Acoustics Speech Processing of Sanskrit Language

Sujay G. Kakodkar
Masters of Engineering
Industrial Automation & Radio Frequency
Goa College of Engineering
Ponda, Goa-India, 403401

Samarth Borkar
Asst. Professor
Goa College of Engineering
Electronics & Telecommunication Department
Ponda, Goa-India, 403401

ABSTRACT
Speech processing (SP) is the latest trend in technology. An intelligent and precise human-machine interaction (HMI) is designed to engineer an automated, smart and secure application for household and commercial application. The existing methods highlight the absence of the speech processing in the under-resourced languages. The novelty of this work is that it presents a study of acoustic speech processing (ASP) using spectral components of Mel frequency cepstrum coefficient (MFCC) of Sanskrit language. A customized speech database is created as no generic database is available in Sanskrit. The processing method includes speech signal isolation, feature selection and extraction of selected features for applications. The speech is processed over a custom dataset consisting of Sanskrit speech corpus. The spectral features are calculated over 13 coefficients providing improved performance. The results obtained highlight the performance of the proposed system with the variation of the lifter parameter.

General Terms
Acoustic Speech Processing, Feature extraction.

Keywords
Speech processing; Human-machine interaction; Mel frequency cepstrum coefficient; Sanskrit language;

1. INTRODUCTION
The verbal communication among human is through speech. Speech has become the basis of textual language, which contrasts in its vocabulary and phonetics from its spoken one. Speech processing is a processing of digital speech signal in conjunction with natural language. With the progressing technology, studies have been worked on mining acoustic speech features and simulating it for various other functions.

SP is a blooming investigative area in a home as well as industrial application. It is employed in e-learning, medicine, law, monitoring, entertainment, marketing etc. [1][2][3].

The ASP finds its usefulness in monitoring the patient to analyse his recovery. The person with autism (struggling to infer emotions) is made to express his emotions through games. In a field of security, the SP is employed in speaker identification. It also discovers its utility in voice navigation over a desktop to navigate to a required window. A presentative style of approach is followed in e-learning by detecting the persons emotions. A music therapy helps a person to relieve stress and tension [4][5][6].

In biological terms, larynx is responsible for the sound production which is a part of respiratory system. The larynx consists of vocal cord which is main part known as voice box. Vocal cords are made up of five layers. Each layer contributes a necessary and unique component to voicing. Epithelium, is a thin skin that acts as a barrier and easily vibrates. A combination of three non-muscular tissues is Lamina propria. The outer and middle layers contain elastin (stretchy fibers) allowing vocal cords to stretch; the innermost layer of the lamina propria has elastin that keeps it from stretching too much. The final part of the vocal fold is a largest and bulky muscle that makes up about three-quarters of the vocal fold. It. can thicken thinner, shorten, lengthen, relax and stiffen to produce different sounds.

Larynx is a passage for inhalation and exhalation i.e. air in and out of lungs. Sound is produced during exhalation process and if a person speaks continuously fast, the person falls out of breath hence inhalation is necessary in sound production. The type of sound generated depends upon the movement of the vocal cords. The vocal cord widens for inhalation. A whispering sound is produced when the vocal cords are narrowed. When the vocal cords brought together a phonation sound is generated. Based on the type of pitch required the vocal cords are tightened or loosened. A high pitch is achieved by tightening the vocal cords whereas loosening them produces a lower pitch.

The larynx (voice box) holding the vocal fold is more extended in males. Hence they tend to have an Adam's apple. When the vocal folds generate a sound wave the wavelength is longer as it has to travel extra along the vocal tract. Longer wavelength makes a lower pitch, reasoning the lower pitch in males. The females having a shorter larynx, a shorter wavelength is a reason for the higher pitch in females. The male pitch is between 80-100Hz, 200-330Hz is a feminine pitch. A male pitch can be feminine between 100-200Hz. Another important thing to note that as the age advances of the people their speech slows down, syllables and words are elongated, and sentences are littered with more recesses for air. Pitch and loudness is be reduced, and tremors can appear.

However only men's larynxes change more than women's. Male voice pitch tends to rise with age, while female voice pitch stays the same, or may lower slightly. Larynx cartilages become harder with age reducing a person's pitch range. Vocal folds become stiffer and thinner, producing higher pitched voice, especially in males. The bulky muscle of the vocal fold shrinks with age, creating a weaker, breathier voice. The respiratory system also tends to work less efficiently as we age, thus speaking is a more difficult task [7].

Sanskrit is an ancient language filled with rich literature and a wide variety of form. It finds its influence over many Indian as well as foreign languages. It helps to decode the various formulae for the developments in different fields of science and technology from ancient times to modern times. With the

changing time, the importance of Sanskrit is overlooked. A research in the field of technology with Sanskrit will help it to regain its lost allure [8][9].

2. LITERATURE REVIEW

In a biomedical research area, Yao *et. al.* investigated in developing Bionic wavelet transform dealing with the energy distribution of auditory system to enhance sensitivity [10]. Improvising on Yao *et. al.*, an online based alliance environment for multiple telemedicine applications in speech therapy was developed by Malandraki *et.al.* [11]. R. Gamasu continued work further and presented a mobile telemedicine system for monitoring patients health by integrating ECG signal processing [12].

P.Y. Oudeyer [13] aimed at the necessity of the robotic pets to recognize the emotion of a human interaction. Koolagudi and Rao extended the research and reviewed on the types of speech corpus utilized for various SER systems [14]. Based on [14], Jamil *et.al.* proposed further relative feature processing an influence of age group in emotion recognition using a spontaneous Malay speech corpus [15]. There is no customary dataset as such followed for SER.

B. Logan presented the foremost features for speaker recognition and applicability in moulding music [16]. I. Trabelsi and D. Ayed followed earlier work and developed speaker recognition with data fusion using telephone speech [17]. Westera *et.al.* presented an online feedback based on vocal intonations and facial expressions [18]. SER finds its application to numerous fields and not only restricted to few fields such as speech recognition, e-learning, emotion recognition, security etc.

Gaikwad *et. al.* reviewed on speech processing techniques based on different types of speech consisting of isolated word, connected word, continuous speech and spontaneous speech [19]. Various feature extraction techniques were also reviewed ranging from MFCC, Linear prediction cepstral coefficient (LPCC) etc. Wiqas Ghai and Navdeep Singh continued the work further based on the different approaches followed i.e. acoustic-phonetic, pattern recognition, Knowledge Connectionist approach etc [20]. Vadwala *et.al.* followed previous work and advanced with speech processing challenges and techniques. The various techniques involved the utterance style, the speaker dependent and independent models, vocabulary, channel variability [21].

A thorough study of the SP systems proposed with the different applications, various features, and different speech corpus reveals that SP with other applications in under-resourced language can be explored [11][14][18]. A hybrid approach of processing with a combination of prosodic and spectral features is followed. A prototype for home as well as the industrial application is designed.

The features are obtained from various feature extraction techniques, such as LPCC, MFCC, fundamental frequency etc. The energy and short energy are intensity measurements, the fundamental frequency is measurement of speech frequency. The temporal features depict the time measurement of duration, timestamp etc. LPCCs are the linear prediction of the cepstral coefficients based on the human articulatory system. MFCCs are human auditory systems cepstral coefficient calculated using frequency wrapping scale (mel scale). MFCCs are computed using combination fundamental frequency, sampling frequency, frame duration, filter bank energies etc. MFCCs have an important role in noise reduction as the MFCC values are normalized when a speech signal contains noise. Hence MFCCs increase the performance of the overall system emphasizing that it a best among other feature extraction methods [18].

The MFCCs calculation depends upon the parameters containing speech signal, sampling frequency, frequency range, pre-emphasis coefficient, number of filters, number of cepstral coefficient and lifter parameter. All other parameters remain unchanged and have not much effect on the performance of the system except lifter parameter. Hence the variation of lifter parameter is studied to enhance cepstrum coefficients.

The novelty of our research is as follows

- In-house database of Sanskrit language has been created.[1]
- The speech is processed using MFCC over a customized Sanskrit database.
- Sanskrit in SP domain is researched.

3. METHODOLOGY USING MFCC

A comprehensive study of SP system exposes the limitations of the system with different classifiers and features providing dissimilar accuracies [2][4]. An improvised methodology is proposed, as shown in Fig 1. The system is improved from existing system with the use of variation in lifter parameters.

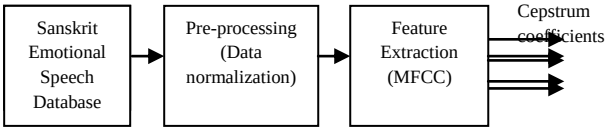

Fig 1: Block diagram of proposed system for sanskrit speech processing [4]

3.1 Sanskrit Speech Database

A Sanskrit speech database is prepared by recording set of 18 sentences each being spoken by 8 subjects, 4 male and 4 female, as shown in Fig 2. The speech corpus is recorded by using ZOOM H4N recorder into 2 channels at 44.1 KHz. A few sets of Sanskrit sentences uttered by different subjects are depicted in fig. below (Fig 2-7).The speech corpus contains normal sentences spoken in daily life.[1].

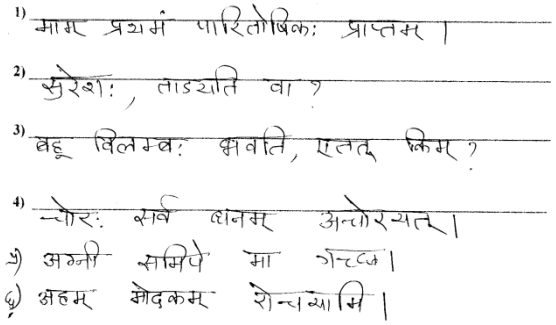

Fig 2: Sanskrit speech set of subject 1

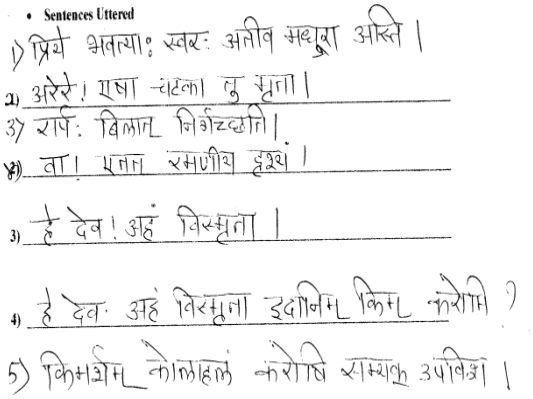

Fig 3: Sanskrit speech set of subject 2

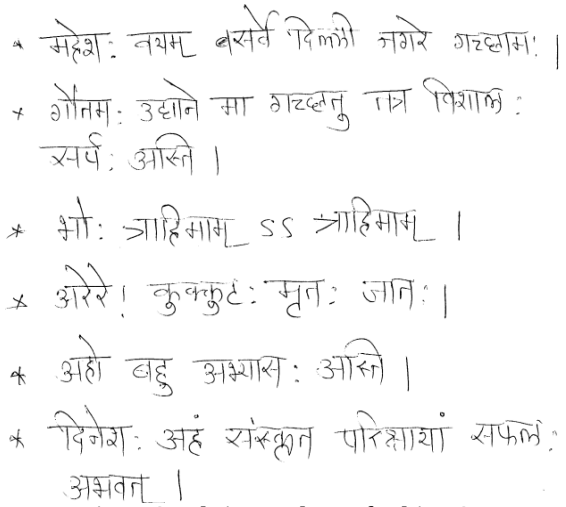

Fig 4: Sanskrit speech set of subject 3

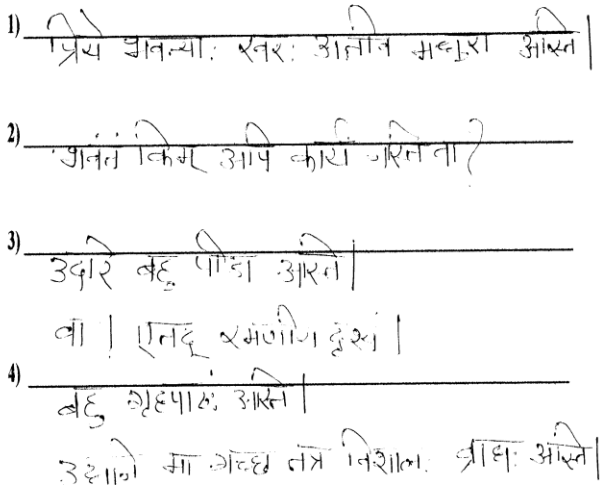

Fig 5: Sanskrit speech set of subject 4

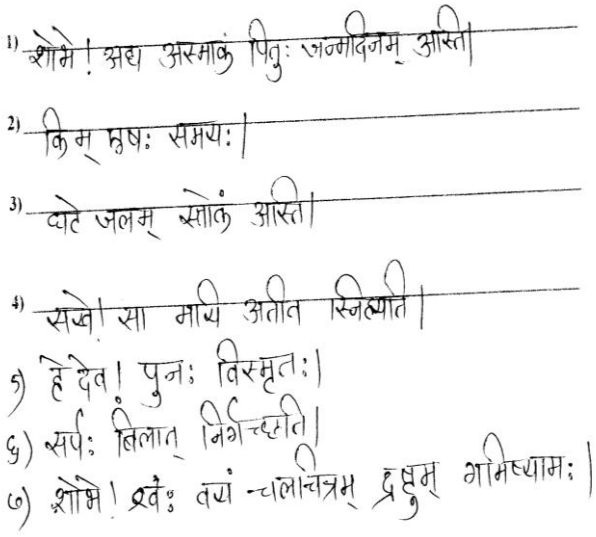

Fig 6: Sanskrit speech set of subject 5

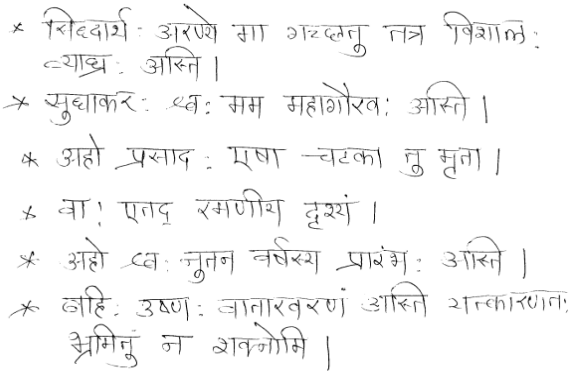

Fig 7: Sanskrit speech set of subject 6

3.2 Pre- processing Techniques

Signal is pre-processed in terms of noise cancellation, normalization, resampling signal etc. before discretizing the signal. The recorded samples of Sanskrit speech corpus is resampled at 16 KHz using Audacity software.[2]

3.3 Feature Extraction

MFCCs computed from speech signal sampled at sampling frequency. The speech signal is windowed using a hamming window function $\omega(n)$ given in (1), as shown in Fig 8.

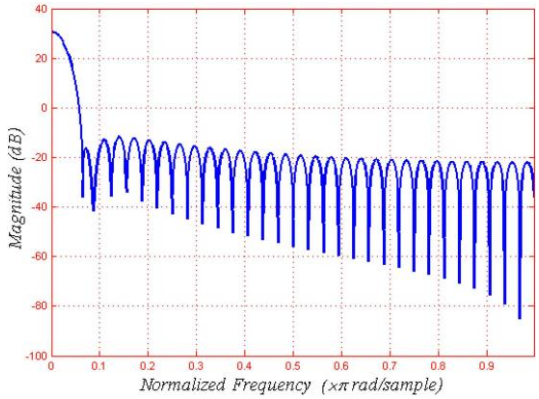

Fig 8: Hamming window - Frequency domain [4]

$$\omega(n) = \begin{cases} 0.54 - 0.46\cos(\frac{2\pi n}{N}); & 0 \le n \le N \\ 0 \quad ; & n \ge N \end{cases} \quad (1)$$

where N = number of samples in each frame, n= nth sample in frame.

Windowed signal is applied to mel filter bank , shown in (2) which sums up the energy in each filter.

$$B(f_l) = a\ln(1+\frac{f}{b}) \quad (2)$$

where B = bandwidth, f = frequency a = 125 and b= 700. The values of parameters a and b are chosen so as to convert the speech signal from frequency domain to mel scale (pitch scale). $B(f_l)$ is the converted speech signal from hertz to mel scale, as given in (2). Taking log and then discrete cosine transform of the filter bank energies (FBE's) MFCC's are calculated.

4. RESULTS

The MFCCs are calculated over 13 coefficients $C_0 - C_{12}$ using Sanskrit speech (Fig 9). The results are plotted in terms of numbers of cepstrum coefficients against the MFCC values corresponding to each speech signal. The maximum number of filter channels in the triangular filter bank is 40. The number of filter channels is more in low frequency than high frequency range. As the frequency increases, the number of filter channels decreases. D1-D10 denotes the filter channels.

In cepstral liftering, the cepstral coefficients are multiplied by the weights. It is corresponding to spectral convolution wherein, the log of power spectral sequence is convolved with impulse response of filter (determined by taking Fourier transform of the weight sequence). The liftering technique involved, results in a band pass (BP) filter.

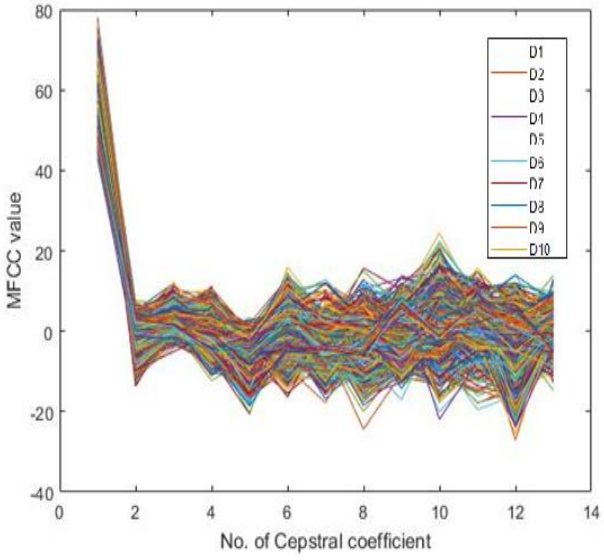

Fig 9: MFCC computation of 13 coefficients with L=22

The variation of the Lifter (L) parameters are studied with varying lifter values. The results below depicts the change in cepstrum coefficient with the change in L (Fig 10, 11, 12, 13).

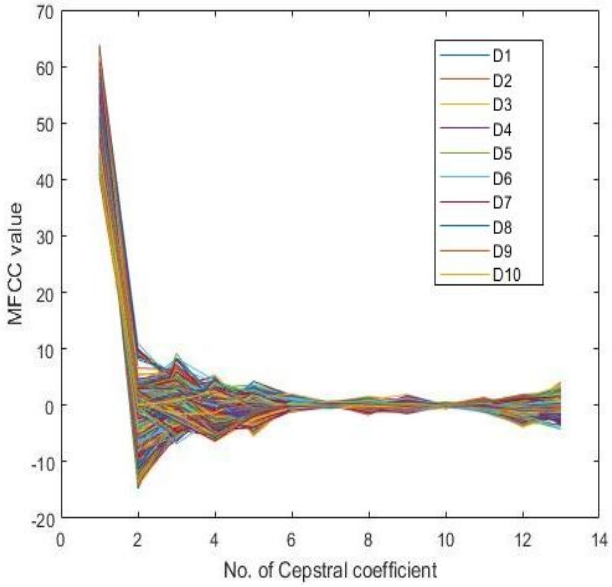

Fig 10: MFCC computation of 13 coefficients with L=5

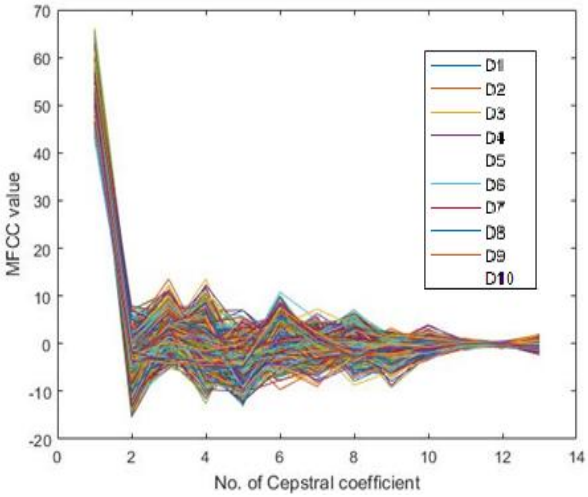

Fig 11: MFCC computation of 13 coefficients with L=10

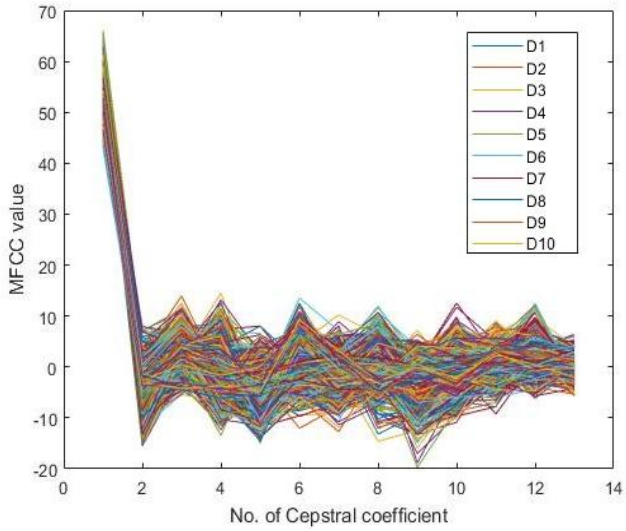

Fig 12: MFCC computation of 13 coefficients with L=15

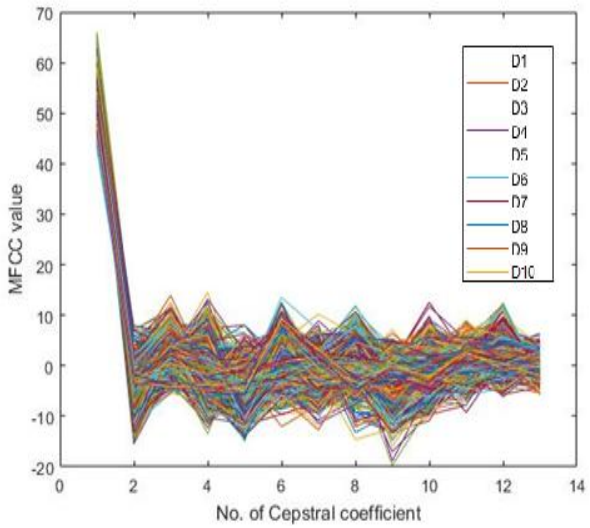

Fig 13: MFCC computation of 13 coefficients with L=20

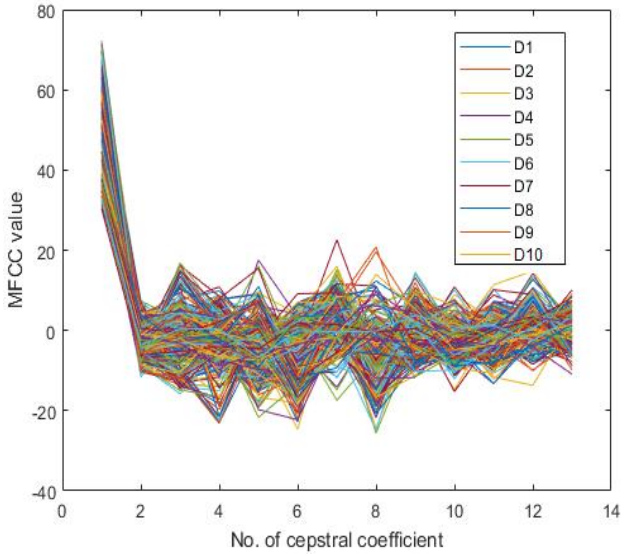

Fig 14: MFCC computation of 13 coefficients with L=25

As the L increases, the less weighted MFCCs are elevated, useful in enhancing the performance of MFCCs. The performance of the system employing MFCCs is hugely benefitting with precise MFCCs.

5. CONCLUSION

In an ever-evolving domain of speech processing, the use of Sanskrit in conjunction with SP has helped to study the spectral coefficients for intelligent applications. The research offers a study of acoustic speech processing on Mel frequency cepstral coefficient spectral components over a custom Sanskrit database. In MFCC, as the frequency increases the spectral resolution becomes lower, which is why the higher frequency information is down-sampled by the mel scale. The resulting MFCCs using Sanskrit speech provide distinguishable feature values useful in various home and industrial applications. The sinusoidal lifter provides greater performance applications for spectral features by providing less weight to the high cepstrum coefficient. A high lifter gives better performance for increasing the utility of MFCC's in various applications. The performance with MFCC will increase by using larger speech dataset. The features obtained by spectral processing using MFCC are used in speaker recognition, speech emotion detection, e-learning, customer relationships management, music therapy, security in home as well industrial areas etc. using several techniques comprising machine learning, artificial neural network, deep learning, etc. The limitation we have encountered so far is that a generic standard Sanskrit database is not yet available, hence a customized database is prepared.

In future, we intend to employ the spectral features in the application involving speech emotion recognition in Sanskrit. The spectral features obtained using MFCC are selected using feature selection technique based on the emotions being classified into different emotions i.e. happy, sad, disgust, fear, angry and excited. We aim at using different machine learning algorithms i.e. k-nearest neighbour (k-nn), support vector machine (svm) etc. for classifying emotions and comparing the accuracy of each algorithm. We also plan on following a hybrid approach to further improve upon the accuracy of classification.

6. REFERENCES

[1] S. Dhonde and S. Jagade, "Significance of frequency band selection of mfcc for text-independent speaker identification", In Proceedings of the International Conference on Data Engineering and Communication Technology, Springer International Publishing, pp 217 - 224, 2017.

[2] A. Benba, A. Jilbab, A. Hammouch, "Detecting patients with parkinson's disease with mel frequency cepstral coefficient and support vector machine", International Journal on Electrical Engineering and Informatics, vol. 7, pp 297-307, 2015.

[3] D. Desai and M. Joshi, "Speaker recognition using mfcc and hybrid model of vq and gmm", Recent Advances in Intelligent Informatics, Springer International Publishing, pp. 53-63, vol. 235, 2014.

[4] S. Casale, A. Russo, and G. Scebba, "Speech emotion classification using machine learning algorithms", IEEE International Conference on Semantic Computing, Santa Monica, 2008, pp. 158-165.

[5] M. Savargiv and A. Bastanfard, "Real-time speech emotion recognition by minimum number of features", IEEE conference on Artificial Intelligence and Robotics (IRANOPEN), Qazvin , 2016, pp. 72-76.

[6] N. Akrami, F. Noroozi, and G. Anbarjafari, "Speech-based emotion recognition and next reaction prediction", 25th Signal Processing and Communications Applications Conference, Antalya, 2017, pp. 1-6.

[7] Zhaoyan Zhang, "Mechanics of human voice production and control", The Journal of Acoustical Society of America, pp. 2614–2635, vol. 140, 2016.

[8] P. Bahadur, A. Jain, D. Chauhan, "Architecture of english to sanskrit machine translation", SAI Intelligent Systems Conference, London, 2015, pp. 616-624.

[9] S. Ladake and A. Gurjar, "Analysis and dissecton of sanskrit divine sound om using digital signal processing to study the science behind om chanting", 7th International Conference on Intelligent Systems, Modelling and Simulation, Bangkok, 2016, pp 169-173.

[10] J. Yao, and Y. Zhang, "Bionic wavelet transform ;A new time-frequency method based on an auditory model", IEEE Transaction on Biomedical Engineering, vol. 48, pp. 856-863, 2001.

[11] C. Pierrakeas, V. C. Georgopoulos and G. A. Malandraki "Online collaboration environments in telemedicine applications of speech therapy", In IEEE Proceedings Engineering in Medicine and Biology, pp 2183 – 2186, Shangai, 2005.

[12] R. Gamasu, "ECG based integrated mobile tele-medicine system for emergency health tribulations", International Journal of Biosci Biotechno, vol. 6, pp. 83-94, 2014.

[13] P.Y. Oudeyer, "The production and recognition of emotions in speech: features and algorithms", International Journal of Human-Computer Studies, vol. 59, pp. 157–183, 2003.

[14] S. Koolagudi and K. Rao "Emotion recognition from speech: a review", International Journal on Speech Technol, vol.15, pp.99-117, 2012.

[15] N. Jamil, F. Apand, and R. Hamzah, "Influences of age in emotion recognition of spontaneous speech a case of an under-resourced language", International Conference on Speech Technology and Human-Computer Dialogue, Bucharest, 2017, pp. 1-6.

[16] B. Logan, "Mel frquency cepstrum coefficient for music modeling", In Proceedings of International Synopsium on Music Information Retrieval, 2000.

[17] I. Trablesi and D. Ayad, "A multi-level data fusion for speaker identification on telephone speech", International Journal of Speech Processing, Image Processing and Pattern Recognition, vol. 6, pp. 33-42, 2012.

[18] W. Westera, K. Bahreini, and R. Nadolski, "Towards real-time speech emotion recognitionfor affective e-learning", Education and Information Technologies, vol. 21, no. 5, pp. 1367–1386, 2016.

[19] S. Gaikwad B. Gawali P. Yannawar, "A review on speech recognition technique", International Journal of Computer Applications, vol. 10, pp. 16-24, 2010.

[20] W. Ghai and N. Singh, " Literature review on automatic speech recognition", International Journal of Computer Applications, vol. 41, pp. 42-50, 2012.

[21] A. Vadwala, K. Suthar, Y. Karmakar and N. Pandya, "Survey paper on different speech recognition algorithm: challenges and techniques", International Journal of Computer Applications, vol. 175, pp. 31-36, 2017.

[22] I. Theodoras, C. N. Anaganostopoulous, I. Giannoukos, "Features and classifiers for emotion recognition from speech: a survey from 2000 to 2011", Artificial Intelligence Review, vol. 43, pp. 155-177, 2012.

[23] S. Wu, T. Falk, and W. Chan, "Automatic speech emotion recognition using modulation spectral features", Science Direct - Speech communication, vol. 53, pp. 768-785, 2011.

[24] Brian C, J. Moore, "The role of temporal fine structure processing in pitch perception, masking, and speech perception for normal-hearing and hearing-impaired people", Journal of the Association for Research in Otolaryngology, vol. 9, pp. 399–406, 2008.

[25] R. Rajoo and C.C. Aun, "Influences of languages in speech emotion recognition: a comparative study using malay, english and mandarin language", IEEE Symposium on Computer Applications & Industrial Electronics, Batu Feringghi, 2016, pp. 35-39.

[26] S. Sahoo N. Das, P. Sahoo "Word extraction from speech recognition using correlation coefficients", International Journal of Computer Applications, vol. 51, pp. 21-25, 2012.

[27] R. Singh, S. Arora "Automatic speech recognition: a review", International Journal of Computer Applications, vol. 60, pp. 34-44, 2012.

[28] J. Nicholson,, K. Takahashi, and R. Nakatsu, "Emotion recognition in speech using neural network", Neural Computing and Applications, Springer, vol. 9, pp. 290-296, 2000.

[29] A. Batliner, J. Buckow, H. Niemann, E. Noth, and Warnke, "Verbmobile Foundations of speech to speech translation". Springer, pp. 122-130, 2000.

[30] A. Fayjie, B. Kachari, M. Singh **"**A survey report on speech recognition system", International Journal of Computer Applications, vol. 121, 2015.

[31] N. Wasvani and S. Sharma, "Speech recognition system: A review", International Journal of Computer Applications, vol. 115, 2015.\

[32] N. Trivedi, S. Ahuja,V. Kumar, R. Chadha, S. Singh, "Speech recognition by wavelet analysis", International Journal of Computer Applications, vol. 15, 2011.