Analysis of Variations in Speech in Different Age Groups using Prosody Technique

Maheshkumar B. Landge
M.Phil. Student
Dept. of CS & IT,
Dr. B.A.M.U. Aurangabad

R.R. Deshmukh
Professor & Head
Dept. of CS & IT
Dr. B.A.M.U. Aurangabad

P.P. Shrishrimal
Ph.D. Student
Dept. of CS & IT
Dr. B.A.M.U. Aurangabad

ABSTRACT
Speech is the vocalized form of human communication. The variations in speech occurred due to vocal tract vibration. The main aim of speech analysis is to derive time varying characteristics from speech. The three features are considered for analysis namely energy, pitch, and formant frequency. It is observed that variations in speech in same and different age groups are minimum.

Keywords
Speech, Speech analysis, Prosodic features, speech analysis methods, Pitch.

1. INTRODUCTION
Speech is the most basic common and efficient form of communication method for people to interact with each other [1]. Speech signal not only transfers language information but also gives information about the speaker i.e. speaker's age, gender, local origin, health, emotional state(mood of speaker) and his unique characteristic(Garvin and Ladefoged,1963; Nolan,1983) [2].

Speech recognition is the capability of a machine or program to recognize words and phrases in spoken language and translate into machine-readable format. There are some applications of speech recognition system available like voice dialing, simple data entry and speech to text. Automatic speech recognition system includes some several components taken from different areas such as statistical pattern recognition, communication theory, signal processing, combinational mathematics and linguistics [3].

Speech is produced as a result of excitation of the time-varying vocal tract system. In speech production, both excitation and the vocal tract change continuously with time. One objective in speech analysis is to derive the time-varying characteristics of the speech production mechanism from the speech signal [4].

2. LITERATURE SURVEY
Speech Analysis Methods

2.1 LPC (Linear Predictive Coding)
It is the most common technique for low-bit-rate speech coding and its popularity derives from its simple computation and reasonably accurate representation of many types of speech signals [5].

The most important characteristic of LPC analysis is to approximation the LPC coefficients from each of the windowed speech waveforms.

2.2 Filter Bank Analysis
Filter bank analysis consists of a set of bank-pass filters. A single input speech signal is simultaneously passed through these filters, each outputting a narrowband signal containing amplitude information about the speech in a narrow frequency range. The bandwidths normally are chosen to increase with center frequency, thus following decreasing human auditory resolution. They often follow the auditory Mel scale, i.e. having equally-spaced, fixed bandwidths below 1 kHz, then logarithmic spacing at higher frequencies. Such a filter bank analysis tries to simulate very simple aspects of the human auditory system, based on the assumption that human signal processing is an efficient way to do speech analysis and recognition.

2.3 Mel-Frequency Cepstral Analysis
The Mel Frequency Cepstral Coefficient is the well-known and most widely used feature extraction method in speech domain. The MFCC is based on the human auditory perception system. The human auditory perception system does not follow a linear scale of frequency. For each tone with actual frequency f measured in Hz, a subjective pitch is calculated known as 'Mel Scale'. The Mel frequency scale is linear frequency spacing below 1000 Hz and logarithmic spacing above 1000Hz. As a reference point, the pitch of a 1 KHz tone, 40 dB above the perceptual hearing threshold is defined as 1000 Mels [6].

2.4 Perceptual Linear Prediction
The Perceptual Linear Prediction analysis method was invented by Hermansky in 1990. The main objective of this method is to define the psychophysics of human hearing more precisely in the feature extraction process.

In contrast to pure linear predictive analysis of speech, perceptual linear prediction modifies the short-term spectrum of the speech by several psychophysically based transformations [7].

2.5 Relative Spectral Transform - Perceptual Linear Prediction
Another popular speech analysis method is Relative Spectral Transform - Perceptual Linear Prediction. Relative Spectral Transform - Perceptual Linear Prediction is a separate technique that uses a band-pass filter to the energy in each frequency sub-band in order to smooth over short-term noise variations and to remove any constant offset resulting from static spectral coloration in the speech channel [8].

3. METHODOLOGY
For the proposed work the text corpus of 10 isolated words in Marathi language is developed. The speech data is collected from different speakers using the developed text corpus. The methodology for the proposed work is shown in figure 1.

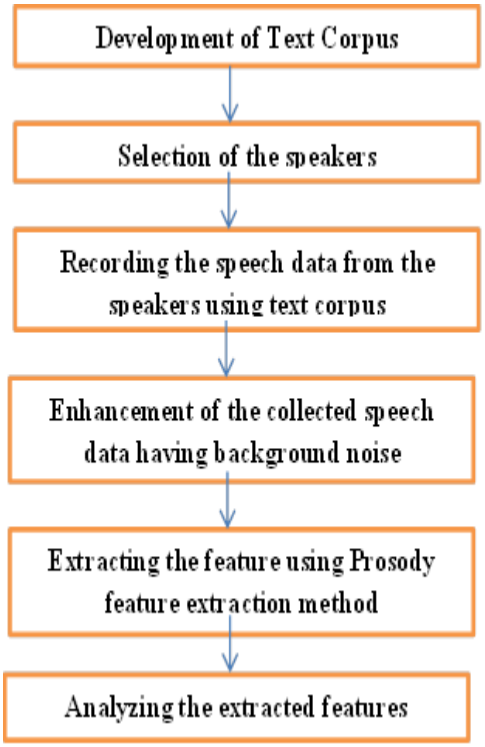

Fig.1 Flow diagram of methodology

In first step a text corpus is developed from Marathi alphabet book. In second step 60 speakers database has been collected. 20 Speakers were selected from each age group out of which 10 were male and 10 female. The selected speakers were according to the age groups. In third step Speech samples were recorded from total 60 speakers from three age groups in normal environment. In fourth step the speech samples having noise were enhanced using low pass filter. In fifth step features were extracted using Prosody feature extraction technique. Finally in sixth step the analysis of the extracted features was carried out.

3.1 Development of speech database

A speech database of different speakers from different age groups is developed to cover the maximum variations of the age groups. Sixty speakers were selected from all over the three age groups. The Speakers were selected such that they were in age groups ranging from 11-20, 21-30, and 31-40. The speakers were selected in the age groups whose native language is Marathi. The literacy was another criterion which was considered during the selection of speaker.

3.2 Data Collection Statistics

The speech data is collected from 3 age groups. From each age group the speech samples are collected from 20 speakers, out of which 10 were Male and 10 were Female. These speakers were also selected according to the age group as mentioned in selection of speaker sub section under the Data Collection section. As the speakers were asked to speak the 10 words with 10 utterances of each word, total 100 speech samples are collected from each speaker. The database contains total 6000 utterances in all of the 10 words from 60 speakers [9].

Table. 1 Data Collection Statistics

Sr. no.	Age group	Male speakers	Female speakers	Number of utterances	Total speakers	Total Utterances= Total speakers* no. of utterances
1	10-20	10	10	10	20	2000
2	21-30	10	10	10	20	2000
3	31-40	10	10	10	20	2000
Total					60	6000

3.3 Feature extraction

Feature extraction is a basic and fundamental preprocessing step to pattern recognition and machine learning problem. It is a special form of dimensionality reduction technique used to reduce the data which is very large to be processed by an algorithm. In feature extraction the given input data is converted into a set of features which provides the relevant information for performing a desired task without the need of the full size data but using the reduced set .

3.3.1 Prosody technique

Prosody of speech is defined in the linguistic literature as the suprasegmental properties of speech. Traditionally, the definition of prosody is thought to include the pitch/F0, loudness/intensity, and rhythm/duration aspects of speech (Brown 2005). Prosody is a collection of components that controls the pitch, loudness, and rate of speaking. The variations or changes of intonation, rhythm, and stress pattern belong to what we call the prosody of a sentence. A sentence can be spoken (uttered) with different prosodic characteristics depends on emotional state of speaker [10].

Prosodic features are the rhythmic and intonational properties in speech like duration, voice intensity, and fundamental frequency [11]. Prosodic features are identified to express various information such as lexical tones, speaking styles, emotional states.

There are following different prosodic features of speech:

Stress: Stress, or emphasis, is easy to use and recognize in spoken language, but difficult to define. A stressed word or syllable is usually preceded by a very slight pause, and is spoken at slightly increased volume [12].

Volume: Apart from the slight increase in loudness to show stress, volume is normally used to show feelings such as fear or anger. In writing, we can show it by the use of an exclamation mark, or typographically with capitals or italics.

Pitch: Pitch is the fundamental frequency of vibration of the vocal cords, which are present at the top of one's trachea. They vibrate quasi-periodically only for voiced phonemes, namely vowel, semivowel and nasal sounds.

So, for unvoiced stops such as /p/, /k/, /t/, /th/, /ch/ and unvoiced fricatives such as /f/, /s/, etc. there is nothing called pitch [13].

Formant frequency: In speech signals the Formant frequencies typically represent the resonance of the vocal tract. Formants are frequency peaks which have a high degree of energy in the spectrum. They are specially projecting in vowels [14].

4. ANALYSIS AND RESULTS

4.1 Analysis of average energy

The table 2 shows the distance matrix of energy values for two male speakers from same age group for first five utterances of Akshar word and similarly table 3 shows the distance matrix of energy values for two male speakers from 31-40 and 11-20 age groups for first five utterances of Akshar word. It is observed that in same and different age groups there are variations in energy value. Unwanted energy variations are caused by many factors, the dominant one being background noise. Background noise drastically changes the sound level of silent segments, and on the other hand slightly changes the sound level of loud segments. Several solutions for this problem are proposed in[15].

Unwanted energy variations can also be caused by: (1) different microphone gains, (2) different microphone placement, (3) variations in loudness levels across different speakers as well as (4) changes in loudness level of a single speaker over time.

Table 2: Distance matrix of Akshar word for two male speakers from 11-20 age groups for calculation of energy

Speaker2 / Speaker1	1	2	3	4	5
1	1.367	1.512	1.369	1.377	1.439
2	1.236	1.381	1.238	1.246	1.308
3	0.796	0.941	0.798	0.806	0.867
4	3.810	3.955	3.811	3.820	3.881
5	1.351	1.496	1.353	1.361	1.422

In table 2 five utterances of Akshar word spoken by two speakers from 11-20 age groups are considered. It is observed that 1st, 3rd, 4th and 5th utterances are having closest range energy values and the 2nd utterance is having different energy value. So there is minimum variation among five utterances.

Table 3: Distance matrix of Akshar word for two male speakers from 31-40 age group and 11-20 age group for calculation of energy

Speaker2 / Speaker1	1	2	3	4	5
1	0.129	0.274	0.131	0.139	0.200
2	1.389	1.534	1.391	1.399	1.460
3	0.106	0.251	0.108	0.116	0.178
4	0.486	0.631	0.487	0.495	0.557
5	0.112	0.257	0.114	0.122	0.184

In table 3, five utterances of Akshar word spoken by two speakers from 31-40 and 11-20 age groups are considered. It is observed that 1st, 3rd and 4th utterances are having closest range energy values and 2nd, 5th utterances are having different range energy values. So there are few variations present in these age groups.

4.2 Analysis of pitch

The table 4 shows the distance matrix of pitch values for two male speakers from same age group for first five utterances of Akshar word and similarly table 5 shows the distance matrix of pitch values for two male speakers from 11-20 and 21-30 age groups for first five utterances of Akshar word. There are some variations in pitch values in both tables. The pitch has aroused the periodicity through vocal cords vibration when madding voiced sound, pitch frequency is a very important parameter using to describe the characteristic of voice excitation source. The variational range of pitch frequency is generally from 50 Hz to 500 Hz, the cycle of the male voice is 50 Hz - 300 Hz, and the female is 100 Hz - 500 Hz. Although each person's different vocal structure lead to different fundamental frequency, because of the pitch frequency's scope is a little small, the gap between different people is little, and the most important is pitch frequency is affected by a lot of factors, such as emotion, tone, it is very difficult to achieve accurate fundamental frequency. The male fundamental frequency is generally lower than the female [16].

Table 4: Distance matrix of Akshar word for two male speakers from 11-20 age group for calculation of pitch extraction

Speaker2 / Speaker1	1	2	3	4	5
1	1.964	399.187	2.964	2.964	-1.627
2	2.223	399.446	3.223	3.223	-1.368
3	3.963	401.186	4.963	4.963	0.372
4	2.772	399.996	3.772	3.772	-0.818
5	2.807	400.030	3.807	3.807	-0.783

In table 4, five utterances of Akshar word spoken by two speakers from 11-20 age groups are considered. We observed that two utterances 2nd and 5th having different pitch value and other 1st, 3rd, and 4th utterances are having closest range pitch value. There are few variations in this age group.

Table 5: Distance matrix of Akshar word for two male speakers from 11-20 age group and 21-30 age group for calculation of pitch extraction

Speaker2 / Speaker1	1	2	3	4	5
1	1.964	0.964	0.964	0.126	0.964
2	2.223	1.223	1.223	0.385	1.223
3	3.963	2.963	2.963	2.125	2.963
4	2.772	1.772	1.772	0.934	1.772
5	2.807	1.807	1.807	0.969	1.807

In table 5, five utterances of Akshar word spoken by two speakers from 11-20 and 21-30 age groups are considered. It is observed that two utterances 1st and 4th having different pitch value and other 2nd, 3rd, and 5th utterances are having same range pitch value. There are more similarities and minimum variations among five utterances.

4.3 Analysis of formant frequency

The table 6 shows the distance matrix of formant frequency values for two male speakers from same age group for first five utterances of Akshar word and similarly table 7 shows the distance matrix of formant frequency values for two male speakers from 11-20 and 21-30 age groups for first five utterances of Akshar word. In these tables the values of formant frequency extraction are having minimum variations. Since, according to gender and age group the formant frequency generated in speech is different for all speakers. Variability in speech occurs due to differences in the rate at which the vocal folds vibrate.

Table 6: Distance matrix of Akshar word for two male speakers from 11-20 age group for calculation of formant frequency extraction

Speaker2 / Speaker1	1	2	3	4	5
1	2.191	2.354	-0.322	7.953	2.134
2	12.443	12.606	9.930	18.204	12.385
3	3.811	3.974	1.298	9.573	3.754
4	6.767	6.930	4.254	12.529	6.710
5	9.878	10.041	7.365	15.640	9.821

In table 6, five utterances of Akshar word spoken by two speakers from 11-20 age groups are considered. It is observed that three utterances 1st, 2nd and 5th are having closest range formant frequency value and other 3rd and 4th utterances are having different range formant frequency value. There are few variations between the speakers from 11-20 age groups.

Table 7: Distance matrix of Akshar word for two male speakers from 21-30 age group and 31-40 age group for calculation of formant frequency extraction

Speaker2 / Speaker1	1	2	3	4	5
1	4.009	3.902	6.055	4.196	2.247
2	4.806	4.699	6.852	4.993	3.044
3	3.461	3.354	5.507	3.648	1.699
4	0.213	0.106	2.259	0.400	-1.549
5	0.808	0.701	2.854	0.995	-0.954

In table 7, five utterances of Akshar word spoken by two speakers from 21-30 and 31-40 age groups are considered. It is observed that three utterances 1st, 2nd and 4th are having closest range formant frequency value; only 3rd and 5th utterances are having different range formant frequency values. There are few variations between the speakers from 21-30 and 31-40 age groups for utterance of Akshar word.

5. ACKNOWLEDGMENTS

We would like to thank to Department of CS & IT, Dr. Babasaheb Ambedkar Marathwada University, Aurangabad for providing us basic facilities to carrying out the research work.

6. CONCLUSION

In this paper the features like average energy, pitch, formant frequency are used to analyze the variations in different age group speakers using Prosody technique. It was observed that there are less differences in the utterances for speakers from different age groups and same age groups. We observed that when we compare energy features of similar age groups there are less variations in energy, similarly observations were for pitch, formant frequency features respectively. Analysis of speech by different age groups leads to develop the state-of-art Automatic Speech Recognition system using developed database. Some more combination of speech signal enhancement techniques and feature extraction techniques will be tried on developed database. Speech analysis will be used to overcome the issue raised due to the variation in the utterance time of a word for increasing the accuracy of the recognition system.

7. REFERENCES

[1] Vimala C., Dr. V. Radha, "A review on speech recognition challenges and approaches", World of Computer Science and Information Technology Journal (WCSIT), ISSN: 2221-0741 Vol. 2, No. 1, 1-7, 2012.

[2] M. Benzeghiba, R. De Mori, O. Deroo, S. Dupont , T. Erbes, D. Jouvet, L. Fissore, P. Laface, A. Mertins, C. Ris, R. Rose, V. Tyagi, C. Wellekens, "Automatic speech recognition and speech variability: A review", M. Benzeghiba, Speech Communication 49, pp. 763–786, (2007).

[3] L. R. Rabiner, B.H. Juang. Fundamentals of Speech Recognition, Prentice-Hall, Inc.,Upper Saddle River, NJ. 1993.

[4] B. Yegnanarayana and P. Satyanarayana Murthy, "Source-system windowing for speech analysis and synthesis", IEEE transactions on speech and audio processing, vol.4, No.2, March 1996.

[5] Li Deng and Jianwu Dang, "Speech analysis: The production perception perspective".

[6] Vibha Tiwari, "MFCC and its application in speaker recognition", International Journal on Emerging Technologies, Vol. 1, No. 1, pp. 19-22 (2010).

[7] Urmila Shrawankar, Dr. Vilas Thakare, "Techniques For Feature Extraction In Speech Recognition System: A Comparative Study",arXiv preprint arXiv:1305.1145(2013).

[8] Hermansky H., Morgan N., Bayya A. & Kohn P.: RASTA-PLP Speech Analysis. Technical Report (TR-91-069), International Computer Science Institute, Berkeley, CA., 1991

[9] P. P. Shrishrimal, R. R. Deshmukh, V. B. Waghmare, "Development of Isolated Words Speech Database of Marathi words for Agriculture purpose", Asian Journal of Computer Science and Information Technology (AJCSIT), vol. 2, No. 7, pp. 217-218 (July 2012)

[10] Thi Duyen Ngo, The Duy Bui, "A Study on Prosody of Vietnamese Emotional Speech", Fourth International Conference on Knowledge and Systems Engineering, 978-0-7695-4760-2/12 © 2012 IEEE, 2012.

[11] Fujisaki, H., Information, Prosody and Modeling – with Emphasis on Tonal Features of Speech. In: Proc. Speech Prosody, pp. 1-4, 2004.

[12] www.litnotes.co.uk/prosodicspeech.htm.

[13] Bratt, H., "Algemy, a tool for prosodic feature analysis and extraction," Personal communication, 2006.

[14] Delattre P. et al., "An experimental study of the acoustical determinants of vowel colour",word, vol.8,pp. 195-210, 1952.

[15] Nikša Jakovljevi´c, Marko Janev, Darko Pekar, and Dragiša Miškovi´c N. Jakovljevi´c,"Energy Normalization in Automatic Speech Recognition" et al.

[16] Qiyue Liu, Mingqiu Yao, Han Xu, Fang Wang ,"Research on Different Feature Parameters in Speaker Recognition",106-110,Journal of Signal and Information Processing, 2013.