Proceedings of the Second International Conference on Research in Intelligent and Computing in Engineering pp. 283–287 DOI: 10.15439/2017R21
ACSIS, Vol. 10 ISSN 2300-5963

Emotion Analysis from Speech of Different Age Groups

Hemanta Kumar Palo
Department of Electronics and Communication Engineering,
Siksha 'O' Anusandhan University
Bhubaneswar, Odisha, India
hemantapalo@soauniversity.ac.in

Mihir Narayan Mohanty
Department of Electronics and Communication Engineering,
Siksha 'O' Anusandhan University
Bhubaneswar, Odisha, India
mihir.n.mohanty@gmail.ac.in

Mahesh Chandra
Department of Electronics and Communication Engineering,
Birla Institute Technology, Mesra,
Ranchi, India
shrotriya69@rediffmail.com

***Abstract*—This Recognition of speech emotion based on suitable features provides age information that helps the society in different ways. As the length and shape of human vocal tract and vocal folds vary with age of the speaker, the area remains a challenge. Emotion recognition system based on speaker's age will help criminal investigators, psychologists and law enforcement agencies in dealing with different segments of the society. Particularly child psychologists, counselors can take timely preventive measures based on such recognition system. The area remains further complex since the recognition system trained for adult users performs poorer when it involves children. This has motivated the authors to move in this direction. A novel effort is made in this work to determine the age of speaker based on emotional speech prosody and clustering them using fuzzy c-means algorithm. The results are promising and we are able to demarcate the emotional utterances based on age.**

Keywords: Emotion Analysis; feature extraction; clustering algorithm; Fundamental frequency, speech rate

I. Introduction

Patterns based on gender and age can be obtained from facial expressions, gestures or verbal communication of individual. Among these modalities, this paper emphasizes on the recognition of emotions based vocal communication. The objective is to determine the speaker's emotional conversation pattern based on his/her age. Determination of these will be beneficial to law enforcement agencies in studying criminal psychology and further investigation. Particularly, the speaker's state of mind and emotional attributes will assist the condition of both victim and the culprit during court hearing and prevent confusion. Identification of intimidating calls, false alarms, kidnapping involving influential people, fanatic religious groups, radicals etc. can be made possible with such systems [1]. Further, the recognition system will help in implementing corrective measures in case negative emotional attributes are manifested among children before it is too late. Utterances of speaker colored with emotion and age detection can also help human robotic interfaces, telecommunications, intelligent tutoring, smart call center application etc.

This work was not supported by any organization

The vocal tract and vocal fold of human speech production mechanism is in a growing stage till a child attains adolescent. Selecting suitable features representing age of the speaker thus remains an ever-growing challenge. Recognition systems trained for adult speakers often proved inefficient when these are trained with children utterances [2]. This is because, the core features representing the speech and emotional contents of an utterances vary with age and gender of the speaker. Especially, the fundamental frequency (F0), formants, speech rate, energy etc. vary drastically between a child and an adult [3]. The acoustic models made for research and business requirement thus become ineffective in case the emotional utterances belong to different age group. Speaker's age and gender have been addressed by different literature during last decades although, these studies little emphasized on emotional contents of speech [4] [5]. These authors attempted the Gaussian weight super-vectors with support vector machine (SVM) classifier for age and gender identification. However, no precise study between different age groups or their emotional states has been made by them. Use of mel-frequency cepstral coefficient (MFCC) with different feature selection algorithm such as PCA (principle component analysis), supervised PCA (SPCA) has been attempted for different age groups in [6]. The prominent prosodic features representing speech emotion of children and adults could not be found in these literatures. Absence of a clear boundary among emotions based on age has motivated the authors to move in this novel effort.

The objective is to cluster the features representing emotional utterances of different age groups. Different clustering approaches such as fuzzy c-means (FCM), hierarchical clustering, Partitioning, Density-Based, Grid-Based, Model-Based, K-means clustering has been applied to recognize human emotions [7] [8]. K-means is a hard clustering algorithm, simple and can solve known clustering problems using unsupervised learning. The algorithm is faster than hierarchical clustering producing tighter clusters. The algorithm proved better when compared with fuzzy c-means classifier for speech emotion recognition using

GMM super vectors [7]. This has been chosen for clustering the desired emotions based on age for our purpose.

Section II of this work describes the database used followed by the feature extraction technique in section III. The clustering algorithm chosen has been explained in section IV and the results are shown in section V. Section VI provides the conclusion.

II. EMOTIONAL SPEECH MATERIALS

Collection of database for speech emotions involving different age group is a tedious task. Most of the available databases are confined to a particular segment of speaker. Emotional database of speakers with age spanning a large range is either unavailable or inaccessible. Further, emotional utterances under real-life scenario in such situation are seldom found. Thus, the database desired for this work has been collected from different sources. The utterances are collected over three months of time by placing recording instruments at different locations. Around one thousand utterances of speakers among eight to forty are obtained in total. Out of these, the emotional contents of speech are taken out for simulation purpose. Eighteen utterances of angry, boredom and sad emotions based on the opinion of linguistic experts have been used for further processing. 16 bit quantization with 16 kHz sampling rate was used for recording the utterances. Average duration of an utterance is of 5 seconds. A mobile set of good quality has been used to record the utterances. Format factory software has been used to transform the mobile data into .wav format. The signal is pre-processed to accommodate speaker variability and noise to the minimum before further processing.

III. FEATURE EXTRACTION

Feature extraction and selection is an important aspect of recognition system [8]-[11]. Uses of acoustic features consisting of speech prosody characteristics have been discussed by researcher in the field of speech emotion recognition [12]-[14]. Among these features, the F0, energy or amplitude, speech rate etc. are few features that vary with age of the speaker hence considered in this work. A brief discussion on these features is made in the following section.

A. Fundamental frequency (F0)

Male speakers tend to have lower F0 as compared to both children and female subjects due to larger vocal cord. Older people have lower fundamental frequency as compared to adults when experiment has been conducted between twenty older and twenty younger adults by the author [15]. The fundamental frequency continues to decrease with age for both genders [16] [17]. When speech is coloured with emotions, it is observed that higher arousal emotions such fear, angry have higher F0 as compared to neutral or lower arousal emotions such as sad or boredom [14] [18]. Among different methods of F0 extraction, autocorrelation and cepstral methods are most popular [13] [19]. Autocorrelation method of F0 extraction has been used in this work. Using this technique, the feature can be extracted directly from the speech waveform. It requires less hardware such as a multiplier and an accumulator than other methods. Further, the method is simple and noise immune. For a signal $x(n)$ delayed by τ, the auto correlation coefficients (ACF) is estimated using the relation

$$S(\tau) = \frac{1}{N}\sum_{n=0}^{N-1} x(n)x(n+\tau) \quad (1)$$

Highest value of ACF can be obtained when the condition $s(n) = s(n+\tau)$, is satisfied and is indicated as $S(0)$. The ACF decreases with increase in the signal delay. Denoting the time period of the signal as T, the ACFs will attain its peak at$\tau = IT$, where I is an integer. From these peak locations F0 can be estimated.

B. Log energy

People speak at higher intensity or energy when aroused by certain emotions such as angry, happy or surprise [14] [18]. These emotions have more energy contents at higher frequencies. Energy or amplitude indicates the volume of the speech. The strength of the voice is automatically raised with significant increase in amplitude when people get excited or agitated. Dull voice related to sad or bore emotions often are of low amplitude or energy. Logarithm of energy remains an important feature of emotion that suits logarithmic nature of hearing mechanism. The log energy can be estimated for a signal $s_k(n)$using the relation

$$e_{log}(n) = 10\ log \sum_{k=1}^{w} |s_k(n)|^2 \quad (2)$$

where, w is the analyzing window.

C. Speech rate

Speech rate is an important feature that provides information on speaker's age, gender, language, demographic and cultural profile [20] [21]. The application domains are speech pathology, speech science, behavioural psychology, emotional analysis, neuropsychology etc. Speaking rate signifies the communication time of a message during conversation. It is an indication of the number of syllables or words or spoken units that is uttered per minute or second. It represents the quickness at which a speaker utters an emotional sentence at certain situation. It is a global feature taken over whole length of the signal. Human being speaks faster when gets excited than in cool mood. Thus, angry, fear or high frequency content emotions are likely to have higher speech rate than neutral or sad or low excited voices. For an utterance, the average speaking rate can be estimated using the relation

$$R(s) = \frac{N_V(s)}{D(u)} \quad (3)$$

Where, $N_V(s)$ and $D(u)$denote the number of vowel segment and utterance duration respectively.

IV. K-MEANS CLUSTERING

K-means is hard clustering algorithm more suitable for exclusive clustering task. The objective of the work is to distinguish speaker's emotion based on features that varies with age. Thus, it will be a supportive approach in this case. Let, there are 'P' numbers of features contain all the states of emotions. Using the algorithm, the features are partitioned into Lclusters each having a cluster center $C_l,\ l = 1, 2, \dots, L$. Each cluster center is associated with the corresponding class. With the help of squared error function, the objective function 'b' is minimized in formation of the clusters. Optimal convergence of 'b' will ensure adequate clustering of the desired emotion. The objective function is represented as

$$b = \sum_{l=1}^{L} \sum_{p=1}^{P} \left\| s_p^{(l)} - C_l \right\|^2 \quad (4)$$

where, $\|\cdot\|$ is norm representing the distance between C_l and the data point $s_p^{(l)}$. K-means algorithm has been performed using following steps

1. From each L feature points, select the centroid.
2. Obtain L cluster by iteratively repeating the procedure. In the process, allot all the source data point to the respective nearest centroid.
3. The centroids are updated by estimating the cluster centers iteratively, until further variation in cluster center is manifested.

V. RESULTS AND DISCUSSION

The variation of F0 of children and adults has been shown in Fig. 1. It is observed that, children and female have higher F0 as compared to the adults due to larger vibration of their short vocal tract. The plot formed is in zigzag fashion as the utterances consist of both genders. The value decreases with age due to increase in vocal tract length owing to growth of facial skeleton and lowering of the larynx. Due to higher excitation level, angry state has shown larger pitch for both adult and children compared to boredom and sadness as shown in the Figure. A comparison on different gender independent prosodic features of adult emotional states attempted in literature is given in Table 1. The features extracted in this work are compared among different age group and is tabulated in Table II.

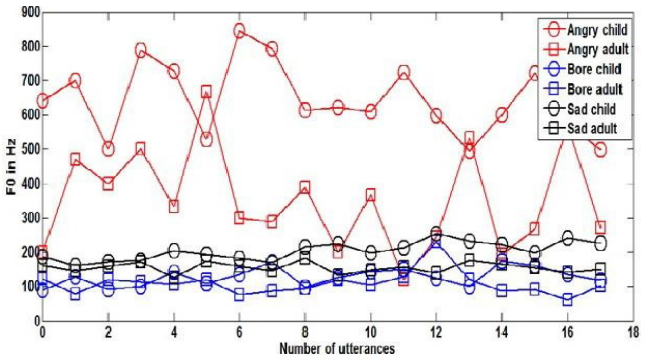

Fig. 1 Variation of F0 with age for different speech emotional states

The observed results indicate a higher value of mean, maximum, and minimum F0 values for children speech as compared to the adult utterances. This is found to be true across all classes of speech emotions chosen in this work. Pitch variation and the mean value have been tested for different gender independent adult emotions [22]-[24]. The pitch mean found to be highest for angry emotion followed by happy and bore state as claimed by these authors.

TABLE I.

Comparative study of the state of art age and gender independent feature extraction techniques

Features	**Emotions**					
	Angry	**Sad**	**Fear**	**Happy**	**Bore**	**Disgust**
Speech rate [21]	↑↑	↓↓	↑↑↑	↑	↓	↓↓↓
F0 mean [22] [23]	↑↑↑	↓	↑	↑↑	↓↓	↓↓↓
F0 Variation [22][23]	↑↑↑	↓	↑	↑↑	↓↓	↓
Energy [9][18]	↑↑↑	↓↓	↑	↑↑	↓↓	↑↑
F1 [12][21]-[24]	↑↑↑	↓↓	↑	↑↑	↓	↑↑
Duration [24]	↓↓	↑↑↑	↑	↑	↑↑	↓
Spectral centroid [24]	↑↑↑	↓↓	↑	↑↑	↓↓↓	↓

↑= increase, ↓= decrease

TABLE II.

Comparison of features based on different age group

Features	**Child**			**Adult**		
	Angry	Sad	Bore	Angry	Sad	Bore
Pitch (mean)	643.5	204.2	127.9	350.7	155.3	109.3
Pitch (max)	844	255	175	667	183	231
Pitch (min)	495	162	89	120	126	61
Speech rate(mean)	0.54	0.28	0.23	0.34	0.17	0.19
Speech rate(max)	0.86	0.36	0.27	0.58	0.24	0.36
Speech rate(min)	0.37	0.24	0.19	0.20	0.14	0.13
Log energy(mean)	32.1	24.6	28.8	18.5	13.9	15.2
Log energy(max)	36.1	29.3	31.7	25.9	16.8	20.2
Log energy(min)	17.8	18.5	24.9	12.7	10.2	10.5

Energy or intensity indicates the arousal level of an emotion. The calculated value indicates higher energy for higher arousal emotional states such as angry, happy, and fear. Bore is found to have the lowest energy among all the states tabulated. The presence of higher frequency components increases the energy level of angry state than that of bore and sad emotion. Computations of spectral energy by different authors are worth noting to support the findings in this work [21] [23]. The log-energy features of both children and adults are plotted for these emotions in Fig. 2. The feature extraction technique is so chosen to approximate human hearing system that acts non-linearly at different bands of the signal. The energy founds to be higher for children than their adult counterpart as observed from the figure. The log-energy in dB is compared in table II for adult and children utterances. The result indicates a larger mean, maximum, and minimum energy for children speech utterances. Children are inherently more excited and enthusiastic to abnormal situations than well matured and

judgmental adults. This makes the children over expressive with larger arousal states than the adults.

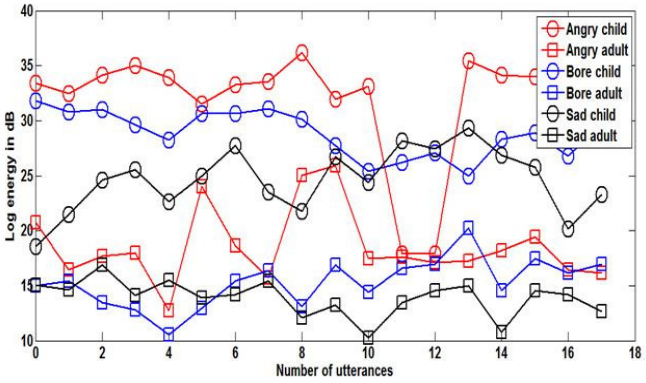

Fig. 2 Variation of log-energy features with age for different speech emotional states

Fig. 3 provides the variation in speech rate of children against that of adult speakers up to 40 years of age. It is observed that, a child takes more time during conversation as compared to adults. These may be attributed to reading disorder and social anxiety that is normally found with children. The neuro-muscular and biological factors are other aspects that tend to decrease the speech rate of children. As child reaches to adulthood, he or she develops the oral-motor skills and linguistic skills like lexical, semantic and phonological parameters. Increase in motor planning specificity of growing children increases the articulation rate. Due to cognitive development with age, the fluency in speech increases. These factors make the speech rate of adults higher than that of children. On contrary, limited exposure to the environment and language makes children to ponder between suitable words or vocabulary during expression of emotions. They invent their own words rather than learned words used by adults using associative skills and imagination to certain situations. This leads to reduction in reaction time and decrease in speaking rate. The emotional utterances are taken in a natural background where, the conversations are task dependent. This may be the reason of variation in speech rate. It has been evidenced with higher mean, maximum, and minimum speech rate for children utterances as shown in table II. This is true across all the emotional classes.

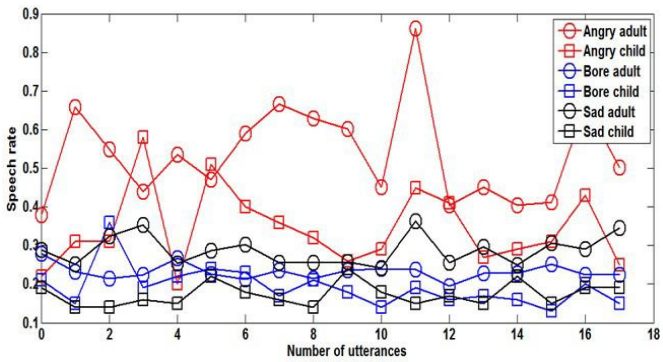

Fig. 3 Variation of speech rate with age for different speech emotional states

The speech rate is highest for fear and angry state than that of sad and bore states as shown in Fig. 3 similar to other observation made in literature [21]. The reason may be attributed to higher energy (or high frequency components) that inherit high arousal emotions. While comparing the duration feature to investigate emotional cues of speakers, few authors provide similar trends [24]. It can be concluded that, due to lower speech rate, the utterance duration tends to be longer for sad emotion followed by bore state. A close observation of duration feature reveals that human being takes lager time to express emotions having lower energy as compared to aggressive states.

An attempt is made in this work to cluster different age group using K-means clustering with the chosen feature sets. K-means is more suitable for exclusive clustering of data. Cluster of angry speech emotion based on three age groups as 8-14 years, 19-24 years and 30-40 years using speech rate features is shown in in Fig. 4. It is observed that, the cluster groups of 19-24 years and 30-40 years are more closure. These groups are similar and thus described by features of similar magnitudes. On the contrary, the clustering of the features representing these groups is widely separated from that of children falling in the age group of 8 to 14 years due to quite distinct feature values.

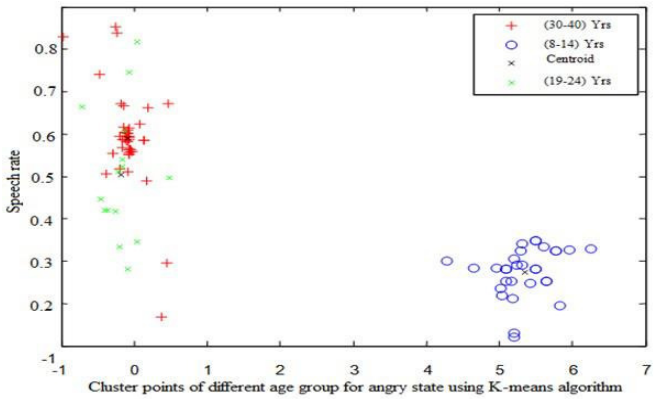

Fig. 4 K-means clustering of angry speech emotion for different age groups using speech rate.

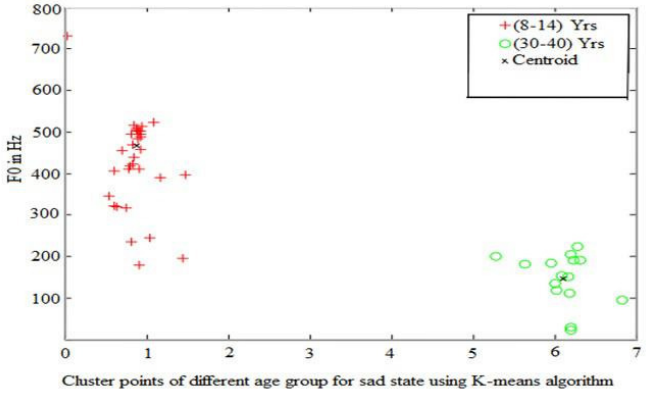

Fig. 5 K-means clustering of sad speech emotion for different age groups using F0 features

A similar comparison with K-means clustering is done using F0 features of sad emotion in Fig. 5. In this case, the older adult group (30-40 years) is compared with the youngest group (8-14 years). A widely separated cluster

has been observed between these two classes using sad emotional state. This may be due to the reason explained in previous paragraph.

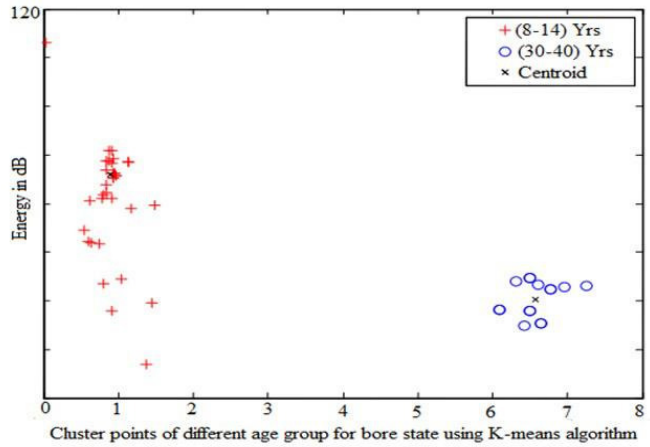

Fig. 6 K-means clustering of bore speech emotion for different age groups using log energy features.

Children and an adult group between 30 to 40 years of age using log energy features have been analyzed. This is shown for bore emotion using the chosen clustering algorithm in Fig. 6. It is found that, the children have more energy as compared to that of adults. Further the clusters are widely separated for the chosen emotional state as shown in this Figure.

VI. Conclusion

A system dividing speaker's emotion based on age may help industries dealing with computer games, on-line tutoring, robotics and multi-media. There have been few speech enabled application such as Windows Phone app, World search, the European Portuguese version of the app have provided manual adjustment by speakers based on age. However, these systems remain cumbersome. Hence, an automatic detection system based on age will provide a new direction in this field. A more entertaining, fun and engaging way can be developed by manufacturers in case the users involved are children. Similarly, more polite and judgmental systems will help multi-media industries concentrating on adults.

References

[1] A. Hämäläinen, H. Meinedo, M. Tjalve, P. Pellegrini, I. Trancoso, and M. S. Dias, "Improving speech recognition through automatic selection of age group–Specific Acoustic Models," *adfa*, Springer, pp. 1, 2011.

[2] D. C. Tanner, and M. E. Tanner, "Forensic aspects of speech patterns: voice prints, speaker profiling, lie and intoxication detection," *Lawyers & Judges Publishing*, 2004.

[3] E. Lyakso, O. Frolova, E. Dmitrieva, A. Grigorev, H. Kaya, A. A. Salah, and A. Karpov, "EmoChildRu: emotional child Russian speech corpus," *Speech and Computer, 17th International Conference*, SPECOM 2015, Athens, Greece, pp. 144–152, 20-24 Sept. 2015.

[4] M. Feld, F. Burkhardt, and C. Müller, "Automatic speaker age and gender recognition in the car for tailoring dialog and mobile services," *In proc. Interspeech*, Japan, pp. 2834–2837, 2010.

[5] R. Porat, D. Lange, and Y. Zigel, "Age recognition based on speech signals using weights supervector," *In proc. Interspeech*, Japan, pp. 2814–2817, 2010.

[6] S. J. Chaudhari, and R. M. Kagalkar, "Automatic speaker age estimation and gender dependent emotion recognition," *International Journal of Computer Applications*, vol. 117, no. 17, pp. 5–10, May 2015.

[7] J. Kaur, and S. Vashish, "Analysis of different clustering techniques for detecting human emotions variation through data mining," *International Journal of Computer Science Engineering and Information Technology Research* (IJCSEITR), vol. 3, iss. 2, pp. 27–36, Jun. 2013.

[8] I. Trabelsi, D. B. Ayed, and N. Ellouze, "Comparison between GMM-SVM sequence kernel and GMM: application to speech emotion recognition," *Journal of Engineering Science and Technology*, 2016 (to be published).

[9] V. M. M. Maca, J. P. Espada, V. G. Diaz, and V. B. Semwal, "Measurement of viewer sentiment to improve the quality of television and interactive content using adaptive content," 2016 *International conference on electrical, electronics, and optimization techniques* (ICEEOT), pp. 4445-4450, Doi: 10.1109/ICEEOT 2016. 7755559.

[10] V. B. Semwal, J. Singha, P. K. Sharma, A. Chauhan, and B. Behera, "An optimized feature selection technique based on incremental feature analysis for bio-metric gait data classification," *Multimedia Tools and Applications*, pp. 1-19, doi:10.1007/s11042-016-4110-y, Dec. 2016.

[11] P. Kumari, and V. Abhishek, "Information-theoretic measures on intrinsic mode function for the individual identification using EEG sensors," *IEEE Sensors Journal*, vol. 15, no. 9, pp. 4950-4960, Sep. 2015.

[12] H. K. Palo, and M. N. Mohanty, "Classification of emotions of angry and disgust," *Smart Computing Review*, vol. 5, no. 3, pp. 151–158, Jun. 2015.

[13] S. G. Koolagudi, and K. S. Rao, "Emotion recognition from speech: a review," *International Journal of Speech Technology*, springer, vol. 15, pp. 99–117, 2012.

[14] M. E. Ayadi, M. S. Kamel, and F. Karray, "Survey on speech recognition: resources, features and methods," *Pattern Recognition*, vol. 44, pp. 572–587, 2011.

[15] B. J. Benjamin, "Frequency variability in the aged voice," *Journal of Gerontology*, vol. 36, no. 6, pp. 722–726, 1981.

[16] B. Das, S. Mandal, P. Mitra, and A. Basu, "Effect of aging on speech features and phoneme recognition: a study on Bengali voicing vowels," *International Journal of Speech Technology*, vol. 16, iss. 1, Springer, pp. 19–31, Mar. 2013.

[17] R. Winkler, "Influences of pitch and speech rate on the perception of age from voice," *Published in Proceeding of ICPhS*, Saarbrücken, pp. 1849–1852, 6-10 August 2007.

[18] H. K. Palo, and M. N. Mohanty, "Performance analysis of emotion recognition from speech using combined prosodic features," *Advanced Science Letters*, vol. 22, no. 2, pp. 288–293 (6), Feb. 2016.

[19] L R. Rabiner, M. Cheng, A. Rosenberg, and C. McGonegal, "A comparative performance study of several pitch detection algorithms," *IEEE Transactions on Acoustics, Speech, and Signal Processing,* vol. 24, no. 5, pp. 399-418, Oct. 1976.

[20] S. Paulmann, M. D. Pell, and S. A. Kotz, "How aging affects the recognition of emotional speech," *Brain and language*, vol. 104, no. 3, pp. 262-269, Mar 2008.

[21] P. Laukka, J. Patrik, and B. Roberto, "A dimensional approach to vocal expression of emotion," *Cognition and emotion*, vol. 19, no. 5, pp. 633-653, Aug. 2005.

[22] X. A. Rathina, K. M. Mehata, and M. Ponnavaikko, "Basic analysis on prosodic features in emotional speech," *International journal of computer science, engineering and applications* (IJCSEA), vol. 2, no. 4, Aug. 2012.

[23] R. Banse, and K. R. Scherer, "Acoustic profiles in vocal emotion expression," *Journal of personality and social psychology*, vol. 70, no. 3, pp. 614-636, Mar. 1996.

[24] D. A. Sauter, F. Eisner, A. J. Calder, and S. K. Scott, "Perceptual cues in nonverbal vocal expressions of emotion," *The quarterly journal of experimental psychology*, vol. 63, no. 11, pp. 2251– 2272, Apr. 2010.