, Speech Communication, vol- 54, year 2012, pp. 1132–1142, www.sciencedirect.com

[16] M.H. Moattar and M.M. Homayounpour, “A review on speaker diarization systems and approaches” Elsevier Speech Communication, vol.54, year 2012, pp. 1065–1103, www.sciencedirect.com

[17] Nilu Singh* and R. A. Khan, “Digital Signal Processing for Speech Signals”, Bilingual International Conference on Information Technology: Yesterday, Toady, and Tomorrow, 19-21 Feburary 2015, pp. 134-138 © DESIDOC, 2015

[18] S. Wu et al. “Automatic speech emotion recognition using modulation spectral features” Elsevier, Speech Communication, vol- 53, year 2011,pp. 768–785

[19] Marcel Kockmann, Lukas Burget and Jan “Honza” Cernocky, “Application of speaker- and language identification state-of-the-art techniques for emotion recognition”, Elsevier, Speech Communication vol-53, year 2011, pp. 1172–1185

[20] Y.Zhou et al. “Deception detecting from speech signal using relevance vector machine and non-linear dynamics features” Elsevier Neuro computing, vol-151, year 2015, pp. 1042–1052

[21] Dmitri Bitouk, Ragini Verma and Ani Nenkova, “Class-level spectral features for emotion recognition” Elsevier, Speech Communication, vol- 52, year 2010, pp. 613–625

[22] Nigel G. Ward, Alejandro Vega and Timo Baumann, “Prosodic and temporal features for language modeling for dialog” Elsevier, Speech Communication, vol- 54, year 2012, pp. 161–174. www.sciencedirect.com

[23] Tomi H. Kinnunen, “Optimizing Spectral Feature Based Text-Independent Speaker Recognition”, UNIVERSITY OF JOENSUU, year 2005, pp. 1-156

[24] Tong Rong, “Automatic Speaker and Language Identification”, March 28, 2006, pp. 1-69.

[25] Susanne Schotz, “Linguistic & Paralinguistic Phonetic Variation in Speaker Recognition & Text-to-Speech Synthesis”, GSLT: Speech Technology 1, Term paper, Department of Linguistics and Phonetics, Lund University

[26] Sheeraz Memon, “Automatic Speaker Recognition: Modelling, Feature Extraction and Effects of Clinical Environment”, RMIT University, June 2010, pp. 1-242.

[27] Utpal Bhattacharjee and Kshirod Sarmah, “SPEAKER VERIFICATION USING ACOUSTIC ANDPROSODIC FEATURES”, Advanced Computing: An International Journal (ACIJ), Vol.4, No.1, January 2013, pp. 45-51

[28] Marc D. Pell, “Reduced sensitivity to prosodic attitudes in adults withfocal right hemisphere brain damage” Elsevier, Brain and Language Vol- 101 Yea r (2007) pp.64–79

Dr. Raees A. Khan, Associate Professor & Head, Department of Information Technology, BBA University, Lucknow (INDIA) Dr. Raees A. Khan done MCA from Punjab Technical University & Ph.D. (Computer Science) from Jamia Millia Islamia (Central University), New Delhi. He has written more than five books in Software Engineering and co-authored book, published numerous articles, several papers in the National and International Journals and conference proceedings. The area of expertise are Software Quality Assurance, Software testing, Software Security. He has done many projects on Software Security. He is a member of various professional bodies.

Nilu Singh received his MCA degree from Rajiv Gandhi Proudyogiki Vishwavidyalaya, Bhopal (M.P), India in 2008 and the M.Tech degree in computer science from Rajiv Gandhi University Itanagar, Arunachal Pradesh, India in 2011. Currently she is a research Assistant in the Babasaheb Bhimrao Ambedkar University (A central university), Lucknow, India. Her research area is Speaker Recognition and digital signal processing.