ADV MATH SCI JOURNAL
Advances in Mathematics: Scientific Journal **9** (2020), no.6, 3963–3970
ISSN: 1857-8365 (printed); 1857-8438 (electronic)
https://doi.org/10.37418/amsj.9.6.77 Spec Issiue on ICAML-2020

# MULTILINGUAL SPEECH TO TEXT CONVERSION - A REVIEW

SALONI[1] AND W. SINGH

ABSTRACT. Speech is the first major primary need and the most convenient means of communication among individuals. Automatic Speech Recognition (ASR) introduces natural phenomena for man-machine communication. Speech recognition systems allow users to use speech as another form of input to communicate easily and efficiently with applications. A detailed study on automatic speech recognition is carried out and this paper offers an overview of the major technological perspective and appreciation of the fundamental progress of multilingual translation of speech-to-text conversion and also provides overview technique developed in each stage of speech-to-text conversion classification. The goal of this review paper is to recapitulate and match different speech recognition systems and approaches for the conversion of multilingual speech to text.

## 1. INTRODUCTION

Speech is the most normal mode of human communication and speech processing has been one of the most exciting research areas for signal processing, [1]. Speech processing is the study of these signals, speech patterns and processing methods. Automatic Speech Recognition provides a medium used by humans and machines for natural communication. The main purpose of speech recognition is to translate to produce a set of words the acoustic signal received

[1]*corresponding author*

2010 *Mathematics Subject Classification.* 68T10, 68T50.

*Key words and phrases.* Automatic Speech Recognition, speech-to-text conversion system (STT), multilingual, end-to-end (E2E) system and feature extraction tools and techniques.

from a microphone or a phone. Language technology can provide solutions in the form of ordinary interfaces so that digital content can reach the masses and promote information exchange between various people who speak different languages, [2].

1.1. **Speech Types:** Speech recognition system can be distinguished by defining what kind of utterances they can identify in different classes, [3]. These are as follows:

Isolated Word, [4], Connected Word [5], Continuous Speech [6], Spontaneous Speech [7].

1.2. **ASR models Based on Speakers:** Due to the unique physical form and personalities all speakers have their own voices. The voice recognition technology is commonly divided into major categories based on speaker types, namely, speaker-dependent and speaker-independent [3, 8].

1.3. **Vocabulary Types:** The size of a speech recognition system's vocabulary determines the system's complexity, processing needs, efficiency and precision. Many programs need just a few terms (e.g. numbers only), others require very broad dictionaries (e.g., machines with direction). The forms of vocabulary may be categorized in ASR systems as Small Vocabulary (10 words), Large Vocabulary (1,000 words), Very-large Vocabulary (10,000 words), Out-of-Vocabulary (Mapping a phrase to unknown word from the vocabulary).

## 2. Automatic Speech Recognition (ASR)

ASR [9,10] performs automatic transformation of an acoustic input speech signal into a transcription of a text. Extraction of the feature is accomplished by changing the speech waveform to a form of parametric representation for subsequent processing and analysis at a relatively minimized data rate. Acceptable classification is therefore derived from performance and consistency characteristics. The speech function extraction techniques are Mel Frequency Cepstral Coefficients (MFCC), Linear Prediction Coefficients (LPCC), Linear Prediction Cepstral Coefficients (LPCC), Linear Spectral Frequencies (LSF), Discrete Wavelet Transform (DWT), and Perceptual Linear Prediction (PLP).

## 3. Multilingual Speech to Text Conversion

Training a traditional automatic speech recognition (ASR) program to support multiple languages is difficult, as the sub-word structure, lexicon and word inventories are usually unique to language. Sequence-to-sequence models, on the other hand, are well adapted for multilingual ASR as they encapsulate an acoustic, pronunciation and language model together within a single network. Multilingual end-to-end (E2E) models showed great promise in extending automatic speech recognition (ASR) coverage of languages around the world.

## 4. Research Process

The study intended to be a systematic review meets Shivakumar rules, [11]. The research process includes finding, discovering, assessing and reviewing the knowledge you need to support your research question and then creating and bringing forward your ideas. An important step in carrying out a thorough research or study is to understand the research process. The following queries about the study are listed as important for our purpose:

1) RQ.1: How many papers in the field of signal processing address the various categories of devices and speech-to-text systems?

To answer this question, table I explains the number of publications in this field.

2) RQ.2: Which types of categories are chosen to complete the evaluation scheme?

To answer this question, section 4 of the discussion explains the related categories that are selected on the concept of facets.

### 4.1. **Strategy for Identification and Screening:**

All the steps that were taken in the systematic review analysis are shown in Figure 1.

1) Information Sources

Using following online resources in this investigation to scan for important studies:

(i) **Sringer** (www.springerlink.com)
(ii) **IEEE explore** (http://ieeexplore.ieee.org)
(iii) **Science Direct** (www.sciencedirect.com)
(iv) **ACM Digital Library** (ww.acm.org)

TABLE 1. Search Selection

| No. | E-Resource | Studies Returned | Excluded based on | | | Keyword used |
|---|---|---|---|---|---|---|
| | | | title | abstract | full text | |
| 1 | ieeexplore.ieee.org | 42 | 16 | 12 | 6 | Speech Recognition, speech to Text, feature extraction tools and technologies |
| 2 | www.acm.org | 28 | 13 | 2 | 4 | Speech Recognition, speech to Text, feature extraction tools and technologies |
| 3 | www.sciencedirect.com | 11 | 3 | 4 | 1 | Speech Recognition, speech to Text, feature extraction tools and technologies |
| 4 | www.springerlink.com | 9 | 3 | 0 | 2 | Speech Recognition, speech to Text, feature extraction tools and technologies |
| 5 | www3.interscience.wiley.com | 30 | 9 | 6 | 5 | Speech Recognition, speech to Text, feature extraction tools and technologies |

(v) **Wiley Interscience** (www3.interscience.wiley.org)

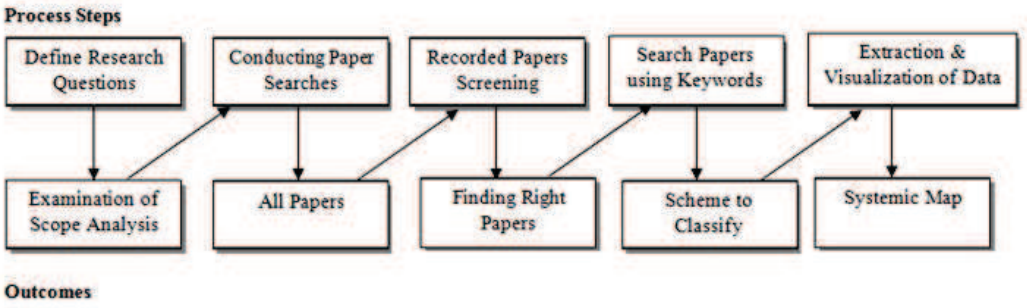


FIGURE 1. Searching Process

**2) Selection of sample**

We are inclined to try to extract the maximum amount of relevant literature to support the completeness of the analysis as possible as shown in table 1. 120 papers are returned during this review, 24 excluded on the basis of abstracts, 44 excluded on the basis of title and 18 excluded on the basis of full text.

## 5. CLASSIFICATION SCHEME

Of all the listed articles, three dimensions are grouped into different categories. The factors are the model for speakers, devices (used in the selected articles), and the method of speech recognition including Feature extraction

and feature classification techniques. The comprehensive description of the categories used in each dimension is given below:

1) Tool Facet

(i) **Hidden Markov Toolkit:** Hidden Markov Model Toolkit (HTK) is a portable toolkit designed to build and manipulate hidden Markov models. This toolkit is used to build speech recognizer based on words consisting of phases of data planning, data training, data processing, and data analysis, using different commands built into this toolkit, [12].

(ii) **Windows Speech Recognition:** In its Windows operating system features, Microsoft Corporation developed Speech Application Program Interface (SAPI) for speech-related works for various languages, [13].

(iii) **CMU Sphinx:** The general term for defining a group of speech recognition systems developed at Carnegie Mellon University is CMU Sphinx, also called Sphinx in short. These include a series of speech recognizers (Sphinx 2-4) and a SphinxTrain acoustic model teacher. The components of a Speech Recognition System are Language Model, Dictionary Acoustic Model, [11].

(iv) **Praat:** PRAAT is a computer program for discourse analysis, synthesis, and manipulation. A PRAAT utility is used to mechanically mark speech files with transcription files, and two folders are generated; clean and incorrect, [14].

(v) **Kaldi:** Kaldi is an open-source speech recognition toolkit for speech recognition and signal processing written in C++, freely available under the Apache License v2.0 which recognizes voice. The Kaldi training and testing process uses deep neural networks for acoustic simulation and with models of Gaussian mixtures, [7].

2) FEATURE EXTRACTION TECHNIQUES

(i) **Linear Predictive Coding:** Linear predictive coding (LPC) is a method mostly used in audio signal processing and speech processing to represent the spectral envelope of a compressed digital voice signal using linear predictive model information. This technique is designed for all poles. It is based upon the fundamental principle of sound production and its output degraded when there is noise, [17].

(ii) **Cepstral Coefficients:** This technique is purely based on Fast-Fourier Transformation (FFT) and is not much compatible with humans because of regular spaced filters.

(iii) **Linear Predictive Cepstral Coefficients:** Cepstral analysis is widely used in speech processing due to its ability to symbolize perfectly speech waveforms and characteristics with a limited size of features. This technique is designed by system pole. It provides smoother spectral envelope and robust representation compared to LPC. The downside of this technique is due to linear frequency spacing, [8].

(iv) **Mel-Frequency Cepstal Coefficients:** In sound processing, the mel-frequency cepstrum (MFC) is a representation of a sound's short-term power spectrum, based on a linear cosine transformation of a log power spectrum at a nonlinear frequency melscale. The principal of this technique is bank coefficients filter. It has knowledge on lower frequencies attributable to mel spaced filter banks is therefore more like a human ear than other techniques, .

3) FEATURE CLASSIFICATION TECHNIQUES

(i) **Support Vector Machines:** This technique comes under Supervised Algorithm category. This technique is beneficial when classifying binary and has poor performance in voice recognition due to its weakness to dealing with fixed-length vectors, [16].

(ii) **Hidden Markov Model:** This is unsupervised Algorithm model, more complex computationally, and require more storage space. This technique requires more data on performance to resolve intersession problems.

(iii) **Vector Quantization:** This is unsupervised Algorithm technique and has feasible storage requirement for application in real time. This technique is less complex in numerical terms, [18, 19].

(iv) **Gaussian Mixture Model:** This is unsupervised Algorithm model. Training and testing data requirement for this model is very less. The de-merit of this model is that it is DTW and HMM compromise, [15].

## 6. Conclusion and Future Work

Speech Recognition is one of the most integrated fields of machine intelligence, as humans conduct a routine speech recognition task. There was a long and winding tradition of speech recognition technology. Nonetheless, today's speech systems like Google Voice, Amazon Alexa, Microsoft Cortana, and Apple's Siri wouldn't be there without the early pioneers paving the way for them today. The conversion of speech to text can appear affective and efficient to its users. For desktop and mobile phone devices, an integrated multilingual speech to text translation program can be introduced according to ease of use. Such devices are useful to naturally deaf and dumb people to communicate with other people. Multilingual speech recognition, faster training and testing development desktop and mobile phone apps with advanced user interfaces can be considered for future work.

## References

[1] X. Huang, L. Deng: *An Overview of Modern Speech Recognition*, Handbook of Natural Language processing, (2010), 339–366.

[2] B. R. Reddy, E. Mahender: *Speech to Text Conversion using Android Platform*, International Journal of Engineering Research and Applications, **3**(1) (2013), 253–258.

[3] S. Das: *Speech Recognition Technique: A Review*, International Journal of Engineering Research and Applications, **2**(3) (2012), 2071–2087.

[4] Y. Kumar, N. Singh: *An automatic speech recognition system for spontaneous Punjabi speech corpus*, International Journal of Speech Technology, **20**(2) (2017).

[5] A. H. Unnibhavi, D. S. Jangamshetti: *A survey of speech recognition on south Indian Languages*, Second ed., International Conference on Signal Processing, Communication, Power and Embedded System (SCOPES), 2016, 1122–1126.

[6] C. Vimala, V.Radha: *A Review on Speech Recognition Challenges and Approaches*, World of Computer Science and Information Technology Journal (WCSIT), **2**(1) (2012), 1–7.

[7] S. K.Gaikwad, B. W.Gawali, P. Yannawar: *A Review on Speech Recognition Technique*, International Journal of Computer Applications, **10**(3) (2010), 16–24.

[8] P. Khilari, V. P. Bhope: *A Review on Speech to Text Conversion*, International Journal of Advanced Research in Computer Engineering & Technology (IJARCET), **4**(7) (2015), 3067–3072.

[9] X. Huang, A. Acero, H. W. Hon: *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development*, Prentice Hall PTR, New Jersey, 2001.

[10] L. Rabiner, L. R. Rabiner, B. H. Juang: *Fundamentals of speech recognition*, Prentice Hall Internacional Inc, 1993.

[11] K. M. SHIVAKUMAR, V. V JAIN, K. P. PRIYA: *A study on impact of Language Model in improving the accuracy of Speech to Text Conversion System*, International Conference on Communication and Signal Processing (ICCSP), Chennai, (2017), 1148–1151.

[12] S. MITTAL, R. KAUR: *Implementation of phonetic level speech recognition system for Punjabi language* , 1st India International Conference on Information Processing (IICIP), (2016), 1–6.

[13] S. SULTANA, M. A. H. AKHAND, P. K. DAS, M. M. H. RAHMAN: *Bangla Speech-to-Text conversion using SAPI* , International Conference on Computer and Communication Engineering (ICCCE), (2012), 385–390.

[14] S. RAUF, A. HAMEED, T. HABIB, S. HUSSAIN: *District names speech corpus for Pakistani Languages* , International Conference Oriental COCOSDA held jointly with 2015 Conference on Asian Spoken Language Research and Evaluation (O-COCOSDA/CASLRE), Shanghai, (2015), 207–211.

[15] R. SINGH, A. SHARMA: *Speech Recognition Method Applied for Different Punjabi Language Accents*, International Journal of Advance Research in Science and Engineering, **6**(1) (2017), 2319-8354.

[16] P. HERACLEOUS, H. ISHIGURO, N. HAGITA: *Visual-speech to text conversion applicable to telephone communication for deaf individuals*, 8th International Conference on Telecommunications, Ayia Napa, (2011), 130–133.

[17] S. TRIPATHY, N. BARANWAL, G. C. NANDI: *A MFCC based Hindi speech recognition technique using HTK Toolkit*, IEEE Second International Conference on Image Information Processing, ICIIP, (2013), 539–544.

[18] Y. LONG, Y. LI, O. ZHANG, S. WEI, H. YE, J. YANG: *Acoustic data augmentation for Mandarin-English code-switching speech recognition*, Applied Acoustics, **161**(2020), 107175.

[19] S. PATIL, M. PHONDE, S. PRAJAPATI, S. RANE, A. LAHANE: *Multilingual Speech and Text Recognition and Translation using Image*, International Journal of Engineering Research and Technology (IJERT), **5**(4) (2016), 85–87.

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
PUNJABI UNIVERSITY
*Email address*: insansaloni@gmail.com

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
PUNJABI UNIVERSITY
*Email address*: williamjeet@gmail.com