# Text Independent Speaker Identification with Prosody Features in Presence of Noise

**S.M. Jagdale, A.A.Shinde, J.S.Chitode**

***Abstract**: Automatic recognition of Meta data of a speaker apart from recognizing only his or her identity is a challenging task. It gives rich behavioral characteristics of a person.Maximum work have been done in speaker recognitionon low level spectral features. Which gives good accuracy with minimum error, but they ignore other information about the speaker. Also in spectral variations, in session variations and in channel variations these features give degraded performance. State-of-the-art systems for text-independent speaker identification use Mel Frequency cepstral coefficients (MFCCs) as main features. Generally this system performs very good under clean conditions and acceptable under matched conditions. Under mismatched conditions, however, performance significantly deteriorates. One of the principal reasons for poor performance in these conditions is because of the nature of low-level features; being spectral, they are susceptible to spectral variations due to noise and channel effects.Prosodic features are used successfully in these variation conditions as well as in presence of noise.In this paper multi SNR environment is considered. Recognition accuracy has been calculated at different SNR levels i.e. 15 dB, 25 dB and 35 dB.Also results are tested at different types of noise such as Traffic noise, cockpit noise, babble noise and fan noise. It has been found that combining prosodic features such as pitch, energy and formants gives improved performance.***

***Keywords: - Prosodic, Spectral, MFCC, SNR***

## I. INTRODUCTION

Human speech contains two types of information. One is physical gives the unique characteristics of an articulator system, other islearned provides the expression's pitch, energy and accentuation.Speech signal contains features which are from low-level to high-level [1, 7]. Speaker recognition comprises two technologies, text dependent and text independent recognition. Text dependent system identifies the person from specific phrase. This system has more accuracy as specific phrase is prompted and simple pattern matching method can be used.

**Revised Manuscript Received on July 25, 2019.**

**S.M. Jagdale, A.A.Shinde, J.S.Chitode**
Ph.D. Research Scholar, Bharati Vidyapeeth (Deemed to be) University COE, Pune, Maharashtra, India.
Department of Electronics, Bharati Vidyapeeth (Deemed to be University) COE, Pune, Maharashtra, India.
Department of Electronics Bharati Vidyapeeth (Deemed to be University) COE, Pune, Maharashtra, India.
Corresponding Author: S.M. Jagdale (sumatijagdale@gmail.com)

A text-independent speaker identification system extracts the acoustic spectral features fromconversational speech signal. This system identifies a person by speaking style, accent etc. of that person.Speaker identificationrefers to identify a person from his or her biometric characteristics. Inaspeakeridentificationsystem, featureextractionandclassifiermodelingarethetwo main units [3]. In this paper, we address the issue of feature extraction methods to enhance the speaker recognition performance under noisy environments. As the speech signal has two types of features, low level and high level features. High level information containsaccent, rhythm, word or phrase usage or pronunciation [18, 19]. The low-level information analyzes basic structure of speech signal i.e. the unique characteristics of a person [4].The former gives details about speaker-specific characteristics, speaking style, accent, pronunciation etc. On the other hand, the latter gives information about physiological properties.Recent work has shown that by combining high-level features of speech into the conventional system, the performance is improved.The recognition accuracy of practical speakerrecognition systems greatly deteriorates when speech signals are corrupted by noise.Therefore, achieving noiserobustnessis an important issue to make these systems robust in real acoustic conditions [8]. Prosodic features are the most common high-level feature used in speaker recognition to improve performance against channel variation, spectral variation and session variation conditions. Prosodic features are pitch, energy, formantsand speaker-specific information like melody, intonation and loudness [1, 5].

## II. SYSTEM OVERVIEW

Speaker identification comprises two parts one is enrollment and other is testing. Figure 1 shows basic overview of speaker recognition system.Speech samples are pre-processed for enhancing purpose. Then prosodic features such as pitch, energy and formants are extracted from speech sample. These features are fused together by concatenating and then classification is done. The result is displayed in terms of accepted and denied. In this paper result is calculated at different SNR levels. As the SNR value increases the recognition accuracy increases [2, 11].


*Published By:*
*Blue Eyes Intelligence Engineering*
*& Sciences Publication*
*Retrieval Number: I30250789S319/2019©BEIESP*
*DOI: 10.35940/ijitee.I3025.0789S319*

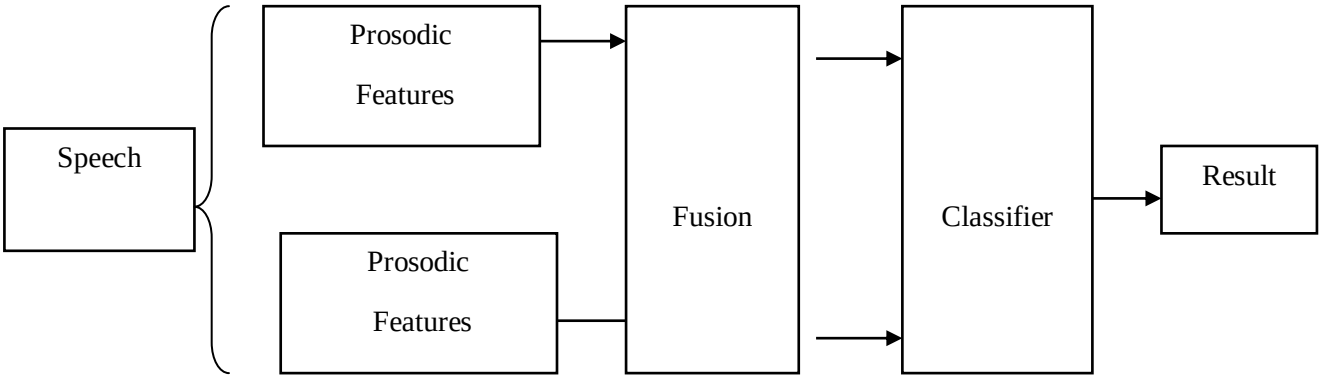


Figure 1. Speaker Recognition system

## III. FEATURE EXTRACTION

Speech signal contains two types of information i.e. acoustic and prosodic information.Acoustic features are extracted with MFCC(Mel Frequency Cepstral Coefficient) algorithm, which gives improved accuracy, but ignores the other level of information of speech signal [22,23].Prosodic features includes fundamental frequency, intensity, speaking rate and disfluencies [5].It spans over long segments like syllables, words,and utterances and reflects differences in speaking style,language background, sentence type, and emotions to mentiona few [6].

### *A. Prosody Features*

Prosody plays an important role in differentiating speakers [17]. Prosodic features carry speaker specific information like melody, intonation and loudness. They are also referred to as source features as they originate at the glottal source.They are maximum/minimum values of pitch, energy etc.Prosodic characteristics such as rhythm, stress and intonation in speech convey some important information apart from identity of the speaker. Also prosody caries mood and emotions of the speaker embedded in the speech utterance.

### *1) Pitch Feature:*

The most important prosodic parameter is the fundamentalfrequency (or F0). Combining F0-related featureswith spectral features has been shown to be effective, especially in noisy conditions.

### *a) Extraction algorithm:*

There are two ways for calculation of pitch, in time domain and in frequency domain.Time Domain methods are Computationally Simple and frequency domain methods are computationally complex.Here pitch is extracted by Autocorrelation algorithm in time domain. Autocorrelation method is simple to implement and robust against noise.Autocorrelation is a correlation of a signal with itself. The maximum of similarity occurs for time shifting of zero. Another maximum should occur in theory when the time-shifting of the signal corresponds to the fundamental period. Following is the equation of Autocorrelation.

$$\emptyset(\tau) = \sum_{n=0}^{N-1} x(n)x(n+\tau) \quad \text{..... 1}$$

Table 1 shows recognition accuracies for pitch feature at different SNR levels. This table shows that the recognition accuracy gradually reduces as the decline of SNR level[8].Table 2 shows recognition accuracies at different types of noise.Figure 2 presents FAR (False Acceptance Rate), FRR (False Rejection Rate) curves. The point of intersection of these curves is the Equal Error Rate(EER) value.

**Table 1. Recognition Accuracies with Pitch**

| Feature | 15 dB | 25 dB | 35 dB |
|---|---|---|---|
| Pitch | 62.67% | 79.82% | 84.21 % |

**Table 2. Recognition Accuracies for different noise types**

| Feature | Babble | Fan | Cockpit | Traffic |
|---|---|---|---|---|
| Pitch | 68.00 % | 80.90 % | 60.20 % | 51.11 % |

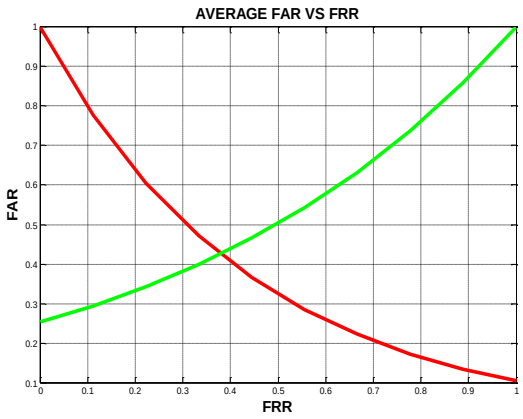


**Figure 2.FAR FRR Curve with Pitch (20 dB)**

### *2) Energy:*

The energy is achieved by averaging energy of each frame. The arousal state of the speaker (high activation versus low activation) affects the overall energy, energy distribution across the frequency spectrum [7].



*Retrieval Number: I30250789S319/2019©BEIESP*
*DOI: 10.35940/ijitee.I3025.0789S319*
*Published By:*
*Blue Eyes Intelligence Engineering*
*& Sciences Publication*

### *3) Formant:*

In voiced speech resonant frequencies of the vocal track are called as formants which are importantparameters for both automatic speaker recognition and speech synthesis.Formants are the spectral peaks of the sound spectrum of the acoustic resonances of the human vocal tract. The sound production can be modeled as a time varying linear system that is excited by a sequence of impulses. The first five vocal resonant frequencies, i.e. formants (F1, F2, F3,F4, F5), during voiced-speech are important for each personand therefore are used more as the speaker features. For voiced-speech, the glottis signal is periodic with a fundamental frequency (i.e. pitch, F0) [21]. There are following methods of estimation of formant frequencies: cepstral analysis method and linear predictive cepstral coefficient (LPCC) method. In this paper LPCC method has been used for recognition purpose. In this method first pre-processing is done. Linear predictive coefficients are found by LPC procedure. The roots of predictor polynomials are calculated to find the peak locations in the spectra of linear predictive filters. Only the roots with positive angles are selected and the angles are converted into frequencies mathematically using equation. The formants are sorted in ascending order with the lowest frequency becomes *F1*, the second lowest becomes *F2* and so forth until the fifth formant *F5*. Table 3 shows recognition accuracies for different SNR levels i.e. 15, 25, and 35dB. As SNR level increases recognition accuracy increases [2]. Table 4 shows recognition accuracies at different noise types. Traffic noise shows more accuracy as compared to other types of noise.

Table 3. Recognition Accuracies with Formants

| Feature | 15 dB | 25 dB | 35 dB |
|---|---|---|---|
| Formant | 19.30 % | 45.08 % | 76.71 % |

Table 4. Recognition Accuracies for different noise types

| Feature | Babble | Fan | Cockpit | Traffic |
|---|---|---|---|---|
| Formant | 31.18 % | 24.74 % | 30.34 % | 40.55 % |

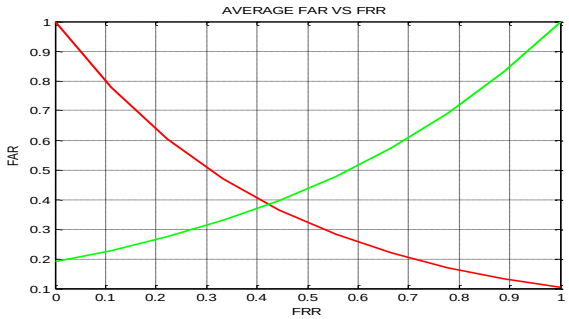


Figure 3.FAR FRR Curve With Pitch (20 dB)

## IV. COMBINED SYSTEM

The recognition accuracy has been increased by concatenation of feature vectors.Integration at feature level offers good recognition rate [13, 14]. Also it Increases the robustness to unexpected failure of subsystems.The feature set holds richer information about speech signal and processing time has been reduced [16].

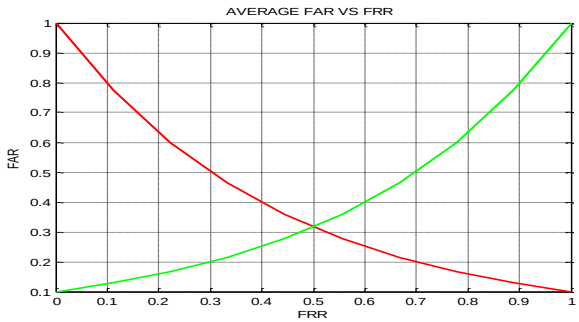


Figure 4.FAR FRR Curve With combined features (20 dB)

Table 5. Recognition Accuracies for combined features.

| Feature | 15 dB | 25 dB | 35 dB |
|---|---|---|---|
| Combined | 67.55% | 88.45% | 100 % |

## V. DATABASE USED

ELSDSR (English Language Speech Database for Speaker Recognition) corpus has been used for the evaluation of automatic speaker recognition system. ELSDSR corpus was designed in Technical University of Denmark. The text language is English, and is read by 20 Danes, one Icelander and one Canadian. The recording work has been done in a chamber of DTU. The equipment for recording is MARANTZ PMD670 portable solid state recorder. The recording type can be stereo, mono or digital, and the file can be recorded into .wav .bwf .mpg or .mp3 format. In this database, the voice messages were recorded into the most commonly used file type-.wav (PCM). The sampling frequency is chosen 16 kHz with a bit rate of 16.

## VI. DISCUSSION

The results shown in Table 1 and Table 2 are recognition accuracies at different SNR levels and at different noise types with pitch feature. This result reflects that as signal quality (SNR value) increases error decreases. The recognition accuracy increases at higher SNR value [20]. Table 4 and 5 shows recognition accuracies with formant features.



Retrieval Number: I30250789S319/2019©BEIESP
DOI: 10.35940/ijitee.I3025.0789S319
Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

With formants also recognition accuracy increases with SNR level. The robustness and performance of a speaker identification system is increased by combining different prosodic features.Table 5 shows recognition accuracies at different SNR levels and the result shows improved accuracies.Figure 2 Figure 3 and figure 4 shows FAR-FRR curves, which reflects the EER value. There are two types of errors, namely, false accept (FA) and false reject (FR). A false accept occurs when an imposter is selected as the genuine speaker. A false reject arises when a true speaker is rejected as an imposter. The equal error rate (EER) of the system is when the FAR and FRR are equal and is used to measure speaker verification system performance [12, 15].

## VII. CONCLUSION

Speaker Recognition is a need in Forensics for voice authentication purpose. Low level features shows low error rates but ignores other levels of information such as learned habits of a speaker. Low level features are susceptible to spectral as well as channel variation and noise present in the system. Prosodic level features shows excellent performance in these variation conditions and in presence of noise.High level features potentially increases the robustness. Fusion improves the performance approximately by 15 to 20 % in speaker recognition system.

## REFERENCES


1. Sithara A, Abraham Thomas, Dominic Mathew, "Study of MFCC and IHC Feature Extraction Methods With Probabilistic Acoustic Models for Speaker Biometric Applications", 8th International Conference on Advances in Computing and Communication, Elsevier publication, 2018, pp.267-276.
2. Furong Yan, Aidong Men, " An Improved Ranking Based Feature enhancement ApprachfroRubust Speaker Recognition", IEEE Access Multidisciplinary Journal, Vol. 4, 2016 , pp.5258-5267.
3. Jia-Ching Wang, Chien-Yao Wang, Yu-Hao Chin, Yu-Ting Liu, En-Ting Chen, Pao-Chi Chang, "Spectral-temporal receptive fields and MFCC balanced feature extraction for robust speaker recognition", Springer publication, 2017, pp.4055-4068.
4. Clark D. Shaver,J.M. Acken"Effect of equipment variation on Speaker Recognition error rate" IEEE conference on Acoustic Speech and Signal Processing,2010,pp.1814-1817.
5. William Wang, FadiBiadsy, Andrew Rosenberg, " Automatic detection of speaker state : Lexical, Prosodic, and Phonetic approaches to level of interest and intoxication classification " Computer speech and Language Journal, 2013, Elsevier Publication, pp.168-189.
6. Tomi Kinnunen, EvgenyKarpov, PasiFranti ;"Real-Time Speaker Identification and Verification", IEEE transaction on Audio, Speech, and Language Processing, Vol. 14,No. 1,pp.277-288.
7. Moataz E Ayadi , Mohamed S.Kamel , FakhriKarray , "Survey on speech emotion recognition: Features, classification schemes, and databases"Elsvier journal on 'Pattern Recognition'2011,pp.572-587.
8. Furong Yan, Aidong Men, " An Improved Ranking Based Feature enhancement ApprachfroRubust Speaker Recognition", IEEE Access Multidisciplinary Journal, Vol. 4, 2016 , pp.5258-5267.
9. Taufiq Hasan, John H.L., " Maximum Likelihood Acoustic Factor Analysis Model for Robust Speaker Verification in Noise", IEEE Transaction on Audio, Speech and Language Processing, Vol. 22, No. 2, 2014, pp. 381-391.
10. Elizabeth Shriberg," Higher-Level Features in Speaker Recognition" Springer-Speaker Classification,2007,pp.241-259.
11. Hanwu Sun, Bin Ma Haizhou Li" An Efficient Feature Selection Method for Speaker Recognition" IEEE conference, 2008, pp. 1-4.
12. R.P.Ramachandrana, K. R.Farrell, R.Mammonec, "Speaker recognition general classifier approaches and data fusion methods",Elsevier pub., Parttern Recognition,2002,pp.2801-2821.
13. Pullella, D.; Kuhne, M.; Togneri,R. "Robust speaker identification using combined feature selection and missing data recognition", IEEE International Conference, signal processing and analysis,2008,pp.4833-4836.
14. Josef P. Campbell, D.A. Reynolds, R.B.Dunn; "Fusing High and Low Level Features for Speaker Recognition", Eurospeech 2003, pp.2665-2668.
15. Marc Ferras, Cheung Leung, C.Barras, J. Gauvain; "Comparison of Speaker Adaptation Methods as Feature Extraction for SVM-Based Speaker Recognition "IEEE Transaction on Audio, Speech and Language Processing,2010,pp.1366-1378.
16. Chakroborty,S.;Roy,A.;Saha,G. "Fusion of a complementary feature set with MFCC for improved closed set Text-Independent Speaker Identification" IEEE International conference on Computing and Processing,2006,pp.387-389.
17. Andre´ Gustavo Adami, "Modeling prosodic differences for speaker recognition",Elsevierpublication,Speech communication, 2007, pp. 277-291.
18. Mangayyagari, TanmoyIslam,R. Sankar;" Enhanced Speaker Recognition based on Intra-Modal Fusion and Accent Modeling", IEEE international conference on pattern Recobnition,2008,pp.1-4.
19. W.M. Campbell, J. P. Campbell, T.P.Gleason, D. A. Reynold;" Speaker Verification using Support Vector Machine and High-Level Features" IEEE transaction on Audio, Speech and Language Processing, 2007, Vol. 15, pp. 2085-2094.
20. K.Rao, A.Vuppala, S.Chakrabarti, L.Dutta, "Robust Speaker Recognition on Mobile Devices", IEEE, 2010, pp.1-5
21. Khaled Daqrouq, Tarek A. Tutunji, "Speaker identification using vowels features through a combined method of formants, wavelets, and neural network classifiers", Elsevier publication, Applied soft computing, 2015, pp. 231-239.
22. 22]. S.M. Jagdale, A.A.Shinde and J.S.Chitode, "Robust Speaker Recognition based on Low Level and Prosodic Level Features", Springer Conference, IDSSA,Delhi, 2019.
23. 23]. S.M. Jagdale,J.S.Chitode, S.Saste, "Automatic Language and Text Independent Speech Emotion Recognition using MFCC and DWT for ATM Security System", International Journal of Pure and Applied Mathematics, Volume 118 No. 18, 2018, pp.3335-3342.




*Retrieval Number: I30250789S319/2019©BEIESP*
*DOI: 10.35940/ijitee.I3025.0789S319*

*Published By:*
*Blue Eyes Intelligence Engineering*
*& Sciences Publication*