Int J Speech Technol (2013) 16:19–31
DOI 10.1007/s10772-012-9147-3

Effect of aging on speech features and phoneme recognition: a study on Bengali voicing vowels

Biswajit Das · Sandipan Mandal · Pabitra Mitra · Anupam Basu

Received: 28 November 2011 / Accepted: 20 May 2012 / Published online: 14 June 2012

Abstract The article studies age related variations of speech characteristics of two age groups, in the Bengali language. The study considers 60 speakers in the each age groups, 60–80 years and 20–40 years, respectively. We have considered different voice source features like fundamental frequency, formant frequencies, *jitter*, *shimmer* and *harmonic to noise ratio*. Cepstral domain feature, Mel Frequency Cepstral coefficients (MFCC) of different voiced Bengali vowels are also analyzed for younger and older adult groups. MFCC feature and Hidden Markov model parameter of different voiced vowels are used to study phoneme dissimilarities measure between two age groups. Age related changes in elderly speech affect the automatic speech recognition performance as was observed in our study, raising the need for specific acoustic models for elderly persons.

Keywords Aging voice · Speech characteristics · Speech recognition · Phoneme similarity measure

1 Introduction

A large section of the population of our world is aged. Automatic Speech recognition (ASR) is an important application for the aged population. ASR based health care and smart home technology are of great help for the aged. It has been observed in different studies that development of robust ASR system for aged population is a difficult task due to many reasons. In general, speech characteristics vary from speaker to speaker. Variation of speech parameters are even more pronounced between young and older adults. Aging causes several physiological changes to take place in human body. Human articulatory system is also affected with aging. Gradual deformation with aging has been observed in different parts of articulatory system. These physiological changes affect different speech parameters. Fundamental frequency (F_0), formant frequencies (F_1, F_2, F_3, ...), *jitter*, *shimmer*, *voice onset time*, *harmonic-to-noise ratio* (HNR) are some of the speech properties affected most with aging. ASR performance decline due to variability (Benzeghiba et al. 2007) with aging. Vocal tract length (VTL) gets modified with aging (Linville and Rens 2001). It differs with gender too. Besides VTL, speaker's physical and mental condition also introduce variability in speech. The goal of this article is to systematically and empirically study the effect of aging on speech parameters of Bengali vowels. As a case study the effect on voicing vowels in the Bengali language, is considered.

Bengali language is mostly spoken in the south eastern Asia. There are nearly 230 million people speak in Bengali in the world, fifth largest among the languages of the world. Effect of aging on voiced vowels has been studied previously in English language. In this paper, we will analyze the effect of aging on voiced Bengali vowels with a corpus of aging Bengali speech.

2 Physiological effects of aging on speech

Natural changes take place in human body with aging. Human articulatory system (velum, pharynx, larynx, vocal

B. Das (✉) · S. Mandal · P. Mitra · A. Basu
Department of Computer Science and Engineering, Indian Institute of Technology, Kharagpur 721302, West Bengal, India
e-mail: biswajit.net@gmail.com

S. Mandal
e-mail: mandal.sandipan@gmail.com

P. Mitra
e-mail: pabitra@gmail.com

A. Basu
e-mail: anupambas@gmail.com

folds, lungs etc.) is also affected with aging. Several deformation of vocal tract and related organs take place with aging (Lindblom 1971). Size of the vocal cavity changes with teeth loss. Changes of acoustic properties and voice quality degradation of dental phonemes with aging leads to poor ASR performance. While vowel pronunciation, an elderly person face problem of producing desired shape of the vocal cavity. Formant frequencies are affected due to this.

Muscle strength of tongue decline with aging (Rother et al. 2002). Utterance of several vowels depends on tongue hump position and height. Forward movement of tongue increase the second formant frequency. It creates a particular shape of vocal tract for a position and height to pronounce a specific vowel. Old population does not move tongue freely due to loss of muscle strength. Elderly people are unable to make the shape of vocal tract for desired phonation. This leads to change in resonance frequencies. If the jaw moves down, it will increase the first formant frequency but aged people can not move jaw freely. It affect the first formant frequency (Harrington et al. 2010) and these changes finally degrade phoneme recognition.

Several changes also take place in pharynx cavity and larynx tube. Larynx shrinks towards lungs resulting increase vocal tract (Xue and Hao 2003). Stiffness of cartilages increases with aging. Rigid cartilages (Paulsen and Tillmann 1998) create problem to easy movement of vocal folds. Aging affects two main aspect of vocal fold (Rodeño et al. 1993) anatomy and function. The muscle of vocal folds loses mass also the flexible tissues which are responsible for vocal fold vibration at the time of voicing become thinner, stiffer and less pliable with aging. Area of larynx tube decreases with aging. It increase the voice harshness for older people. Instability in vocal folds vibration increase the jitter (Wilcox and Horii 1980). Improper closure of vocal folds causes turbulent airflow during phonation. Airflow introduces additive noise (aperiodic signal) to the speech signal which decreases HNR (Yumoto et al. 1984; Krom 1993; Hillenbrand et al. 1994). Strength of diaphragm muscles also reduce with aging. Volume of outgoing air from lungs decreases due to loss of muscle strength of diaphragm (Tolep et al. 1995).

Changes in articulatory organs vary from person to person with aging. Speech quality of some subjects are preserved at even at their 80s, whereas others begin to sound old at their 50s. Degradation of voice quality depends on many factors, like food habits, smoking, alcoholism (Gorham-Rowan and Laures-Gore 2006; Linville 2001). Another important factor for voice quality degradation is hereditary traits of the family. It has been observed that are even other deformation take place with aging such as sensory feedback reduction, speed and accuracy of motor control degradation. Speaking and reading rate also reduces with aging; due to cognitive decline (Ulatowska 1985).

It has been observed in literature that voice source parameters changes (Vipperla et al. 2010; Barlow 2009; Ramig and Ringel 1983) with aging. Some of these parameter increases in value and some of them decrease depending upon gender. Vocal tract lengthening with aging alters vowel articulation (Xue and Hao 2003). Physical changes impact on speech perception and cognition linguistics (Liss et al. 1990). Aged population shows slow speaking rate (Markus and Walter 2003). Vowel duration decreases with slow to fast speech (Hisao 1997).

In the next section, we discuss some of the phonetic characteristics of Bengali speech, before studying the effect of aging on them.

3 Phonetic characterization of Bengali speech

Rarh, Banga, Kamarupa and Varendra are four type of well known dialects in Bengali. Each dialect has unique pronunciation style, stress and intonation. ASR system with speech samples of one dialect does not perform good recognition with another dialect. We have considered the standard colloquial Bengali language of south-western region in our study. In Table 1, we have showed example of different dialect of Bengali language.

We found that forty seven phonemes can be used to represent all possible utterances across Rarh Bengali. There are six vowels (in Bengali "অ, আ, ই, এ, ও, উ" and their phonetic representation /a/, /A/, /i/, /e/, /o/, /u/) , six nasal vowels (Bengali letters "ঁঅ, ঁআ, ঁই, ঁএ, ঁও, ঁউ" and their phonetic symbols are /ˆa/, /ˆA/, /ˆe/, /ˆi/, /ˆo/, /ˆu/), three diphthongs (phonetic representation /oi/, /ou/ and /E/ and in Bengali "ঐ, ঔ, অ্যা"), three semivowels and twenty nine consonants in the phoneme list. The Bengali phonetic properties (Chatterji 1921; Tanmay 2000) are described as place of excitation, manner of excitation and type of excitation, and on the basis of the articulatory system phonemes are designated. In the English phonetic structure, vowels are categorized as long and short in duration. In Bengali, there is no such classification. Linguistic differences of vowels and consonants between Bengali and English are described in the study of Barman (2011). Manner of vowel pronunciation are different with respect to accent, dialect and intonation.

Table 1 Example of different dialect of Bengali

Dialect	Example
Standard colloquial dialect	একজনের দুই ছেলে ছিল
Rarh	একলোকের দুটা বেটা ছিল
Banga	একলোকের দুইটা পোলা আসিল
Kamrupi	একমানুষের দুই ছাওয়া আসিল
Barendri	একজনা মানষির দুই কোনা বেটা আছিল

Phonemes can be divided as consonant, vowel and semi-vowel. Phone classification depends on the following questions. Duration of the phones are short or long, position of articulation, voiced or unvoiced, open or close and front or back. All vowels are voiced in characteristics because these are generated with the vibration of vocal folds. Different manner of articulation like nasal, fricative, flap, affricative or stop. Point of articulation also differentiate the phonemes like labial, dental, bilabial, velar, alveolar or alveolar-dental.

4 Speech features considered in our study

Speech signal can be represented by various parameters of interest. In this paper, fundamental frequency (F_0), formant frequencies (F_1, F_2, F_3), jitter, shimmer and harmonicity has been considered to analyze the aging effect on Bengali vowels. The pattern of changes of these features are different for male and female.

Fundamental frequency is the measure of number of excitation in a time period. It varies with the type of voiced phoneme utterance pronunciation. It has been observed in different studies that F_0 is higher for children and female subjects as compared to male subjects (Traunmuller 1984). Jitter is the time variation of periodic signal and it is cycle to cycle variation in pitch period. Shimmer is the amplitude variation from period to period in speech signal. Harmonic to noise ratio is the another index of voice quality. In Table 2, trends of variation of speech source features (Barlow 2009; Vipperla et al. 2010; Reubold et al. 2010; Baken 2005) with aging are described.

Mel Frequency Cepstral Coefficient (MFCC) features are most commonly been used for automatic speech recognition. MFCC features are naturally related to our auditory system. Human auditory system resolves frequencies non-linearly. Speech is a non-stationary signal. Stationary may be assumed for small segment of speech signal. Speech signal is segmented into 25 msec frame with Hamming window and frame shift is chosen 15 msec. We converts each frame to frequency domain with fast Fourier transform. After extracting spectral features from speech signal, frequencies are converted to Mel frequencies. 26 triangular band-pass filter are then placed in mel-frequency scale within a band. MFCC features are computed with discrete cosine transform (DCT) using the filter output amplitude. Each frame is parametrized to their feature vector. It consist of base MFCC features plus their first and second order time derivatives. It has the benefit of being capable of capturing the phonetically important characteristics of speech. We have extracted 39 dimensional MFCC features from each frame of speech signals of our corpus.

Table 2 Effect of aging on different speech parameter

F_0	There is some disagreement about changes of F_0 with aging. Some study states F_0 increases for male but decreases for female (Benjamin 1981; Hollien and Shipp 1972; Ramig 2001) whereas other studies (Deliynski and Xue 2001; Endres et al. 1971) shows decrease in F_0 for male and female
Formant frequencies	In average it decreases for male and female
Jitter (Local)	Increases for both male and female
Shimmer (Local)	Increases for both male and female
HNR (dB)	Decreases for both male and female

4.1 ASR features

We have used monophone and triphone HMM acoustic model parameters to study of aging effect on Bengali phoneme recognition. To get the model parameters, we have estimated phoneme and triphone model parameters by expectation maximization (EM) algorithm. HMM of each phone is constructed with 5 states and each state is represented with single Gaussian component. State 2, state 3 and state 4 of HMM model of a phone are considered for studying the statistical differences of voice parameters between young and elderly. First and last states are non-emitting states, these are used only to concatenate two phone HMM model.

The mixture model approximates the data distribution by fitting k component density functions $f_n, n = 1, 2, \ldots, k$ to a data set D having m patterns and d features. Let $x \in D$ be a pattern, the mixture model probability density function evaluated at x is:

$$p(x) = \sum_{n=1}^{k} w_n f_n(x|\varnothing) \tag{1}$$

The weights w_n represent the fraction of data points belonging to model n, and they sum to one. The functions $f_n(x|\varnothing), n = 1, \ldots, k$ are the component density functions modeling the points of the nth cluster. For continuous data, Gaussian distribution is the most common choice for component density function. This is motivated by a result from density estimation theory stating that any distribution can be effectively approximated by a mixture of Gaussians (Scott 1992). The multivariate Gaussian with d-dimensional mean vector μ_n and $d \times d$ co-variance matrix Σ_n is:

$$f_n(x|\mu_n, \Sigma_n) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(x-\mu_n)^T|\Sigma_n|^{-1} \times (x-\mu_n)\right\} \tag{2}$$

The quality of model parameters $\varnothing = (w_n, \mu_n, \Sigma_n), n = 1, \ldots, k$ determined by how well the corresponding mixture

model fits the data. This is quantified by the log-likelihood of the data, given the mixture model:

$$L(\varnothing) = \sum_{x \in D} \left(\sum_{n=1}^{k} w_n f_n(x \mid \mu_n, \Sigma_n) \right) \tag{3}$$

Expectation maximization (EM) algorithm iteratively update $\varnothing$ such that $L(\varnothing)$ is non-decreasing. In EM algorithm, D dataset with m patterns and d continuous feature are provided to calculate model parameters. Convergence ratio, $\epsilon < 0$ is set for stopping the iteration of EM algorithm. If $\varnothing_j$ is the mixture parameters at iteration j, $\varnothing_{j+1}$ at iteration $j+1$ can be computed with following steps.

Membership probability of x which belongs to detaset D is computed for each mixture component $n = 1, \ldots, k$ as follows:

$$w_n^j(x) = \frac{w_n^j f_n(x|\mu_n^j, \Sigma_n^j)}{\sum_i w_i^j f_i(x|\mu_i^j, \Sigma_i^j)} \tag{4}$$

In the next step, mixture model parameters are updated at iteration $j+1$ as:

$$w_n^{j+1}(x) = \sum_{x \in D} w_n^j(x) \tag{5}$$

$$\mu_n^{j+1}(x) = \frac{\sum_{x \in D} w_n^j(x) x}{\sum_{x \in D} w_n^j(x)} \tag{6}$$

$$\Sigma_n^{j+1} = \frac{\sum_{x \in D} w_n^j(x)(x - \mu_n^j)(x - \mu_n^j)^T}{\sum_{x \in D} w_n^j(x)} \tag{7}$$

Iteration will stop when $|L(\varnothing^{j+1}) - L(\varnothing^j)| \leq \in$ condition is satisfied.

5 Bengali speech corpus of elderly speakers

To the best of our knowledge, there is no speech corpus of aged people in Bengali. A goal of our research was to create such a corpus. We have collected speech signal from 60 speakers. There are 40 male speakers and 20 female speakers in this Bengali speech corpus. Though there are so many discrepancies about the selection of starting age to be called elderly, we have selected range of age of speakers between 60 to 80 years. Age distribution of older people are shown in Fig. 1. Details of speaker's education level, profession and signal quality are provided in Table 3. Speakers are mentally and physically fit. Some speakers have low vision. More than 50 % speakers don't have their teeth. No one of them has case history of neurological disorder or suffering from lungs problem.

Speech data collection from elderly people is a challenging task. Older people are unable to record speech for long sessions. Most of the older people do not feel motivated to record their voice. As voice quality changes in different session of a single speaker due to different mental and physical states, speech data has been recorded in two sessions. All the recording has been done at room environment. Mother tongue of all speakers is Bengali.

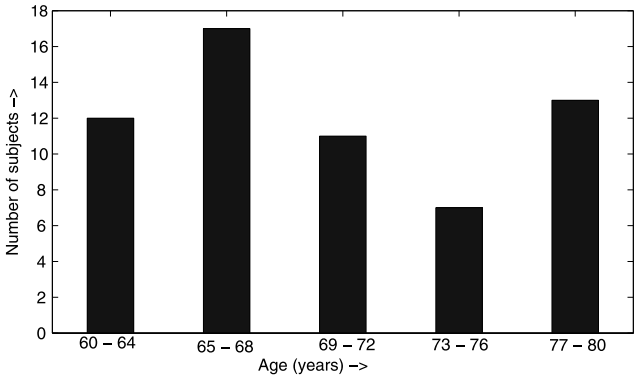

Fig. 1 Distribution of age of speakers in the corpus

Table 3 Speaker's details of the aged speech corpus

	Corpus details
Speaker's education level	Secondary—27.16 %, HS—48.14 %, Graduation—22.22 % and >Secondary—2.4
Speaker's profession	Business—12.5 %, House wife—15.85 %, Service—69.51 %, and Teacher—2.43 %
Average signal-to-noise ratio	84.11 dB

We have selected seven thousand five hundred text sentences for recording. These are sourced from the Bengali News paper Anandabazar patrika and Bengali literature. Sentences are selected with optimal text selection procedure, where a process of balanced phoneme and triphone selection in the text corpus is adopted. Each sentence is recorded with sample frequency 16000 Hz in mono channel. Speech signals are encoded with 16 bit Pulse Code Modulation. Sony FV-220 microphone and Emu speech toolkit (Cassidy and Harrington 2001) has been used for speech recording. We have maintained 15 cm distance from mouth to microphone for each speakers at the time of voice recording. Each speaker has recorded 200 sentences. There are 19500 words in the phonetic dictionary. The Bengali phonetic dictionary has been created with grapheme to phoneme conversion procedure on the words. Finally, the dictionary is corrected manually.

We have created another speech corpus of younger adults using the same set of sentences. There are 40 male speak-

ers and 20 female speakers in this corpus. Speech recording configurations are same as stated above. Young speakers are selected with age between 20 to 40 years. This corpus has been used for comparative purpose.

We have applied perceptual evaluation speech quality (PESQ) method to the current study. Determining the subjective speech quality of a speech signal has always been an expensive and laborious process. PESQ is an objective measurement method that predicts the results of subjective listening tests on speech signals. Speech signals of young people are considered as a reference signal. Reference signal and speech sample having same context from aged population are provided to compute mean opinion score (MOS) and predicted signal distortion (PSD). Average MOS score for objective evaluation with respect to reference signal is 2.86 out of 5. Average PSD with respect to reference speech signal is 2.14 out of 5. It can be concluded from MOS and PSD analysis that perceptual quality degrades with aging.

6 Methodology for studying variation of speech parameter with aging

We have extracted basic source characteristics of speech signal of different voiced Bengali vowels with the Praat speech toolkit (Boersma and Weenink 2011). Characteristics of six vowels and three diphthongs are analyzed in this study. Source features of six vowels are computed from consonant-vowel-consonant (CVC) segment of a word, which is the part of a continuous speech sample. F_0, formant frequencies (F_1, F_2, F_3), jitter, shimmer and harmonicity are computed from labeled speech corpus. Basic speech source characteristics are calculated for both young and older adults.

After extracting formant frequencies of each vowels, the mean and standard deviation was calculated for F_1, F_2, and F_3 values. We have selected approximately 20 frequency points for each formant from each vowel duration. If there are less than 20 frequency components in F_1, F_2, and F_3, frequency components are interpolated to make it 20 components. Mean fundamental frequency of each vowels has also been computed to check age related variation among speakers.

All these experiment has been performed for both young and old subjects. The divergence of the distribution of speech features mentioned above for the old and young population are studied using statistical distance measurements. We have measured phoneme similarity of two age groups. Probability distribution of different phoneme over the population has been incorporated in distance calculation. Hidden Markov Model Toolkit (HTK) (Young et al. 2000) was used for monophone and triphone modeling.

In our study, we have computed statistical distance of phone parameter distribution of each voiced vowels taken from two different user groups. We have used the following statistical measures for distance computation.

– Kullback-Leibler divergence (Ghosh et al. 1987): It measures the divergence between two probability distribution X and Y. It is a non symmetric measure. It can be defined as:

$$Dis_{KL}(X||Y) = \sum_i X(i) \log \frac{X(i)}{Y(i)} \tag{8}$$

In case of Gaussian distribution, Kullback-Leibler divergence can be expressed as

$$Dis_{KL} = \frac{1}{2}(\mu_2 - \mu_1)^T \left[\Sigma_2^{-1} + \Sigma_1^{-1}\right](\mu_2 - \mu_1) + \frac{1}{2}\mathrm{tr}\left(\Sigma_1^{-1}\Sigma_2 + \Sigma_2^{-1}\Sigma_1 - 2I\right), \tag{9}$$

where μ_1 and μ_2 are mean vector of first and second Gaussian distribution respectively whereas Σ_1 and Σ_2 are co-variance matrix of two different distribution. I is an identity matrix.

7 Results and discussion: feature variability with aging

We have divided the experiment in two parts. In the first part, basic source features are examined. In the next part, cepstral features of different phoneme are observed for two user groups namely young and elderly.

7.1 Voice source feature

F_0 is affected significantly with aging. Mean F_0 of male and female were computed of each age groups. Mean F_0 of young men is 158 Hz whereas mean F_0 of older men is 150 Hz. F_0 decreases with aging for women. Mean F_0 of young women is 258 Hz whereas mean F_0 of older women is 200 Hz. F_0 is less sensitive for male with aging. Physiological effect on F_0 is significant ($p < 0.05$ with Manwitney-Wilcox test) for male as well as female speakers. Analysis of changes of F_0 correlates to the studies (Vipperla et al. 2010; Reubold et al. 2010; Barlow 2009) which are conducted in English language. There are mixed interpretation about changes of F_0 for male and female subjects with aging. Female subjects are more sensitive on F_0. F_0 decreases in average for females with aging. One of the reason of sharp drop of F_0 is hormonal chanes due to menopause described in the study (Linville 1996). In Table 4, analysis of F_0 across two age groups has been shown. Effect of aging on formant frequencies of different vowels are shown in Figs. 2–13.

In Tables 4 and 5, mean and standard deviation of F_1, F_2, and F_3 are provided. Formant frequency (F_1) decreases to the tune of 30–40 Hz consistently whereas F_2 and F_3 changes up to 50–150 Hz for aged males. However such

Table 4 Formant frequencies of vowel A, a, e, i, o and u of male population

	F_1		F_2		F_3	
	Mean	STD	Mean	STD	Mean	STD
(a) Vowel A						
Young_male	682.03	106.12	1343.41	208.98	2526.89	327.1
Old_male	628.52	120.65	1390.89	281.88	2644.75	409.14
(b) Vowel a						
Young_male	561.55	107.37	1144.87	338.76	2560.89	343.33
Old_male	508.32	87.44	1073.07	83.8	2609.17	322.12
(c) Vowel e						
Young_male	417.52	64.1	1991.32	225.13	2649.54	219.87
Old_male	404.97	60.72	1843.85	215.54	2632.89	369.61
(d) Vowel i						
Young_male	310.18	43.44	2171.2	251.69	2743.03	276.91
Old_male	300	38.08	2049.42	237.34	2664.93	350.95
(e) Vowel o						
Young_male	450.34	117.92	1167.55	402.16	2689.02	314.57
Old_male	414.85	74.95	1061.68	273.8	2662.32	340.84
(f) Vowel u						
Young_male	395.42	149.3	1370.36	612.93	2722.13	470.55
Old_male	341.37	96.34	1249.59	381.9	2596.25	375.9

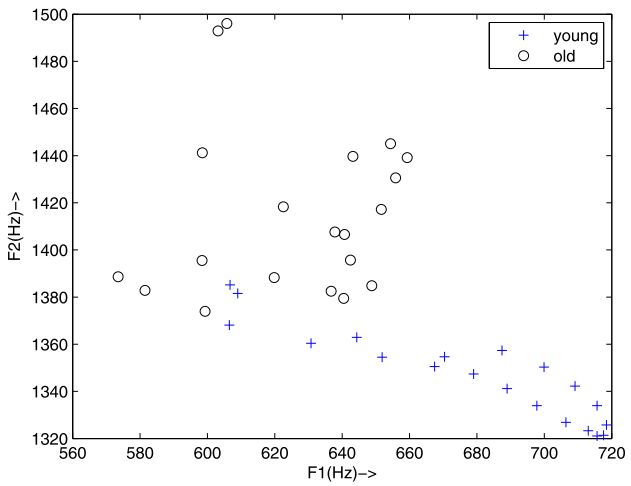

Fig. 2 Distribution of F_1 and F_2 of vowel "A" for young and old male population

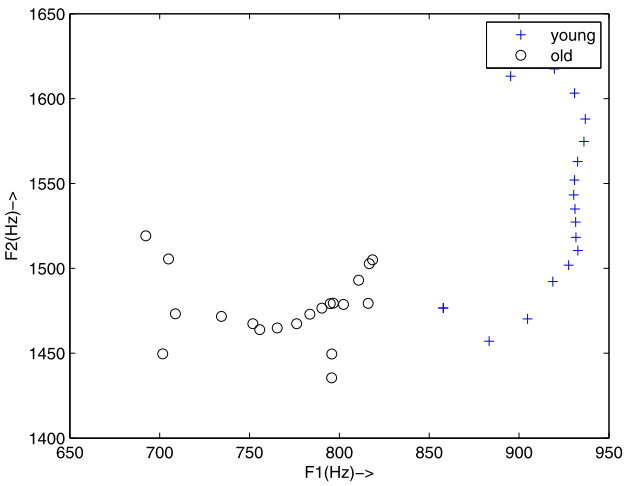

Fig. 3 Distribution of F_1 and F_2 of vowel "A" for young and old female population

consistent trends observed for female subjects also. From the figures, we can conclude that particular vowel pronunciation from young and aged subjects can be discriminated with F_1–F_2 plot. One may be confused to differentiate only for vowels /o/ and /u/ of female young and aged subjects because F_1–F_2 plot of /o/ and /u/ for young and aged speakers displays overlapped nature with each other. We have not shown the analysis of F_0, F_1, F_2 and F_3 of nasal vowels in this study but these vowels are also affected similarly as normal vowels.

We have used Mann-Whitney-Wilcoxon (MWW) (Mann and Whitney 1947) non parametric unpaired test for measuring statistical significance of the increase or decrease of mean frequency values with aging. We have considered degree of freedom 19 for this statistical significance test. Twenty dimensional vectors of each formant frequency

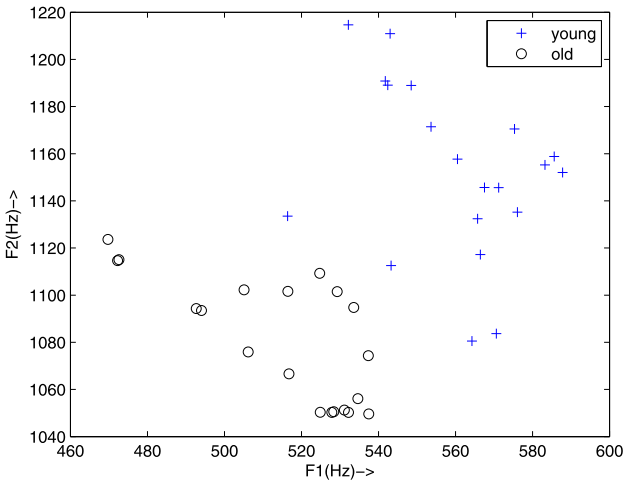

Fig. 4 Distribution of F_1 and F_2 of vowel "a" for young and old male population

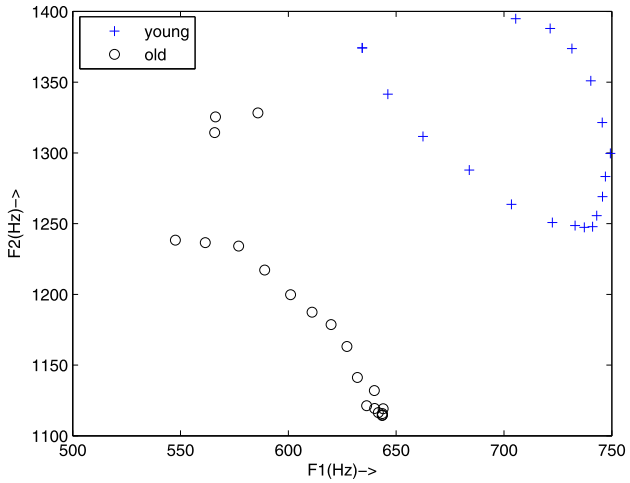

Fig. 5 Distribution of F_1 and F_2 of vowel "a" for young and old female population

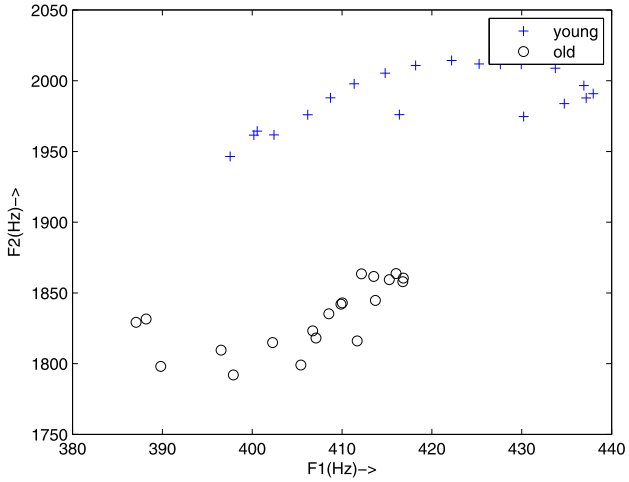

Fig. 6 Distribution of F_1 and F_2 of vowel "e" for young and old male population

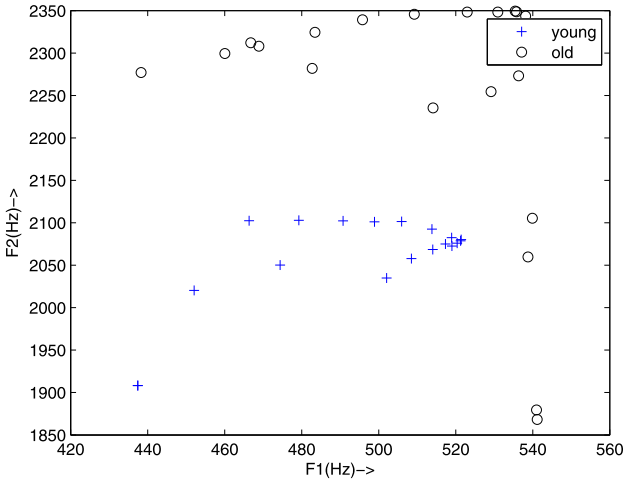

Fig. 7 Distribution of F_1 and F_2 of vowel "e" for young and old female population

F_1, F_2 and F_3 from two age groups are taken as input in MWW test to calculate P-value. All the P-value has been calculated as 1-tailed. Formant frequency F_1 of Vowel /A/, /a/, /o/, /u/ and /o/ are statistically significant (P-value < 0.001 (1-tailed)) for male population of young and old but F_1 of vowel /e/ is insignificant at $P < 0.001$. Differences F_2 and F_3 are statistically significant for all vowels at $P < 0.001$ but F_3 of /A/, /i/ and F_2 of /u/ are insignificant at $P < 0.001$. Changes of formant frequencies with aging for male speakers are shown in Table 2.

Changes in formant frequencies of /A/, /a/, /e/, /u/ and /i/ are statistically significant for female speakers at $P < 0.01$. We have observed that formant F_1 of /o/ is insignificant for women speakers at $P < 0.01$ but F_2 and F_3 are significant at same P value. Deviation of formant frequencies for female speakers with aging has been shown in Table 5.

Jitter and *shimmer* increases with aging as shown in Table 6. In the study (Ferrand 2002), deviation of F_0, jitter and HNR Mean are extensively studied of different age groups. Jitter of young men is 1.32 % but for older men is 1.48 %. Mean jitter of young women is 1.18 % but for older women is 0.98 %. Shimmer of young men and women is 7.42 % and 5.58 % respectively. Mean shimmer of older men and women is 8.47 % and 5.06 %. This analysis for Bengali correlates with that study. Changes of jitter and shimmer for male speakers are however statistically insignificant ($P < 0.05$), changes in all the parameters (in Table 6) are statistically significant ($P < 0.05$). As unwanted noise signal increases in the speech signal with aging due to incomplete closure of vocal folds, *harmonic to noise ratio* decreases for older people. *Harmonic to noise ratio* of young men and women are 13.58 dB and 18.23 dB respectively. *Harmonic*

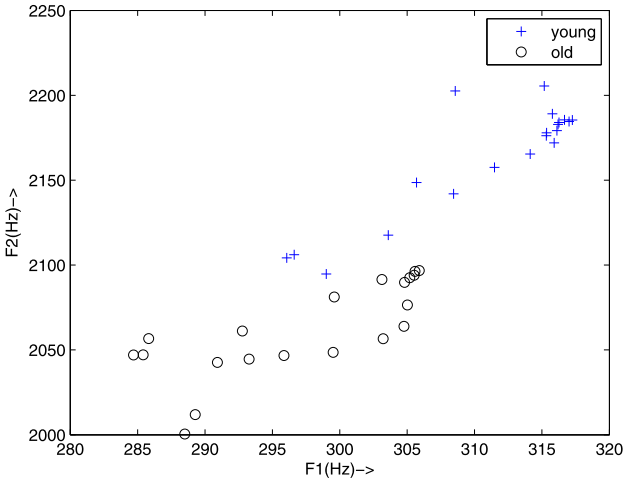

Fig. 8 Distribution of F_1 and F_2 of vowel "i" for young and old male population

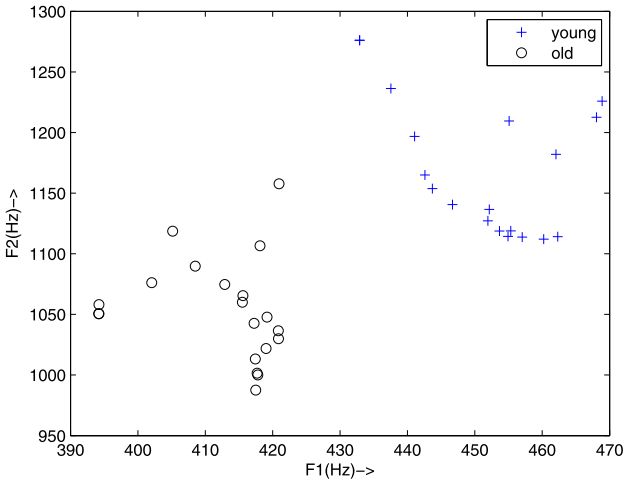

Fig. 10 Distribution of F_1 and F_2 of vowel "o" for young and old male population

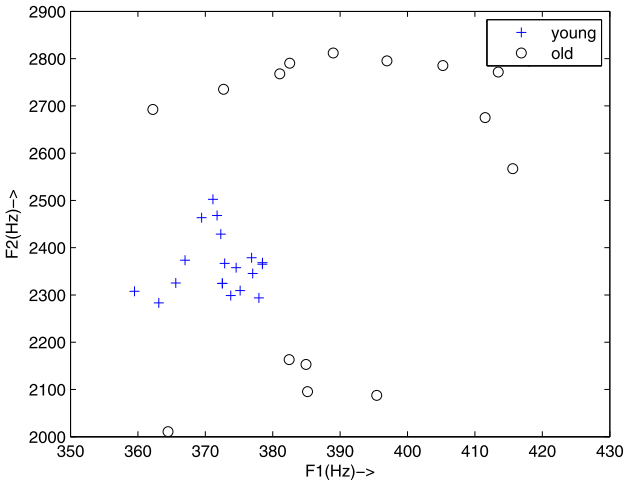

Fig. 9 Distribution of F_1 and F_2 of vowel "i" for young and old female population

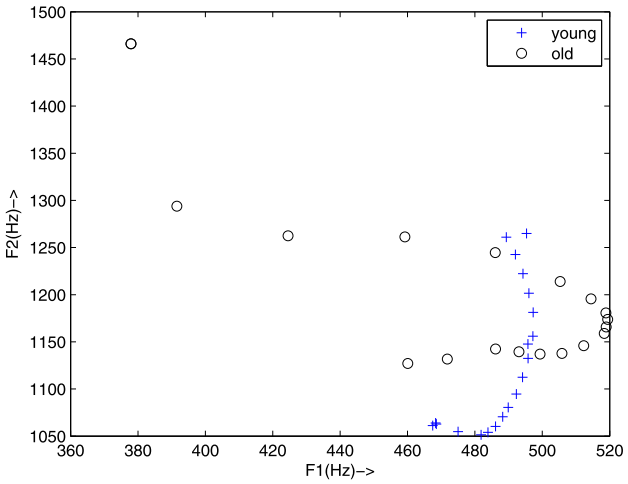

Fig. 11 Distribution of F_1 and F_2 of vowel "o" for young and old female population

to noise ratio of older men and women are 12.34 dB and 15.31 dB. In Table 6, mean and standard deviation over all vowels are tabulated.

7.2 HMM parameters of vowels

We have compared HMM model parameters of each phone obtained by training on the young and old population respectively. Mean and variance vector of the Gaussian for each state has been used for phone divergence computation between young and older subjects. We have measured the phone model divergence for both context independent and context dependent phoneme. Monophone HMM model parameters (mean and variance) are used for context independent phoneme utterances. Triphone HMM model parameters has also been used for context dependent phoneme to alleviate co-articulation effect. Note that, phoneme are influenced by it's neighbors at the time of phonation. Triphone model parameters of each vowels are used to measure phone model divergence between two age groups. As we have used five states HMM model, state specific divergence has been computed for each phone between two user groups. In the five state HMM model first and terminal states are starting and emitting state respectively. These two states are used for concatenation of two phone HMM model. In this study, we have used mixture components of Gaussian distribution of states 2, 3 and 4 of each vowel for distance measurement between young and elderly. We have measured divergence of voiced vowels by well known method like Kullback-Leibler technique.

It is clear from Table 7 that vowels /A/, /a/, /e/, /E/ and /i/ displayed more mismatch than others among all the states.

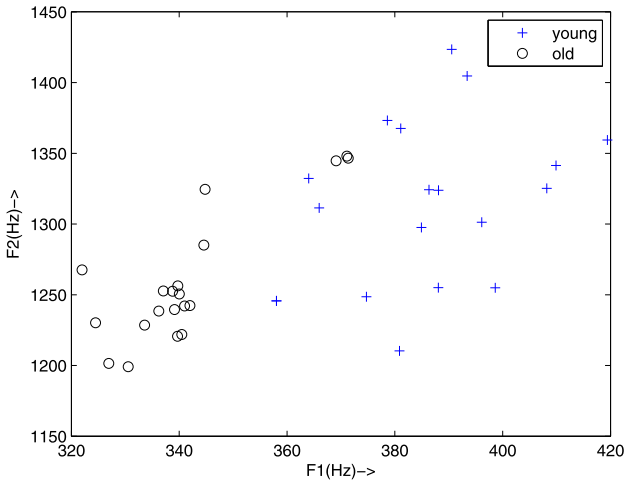

Fig. 12 Distribution of F_1 and F_2 of vowel "u" for young and old male population

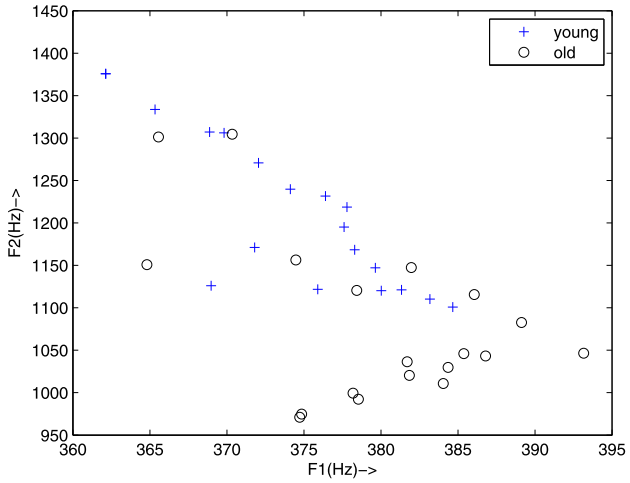

Fig. 13 Distribution of F_1 and F_2 of vowel "u" for young and old female population

Divergence of Gaussian distribution of diphthongs like /oi/ and /ou/ are higher. As we have discussed in Sect. 2 that muscle strength of tongue decreases with aging. This affects central and front vowels i.e. /A/, /e/ and /i/.

Diphthong pronunciation are more prone to be diverse between young and older subjects. From the distance measurement result, it can be concluded that divergence varies for individual states for each phone. All the distance measuring method displays most mismatch for phoneme, /A/, /e/, /i/, /E/, /ou/ and /oi/ in state 3.

We have used specific triphone to observe the effect of aging on context dependent vowels and diphthong utterances. Triphones are represented as root phone at the middle and it's neighbor at his left and right side. In this study, /k-A+l/, /k-a+T/, /ph-E+m/, /k-e+ch/, /k-o+ch/, /ch-u+p/, /m-oi+n/ and /b-ou+m/ are chosen as context dependent phoneme for /A/, /a/, /E/, /e/, /o/, /u/, /oi/ and /ou/ respectively. In Table 8, statistical distance of different vowels and diphthongs of young and aged population are provided. From these result, vowel /A/, /u/, /E/ and /oi/ display more mismatch. We have also found that nasal vowels are less affected with aging. Statistical distance of nasal vowels are very less.

We have done Mann-Whitney-Wilcoxon non parametric test over mean and variance vector of each vowels. Thirty nine dimensional mean and variance vector has been used as input to the probability significance test. Differences of mean vectors of context independent and dependent vowels between two age groups are statistically significant at $P(1\text{-tailed}) < 0.5$ with degree of freedom 38.

8 Results and discussion: ASR performance

We have created another acoustic model with mixed (both young and old subjects) training data besides the acoustic model of elderly and young subjects. We have tried to observe vowel recognition performance degradation or improvement in all possible ways. ASR performance has been tested with three types of test data. Test sets encompass 10 younger (20–40 years) and 10 older (60–80) people. Another test set has 10 people (more than 75 years of age). There are no speech sample of these speakers in the training data set. We have also selected 20 sentences which are not present in training text corpus.

We have evaluated the ASR system, trained with young, elderly and mixed training set. Test data of young and aged subjects are tested with ASR system trained on young population. Test data set of aged people has been tested with the acoustic model of aged and mixed subjects. In this work, we have used trigram language models with acoustic model trained with young and old data to achieve good ASR performance. To compare the ASR performance of two age groups, we have created five state triphone HMM model. Each state is represented with eight Gaussian mixture component. Phoneme recognition has been obtained with phone loop decoder. It is clear from the phoneme recognition accuracy, some phoneme are affected strongly.

In Tables 9 and 10, phoneme recognition accuracy of first two columns are obtained from the young speakers acoustic model which is evaluated with test data set from young and older population. Older population test data gives poor performance in this case. It has been observed from the result that /A/, /a/, /e/ and /i/ are affected more. Recognition performance of nasal vowels are also affected for aged. Nasal vowels are confused with their base vowels e.g. $\tilde{A} \rightarrowtail A$, and $\tilde{e} \rightarrowtail e$ etc. In the third column, data model trained on training data of aged population is also tested on test data of aged. Accuracy improves in this case. In the fifth column, we

Table 5 Formants of vowel A, a, e, i, o and u of female population

	F_1		F_2		F_3	
	Mean	STD	Mean	STD	Mean	STD
(a) Vowel A						
Young_female	920.77	145.99	1537.04	189.35	2632.53	488.51
Old_female	779.04	137.67	1481.01	194.32	2808.57	419.68
(b) Vowel a						
Young_female	712.43	108.13	1297.04	227.25	2642.57	435.72
Old_female	618.75	134.17	1131.6	273.14	2859.63	449.37
(c) Vowel e						
Young_female	495.02	80.12	2011.76	375.51	2707.89	408.58
Old_female	528.26	169.07	2240.5	452.55	3051.13	419.57
(d) Vowel i						
Young_female	361.58	62.65	2341.72	403.55	2892.45	348.15
Old_female	393.59	57.11	2605.33	632.18	3233.52	480.23
(e) Vowel o						
Young_female	476.59	69.54	1078.23	227.52	2827.22	392.27
Old_female	481.54	124.18	1175.76	312.95	2971.75	433.65
(f) Vowel u						
Young_female	371.03	86.91	1161.68	379.54	2643.52	417.32
Old_female	370.03	65.92	1035.92	262.85	2933.29	521.66

Table 6 Speech parameter values: Means (standard deviation) over populations

	Young male	Older male	Young female	Older female
F_0	158.96 (34.11)	150.19 (37.38)	258 (42.27)	200 (18.8)
Jitter (local)	1.32 (1.06)	1.48 (1.07)	0.98 (0.43)	1.18 (0.9)
Shimmer (local)	7.42 (4.87)	8.47 (4.72)	5.06 (1.69)	5.58 (1.95)
HNR (dB)	13.58 (3.11)	12.34 (3.99)	18.23 (2.68)	15.31 (3.55)

Table 7 Kullback-Leibler distance between models of HMM sates for the young and old population: Gaussians for states 2, 3 and 4

	State1	State2	State3
A	9.607	17.213	6.543
a	8.240	14.786	9.477
e	10.412	16.160	7.119
i	10.776	14.782	6.376
o	8.220	12.847	5.972
u	7.899	11.398	7.257
E	7.276	19.67	11.802
oi	7.550	17.44	14.33
ou	17.647	18.403	11.273

Table 8 Kullback-Leibler distance between triphone models of HMM sates for the young and old population: Gaussians for states 2, 3 and 4

	State1	State2	State3
A	18.92	19.63	28.42
a	19	15.04	23.27
i	13.17	21.69	11.5
e	16.85	16.23	9.16
u	18.96	28.26	10.69
o	20.48	19.67	25.64
E	14.03	21.22	31.97
oi	19	18.27	14.71
ou	12.37	14.69	9.59

have shown vowel recognition performance of test data from more than 75 years of age speakers with acoustic model of aged population. Labial, Retroflex and dental phoneme are affected more with respect to recognition accuracy.

In Table 11, we have provided overall phoneme accuracy of different acoustic models and test sample. Phoneme recognition accuracy is very much poor when test samples of aged population are applied to acoustic model of young

Table 9 Bengali vowel recognition accuracy (% correct) of young acoustic model and young test data (YY), young acoustic model and old test data (YO), old acoustic model and old test data (OO), mixed acoustic model and old test data (MO) and old acoustic model and old test data (>75 yrs) (O_O >75)

	YY	YO	OO	MO	O_O >75
A	97.2	72.5	89.6	82.3	88.4
ˆA	75.3	50.8	73.7	64.5	63.8
a	88.9	69.3	83.1	70.5	74.5
ˆa	82.3	63.6	77.6	69.7	71.5
E	83.3	60.6	78.9	75.7	76.3
ˆE	79.4	45.4	68.7	61.3	62
e	95.9	78.6	90.7	83.8	85.4
ˆe	85.2	45.2	80	50	69.8
i	93.6	77.8	92.9	83.9	84.6
ˆi	75.7	36.7	68.5	61.5	60.8
o	91.7	85.5	88.1	85.8	85.2
ˆo	89.4	50.2	83.3	60.4	75.8
oi	92.2	56.2	86.9	80	79.8
ou	90.4	55	80.5	75	75.4
ˆou	81.3	55.4	70	59.4	62.1
u	90.5	75.6	86.1	78.2	89.3
ˆu	79.3	40	66.7	53.3	59.6

population. Accuracy improves when test samples are applied to the acoustic model of aged population.

In Table 12, deviation of speech parameters of more affected vowels between two age groups are provided. Recognition accuracy deviation of acoustic model of young population tested by test data of young and aged population has been provided in Table 12. We have considered two vowels and three diphthongs whose recognition accuracy is more affected. It has been observed that other features e.g. F_0, formant frequency, jitter, shimmer and harmonic-to-noise ratio are also highly affected in these case. Recognition accuracy degrades due to speech quality loss owing to various reasons. However, we have observed jitter and shimmer features do not contribute to recognition accuracy significantly. Probable reasons for each vowel may be as follows. For the vowel /A/, /a/ and /E/, need more lip opening and jaw movement but aged speakers suffers to do the same. /A/ and /a/ are central and back vowels respectively. Tongue position also plays an important role on F_0 and formant frequencies deviation. It can be conclude from Table 12 that recognition accuracy depends on F_0, formant frequencies and harmonic-to-noise ratio. Differences of these acoustic features of /A/, /a/ and /E/ are the controller of recognition accuracy. Another possible reason of poor recognition of diphthongs, /oi/ and /ou/ are the incapability of instant movement of different articulatory parts. For diphthong pronunciation, speakers has to change the shape of vocal tract in a moment. Aged population suffer to do that. These are the possible reasons for high differences of acoustic features between two age groups.

Table 10 Bengali consonant recognition accuracy (% correct) of young acoustic model and young test data (YY), young acoustic model and old test data (YO), old acoustic model and old test data (OO), mixed acoustic model and old test data (MO)

Consonant	Phoneme	YY	YO	OO	MO
Labial	b	92.3	63.5	76.7	66.9
	m	96.7	67.5	84.1	73.8
	bh	71.4	50	69.3	69.2
	p	87.5	70.5	79.1	72.2
	ph	76.4	60.7	66.7	62.5
Palato-Alveolar	ch	73.3	37.7	61.8	53.7
	chh	95.2	78.6	92.9	84.7
	j	89.6	71.6	87.1	75
	jh	68.3	43.3	55.6	50.3
Retroflex	D	75.6	40	66.7	45
	Dh	79.2	50.1	69.8	63.5
	T	80.6	40.8	73.6	59.2
	Th	66.7	34.8	63.6	60.9
	R	87.5	38.5	65.5	42.6
Dental and Alveolar	d	91.7	50	87.4	73.7
	dh	78.8	43.3	67.9	67.9
	t	94.3	62.4	86.9	83
	th	87.8	60	84.4	62.5
	n	95.9	82.3	95.5	86.7
	r	90.5	74.4	87.1	73.7
	l	90	75.3	83.4	78.4
	s	99.2	86.8	98.9	94.2
Velar	g	85	37.8	84.8	54.8
	gh	60	40	58.3	58.3
	k	97.5	67	88	76.3
	kh	80.4	59.6	78.4	70.5
	Y	85.6	51.2	79.5	80
	ˆn	95.2	69.6	92.3	73.1
Glottal	h	86.4	66.7	84.1	85.4

Table 11 Overall phoneme accuracy by employing different acoustic model and test data of young and aged population

	YY	YO	OO	MO	O_O (>75 years)
Accuracy (%)	92.42	56.29	72.6	64.11	65.90

Table 12 Average difference (Δ) in speech parameters of each vowels between old and young speakers

Vowels	ΔF_0 (Hz)	ΔF_1 (Hz)	ΔF_2 (Hz)	ΔF_3 (Hz)	ΔJitter (%)	ΔShimmer (%)	ΔHNR (dB)	ΔRecognition (%)
A	23.5	97	50.15	147	0.81	3.73	4.9	24.7
a	31.4	73.5	118.5	133	0.10	3.04	3.02	19.6
E	21.3	80.7	70.5	146.8	0.3	4.1	3.3	22.7
oi	33.2	87	120	139	0.35	3.9	4.1	36
ou	40.6	79.2	125.7	137.5	0.49	3.7	5.1	35.4

9 Conclusion

In this paper, we report empirical studies on changes in different speech characteristics and cepstral features with aging for Bengali vowels. Fundamental frequency changes in varying ways for male and female speakers. We have observed that formant frequencies also changes with aging. On average, *jitter* and *shimmer* increases with aging but it varies in degree from speaker to speaker. There are some younger speaker too who has higher jitter and shimmer. It is widely observed that HNR decrease with age.

In ASR, MFCC features are used for phone modeling. We have extracted the MFCC features of different voiced vowels. We have used mean and variance of HMM model parameter of each phone for measuring the divergence between two age groups. Changes are clearly evident in the results.

We have used triphone HMM model for ASR evaluation. We have applied triphone language model. Triphone language model will incorporate linguistic information to improve the recognition accuracy. It is seen that there is a mismatch in acoustic models of young and older groups. Specific model for younger and older can thus improve accuracy.

ASR performance can be improved further with speaker normalization. Vocal tract length normalization and maximum likelihood linear regression are two well known speaker normalization techniques.

Acknowledgements Author wish to acknowledge financial support from the Technology Intervention for Elderly, Department of Science and Technology, Government of India, project Elderly speech recognition with applications. I will remain grateful to Communication Empowerment Laboratory, Indian Institute of Technology, Kharagpur for providing me all the facilities. Finally, I wish to thank all the individuals who participated in this study.

References

Baken, R. J. (2005). The aged voice: a new hypothesis. *Journal of Voice*, *19*, 317–325.

Barlow III, J.A. (2009). Age-related changes in acoustic characteristics of adult speech. *Journal of Communication Disorders*, *42*(5), 324–333.

Barman, B. (2011). A contrastive analysis of English and Bangla phonemics. Dhaka University. *Journal of Linguistics*, *2*(4), 19–42.

Benjamin, B. J. (1981). Frequency variability in the aged voice. *Journal of Gerontology*, *36*(6), 722–726. doi:10.1093/geronj/36.6.722.

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

Boersma, P., & Weenink, D. (2011). Praat: doing phonetics by computer (version 5.2.16). (Computer program): Retrieved February 20, 2011. http://www.praat.org.

Cassidy, S., & Harrington, J. (2001). Multi-level annotation in the emu speech database management system. *Speech Communication*, *33*(1–2), 61–77.

Chatterji, S. K. (1921). Bengali phonetics. *Bulletin of the School of Oriental Studies, University of London*, *2*(1), 1–25.

Deliyski, D. & Xue, S. A.: (2001). Effects of aging on selected acoustic voice parameters: preliminary normative data and educational implications. *Educational Gerontology*, *27*(2), 159–168.

Endres, W., Bambach, W., & Flösser, G. (1971). Voice spectrograms as a function of age, voice disguise, and voice imitation. *The Journal of the Acoustical Society of America*, *49*(6B), 1842–1848.

Ferrand, C. T. (2002). Harmonics-to-noise ratio: an index of vocal aging. *Journal of Voice*, *16*(4), 480–487.

Ghosh, S., Burnham, K. P., Laubscher, N. F., Dallal, G. E., Wilkinson, L., Morrison, D. F., Loyer, M. W., Eisenberg, B., Kullback, S., Jolliffe, I. T., & Simonoff, J. S. (1987). Letters to the editor. *The American Statistician*, *41*(4), 338–341.

Gorham-Rowan, M. M., & Laures-Gore, J. (2006). Acoustic-perceptual correlates of voice quality in elderly men and women. *Journal of Communication Disorders*, *39*(3), 171–184.

Harrington, J., Palethorpe, S., & Watson, C. I. (2010). Age-related changes in fundamental frequency and formants: a longitudinal study of four speakers. In *Interspeech* (pp. 2753–2756).

Hillenbrand, J., Cleveland, R. A., & Erickson, R. L. (1994). Acoustic correlates of breathy vocal quality. *Journal of Speech, Language, and Hearing Research*, *37*(4), 769–778.

Hisao, K. (1997). Acoustic and perceptual properties of phonemes in continuous speech as a function of speaking rate. In *EUROSPEECH* (pp. 1003–1006).

Hollien, H., & Shipp, T. (1972). Speaking fundamental frequency and chronologic age in males. *Journal of Speech and Hearing Research*, *15*(1), 155–159. http://jslhr.asha.org/cgi/content/abstract/15/1/155.

Krom, G. d. (1993). A cepstrum-based technique for determining a harmonics-to-noise ratio in speech signals. *Journal of Speech, Language, and Hearing Research*, *36*(2), 254–266.

Lindblom, B. E. F. (1971). Acoustical consequences of lip, tongue, jaw, and larynx movement. *The Journal of the Acoustical Society of America*, *50*, 1166–1179.

Linville, S. E. (1996). The sound of senescence. *Journal of Voice*, *10*, 190–200.

Linville, S. E. (2001). *Vocal aging*. San Diego: Singular Publishing Group.

Linville, S. E., & Rens, J. (2001). Vocal tract resonance analysis of aging voice using long-term average spectra. *Journal of Voice*, *15*(3), 323–330.

Liss, J. M., Weismer, G., & Rosenbek, J. C. (1990). Selected acoustic characteristics of speech production in very old males. *Journal of Gerontology*, *45*(2), 35–45.

Mann, H. B., & Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *Annals of Mathematical Statistics*, *18*(1), 50–60.

Markus, B., & Walter, S. (2003). Aging female voices: an acoustic and perceptive analysis. In *VOQUAL* (pp. 163–168).

Paulsen, F. P., & Tillmann, B. N. (1998). Degenerative changes in the human cricoarytenoid joint. *Archives of Otolaryngology, Head of Neck Surgery*, *124*, 903–906.

Ramig, L. A., & Ringel, R. L. (1983). Effects of physiological aging on selected acoustic characteristics of voice. *Journal of Speech, Language, and Hearing Research*, *26*(1), 22–30.

Ramig, L. O., Gray, S., Baker, K., Corbin-Lewis, K., Buder, E., Luschei, E., Coon, H., & Smith, M. (2001). The aging voice: a review, treatment data and familial and genetic perspectives. *Folia Phoniatrica et Logopaedica*, *53*(5), 252–265.

Reubold, U., Harrington, J., & Kleber, F. (2010). Vocal aging effects on F_0 and the first formant: a longitudinal analysis in adult speakers. *Speech Communication*, *52*(7–8), 638–651.

Rodeño, M. T., Sánchez-Fernández, J. M., & Rivera-Pomar, J. M. (1993). Histochemical and morphometrical ageing changes in human vocal cord muscles. *Acta Oto-Laryngologica*, *113*, 445–449.

Rother, P., Wohlgemuth, B., Wolff, W., & Rebentrost, I. (2002). Morphometrically observable aging changes in the human tongue. *Annals of Anatomy - Anatomischer Anzeiger*, *184*(2), 159–164.

Scott, D. W. (1992). *Multivariate density estimation*. New York: Wiley.

Tanmay, B. (2000). Bangla (Bengali). In *Gary, Jane; Rubino, Carl, Encyclopedia of World's languages: past and present (facts about the World's languages)*.

Tolep, K., Higgins, N., Muza, S., Criner, G., & Kelsen, S. G. (1995). Comparison of diaphragm strength between healthy adult elderly and young men. *American Journal of Respiratory and Critical Care Medicine*, *152*, 677–682.

Traunmuller, H. (1984). Articulatory and perceptual factors controlling the age and sex-conditioned variability in formant frequencies of vowels. *Speech Communication*, *3*(1), 49–61.

Ulatowska, H. K. (1985). *The aging brain: communication in the elderly*. San Diego: College-Hill Press.

Vipperla, R., Renals, S., & Frankel, J. Ageing voices: the effect of changes in voice parameters on asr performance. *EURASIP Journal on Audio, Speech, and Music Processing*, *2010*, 41–50 (2010). doi:10.1155/2010/525783.

Wilcox, K. A., & Horii, Y. (1980). Age and changes in vocal jitter. *Journal of Gerontology*, *35*(2), 194–198.

Xue, S. A., & Hao, G. J. (2003). Changes in the human vocal tract due to aging and the acoustic correlates of speech production: a pilot study. *Journal of Speech, Language, and Hearing Research*, *46*(3), 689–701.

Young, S., Kershaw, D., Odell, J., Ollason, D., Valtchev, V., & Woodland, P. (2000). *The HTK book version 3.0*. Cambridge: Cambridge University Press.

Yumoto, E., Sasaki, Y., & Okamura, H. (1984). Harmonics-to-noise ratio and psychophysical measurement of the degree of hoarseness. *Journal of Speech, Language, and Hearing Research*, *27*(1), 2–6.