INDIAN JOURNAL OF SCIENCE AND TECHNOLOGY

*Corresponding author.

naveenkumartr@rcub.ac.in

KHiTE: Multilingual Speech Acquisition to Monolingual Text Translation

Naveenkumar T Rudrappa[1]*, Mallamma V Reddy[2], M Hanumanthappa[3]

1 Research Scholar, Department of Computer Science, Rani Channamma University, Belagavi, Karnataka, India
2 Assistant Professor, Department of Computer Science, Rani Channamma University, Belagavi, Karnataka, India
3 Professor of Computer Science and Applications, Jnanabharati Campus, Bangalore University, Bangalore, Karnataka, India

Abstract

Objectives: To develop a system that accepts cross-lingual spoken reviews consisting of two to four languages, translate to target language text for Indic languages namely Kannada, Hindi, Telugu and/or English termed as cross lingual speech identification and text translation system. **Methods:** Hybridization of software engineering models are used in natural languages for pre-processing such as noise removal and speech splitting to obtain phonemes. Combinatorial models namely Hidden-Markov-Model, Artificial Neural Networks, Deep Neural Networks and Convulutional Neural Networks were deployed for direct and indirect speech mapping. Trained corpus consisting of thousand phonemes in the form of wave files for each language considered is named as KHiTEShabdanjali. The basic parameters cosidered for training dataset are pause, pitch, sampling frequency, threshold etc. **Findings:** The research has resulted in the development of mono-lingual and multi-lingual speech identification, tool for processing of cross-lingual speech and language identification, mono-lingual, bi-lingual, tri-lingual and quad-lingual speech to monolingual text translation for the four languages. It is a generic approach and can be used for other regional languages of India by training the corpus with the selected language. **Novelty:** Cross-lingual speech identification and text translation system helps users in e-shopping by reducing the time incurred in making a decision to purchase a product having enough features at an economical price, e-tutoring, e-farming activities, digitizing, defence etc.

Keywords: Artificial Neural Networks; Convulutional Neural Networks; Deep Neural Networks; HiddenMarkovModel; Speech Processing

1 Introduction

Exponential elevation in the number of World Wide Web users each year has led to an enormous amount of speech and text information to be deposited online every second. This requires shortening of the information without altering the genuine

meaning. Hence processing of natural languages is carried out through speech and text summarization. This is accomplished by utilizing subfields of natural language processing namely Artificial Intelligence (AI), Machine Learning (ML) and Deep Learning (AL).

Purchasing of products online offers a wide variety of options at a click of a button. The simplicity of requesting a commodity online either orally or typed and making it reach our home at a suitable date, time and location has attracted a number of individuals. The concessional offers introduced by web based shopping portals are making more individuals join online shopping. Most of them refer to the product reviews those are likely to be speech or text reviews or both prior to purchasing, analyze and then choose the best commodity having enough features at economical price. Let us assume that a customer wishes to purchase a smart phone. He goes through a wide variety of smart phones within his price range, visualizes various reviews for each smart phone and then selects the best. It is a complex and tedious task as some customer reviews are long and the significant meaning of it can only be justified after listening or reading to a complete review. These mandates minimizing each review to a shorter oral or textual or both which portrays similar meaning as that of the original review. This mandates understanding, processing and studying of spoken languages[1] like Kannada, Hindi, Telugu and English referred to as (KHiTE) to build AI applications. Many of the knowledge illustration and inference mechanisms like speech summarization, text summarization[2] have been applied. The research is an effort to illustrate multilingual speech processing challenges, challenges in text translation, techniques and applications to solve real-world speech problems for Indian languages and English[3].

Currently it is necessary to manage language away from individual statements like usage of more than one language in public speeches. Processing of spoken languages between man and machine involves a conversation in which control switches between the two entities. This leads to a development of an application involving a mixture of man-machine conversation like speech to text. Understanding of spoken languages involves parallel processing. Merging the two techniques from theoretical and practical aspects, a new system emerged to complete human-computer-human or computer-human-computer dialogue. This paper provides a glimpse on the theoretical and practical importance of speech production, challenges faced by researchers, applications and illustration of generic NLP tools.

Linguistics is the analysis of spoken or written language. It studies the three viewpoints of a language namely formation, meaning and the context of usage. Some of the terminologies in linguistics are Phonetics[4] is a study of acoustic and articulatory properties of speech, spoken speech is categorized into set of vowels and consonants, vowels[5] are sounds produced with an open vocal tract, no air pressure in glottis and consonants are sounds produced in vocal tract after reducing air in and out of lungs.

Some of the human speech[6] production terminologies are respiration that is inhaling and exhaling air to control vocal intensity and loudness, phonation is the determination of how voiced sounds are produced, articulation is the restriction of airflow in the vocal tract producing a word, oral/nasal resonance is the sound produced as it goes through the mouth/nose, prosody is the tone of speaker utterance that may be a question/command/irony/emphasis/inference, phoneme[7] smallest component that may cause a change in meaning, phone[8] is a basic unit of sound utterance, phonological responsiveness is capability of the audience to perceive words, phonemic responsiveness is the capability of the audience to recognize phonemes, phonemic transcripts use very few symbols of phonetics, a single phoneme for each, phonetic spelling is the confirmation of pronunciation of each single letter as a word, phonetic transcriptions are the illustrations of spoken speech sounds, phonetics[9] involves detailed analysis of human speech and its perception, acoustic phonetics[10] is the analysis of transmission of sounds from narrator to listener, phonology[11] is a subdivision of phonetics that analyzes phoneme pronunciation, articulatory phonetics[12] is a study of building spoken speech sounds by narrator, auditory phonetics[13] is the treatment and sensitivity of verbal communication, stress may be given to a syllable in a word for some words in a sentence, phonics is a subdivision of linguistics that is concerned with the spoken oral sound, phonotactics[14] is a branch for specifying the rules for a phoneme and articulators[15] are the movable speech organs in the speech production.

Some of the terminologies related to text[16] pre-processing are character unigram is a single letter, bigram is a two-letter chain sequence, trigram is a unique three-letter chain, n-gram is unique n-character long sequence of letters and n-gram frequency illustrates how frequently an n-gram chain repeats in some sample text.

Multilingual KHiTE processing of speech utterance involves challenges in acquisition, storage, retrieval and transmission as illustrated below:

1.1 Efficiency and performance of the input system for audio processing

Noise from background sources affects capacity of fault tolerance, reliability, accuracy, efficiency and performance of the audio input processing[17] system thus resulting in low output. Another issue is the synchronization of the user reaction with that of ready audio input device. Ex Users start speaking before the device is all set to accept voice input. Language syntax, separation of audio word boundaries and pronounced word meaning ambiguity also plays a major role as a word varies in meaning depending

upon the context in which it is spoken. Construction of a system that can understand and analyze the KHiTE languages is often challenging and difficult to achieve.

1.2 Memory Constraints in speech storage and retrieval

A simple wave file which consists of English alphabets starting from A to Z consumes 6176328 bytes or 5.88 megabytes of space on the file system. Similarly other languages considered for research like Kannada have ಸ್ವರಗಳು (Swaragalu), ವ್ಯಂಜನಗಳು (Vyanjanagalu), ವತ್ತಕ್ಷರಗಳು (Wattaksharagalu) and ಕಾಗುಣಿತಗಳು (Kaagunithagalu) and storage of these speech file increases exponentially as we move from ಕನ್ನಡ ವರ್ಣಮಾಲೆ (Kannada Alphabets) to ಶಬ್ದಗಳು (Shabdagalu, Words), ಶಬ್ದಗಳು (Shabdagalu, Words) to ವಾಕ್ಯಗಳು (Vakyagalu, Sentences), ವಾಕ್ಯಗಳು (Vakyagalu, Sentences) to ವಾಕ್ಯವೃಂದ (Vakyavranda, paragraphs) and from ವಾಕ್ಯವೃಂದ (Vakyavranda, paragraphs) to an entire ಸಂದರ್ಭ (Sandarbha known as a narrative context). Telugu has అక్షరమాల (Aksharamaala) which consists of అచ్చులు (acchulu vowels), హల్లులు (hallulu consonants) and storage of these speech file increases exponentially as we move from మాటలు (Maathulu words) to వక్యయలు (vakyalu sentences), to పేరలు (Pyaralu, Paragraphs) to an entire సందర్భం (sandarbham narrative context) and storage of Hindi speech file increases exponentially as we move from Hindi अक्षर (akshar, alphabets) to शब्द (shabd, words), शब्द (shabd, words) to वाक्य (vaakya, sentences), वाक्य (vaakya, sentences) to उपधारा (upadhara, paragraph) and उपधारा (upadhara, paragraph) to प्रसंग (prasanga, context). All the KHiTE languages considered for research require a large memory space to store spoken and textual words in their corresponding repositories in the file system and hence have a drawback of space complexity for limited memory devices.

1.3 Understanding of the Natural Languages KHiTE

 Building of a system that reads and understands the text is similar to a human-being who needs illustration of text through the usage of objects, agents, relationships, goals, beliefs and desires. Legitimate goal is to build and improve the system's ability to perform a pattern or string matching, incorporating common sense, prediction and lifelong learning capabilities. Ex: Self inferring common sense modules can be embedded into the model to detect if a vehicle has 2 wheels and a registration number then it has to be a bike or a scooter. The task of natural language understanding (NLU)[18] is a precondition for the generation of natural language. Incorporation of a psychological factor like emotion, understanding a methodology and to evaluate them is difficult unless we have a deeper understanding of a dialect and its basics. To make the models think one should imbibe the knowledge of neuroscience and cognitive science into our modules. Better systems can be obtained by performing interdisciplinary work. We can merge methodologies to obtain better results. Methodologies can retrieve data from well-structured sources like wikipedia with the overhead of larger computational power. Research improvements on textual emotion detection have lead to the development of emotion recognition systems having sensors. Ex: Smiley's in messaging.

1.4 Impact of low-resource availability on processing of Natural Language

There are many languages in India which are referred to as low-resourced[19] languages. Majority of the people around the globe speak a language that is low in resource where scripted data is not available. Scarcity of resources for processing of natural language data is an issue of concern where there is no availability of training or testing data. The challenge here is to build a generalized tool for multiple languages or a specific tool for each language as there are some common features between languages. There are a handful of researchers who work on low-resourced languages and multi-lingual platforms due to the non availability of datasets. Mono lingual data or word translation pairs are needed to evaluate the results and efficiency of multi-lingual platforms. These platforms do well in case of coarse-grained tasks like topic classification, but underperform in case of fine-grained tasks such as machine translation and hence they are the basic building blocks of unsupervised machine transformations. Question answering systems need a huge set of training and testing data for learning and hence development of low-resourced languages is at a slow pace. Educational growth does not rely on low-resourced languages but if multi-lingual systems become more persistent it may amplify the growth of low-resourced languages. Creating datasets for low resourced languages and making them freely available on an Open Source Platform would motivate people to lower the barrier. Testing and training data should be made available in different languages as they contribute to the evaluation of multi-lingual models. Researchers focus more on high-resourced languages as they are taught throughout the world. Ex: English.

1.5 Speech file analysis and text file analysis

Existing tools depend on repetitive neural networks, which cannot address lengthier contexts well. Another issue is analysis of multiple documents[20]. Ex Narrative story question answering dataset. Reading of a complete movie script or a book requires analysis and understanding of naturally spoken languages which mandates scaling up of existing resources. Difficulty lies in

whether we need better models or just train the existing model with more data. The issue is that supervision of large documents is sparse and costly to achieve. We could visualize a document level unaided task that requires anticipation of the subsequent passage or subsequent chapter of a book or a decision regarding the chapter that occurs next. Tracking of relevant information while reading a document and to materialize a context effectively is another matter of concern. Ex: Multi-document question answering and summarization. Development of a language model with long-lasting memory and learning strategies is the need of the hour.

1.6 Dataset, problem and valuation

The most serious issue is to appropriately characterize the actual problem of constructing datasets[21] and assessment methodology that are proper to quantify progress. The actual difficulty is to handle the low-resource issue making translation in education sector to enable people access information in their regional language. The research has utilized Shabdanjali an online bilingual dictionary for English to Hindi translation and developed a benchmark in creation of KHiTE multilingual vocal data set for ringing physical temple bell through multilingual vocal command.

2 Methodology

The proposed methodology[22] has resulted in the development of an incremental prototype named as cross lingual speech identification, retrieval and translation (CLSIRT) by utilizing readily available open source software and hardware. The sub-system performs language identification by utilizing a combinatorial model consisting of Hidden-Markov-Model (HMM), Gaussian-Mixture-Model (GMM), Artificial-Neural-Networks (ANN), Deep-Neural-Networks (DNN) for improving speech matching accuracy. HMM-GMM performs phoneme recognition, emission distribution is modeled using GMM by calculating mean and covariance which increases the probability of fetching a sequence of phonemes. HMM-ANN: HMM is used to obtain the probability of the data under observation for an HMM state that corresponds to a specific sound. ANN training produces posterior probabilities of HMM state given the speech data. HMM-DNN: HMM is used for phoneme recognition while training of a DNN is carried out in two phases, phase 1 which consists of unsupervised pretraining and phase 2 which consists of supervised fine-tuning. MLSI imbibes a Hybrid/HiTEK dictionary consisting of multi lingual Speech/Text. The output of the developed methodology is identification of spoken speech language in real time and in a recorded wave file as shown in Figure 1.
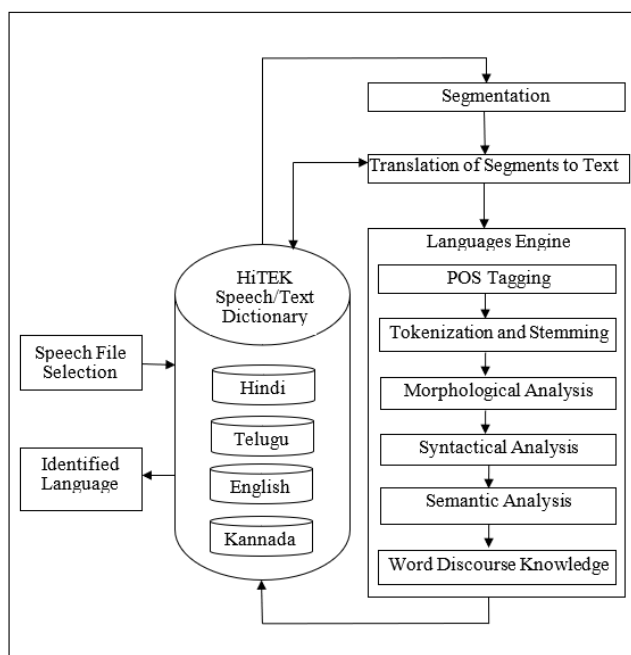

Fig 1. Block diagram of speech language identification

The developed sub-system provides a user-friendly graphical interface[22] to record and store small, medium and large speech files of user defined duration on a large storage media in the form of wav file with .wav file extension and can be utilized for further processing. This graphical user interface has laid a benchmark in the field of NLP to capture human speech for a variety of research applications for upcoming researchers as shown in Figure 2.
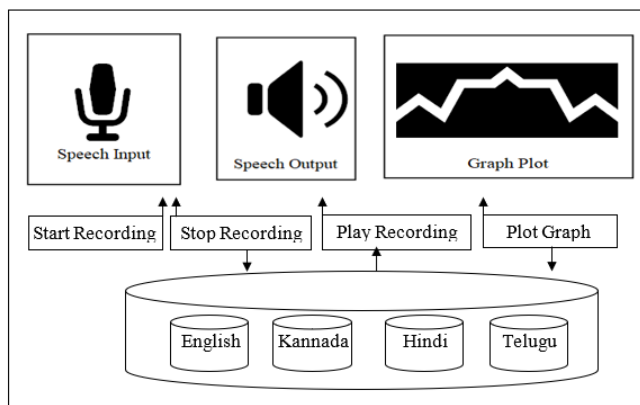

Fig 2. FOSS framework for speech acquisition and storage

The developed sub-system performs monolingual and multilingual speech language identification and translates it to a selected target language. It is a hybridized mixture of a software engineering methodology called the speech information retrieval cycle, dictionary based translation, cognate matching, document translation to speech, interlingual techniques that incorporate text converted to speech and spoken speech query, query translation due to the development and enhancement of multilingual HiTEK dictionaries from scratch named as (MLCLSRT)[23] which inculcates word co-occurrence pattern and a greedy algorithm for phone translation using cohesion technique. It utilizes multilingual dictionary to implement cross language speech retrieval. Dictionary based speech query translation increases the machine readability for multilingual phones. Dictionary based translation suffers with the selection of the best translated phones from the dictionary, if the required phone is not present in the dictionary, may result in inaccurate translation and hence the phone need to be transliterated for accurate results as shown in Figure 3.
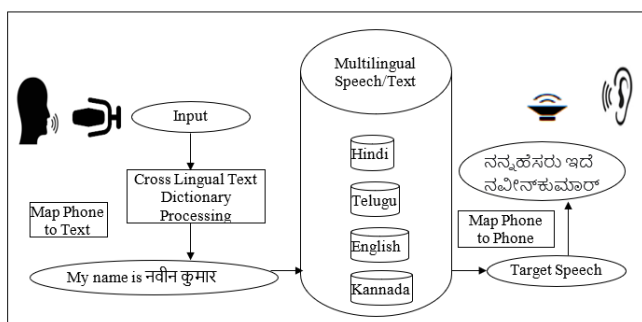

Fig 3. Input Speech extraction and phone-to-phone translation

3 Results and Discussion

World Wide Web earlier consisted of only textual data. Availability of internet at remote geographical locations has resulted in the hosting and updating a large quantity of audio and text on the web which needs language expertise in processing by humans and machines. This lead to the development of cross lingual speech to text translation system (CLSTT) as shown in Figures 4, 5, 6, 7 and 8.

Some of the applications of the developed system are analysis of oral survey, voice interaction chat-bots, vocal news classification, farmer query system, spelling and grammar checking, language translator, biometric speech authentication,

Fig 4. Screenshot of developed graphical user interface for HiTEK speech to text system

Fig 5. Screenshot output of text language identification

search auto correct and auto complete, social media monitoring, text summarization. The evolution of the system resulted in a byproduct named as physically operated vocal temple bell for COVID-19 [24] pandemic. This real time application deploys code switching technique, where the devotees of the temple can ring the bell by uttering the command in monolingual or cross lingual speech restricted to KHiTE set of languages as shown in Figure 9.
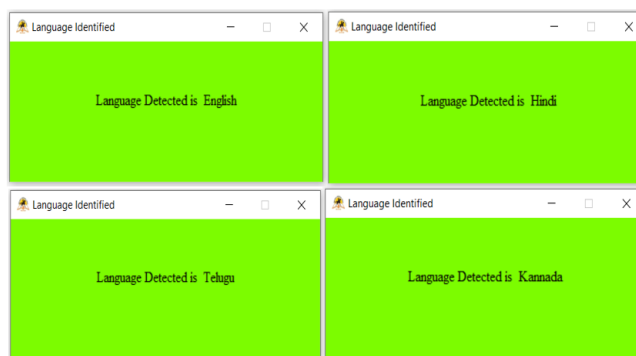

Fig 6. Screenshotof speech input and text output for monolingual identification system
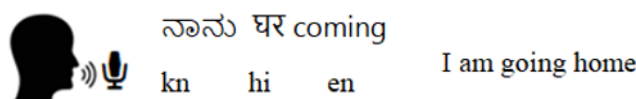

Fig 7. Trilingual speech to text where kn is Kannada, hi is Hindi and en is English
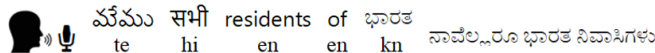
Fig 8. Quadlingual speech to text where te is Telugu, hi is Hindi, en is English and kn is Kannada

Fig 9. Speech commands to vocal temple bell, utilizationof cross lingual speech acquisition

4 Conclusion

Geographical linguistic diversity in India has escalated the usage of internet with a need to deliver audio data specifically for illiterates in native languages. This paper focuses on text and speech analysis, understanding, generation, preprocessing and storage of grammar and vocabulary for different domains of four languages. The recent developments in speech processing, its significance, challenges, applications and tools used were described. The developed Cross lingual speech to text translation tool has leveraged free open source software development producing a benchmark prototype for speech researchers across various fields like medical analysis of speech impaired children for correction, business sections to generate high profit making customer interactions, improve and fortify client relationships, elevate company sales, enhance verbal communication, employee motivation and leadership qualities like pubic speaking. The system can be enhanced for other regional languages of India and across the World.

Acknowledgement

References

1) Locke S, Bashall A, Al-Adely S, Moore J, Wilson A, Kitchen GB. Natural language processing in medicine: A review. *Trends in Anaesthesia and Critical Care*. 2021;38:4–9. Available from: https://doi.org/10.1016/j.tacc.2021.02.007.

2) Boorugu R, Ramesh G. A Survey on NLP based Text Summarization for Summarizing Product Reviews. In: 2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA). IEEE. 2020;p. 352–356. Available from: https://doi.org/10.1109/ICIRCA48905.2020.9183355.

3) Peguda J, Santosh VSS, Vijayalata Y, Mounish V. Speech to Sign Language Translation for Indian Languages. In: 2022 8th International Conference on Advanced Computing and Communication Systems (ICACCS). IEEE. 2022;p. 1131–1135. Available from: https://doi.org/10.1109/ICACCS54159.2022.9784996.

4) Zue V, Glass J, Goodine D, Leung H, Phillips M, Polifroni J, et al. Integration of speech recognition and natural language processing in the MIT VOYAGER system. In: ICASSP 91: 1991 International Conference on Acoustics, Speech, and Signal Processing;vol. 91. IEEE. 1991;p. 713–716. Available from: https://doi.org/10.1109/ICASSP.1991.150439.

5) Li S, You J, Zhang X. Overview and Analysis of Speech Recognition. In: 2022 IEEE International Conference on Advances in Electrical Engineering and Computer Applications (AEECA). IEEE. 2022;p. 391–395. Available from: https://doi.org/10.1109/AEECA55500.2022.9919050.

6) Zhu C, An K, Zheng H, Ou Z. Multilingual and Crosslingual Speech Recognition Using Phonological-Vector Based Phone Embeddings. In: 2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU). IEEE. 2021;p. 1034–1041. Available from: https://doi.org/10.1109/ASRU51503.2021.9687966.

7) Kumar PS, Jayanna HS. Performance Analysis of Hybrid Automatic Continuous Speech Recognition Framework for Kannada Dialect. In: 2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT). 2019;p. 1–6. Available from: https://doi.org/10.1109/ICCCNT45670.2019.8944767.

8) Sharma RS, Paladugu SH, Priya KJ, Gupta D. Speech Recognition in Kannada using HTK and Julius: A Comparative Study. In: 2019 International Conference on Communication and Signal Processing (ICCSP). IEEE. 2019;p. 68. Available from: https://doi.org/10.1109/ICCSP.2019.8698039.

9) Dong L, Guo ZQ, Tan CH, Hu YJ, Jiang Y, Ling ZH. Neural Grapheme-To-Phoneme Conversion with Pre-Trained Grapheme Models. In: ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE. 2022;p. 6202–6206. Available from: https://doi.org/10.1109/ICASSP43922.2022.9746447.

10) Khurana D, Koli A, Khatter K, Singh S. Natural language processing: state of the art, current trends and challenges. *Multimedia Tools and Applications*. 2023;82(3):3713–3744. Available from: https://doi.org/10.1007/s11042-022-13428-4.

11) Zhang X, Peng Y, Xu X. An Overview of Speech Recognition Technology. In: 2019 4th International Conference on Control, Robotics and Cybernetics (CRC). IEEE. 2019;p. 81–85. Available from: https://doi.org/10.1109/CRC.2019.00025.

12) Brennan A, Mcdonagh T, Dempsey M, Mcavoy J. Cosmic Sounds: A Game to Support Phonological Awareness Skills for Children With Dyslexia. *IEEE Transactions on Learning Technologies*. 2022;15(3):301–310. Available from: https://doi.org/10.1109/TLT.2022.3170231.

13) Ahmed T, Suffian M, Khan MY, Bogliolo A. Discovering Lexical Similarity Using Articulatory Feature-Based Phonetic Edit Distance. *IEEE Access*. 2022;10:1533–1544. Available from: https://doi.org/10.1109/ACCESS.2021.3137905.

14) Chodroff E, Wilson C. Acoustic–phonetic and auditory mechanisms of adaptation in the perception of sibilant fricatives. *Attention, Perception, & Psychophysics*. 2020;82(4):2027–2048. Available from: https://doi.org/10.3758/s13414-019-01894-2.

15) Romero D, D'haro LF, Estecha-Garitagoitia M, Salamea C. Phonotactic Language Recognition Using A Universal Phoneme Recognizer and A Transformer Architecture. In: ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE. 2022;p. 6872–6876. Available from: https://doi.org/10.1109/ICASSP43922.2022.9746459.

16) N P, Kini S. Phoneme based Kannada Speech Corpus for Automatic Speech Recognition System. In: 2022 IEEE International Conference on Distributed Computing and Electrical Circuits and Electronics (ICDCECE). 2022;p. 1–5. Available from: https://doi.org/10.1109/ICDCECE53908.2022.9793010.

17) Yu F, Yao Z, Wang X, An K, Xie L, Ou Z, et al. The SLT 2021 Children Speech Recognition Challenge: Open Datasets, Rules and Baselines. In: 2021 IEEE Spoken Language Technology Workshop (SLT). IEEE. 2021;p. 1117–1123. Available from: https://doi.org/10.1109/SLT48900.2021.9383608.

18) Xie S, Li L. Improvement and Application of Deep Belief Network Based on Sparrow Search Algorithm. In: 2021 IEEE International Conference on Advances in Electrical Engineering and Computer Applications (AEECA). IEEE. 2021;p. 705–708. Available from: https://doi.org/10.1109/AEECA52519.2021.9574138.

19) Nadimpalli VLV, Kesiraju S, Banka R, Kethireddy R, Gangashetty SV. Resources and Benchmarks for Keyword Search in Spoken Audio From Low-Resource Indian Languages. *IEEE Access*. 2022;10:34789–34799. Available from: https://doi.org/10.1109/ACCESS.2022.3162854.

20) Black D, Rapos EJ, Stephan M. Voice-Driven Modeling: Software Modeling Using Automated Speech Recognition. In: 2019 ACM/IEEE 22nd International Conference on Model Driven Engineering Languages and Systems Companion (MODELS-C). IEEE. 2019;p. 252–258. Available from: https://doi.org/10.1109/MODELS-C.2019.00040.

21) Sulaimani SA, Starkey A. Short Text Classification Using Contextual Analysis. *IEEE Access*. 2021;9:149619–149629. Available from: https://doi.org/10.1109/ACCESS.2021.3125768.

22) Rudrappa NT, Reddy MV, Hanumanthappa M. Standardized GUI Framework Using Python for Speech Processing: NLP. *2021 International Conference on Electrical, Computer and Energy Technologies (ICECET)*. 2021;p. 1–4. Available from: https://doi.org/10.1109/ICECET52533.2021.9698522.

23) Rudrappa NT, Reddy MV, Hitek. HiTEK: Multilingual Speech Identification Using Combinatorial Model. In: International Conference On "Advances in Computer Vision and Artificial Intelligence Technologies ACVAIT. 2022.

24) Reddy MV, Rudrappa NT. Advanced Code Switching Application-COVID-19. *SSRN Electronic Journal*. 2020. Available from: http://dx.doi.org/10.2139/ssrn.3735927.