Voice Recognition for Gujarati Dialects: An in-depth Survey

Meera M. Shah
Department of Computer Science, Atmiya university, Rajkot, 360005, Gujarat

Hiren R. Kavathiya, PhD
Department of Computer Science, Atmiya university, Rajkot, 360005, Gujarat

ABSTRACT
Voice recognition technology nowadays is gaining so much importance, and plenty of work has been done on it for different languages like English, Arabic, Hindi, Chinese, etc. But when we talk about a language like Gujarati, we find a particular lack of work. In this paper, we examined the process of voice recognition in Gujarati. The systematic literature review for voice recognition has been shown here. This paper mainly focuses on the problems that can be found in voice recognition systems for Gujarati.

Keywords
Voice Recognition, Speech Processing, Gujrati, Feature Extraction, MFCC, HMM

1. INTRODUCTION
A user-friendly interface is provided by voice recognition systems to the user. Having it in natural languages will make it more beneficial. Voice recognition software allows people with impairments and those who are less at ease using machines due to lack of expertise or a language barrier to use technology. The user benefits from a convenient, hands-free environment thanks to voice recognition in native languages. The voice recognition approach is frequently used to address practical problems. The effectiveness and performance of speaker recognition systems are influenced by numerous factors. The task of creating an autonomous speaker recognition system is difficult because of the many grammatical conventions, noisy surroundings, and speaker pronunciations.

2. ANALYSIS OF RESEARCH WORK IN DIFFERENT LANGUAGES
The work done on speaker recognition for several languages in different situations, along with multiple feature extraction techniques, is summarized in the tables below.

Table 1: Analysis of Voice Recognition in different languages

Author	Dataset/ Language	Performance	Major Findings
[2]	Real Speech Dataset English	93.33% accuracy	Noisy Speech Signals Affects the performance
[3]	Voxceleb (English Language)	Speaker Identification 89.1% Accuracy & Speaker Verification EER 5.5%	Accuracy can be improved in terms of speech identification
[4]	Real Speech Dataset Hindi (10 M /5 F & 17 trails)	Text Independent : MFCC-VQ-77.64% / MFCC-GMM-86.27% Text Dependent : MFCC-VQ-85.49% / MFCC-GMM-94.12%	Accuracy can be improved in terms of text independent speech recognition
[5]	TIMIT (English)	6.94% EER	used Low dimensional Feature vectors
[1]	English(Voxceleb Dataset)	3.48% EER	—
[6]	Dataset: LDC-IL	96.2% for Speech Identification	Dataset comprising single words & phrases of adults.
[7]	Manually Collected Dataset of 30 speakers (10 F & 20 M)	Accuracy rate is 1% higher than traditional MFCC+GMM approach	Accuracy can be improved in terms of new approach
[8]	VoxCeleb2(6000 speakers' dataset)	Obtain 3.48% Equal Error Rate.	Determine if two given uncontrolled utterances originate from the same speaker or not.
[9]	Fisher (English, Arabian, Chinese) 4000 Speakers & 343 Hours speech signal	76.9% accuracy rate for individual voice segments, and 99.5% for each	Doubling the dataset can lead to accuracy improvement for the

		speaker as a bundle.	BiLSTM model.
[10]	sepedi Home Language	Accuracy: MLP: 97% RF: 99.9%	It was observed that MLPs performed well on the given dataset, however, Auto-WEKA selected Random Forest as the best algorithm
[11]	Manually Collected Database in noisy Environment	82% accuracy using MFCC model	In this paper, an automatic speech-speaker recognition system is implemented in a real time noisy environment.
[12]	THUYG-20	EER is 4.01%	Speaker identification for short utterances using English manual and THYUG dataset.
[13]	Dataset: Arabic	98.38% recognition rate	Dataset were recorded in office environment and only 5 fixed sentences are considered
[14]	2 different corpora of British English (adult, children)	17% and 31% relatively improvement over baseline	In this, the effectiveness of prosody-modification technique based on fuzzy classification of spectral bins is studied.
[15]	Dataset: Studio Recordings & Dialogs of Indian Language	Overall, 96.2% accuracy	ERIL is a multilingual emotion classifier, it is independent of any language
[16]	Manually Collected Audio-Visual dataset: 154 identities, 3 language annotations	Performance Degradation	Audio-Visual Speaker Recognition
[17]	Kaggle & Urdu Corpus (Regional Language of Pakistan)	vectors method demonstrates 80.4% FSR accuracy. With AC, it achieves 85.4% accuracy. With LI, its accuracy is 90.2%. Whereas by combining AC and LI it obtains 95.1% accuracy.	This new method is based on extracting accent and language information from short utterances.
[18]	Indian Languages	Accuracy Rate 98.34%	Identify not only frequency of voice but also textual features to improve accuracy I.e., a person can be angry with a slow voice.
[19]	English: Voxceleb1 (153516 audio files extracted from 1251 speakers)	95% accuracy (Top - I accuracy rate for voxceleb1.)	——

3. ANALYSIS OF RELATED WORK IN GUJARATI LANGUAGE

A Systematic survey has been done for a speaker recognition system developed for Gujarati language. the summary of this survey discussed here:

A model has been purposed for automatic speaker identification in Gujarati. Model includes two major processes: voice verification of speakers and identification. At registration phase voice model generated and stored on smart card that will be further used for verification. There also listing of some verification errors which include stress, time varying, aging, sickness etc. model presents phases that include feature selection and measures and pattern matching. Two types of models are included stochastic and template. The proposed model is implemented using MARF (Modular Audio Recognition Framework) which includes a collection of algorithms of sound, speech and natural language processing[20].

An algorithm that worked in different 13 languages like Gujarati, Assam, Punjabi, etc. includes the speaker recognition model like HMM and CNN and has an accuracy of 95.21% in both the environment text-dependent & text-independent but found language mismatch more pronounced in speaker verification.

The Speaker-independent systems that accept telephony commands in Gujarati that used speaker independent 29 words for the experiment. Implementation is done using HMM based speech recognizer Sphinx4 toolkit. In the whole experiment 20 speakers are involved, 10 female and 10 males with the age criteria between 20 to 30 years. Total 29 words were used for testing as a command. During the experiment, Average accuracy for female speakers was 83.79%. Average accuracy for male speakers found 80%. The minimum accuracy of the system is 72.41% and after taking extra care the system can achieve highest accuracy up to 96.55%. Average accuracy rate of all the words found 83.62%[21].

A multilingual model including Gujarati SPEAKERSTEW for identifying a speaker. The model capable of working with different 46 languages at the same time. That model work for text-independent speaker recognition systems with 73% accuracy[22].

The impact of online speaker adaptation on the performance of a speaker independent, continuous speech recognition system for Hindi language also been listed as a part of speaker recognition system. The speaker recognition is executed using the Maximum Likelihood Linear Regression (MLLR) transformation approach. The MLLR transform based speaker adaptation technique is found to significantly improve the accuracy of the Hindi ASR system by 3%. After the experiment they have concluded that MLLR transform based speaker adaptation of Hindi speech models indeed decreases the recognition error by a factor of 0.19.

A multilingual model for speaker recognition that includes all the Indian languages they used the MFCC method for feature extraction and built their own model for identifying the speaker. Modal also worked for emotions recognition for a person like a person can be angry while having a slow voice tone. The model provides 98.34% accuracy [18].

Based on the analysis, it can be inferred that numerous models have been created to identify various speech parameters, such as emotions, voice, and pitch, across different languages. However, when considering models specifically designed for recognizing speakers in vernacular Gujarati dialects, a notable deficiency in accuracy becomes apparent.

Despite significant efforts in speaker recognition for the Gujarati language, there remains an unexplored territory regarding accuracy in specific recognition across different environments and accents. It is emphasized that each region has its unique speech accents, which demand focused attention for achieving precise results. Furthermore, the challenges posed by the limitations and complexities of the Gujarati Framework underscore the extensive research opportunities in implementing speaker recognition systems.

4. CONCLUSION & FUTURE ENHANCEMENT

This paper presents the research activities done in the area of Gujarati voice Recognition using different platforms and experimenting the same on various models along with different sample sizes. The activity of Voice recognition involves taking sound in the form of input and providing text, which exactly matches the sound. Speaker recognition process deals with variability in an individual's speech, range, pitch, accent, style of speaking etc. The previous model might provide a higher accuracy rate for multi languages but the model specifically worked for Gujarati language has not had that much of accuracy rate. There is therefore an opportunity to design a Voice recognition system for Gujarati.

5. ACKNOWLEDGMENTS

I would like to thank my guide Dr. Hiren R. Kavathiya and all my colleges that helped me in my work.

6. REFERENCES

[1] Xu, J., Wang, X., Xu, S., & Liu, W. (2020). Deep multi-metric learning for text-independent speaker verification. *Neurocomputing*, *410*,394–400.https://doi.org/10.1016/j.neucom.2020.06.045

[2] Devi, K. J., Singh, N. H., & Thongam, K. (2020). Automatic Speaker Recognition from Speech Signals Using Self Organizing Feature Map and Hybrid Neural Network. *Microprocessors and Microsystems*, *79*, 103264. https://doi.org/10.1016/j.micpro.2020.103264

[3] Bian, T., Chen, F., & Xu, L. (2019). Self-attention-based speaker recognition using Cluster-Range Loss. *Neurocomputing*, *368*, 59–68. https://doi.org/10.1016/j.neucom.2019.08.046

[4] Maurya, A., Kumar, D. P., & Agarwal, R. (2018). Speaker Recognition for Hindi Speech Signal using MFCC-GMM Approach. *Procedia Computer Science*, *125*, 880–887. https://doi.org/10.1016/j.procs.2017.12.112

[5] Kinnunen, T., Karpov, E., & Fränti, P. (2006). Real-time speaker identification and verification. *IEEE Transactions on Audio, Speech, and Language Processing*, *14*(1), 277–288. https://doi.org/10.1109/tsa.2005.853206

[6] Gupta M, Singh RK, Singh S. G-Cocktail: An Algorithm to Address Cocktail Party Problem of Gujarati Language using CatBoost. Research Square; 2021. DOI: 10.21203/rs.3.rs-305722/v1.

[7] Patel, J. A., & Nandurbarkar, A. B. (2015). Development and Implementation of Algorithms for Speaker recognition for Gujarati Language. *International Research Journal of Engineering and Technology (IRJET)*.

[8] Xu, J., Wang, X., Xu, S., & Liu, W. (2020b). Deep multi-metric learning for text-independent speaker verification. *Neurocomputing*, *410*, 394–400. https://doi.org/10.1016/j.neucom.2020.06.045

[9] Hanifa, R. M., Isa, K., & Mohamad, S. (2021). A review on speaker recognition: Technology and challenges. *Computers & Electrical Engineering*, *90*, 107005. https://doi.org/10.1016/j.compeleceng.2021.107005

[10] Mokgonyane, T. B., Sefara, T. J., Modipa, T. I., Mogale, M. M., Manamela, M. J., & Manamela, P. J. (2019). Automatic Speaker Recognition System based on Machine Learning Algorithms. *2019 Southern African Universities Power Engineering Conference/Robotics and Mechatronics/Pattern Recognition Association of South Africa (SAUPEC/RobMech/PRASA)*. https://doi.org/10.1109/robomech.2019.8704837

[11] Kakade, M. N., & Salunke, D. B. (2020). An Automatic Real Time Speech-Speaker Recognition System: A Real Time Approach. *Lecture Notes in Electrical Engineering*, 151–158. https://doi.org/10.1007/978-981-13-8715-9_19

[12] Tiwari, V., Hashmi, M. S., Keskar, A. G., & Shivaprakash, N. C. (2019). Speaker identification using multi-modal I-vector approach for varying length speech in voice interactive systems. *Cognitive Systems Research*, *57*, 66–77. https://doi.org/10.1016/j.cogsys.2018.09.028

[13] Ghoniem, R. M., & Shaalan, K. (2017). A Novel Arabic Text-independent Speaker Verification System based on

Fuzzy Hidden Markov Model. *Procedia Computer Science*, *117*, 274–286. https://doi.org/10.1016/j.procs.2017.10.119

[14] Shahnawazuddin, S., Adiga, N., Sai, B. T., Ahmad, W., & Kathania, H. K. (2019). Developing speaker independent ASR system using limited data through prosody modification based on fuzzy classification of spectral bins. *Digital Signal Processing*, *93*, 34–42. https://doi.org/10.1016/j.dsp.2019.06.015

[15] Mehra, P., & Jain, P. (2021). ERIL: An Algorithm for Emotion Recognition from Indian Languages Using Machine Learning. *Wireless Personal Communications*. https://doi.org/10.21203/rs.3.rs-449758/v1

[16] Nawaz, S., Saeed, M., Morerio, P., Mahmood, A., Gallo, I., Yousaf, M. H., & Del Bue, A. (2021). Cross-modal Speaker Verification and Recognition: A Multilingual Perspective. *Computer Vision and Pattern Recognition*. https://doi.org/10.1109/cvprw53098.2021.00184

[17] Saleem, S., Subhan, F., Naseer, N., Bais, A., & Imtiaz, A. (2020). Forensic speaker recognition: A new method based on extracting accent and language information from short utterances. *Forensic Science International: Digital Investigation*, *34*, 300982. https://doi.org/10.1016/j.fsidi.2020.300982

[18] Mehra, P., & Verma, S. B. (2022). BERIS: An mBERT-based Emotion Recognition Algorithm from Indian Speech. *ACM Transactions on Asian and Low-Resource Language Information Processing*, *21*(5), 1–19. https://doi.org/10.1145/3517195

[19] Farsiani, S., Izadkhah, H., & Lotfi, S. (2022). An optimum end-to-end text-independent speaker identification system using convolutional neural networks. *Computers & Electrical Engineering*, *100*, 107882. https://doi.org/10.1016/j.compeleceng.2022.107882

[20] Patel H., Virparia P., - "Generic Model for Text Dependent Automatic Gujarati Speaker Recognition", International Journal of Emerging Trends & Technology in Computer Science, Vol. 1, Issue 3, September – October 2012

[21] Patel J., Patel P., and Virparia P., - "Voice Enabled Telephony Commands using Gujarati Speech Recognition", International Journal of Advanced Research in Computer Science and Software Engineering", Vol. 3, Issue 10, October 2013

[22] Chojnacka, R., Pelecanos, J., Wang, Q., Moreno, I.L. (2021) SpeakerStew: Scaling to Many Languages with a Triaged Multilingual Text-Dependent and Text-Independent Speaker Verification System. Proc. Interspeech 2021, 1064-1068, doi:10.21437/Interspeech.2021-646

[23] Purnima P., Bhatt S, - "Automatic Speech Recognition of Gujarati Digits using Dynamic Time Warping", International Journal of Engineering and Innovative Technology, Vol. 3, Issue 12, June 2014

[24] Rania M. Ghoniem, Khaled Shaalan, (2017), A Novel Arabic Text-independent Speaker Verification System based on Fuzzy Hidden Markov Model, Procedia Computer Science,Volume 117.

[25] Kharibam Jilenkumari Devi, Nangbam Herojit Singh, Khelchandra Thongam,(2017),Automatic Speaker Recognition from Speech Signals Using Self Organizing Feature Map and Hybrid Neural Network,Microprocessors and Microsystems,Volume 79,(2020).

[26] Tengyue Bian, Fangzhou Chen, Li Xu, (2019), Self-attention based speaker recognition using Cluster-Range Loss,Neurocomputing,Volume 368.

[27] Ankur Maurya, Divya Kumar, R.K. Agarwal, (2018), Speaker Recognition for Hindi Speech Signal using MFCC-GMM Approach,Procedia Computer Science,Volume 125.

[28] Shabnam Farsiani, Habib Izadkhah, Shahriar Lotfi, (2022), An optimum end-to-end text-independent speaker identification system using convolutional neural network,Computers and Electrical Engineering,Volume 100

[29] M. M. Kabir, M. F. Mridha, J. Shin, I. Jahan and A. Q. Ohi, "A Survey of Speaker Recognition: Fundamental Theories, Recognition Methods and Opportunities," in IEEE Access, vol. 9, 2021.

[30] Mohammad K. Nammous, Khalid Saeed, Paweł Kobojek,Using a small amount of text-independent speech data for a BiLSTM large-scale speaker identification approach, Journal of King Saud University - Computer and Information Sciences,Volume 34, Issue 3,2022.

[31] Sajid Saleem, Fazli Subhan, Noman Naseer, Abdul Bais, Ammara Imtiaz,Forensic speaker recognition: A new method based on extracting accent and language information from short utterances,Forensic Science International: Digital Investigation,Volume 34,2020.

[32] Monika Gupta, R K Singh, Sachin Singh et al. G-Cocktail: An Algorithm to Address Cocktail Party Problem of Gujarati Language using CatBoost, 17 March 2021.

[33] B. Pandey, A. Ranjan, R. Kumar and A. Shukla, "Multilingual speaker recognition using ANFIS," 2010 2nd International Conference on Signal Processing Systems, 2010.

[34] T. Kinnunen, E. Karpov and P. Franti, "Real-time speaker identification and verification," in IEEE Transactions on Audio, Speech, and Language Processing, vol. 14, no. 1, pp. 277-288, Jan. 2006.