International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 09 Issue: 04 | Apr 2022 www.irjet.net p-ISSN: 2395-0072

DATABASES, FEATURES, CLASSIFIERS AND CHALLENGES IN AUTOMATIC SPEECH RECOGNITION: A REVIEW

Mahadevaswamy[1]

[1]Associate Professor
[2]Dept. of ECE, VVCE, Mysuru,
Visvesvaraya Technological University, Belagavi, Karnataka, India

---

Abstract— Speech research is marked as a most challenging areas among the several challenging research areas. The present literature is analyzed and projected to aid future investigations of the global speech research community. The problems concerned with the speech corpora, front end algorithms yielding to efficient speech representation, back end engines which outright chore of recognition, are projected in this study. Thirteen speech corpora are analyzed from the perspective of languages, time length, conviction of development and accessibility. The powerful techniques tending to rich features and robust speech recognition are enlightened in this research review.

Keywords— *Speech Databases; Energy; Fundamental frequency; Formants; Deep Neural Networks.*

Introduction

The pauses between the speech sounds of a speech signal portrays its unique characteristic distinguishes it from all other signals. The speech database created in uncontrolled conditions of the environment must be processed to implement a robust speech recognition engine. Human speech is an important and efficient tool of communication. The speech research drawn the attention of infinite researchers and has evolved as one of most fascinating research domain during the past few decades. Speech recognition is the technology of identifying the spoken word irrespective of the speaker. The speech recognition has been performed over the several languages. The UNESCO 2009 report presented that about 197 Indian languages are in critical situation of being extinct. According to Indian census report Percentage of people speaking local languages has drastically reduced [10]. The content flow of the paper is provided as follows, the section II rewards the readers with brief literature survey about traditional speech recognition methods. Section III describes about the acoustic parameters of speech. Section IV describes about the database. Section V describes about the feature extraction techniques. The classifiers described in section VI. The last section describes the challenges encountered by different class of speech recognition paradigms.

Review of Literature

Voice recognition system has been implemented for Assamese Language. The vocabulary size is 10 Assamese words, task of speech recognition is achieved by HMM classifier and I-vectors. A 39-dimensional feature vectors are derived using MFCC, first derivative and second derivative. The Novel Fusion technique outperforms the conventional techniques such as I-Vectors and HMM by achieving speech recognition accuracy of 100% [10]. Automatic speech recognition system is developed and evaluated using a moderate Bengali speech. Then 39-dimensional features are extracted and used to train triphone based HMM technique. The system was able to achieve an accuracy of 87.30% [17]. The speech recognition system is developed for Bangla accent. The Mel LPC features and their delta. The HMM modeling, lead to 98.11% recognition accuracy [18]. A Hindi isolated word recognition system is realized with LPC features and HMM Modeling and an accuracy of 97.14% was achieved corresponding to the word "teen" [19]. Another isolated word recognition system was realized with MFCC features and HTK Toolkit for Hindi language. An accuracy of 94.63% and a WER of 5.37% was achieved [20]. A speech system architecture is proposed for recognition of connected words in Hindi using MFCC features and HTK Toolkit. An accuracy of 87.01% was achieved [21]. An isolated digit recognition system was designed using MFCC features and HTK Toolkit for Malayalam isolated words to achieve an accuracy of 98.5% [22]. LPCC, MFCC, Delta-MFCC, Acceleration coefficients and vector quantization is utilized to build a speaker identification system to yield an accuracy of 96.59%. There was a boost in the performance of the system by 3.5% accuracy during testing stage with a consideration to text dependent system [23]. An automatic language identification task is achieved among five Indian languages. The languages selected were Hindi, Kannada, Telugu and Tamil. All the utterances are created from a subject set of five pairs consisting of equal male and female speakers. The Mel cepstral feature vectors are derived and vector quantization with codebook concept is used to achieve the task of classification. The system achieved 88.10% recognition performance with Kannada speech [24]. A speech system is implemented for Punjabi isolated words. The LPC feature vectors were extracted from speech signals. The DTW, VQ were employed for achieving speech recognition. Experiments were carried out for different code book sizes from 8 to 256. The system achieved 94% classification performance [25]. A speaker recognition system was developed for two speech databases. One speech database is created using microphone speech and other

speech database is telephone speech. MFCC features are used with the Linear discriminant Analysis technique, Co-variance Normalization. Support vector machines along with cosine distance scoring were used for the task of classification [26]. The speech signal is a complex signal has information of vocal tract and the source signal. To extract the source signal the speech signal is subjected to Linear Prediction and residual purely contains source signal information. The LP residual, Phase and Magnitude and Phase signals are subjected to signal processing operations at different levels such as segmental, subsegmental and suprasegmental to derive the source information. The Gaussian Mixture Models were able to perform the classification task [31]

Acoustic study of human speech

Human speech parameters fall into following classes, i.e., Acoustic parameters, auditory parameters and articulatory parameters. Human speech has variations in its quality. The causes for these quality variations are movement and position of articulatory parameters, such as lips, tongue, teeth's, vocal cord and actions related to lungs. The transmission of speech is governed by the acoustic features, pitch, formant frequencies, prolonged durations of the speech sounds and energy. The listeners task of reception and perception of speech belongs to auditory parameters [1].

Fundamental frequency

The speech production by human beings is a physiological process occurs due to the opening and closing action of vocal cords at a specific value of frequency called the fundamental frequency. This opening and closing action leads to the production of speech sounds. However, the fundamental frequency is different for different speech sounds uttered by the same speaker. It ranges between 85-155Hz for men, 165-255Hz for Women and approximately 300Hz for kids [1], [2], [3] The Figure 1 is the waveform representation of vowel sound 'a' in time domain. The speech sound is defined for about 0.29 secs. The Figure 2 is pitch of speech sound 'a'.

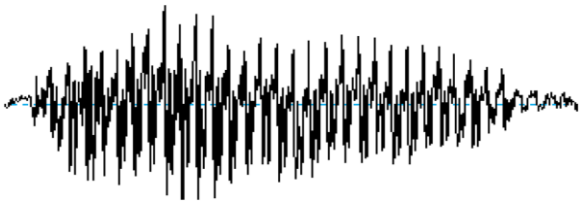

Fig. 1: Normal speech sound of vowel 'a'

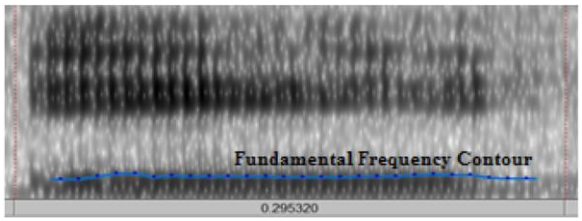

Fig. 2: Pitch of speech utterance 'a'.

Formant frequency

The air enters the lungs during breathing mechanism. The respiration system pushes air out of lungs. The air passes through the bronchi or trachea and strikes the vocal cords. If the vocal cords are in the tensed condition the vibration of these lead to the generation of quasi-periodic pulses of air, resulting into voiced sounds. If the vocal cords are in the relaxed condition the flow of air leads to the generation of unvoiced speech sounds. The vocal cords do not vibrate for the production of unvoiced speech sounds [4]. The frequency at which the vocal cords vibrate is called the formant frequency. The first few formant frequencies are sufficient to provide information of the uttered speech sounds [2], [5].

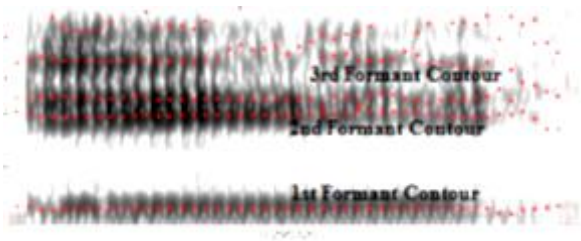

Fig. 3: Formants of sound 'a'

Energy of a human voice signal

The energy of the speech sound is the average energy [1]. Intensity of speech sound is the power per unit area. Power of the speech sound is the amount of signal energy utilized over given value of time. The power of the speech sound increases the speech intensity also increases and vice versa. Therefore, the energy directly proportional to the speech intensity. The average energy 62.65dB contained in the speech sound 'a' [1]

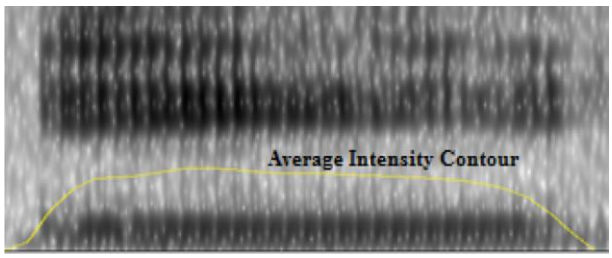

Fig. 4: Power per unit area of utterance 'a'

Databases

The databases play vital role in the speech research and development.

Dysarthric speech databases

The alphabets, digits [0-9] and the continuous speech consists of speech sounds recorded by reading paragraphs, passage and sentences [1]. The following are dysarthric

speech databases. Disordered speech data is collected from clients suffering from dysarthria of varying levels in Indian accent [1]. Nemours database is the same category of data in American English [7]. The UARD speech corpus is created for English language in American accent [8], [9]

LDC-IL

The CIIL is a central government organization located in Mysuru, Karnataka, India. The LDC-IL is a part of CIIL has created speech databases for several different Indian languages. Some of the languages selected for developing the speech corpus are, Assamese, Bengali, Bodo, Dogri, Gujarati, Hindi, Kannada, Kashmiri, Konkani, Maithili, Malayalam, Manipuri, Marathi, Nepali, Odia, Punjabi, Tamil, Telugu, Urdu. The speech corpus is created into speech corpora in segmented form and the speech corpora in annotated form. The speech corpora also has the pronunciation dictionaries. Few samples of several languages are publicly available [27]

LDC

The LDC is an open community of universities, corporate and governmental speech laboratories. LDC was established in the year 1992, to alleviate the problems of limited data for Research and Development activities pertaining to language technology. Some of the most popular databases in different foreign languages are available through formal license procedure through online purchase order. Few databases used for speech recognition research are, TIMIT, Web 1T 5gram Version 1, OntoNotes Release 5.0, CELEX2, Tree bank-3, The Newyork Times Annotated Corpus, TIDIGITS, Switchboard-1 Release 2, English Giga Word Fifth Edition and TIPSTER complete [28].

IITKGP-MLILSC

The IITKGP-MLILSC has speech data collected in 27 different Indian Regional Languages basically developed for language identification task. The data of language is about an hour duration has equal contribution by both male and female speakers [29].

OGI-MLTS

This speech corpus consists of the voice data of the following languages such as, English, French, Farsi, Mandarin Chinese, Vietnamese, Japanese, Korean, German, Hindi, Tamil and Spanish [30].

Indic speech

The IIIT-Hyderabad INDIC speech dataset [32] developed to motivate the speech language research. This database has speech signal files for few languages of Indian subcontinent like Kannada, Telugu, Tamil, Bengali, Hindi, Marathi and Malayalam. The complete description about this speech data is provided by Table 1

Table 1: Description of IIIT-H INDIC data

Sl. No.	Language	Region	No. of utterances	No. of speakers	Duration (min)
1	Kannada	Karnataka	1000	5	95
2	Telugu	Andra Pradesh	1000	5	110
3	Malayalam	Kerala	1000	5	103
4	Tamil	Tamilnadu	1000	5	92

The NITK Speech Database [34] has been developed for the task of Dravidian language classification. The complete details of this database are shown below in Table 2.

Table 2: Details of NITK speech dataset

Sl. No.	Language	Region	No. of utterances	Duration (min)
1	Kannada	Karnataka	1500	250
2	Telugu	Andra Pradesh	1500	250
3	Malayalam	Kerala	1500	250
4	Tamil	Tamilnadu	1500	250

Indic speech database is public-ally available for free to use in the academic research and development. The Indian Institute of Technology has speech corpus called Indic Speech Database for few languages Kannada, Telugu and Tamil. The IIT-M also has Indic dataset. The dataset has the recordings of speech utterances in seven languages such as Telugu, Hindi, Tamil, Kannada, Malayalam, Marathi and Bengali. The speech data was collected from four to five native speakers [32]. The Table 3, Table 4, Table 6 gives description of few databases.

FEATURES

Choosing relevant features for developing any automatic speech recognition system is most important, critical and challenging task. The choice of features is made in such a way that they represent the desired information. The speech information such as speaker, language, emotions, health state are represented by different features. Hence the speech researchers choose features on experimental means and also sometimes based on mathematical techniques like PCA (Principal Component Analysis). The speech features commonly used for ASR systems are spectral features.

The source features are derived by processing the residual signal extracted by using linear prediction technique. The vocal tract features are represented by the spectral features. The Table 5 indicates few important features.

Source feature vector

These are extracted from source information. The source information signal is isolated from the response of vocal apparatus in two steps: namely, Prediction of vocal tract signal, Filtering of vocal tract signal. In the first step the vocal tract signal is predicted form the filter coefficients and the speech signal, the vocal tract signal is then subjected to inverse filtering. The output of the inverse filter is linear prediction residual and this mostly contains the excitation source information [35], [36], [42]. The features derived from this error signal are known as sub-segmental features. These features are derived to carry out a study on opening and closing instants of glottis. The excitation source signal contains several information's such as speaker, meaning, language, emotional state of the speaker. The significance of excitation source information and Glottal activity detection are inferred [37]. Few good references which have used the sub-segmental features for implementing several speech systems are mentioned here. Extraction of the excitation source information contained in the speech signal is achieved using inverse filtering. Then by processing LP residual, phase and magnitude at different durations like segment level, sub-segment level and suprasegmental level the language specific excitation source information is computed. Residual Phase, Hilbert envelope, LP Residual, MFCCs are the input features and the GMMs are used to perform the classification task [31]. The Speaker Recognition task is achieved with the Pitch information derived from the LP Residual [38]. The task of Speaker Recognition is successfully completed using Energy of LP Residual [39]. The cepstral coefficients of the LP Residual are used to achieve the task of speaker identification [40]. Higher orders relations from LP Residual samples are employed for audio classification [41]. Though these were given less importance during the earlier days but this review study conveys that source feature vectors are very useful nowadays.

Segmental features

The speech signal has several important acoustic parameters like fundamental frequency, formants, spectral slope and energy. To create a room for better analysis of spectral features of the speech signal, it is subjected to Short Time Fourier Transform. The standard duration of a speech segment subjected to Short time Fourier transform lies between 20-30ms. The Fourier transformation of the logarithm of the magnitude spectrum of the speech signal is referred to as cepstrum [4]. A novel feature extraction technique based on fusion technique is proposed for isolated word recognition of Assamese language [10]. The predominant algorithms used to derive input signal vectors applied to train classifiers are spectral feature vectors such as LPCC, MFCC and PLP.

MFCC

MFCC features are employed to train Models [10]. MFCC feature extraction involves the few important steps.

- Pre-emphasizing: Human speech is allowed to pass through a filter that allows only high frequencies to pass through it.
- Framing: Speech data is segmented into overlapping frames.
- Windowing – To remove the tapering effect windowing operation is performed.
- MELSPEC – Mapping of linear frequency scale into Mel frequency scale is done.
- DCT is applied on logarithmic Mel frequencies to get MFCC features.

Classifiers

The task of classification in speech recognition depends on Several classifiers are used in speech research. Some of the classifiers used are Support Vector Machines, Sub-space Gaussian Mixture Models (SGMM), Hidden Markov Models (HMM), Deep Neural Networks (DNN), Hybrid classifiers like GMM-HMM, DNN-HMM so on, are also reported in the literature. Speech recognition models have been created using Vector quantization and I Vector techniques. The block diagram of word recognition model [10] in shown in Figure 5

HMM

The HMM is a sequential framework, in which output is a sequence of observations or acoustic signals and the sequence of words as hidden state sequence. The MFCC features of each word are applied as input to HMM. After all the HMM models are initialized Baum welch algorithm performs repeated computations till an optimal threshold is attained [10], [11], [13].

Vector Quantization (VQ)

A huge number of features are derived to create a training set of vectors. This is technique of mapping a very large vector space into compact space. The central point called codeword is calculated. A codebook consisting all the codewords representing distinct words is created. For each test vector a classification algorithm is utilized to find best match to testing vectors [10], [12], [14].

I-Vector Techniques

The advanced technique of I-vectors is due to the result of alteration to JFA [15], lower dimensional space is utilized to constitute the talker as well as channel variations [10].

Conclusion

The review carried out in this article clearly indicates that

- Speech recognition research community has attained ultimate success with the amalgamation of several front-end speech features with deep neural networks and probabilistic statistical models.
- Future ASR systems demand for reliable and consistent performance to unsolved challenges of local parlance, native language influence, accent, health conditions like dysarthria, emotional state of the speakers and noisy environmental conditions.
- Speech recognition architectures based on neural networks demand large time length speech data and
- Only few attempts are made to develop hybrid classifiers.

REFERENCES

[1] Giri, M.P. & Rayavarapu, N., Assessment on impact of various types of dysarthria on acoustic parameters of speech, Int J Speech Technol (2018) 21: 705. https://doi.org/10.1007/s10772-018-9539-0

[2] Patel. Identifying information bearing prosodic parameters in severely dysarthric speech (PhD thesis). Department of Speech Language Pathology, University of Toronto, 2000

[3] Tolba, H., & ElTorgoman, A. S. Towards the improvement of automatic recognition of dysarthric speech. In 2009 2nd IEEE International Conference on Computer Science and Information Technology, 2009, pp. 277–281.

[4] Rabiner, L. R., & Juang, B. H. Fundamentals of speech recognition. Englewood Cliffs: Prentice-Hall, 1993

[5] Mekyska, J., Rektorová, I., & Smékal, Z. (2011). Selection of optimal parameters for automatic analysis of speech disorders in Parkinson's disease. In 34th International Conference on Telecommunications and Signal Processing (TSP), pp. 408–412.

[6] Hosom, J., Kain, A., Mishra, T., Santen, J. P., Fried-Oken, M., & Staehely, J. Intelligibility of modifications to dysarthric speech. In IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2003, pp. 924–927.

[7] Menéndez-Pidal, X., Polikoff, J. B., Peters, S. M., Leonzio, J. E., & Bunnell, H. T. (1996). The Nemours Database of Dysarthric Speech. In Proceeding of Fourth International Conference on Spoken Language Processing- ICSLP'96, pp. 1962–1965.

[8] Kim, H., Hasegawa-Johnson, M., Perlman, A., Gunderson, J., Huang, T. S., Watkin, K., & Frame, S. Dysarthric speech database for universal access research. In Annual Conference of the International Speech Communication Association, INTERSPEECH 2008, pp. 2876–2879.

[9] Rudzicz, F., Namasivayam, A. K., & Wolff, T. The TORGO database of acoustic and articulatory speech from speakers with dysarthria. Language Resources and Evaluation, 2012, 46(4), 523–541.

[10] Bharali, S.S. & Kalita, S.K. "Speech recognition with reference to Assamese language using novel fusion technique", Int J Speech Technol 2018, 21: 251. https://doi.org/10.1007/s10772-018-9501-1

[11] Rabiner, L. R., & Juang, B. H. An introduction to hidden Markov Models. IEEE ASSP Magazine, 1986, 3(1), 4–1

[12] Rabiner, L. R., Levinson, S. E., & Sondhi, M. M. On the application of vector quantization and hidden Markov models to speaker-independent, isolated word recognition. Bell System Technical Journal, 1983, 62(4), 1075–1105.

[13] Rabiner, L. R. A tutorial on hidden Markov Models and selected applications in speech recognition. Proceedings of the IEEE, 1989, 77(2), 257–286.

[14] Soong, F. K., Rosenberg, A. E., Juang, B. H., & Rabiner, L. R. Report: A vector quantization approach to speaker recognition. AT&T Technical Journal, 1987, 66(2), 14–26.

[15] Verma, P., & Das, P. K. i-Vectors in speech processing applications: A survey. International Journal of Speech Technology, 2015, 18(4), 529–546.

[16] Dehak, N., Dehak, R., Kenny, P., Brümmer, N., Ouellet, P. Dumouchel, P. "Support vector machines versus fast scoring in the low-dimensional total variability space for speaker verification", vol. 9. In Interspeech 2009, Brighton.

[17] Hassan, F., Khan, M. S. A., Kotwal, M. R. A., & Huda, M. N. Gender independent bangia automatic speech recognition. In International Conference on Informatics, Electronics & Vision (ICIEV-2012).

[18] Muslima, U., & Islam, M. B. Experimental framework for melscaled LP based Bangla speech recognition. In 2013 IEEE 16th international conference on computer and information technology (ICCIT), Khulna 2014, (pp. 56–59).

[19] Pruthi, T., Saksena, S., & Das, P. K. Swaranjali: Isolated word recognition for Hindi language using VQ and HMM. In international conference on multimedia processing and systems (ICMPS), Chennai, 2000.

[20] Kumar, K., & Aggarwal, R. K. Hindi speech recognition system using HTK. International Journal of Computing and Business Research, 2011, 2(2), 2229–6166.

[21] Kumar, K., Aggarwal, R. K., & Jain, A. A Hindi speech recognition system for connected words using HTK. International Journal of Computational Systems Engineering, 2012, 1(1), 25–32.

[22] Kurian, C., & Balakrishnan, K. Speech recognition of Malayalam numbers. In IEEE World Congress on Nature & Biologically Inspired Computing, 2009. NaBIC 2009, Coimbatore (pp. 1475–1479).

[23] Bansal, P., Dev, A., & Jain, S. B. Automatic speaker identification using vector quantization. Asian Journal of Information Technology, 2007, 6(9), 938–942

[24] Balleda, J., Murthy, H. A., & Nagarajan, T. Language identification from short segments of speech. In Interspeech 2000, Beijing.

[25] Kumar, R., & Singh, M. Spoken isolated word recognition of Punjabi language using dynamic time warp technique. In Information systems for Indian languages. Berlin: Springer, 2011, (pp. 301–301)

[26] Senoussaoui, M., Kenny, P., Dehak, N., & Dumouchel, P. An I-vector extractor suitable for speaker recognition with both microphone and telephone speech. In Odyssey, Brno 2010.

[27] www.ldcil.org

[28] www.ldc.uppen.edu

[29] Maity, S., Vuppala, A.K., Rao, K.S., Nandi, D., IITKGP-MLILSC speech database for language identification. In: National Conference on Communication, 2012.

[30] Muthusamy, Y.K., Cole, R.A., Oshika, B.T., The OGI multi-language telephone speech corpus. In: Spoken Language Processing, pp. 895–898,1992.

[31] Dipanjan Nandi, Debadatta Pati, K. Sreenivasa Rao, Implicit processing of LP residual for language identification, Computer Speech & Language, Volume 41, 2017, Pages 68-87, ISSN 0885-2308, https://doi.org/10.1016/j.csl.2016.06.002.

[32] Kishore Prahallad , E. Naresh Kumar , Venkatesh Keri , S. Rajendran , Alan W Black, The IIIT-H Indic Speech Databases, 2012, http://speech.iiit.ac.in

[33] Dynamic Time Warping. In: Information Retrieval for Music and Motion. Springer, (2007), Berlin, Heidelberg

[34] Shashidhar G. Koolagudi, Akash Bharadwaj, Y. V. Srinivasa Murthy, Nishaanth Reddy, Priya Rao. "Dravidian language classification from speech signal using spectral and prosodic features", International Journal of Speech Technology, Volume 20 Issue 4, December 2017 Pages 1005-1016

[35] Makhoul, J. Linear prediction: A tutorial review. Proceedings of the IEEE, 1975, 63(4), 561–580.

[36] Krothapalli S.R., Koolagudi S.G, Speech Emotion Recognition: A Review. In: Emotion Recognition using Speech Features. Springer Briefs in Electrical and Computer Engineering (Springer Briefs in Speech Technology). Springer, New York, NY, 2013

[37] Kodukula, S. R. M. Significance of excitation source information for speech analysis. PhD thesis, Dept. of Computer Science, IIT, Madras. 2009.

[38] Atal, B. S. Automatic speaker recognition based on pitch contours. The Journal of the Acoustical Society of America, 1972, 52(6), 1687–1697.

[39] Wakita, H. Residual energy of linear prediction to vowel and speaker recognition. IEEE Transactions on Acoustics, Speech, and Signal Processing, 1976, 24, 270–271.

[40] Thevenaz, P., & Hugli, H. Usefulness of LPC residue in text independent speaker verification. Speech Communication, 1995, 17, 145–157.

[41] Bajpai, A., & Yegnanarayana, B. Exploring features for audio clip classification using LP residual and AANN models. In The international conference on intelligent sensing and information processing 2004 (ICISIP 2004), Chennai, India, Jan. 2004 (pp. 305–310).

[42] Springer Briefs in Electrical and Computer Engineering, 2015

Biography

Dr. Mahadevaswamy completed his BE degree from Visvesvaraya Technological University in the year 2010. M.Tech in Digital Signal Processing from Jain University in the year 2012 and Ph.D degree in the year 2021 from Visvesvaraya Technological University, Belagavi. He is presently working as Associate Professor in Department of Electronics and Communication Engineering, Vidyavardhaka College of Engineering, Mysuru. His area of research includes Digital Signal Processing and Speech Processing, Communication.

Table 3: Description of few important datasets

Type of Database	Languages	Duration in terms of no. of hours	Advantages	Disadvantages
Normal Eg.: IIITH-Indic Speech Dataset [32]	Bengali	1:39:00	• Useful for the speech community within India and abroad towards the development of speech systems in Indian languages • Developed for speech synthesis and can be used for speech recognition.	• Duration is less and hence difficult to train Deep neural networks (DNN).
	Hindi	1:12:00		
	Kannada	1:41:00		
	Malayalam	1:37:00		
	Marathi	1:56:00		
	Tamil	1:28:00		
	Telugu	1:31:00		
Dysarthric Speech (Spastic) [1]	Indian English (Isolated Words)	00:06:30	• Dysarthric Speech Research and Development, Eg: Improving the quality of Dysarthric speech using signal processing algorithms via study of acoustic parameters of Dysarthric speech.	• Data collection process needs training in Speech & Hearing Centers • Data collection process is constrained by factors like client health condition, client consent and guardian's consent, clients emotional state
Dysarthric Speech (Hypokinetic) [1]	Indian English (Isolated Words)	00:04:48		
Dysarthric Speech (Flaccid) [1]	Indian English (Isolated Words)	00:04:10		
Dysarthric Speech (Spastic) [1]	Indian English (Continuous speech)	00:08:20		
Dysarthric Speech (Hypokinetic) [1]	Indian English (Continuous speech)	00:03:45		
Dysarthric Speech (Flaccid) [1]	Indian English (Continuous speech)	00:03:00		
The Nemours Database of Dysarthric Speech [7]	English (American Accent)	814 sentences from 11 speakers with each speaker speaking 74 sentences	• Word level labelling is available • Useful for Dysarthric speech recognition	• Useful only for Dysarthric speech recognition • Limited speakers
Universal Access Research Database [8]	English (Foreign Accent) Isolated Words	765 isolated words per speaker (19 speakers)	• Freely available • Useful for Dysarthric speech recognition of people with neuromotor disability.	• Limited speakers
Linguistic data consortium for Indian language (LDC-IL) [27]	Assamese	28:18:56	• Speech Corpora Annotated Data for development of speech applications in Indian languages	• The data needs to segmented and labeled for developing ASR systems
	Bengali	36:24:39		
	Bodo	30:45:56		
	Gujarati	02:39:39		
	Hindi	80:01:48		
	Kannada	62:13:07		
	Kashmiri	06:28:25		
	Konkani	37:00:00		
	Maithili	30:10:40		
	Malayalam	92:40:43		
	Manipuri	109:48:27		
	Nepali	12:23:51		
	Odia	62:33:15		
	Punjabi	47:07:13		
	Tamil	58:06:16		
	Urdu	23:48:27		

Table 4: Description of LDC AND IITKGP-MLILSC

Type of Database	Languages	Duration in terms of no. of hours	Advantages	Disadvantages
Linguistic data consortium (LDC) [28]	TIMIT	04:00:00	• 6300 Sentences from 630 speakers • Phonetically rich sentences along with the label files	• Not freely available
	Chime3	342:00:00	• Noisy speech	• Not freely available
IIITKGP-MLILSPC	Arunachali	01:12:00	• Speech data of 27 Indian languages developed by IITKGP	• Limited time duration
	Assamese	01:07:33		
	Bengali	01:09:00		
	Bhojpuri	00:59:00		
	Chhattisgarhi	00:01:10		
	Dogri	00:01:10		
	Gojri	00:44:00		
	Gujarati	00:48:00		
	Hindi	02:14:00		
	Indian English	01:21:00		
	Kannada	01:09:00		
	Kashmiri	00:59:00		
	Konkani	00:50:00		
	Malayalam	01:21:00		
	Manipuri	01:00:00		
	Marathi	01:14:00		
	Mizo	00:48:00		
	Nagamese	01:00:00		
	Nepali	00:54:00		
	Oriya	00:59:00		
	Punjabi	01:20:00		
	Rajasthani	01:00:00		
	Sanskrit	01:10:00		
	Sindhi	00:50:00		
	Tamil	01:10:00		
	Telugu	01:13:00		
	Urdu	01:26:00		

Table 5: Review of existing features

Sl. No.	Features	Purpose and approach	Ref.
1	Fusion Technique	Recognition of Assamese Language Data	Sruti Sruba Bharali et al. 2018
2	MFCC features	Recognition of Bangla speech, gender independent task	F. Hassan et al. 2012
3	MEL scaled LPC features	Recognition of Bangla speech	U. Muslima et al. 2014
4	LPC features	Recognition of Hindi speech, speaker dependent task	Pruthi T et al. 2000

Table 6: Description of speech database

Sl. No.	Database	Languages	Duration in terms of no. of hours	Advantages	Disadvantages
1	Oregon Graduate Institute (OGI) Multi-Language Telephone-based Speech (MLTS)	English, Farsi, French, German, Japanese, Korean, Mandarin Chinese, Spanish, Tamil, and Vietnamese	01:23:00	• Corpus has Foreign Speech data • Language identification and multi-lingual speech recognition • Freely available	• Has data from few languages only • Only one Indian language data is available
2	CSLU: Kids' Speech Version 1.1	English	146:00:00	• Spontaneous and prompted speech • Can be used to train deep neural networks	• Only has children's speech
3	TI 46-Word	English	The 46-word vocabulary	• Isolated-word speaker-dependent technology	• Can be useful to build speaker dependent systems only • Not suitable for deep neural networks
4	TI-digits	English	326 speakers	• Digit and word recognition	• Can be used for isolated digit and word recognition only • Not appropriate for Deep Neural Networks
5	Switchboard-1 Release 2	English	260:00:00	• Speech recognition, speaker identification • Freely available	• Speech corpus has normal speech only