Extraction of Prosodic Features for Speaker Recognition Technology and Voice Spectrum Analysis

Authors: Nilu Singh[1] , R. A. Khan[1]
[1]SIST-DIT, Babasaheb Bhimrao Ambedkar University (Central University), Lucknow, UP, India
E-mail: nilu.chouhan@hotmail.com

Abstract:

The objective of this paper is to provide information and overview of prosodic features and spectral analysis of a speech signal. Speaker Recognition System is the make use of a machine to recognize the people from a spoken words. The majority in progress the highest level of development in the Automatic Speaker recognition System, done by using short term spectral information this approach disregard long-term selective information that can transmit supra segmental information such as prosodic and speaking style. We discussed in detail that what is prosody and its feature extraction technique including their mechanism and functionality. The goal of this paper is to provide an overview of prosodic feature extraction technique which helps people in Speaker Recognition Area. There are several characteristic in human speech prosody such as intonation, rhythm and stress, using these characteristic Speaker recognition can be done.

Index Term: Introduction of web 2.0, History of web 2.0, tools & technology, why web 2.0

1 INTRODUCTION

It is well known that the speech/voice of human being is the most natural way for the communication. Speech is the good medium to identify/recognize the people and the reason because it conveys information to the listeners. As we know the tone of the people is unique as their native place, it is possible to mimic voice but not exactly the tone if the people does not belong to the same native place. Speaking styles of different peoples will appear differently because the accent belongs to their native places. For example English speaking style is dissimilar for Indian people or any other people if their native place is different because there is a touch of native dialect in their voice/accent.

As per speech production system, the speech signal conveys Linguistic information i.e. related to language and speaker information. As of the speech awareness point of view it conveys information concerning the environment in which the speech was formed and transmitted. In general human can naturally make sense of most of this information; this skill of human has encouraged researchers to understand speech production and becoming aware of something via sense (perception) for developing the system that

automatically extract and process the prosperity of information in speech [1].

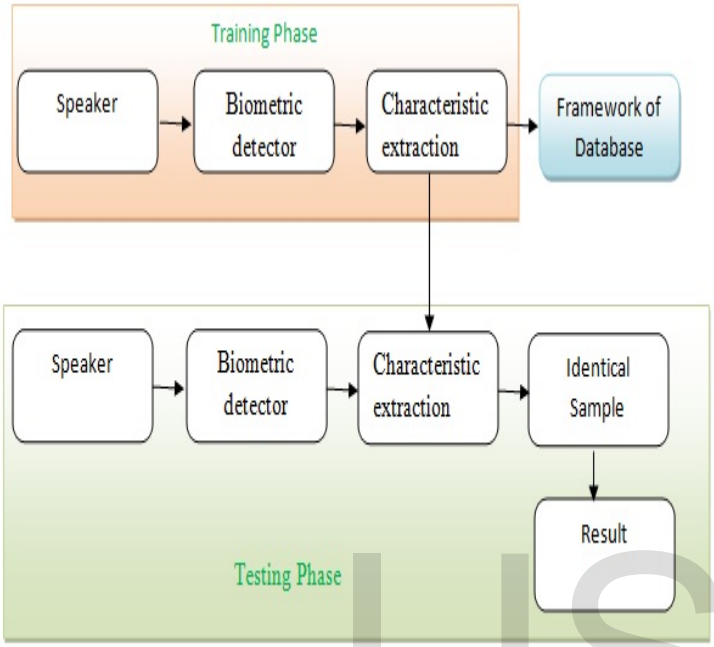

Figure 1: structural design of a usual biometric Recognition System

Automatic Speaker recognition is a machine proficient to identify an individual from a spoken words/sentence. Nowadays this technology used mostly in various areas such as forensic labs, access control, transaction authentication and many other areas. If we talk about Automatic Speaker Recognition then a question arise that why take speech signal? The suitable answer for this question that the speech signal conveys several information's about the speaker so it is used mostly for identifying a people. As discussed in [1][8] Automatic speaker recognition technology has wide area of applications where it can be used such as it enables systems to use a person's voice to control the access to restricted services e.g. automatic banking services, telephone access to financial transactions or some other areas. Automatic Speaker recognition technology also allows detection of speaker for in case accent based information recovery within audio collection, recognition of performer in forensic analysis and personalization of user device.

As discussed in [9] the basic prosodic includes Speaking rate, pause rate, timing and pitch f0 where pitch include melody, rate of change, global regrets. Prosodic feature are useful for assigning meaning, detecting sentence and topic boundaries also for speaker identification. As many studies say that the speech signal conveys the linguistic environment of the speaker. The fundamental frequency (f0) of the speech signal conveys the gender of speaker for example f0 is usually lower for male speakers, reason behind this the usually male have longer vocal cords [10]. Sound spectrum adopted the different frequencies present in a sound signal; it is an illustration of a segment of sound signal in terms of the quantity of vibration at every entity of the frequency. Spectrum of a speech signal generally existing as a graph of frequency. The aspect of measurement of spectrum there are many ways such as using a computer, using a microphone and an analog-digital converter as a function of time etc [11].

2 VOICE SPECTRUM ANALYSIS

As discussed in [9] the basic prosodic includes Speaking rate, pause rate, timing and pitch f0 where pitch include melody, rate of change, global regrets. Prosodic feature are useful for assigning meaning, detecting sentence and topic boundaries also for speaker identification. As many studies say that the speech signal conveys the linguistic environment of the speaker. The fundamental

frequency (f0) of the speech signal conveys the gender of speaker for example f0 is usually lower for male speakers, reason behind this the usually male have longer vocal cords [10]. Sound spectrum adopted the different frequencies present in a sound signal; it is an illustration of a segment of sound signal in terms of the quantity of vibration at every entity of the frequency. Spectrum of a speech signal generally existing as a graph of frequency. The aspect of measurement of spectrum there are many ways such as using a computer, using a microphone and an analog-digital converter as a function of time etc [11].

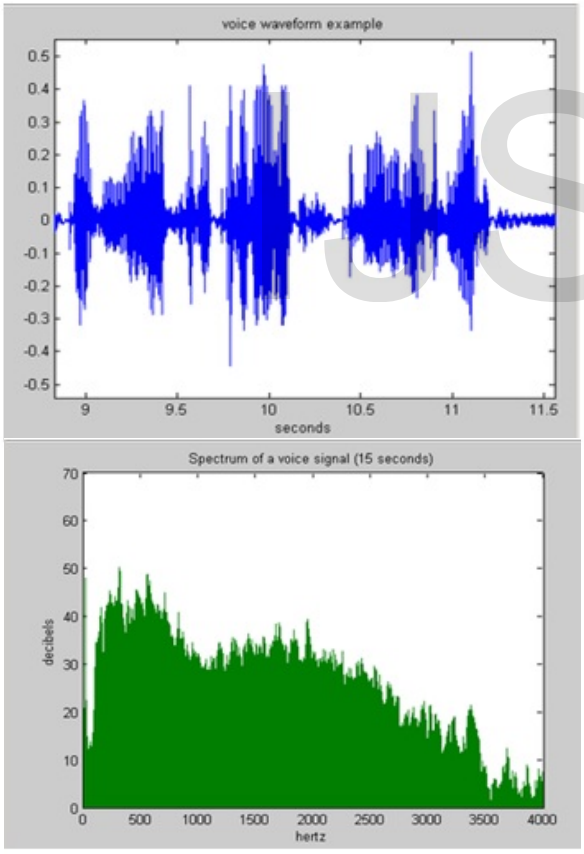

Fig 2: voice waveform and voice Spectrum (created using MATLAB)

Spectrum also defined as that it is a connection characteristically represented by a plot of the magnitude of a quantity of factor beside frequency; spectrum is a segment of speech signal. For speaker recognition, sound spectrum is used for to break a speech signal into small blocks. Spectrum analysis is a preface measurement to carry out reduction of speech bandwidth and supplementary acoustic processing [2][10]. As discussed in [10] using the spectrum of the speech signal we can be able to obtained most of the parameters which used for recognition process. Speech signal spectral moderate information regarding the vocal tract as well as the excitation source in the glottis by resources of the formants and the fundamental frequency. For speaker recognition technique, the parameters obtained by spectral typically the equivalent as the ones used in speech recognition technique.

3 PROSODIC FEATURES FOR SPEAKER RECOGNITION

Prosodic described as it use relating to the rhythmic characteristic of language or to the suprasegmental phonemes of pitch and stress and stage and nasalization i.e. the utterance of sounds modulated by the nasal resonators and voicing. Prosodic features of a speech signal can be used for to confine speaker specific information about variation in intonation, timing & loudness. Since prosodic features are Supra- Segmental long term features, they can provide corresponding information to systems/machines based on phonetic features/frame-level features [2]. One of the most considered features of speech is pitch/fundamental frequency which reflects vocal folds vibration rate, vibration rate is affected by

different physical properties of vocal fold. As discussed in [1] the experimental results specify that prosodic features are valuable and make available new information for speaker recognition, prosodic features have been used for speaker recognition for a long time. There are two approaches to exploring the prosodic features first is pitch and energy sharing here a feature vector consisting of per- frame log pitch, log energy and their first derivatives was used for speaker verification. Second is pitch and energy track dynamics, in this pitch and energy gestures use by modeling the joint slope dynamics of pitch and energy contours. Pitch and energy slope states i.e. rising & falling, describe as segment duration and phoneme or word context is use to train an n-gram classifier.

The most common features of prosodic are pitch, energy & duration, the value of pitch and energy are the average value and standard deviation for all frames and for rising or falling frames in other words says that number of frames where pitch is rising. The term duration described as the average & standard deviation of words and silence lengths in frames. Syllable- based prosodic features are more effective for speaker recognition. In case of prosodic system the term prosody stand for the patterns of stress and intonation in a language or in other words we can say that it lay out a collection of characteristics such as intonation, stress and timing, for the most part expressed using variation in pitch energy and duration at various levels of speech. Prosody may speculate various features of the utterance or also look up the gotten through environmental forces speaking habits as a person and hence it put up for speaker Recognition [4]. Extraction of prosodic feature can be categorized into two methods first is using the automatic speech recognizer, in this approach syllabic boundaries are obtained with the help of Automatic Speech Recognizer in this method variation points and start and end of articulation are used to segmented the speech signal. These segmented trajectories are then approximated and labeled into a small set of classes that describes the dynamics of f0 contour and energy contour. The second is the complimentary of automatic speech recognizer; here segments boundaries of articulation are estimated using discriminative information derived from the speech signal [5]. In [1][6] describe that a collection of prosodic features from duration and pitch related features such as mean and variance of pause duration and F0 values per word, extracted from each conversation face. In this study the experimental result show that prosodic features are valuable and make available new information for speaker recognition technology.

4 SPAEKER SPECIFIC FEATURES OF PROSODY

Prosody for the linguistics reflects various features of the speaker/utterance and it contains information regarding rhythm, stress and intonation of the speech. The speaker

communication manner in a dialogue features are analyzed to see that the speaker communication style observed in conversation, contained useful information to the speaker recognition. The concept about prosody that the speaker information might be found in both static and dynamic forms, and speech production possibly initiate from anatomical, physiological/behavioral each & every individual in nature hence speaker characteristics varying in nature [5][7]. The differences in physiological uniqueness occurred due to the shape and size of oral tract, nasal tract, vocal folds and trachea it can also go ahead to differences in vocal tract dynamics and excitation distinctiveness. The values of fundamental frequency f0 vary with speakers because of differences in the physical structure of the vocal folds of persons. As aerial discussed that speaker uniqueness also prejudiced by the speaking style of speaker. The speaking style of speaker is habitually determined by the persons/speakers source of revenue surroundings and the native language also. As studies of [7] say that the prosodic features are demonstrated in speech signal giving significant information concerning the speaking style of speakers.

5 ROBUSTNESS OF PROSODIC BASED SYSTEMS

As many studies [5][8][9] results show that the prosodic based features can be used to efficiently improve the performance of Automatic Recognition System and also add robustness to these systems. As discussed in [7] the majority in progress the highest level of development in the Automatic Speaker recognition System be dependent on the spectral features which is derived from short-term spectral analysis (MFCC) of the speech signal. Since the scale of the short-time spectrum encodes information about vocal tract shape for this reason spectral features are extensively used for speaker recognition technology. Since prosodic features derived from pitch, energy and duration which is relatively less affected by channel variation and noise as compared to spectral features, including all these aspects the conclusion is that prosodic based systems are more robust.

6 CONCLUSION

In this paper we try to explain about prosody and prosodic feature for Speaker Recognition and speech signal, mainly work of prosodic features based on modeling of pitch f0 statistics and early on work enlarged feature vector with raw f0. More recent work on prosodic modeled f0 separately and include other factors such as pause and voice duration. This paper has presented an Automatic Speaker Recognition System using prosodic features derived from pitch, energy and duration based parameters. Consistent pitch detection is especially significant to the statistical modeling of speech prosody. Pitch estimation of speech made natural mistakes due to acoustic noise and channel distortion, pitch halving and repetition errors.

REFERENCES

1. Jin, Qin, and Thomas Fang Zheng. "overview of Front-end Features for Robust Speaker Recognition." *APSIPA ASC 2011 Xian*. n. page. Print.
2. Shriberg, Elizabeth . "Higher-Level Features in Speaker Recognition." *Springer-Verlag Berlin Heidelberg 2007*. Speaker Classification I, LNAI 4343. (2007): 241-259. Print.
3. Mary, Leena. "Prosodic feature for speaker recognition." Trans. Array. Forensic Speaker RecognitionSpringer, 365-370. Print.
4. G.Adami, Andre, radu Mihaescu, Douglas A.Reynolds, and Reynolds J.Godfrey. "Modeling Prosodic Dynamics For Speaker Recognition." *ICASSP 2003,IEEE*. IV. (2003): 788-791. Print.
5. B. Peskin, J. Navratil, J. Abramson, D. Jones, D. Klusacek, D. Reynolds, B. Xiang, "Using Prosodic and Conversational Features for High-performance Speaker Recognition: Report from JHU WS'02," *ICASSP 2003*.
6. Mary, Leena, and B. Yegnanarayana. "Extraction and representation of prosodic features for language and speaker recognition." *ELSEVIER*. Speech Communication 50. (2008): 782–796. Print.
7. Dahak, Najim, Pierre Dumouchel, and patrick Kenny. "Modeling Prosodic Feature with joint Factor Analysis for Speaker Verification." *IEEE Transaction on Audio, Speech And Language Processing*. (2007): n. page. Print.
8. Laskowski, Kornel, and Qin Jin. "Modeling prosody for speaker recognition: Why Estimating Pitch May Be A red Herring." *Odyssey 2010, The speaker and language Recognition Workshop*. (2010): n. page. Print.
9. P. Heck, Larry. "Integrating High-Level Information for Robust Speaker Recognition." CLSP Workshop 2002. The Johns Hopkins University. Nuance Communications.
10. Farr´us i Cabeceran, Mireia. "FUSING PROSODIC AND ACOUSTIC INFORMATION FOR SPEAKER RECOGNITION." TALP Research Center, Speech Processing Group Department of Signal Theory and Communications Universitat Polit`ecnica de Catalunya. Barcelona, 07 2008.
11. "What is a Sound Spectrum?." *school of physics*. UNSW. Web. 18 Feb 2014.
12. HERMANSKY, HYNEK. "Speech recognition from spectral dynamics." *Sa¯dhana¯, Indian Academy of sciences*. Vol. 36. Part 5 (October 2011,): 729–744. Print.
13. "Spectral Analysis of Speech Signals." . N.p.. Web. 18 Feb 2014.
14. Weenink, David . *Speech Signal Processing with Praat*. ISBN-13. January 20, 2014 . 1-328. Print.