EasyChair Preprint

№ 3064

Multilingual Speech to Text Conversion – A Review

Saloni and Williamjeet Singh

March 29, 2020

Multilingual Speech to Text Conversion – A Review

Saloni[1] and Dr. Williamjeet Singh[2]

Dept of Computer Science and Engineering, Punjabi University, Patiala
lncs@springer.com

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. A great deal of work on various aspects of speech recognition and its implementations has been conducted for more than three decades. 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. It is possible to build an automated application to resolve the language barrier between countries and states within the world. The specification must include 4 modules of voice recognition, translation and speech synthesis, and the translated language text is provided. 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.

Keywords: Automatic Speech Recognition, speech-to-text conversion system (STT), multilingual, end-to-end (E2E) system and feature extraction tools and techniques.

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. In a digital representation, the signals are typically interpreted, and speech processing can be viewed as a special case of digital signal processing applied to speech signal. 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 from a microphone or a phone. We have to employ computers or electronic circuits to retrieve and determine the linguistic information conveyed by a speech stream. For a few who can read or understand a scrupulous language, most of the knowledge in the digital world is available. 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].In multi-lingual societies like India, which has about 1652 dialects / native languages, these technologies play a vital role.Languages, on which automatic speech recognition systems have been developed so far, are only a fraction of around 7300 known languages in total. Notable among them are Russian, Portuguese, Chinese, Vietnamese, Japanese, Spanish, Filipino, Arabic, English, Bengali, Tamil, Malayalam, Sinhala, Hindi. English is the language for which the most research work is done.

1.1 Speech Types

Speech recognition system can be distinguished by defining what kind of ullerances they can identify in different classes [3]. These are listed as follows:

- **Isolated Word:** Isolated word recognizes typically allow that each utterance has silence on both sides of the sample windows. This allows single words or single pronouncements at a time. "Listen and do not listen" This class might be better named for isolated utterance [4].
- **Connected Word:** Connected word system is similar to isolated words which allows for the division or separation of sound to be "run a minimum pause between them [5].

- **Continuous Speech:** Continuous speech recognizers allow users to talk almost naturally, while the content is decided by the machine. Recognizer with continuous speech capabilities are among the hardest to construct because they use unique sound and different methods to establish the limits of utterance [6].
- **Spontaneous Speech:** It can be thought of at a basic level as an expression that is natural sounding and not rehearsing. An ASR device with spontaneous speech capacity should be able to handle a different wording and a range of natural speech features such as words running together [7].

1.2 Type of ASR model, 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].

- **Speaker Dependent Model:** Dependent speaker systems are designed for a single speaker. Generally such systems are easier to develop, cheaper and more reliable but not as versatile as adaptive speakers or independent speaker systems. For the same speaker, they are usually more reliable, but much less accurate for other speakers.
- **Speaker Independent Model:** Independent Speaker systems are designed for a range of speakers. This identifies one broad group of people's speech patterns. This system is the hardest to build, the most costly and offers less accuracy than speaker-dependent systems. We are however more versatile [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 as below in ASR systems.

- **Small Vocabulary -** 10 words
- **Medium Vocabulary -** 100 words
- **Large Vocabulary -** 1,000 words
- **Very-large Vocabulary -** 10,000 words
- **Out-of-Vocabulary -** Mapping a phrase to unknown word from the vocabulary.

Besides the above features, the variation of the environment the variability of the networks, the style of the speakers, sex, age, speed of speech also make the ASR system more complex.

2 Automatic Speech Recognition

Speech Recognition is one of the key fields of speech processing science, and is also known as Automatic Speech Recognition (ASR).Speech recognition technology has been widely studied for the past three decades as the degree of acceptance is strong for such systems. Spoken language is considered one of the simplest and most natural means of communication. It is an important way of recognizing a person through voice. It also represents an important component of biometric authentication. In order to share any information between man and machine keyboards, pointing devices are needed which are not convenient for a layman since special skills are required. Hence, speech provided a great platform for solving this problem. Speech recognition is the mechanism where the human speech is inserted into the device in analog form and the computer translates it into digital form to make it comprehensible. It is the method of translating a speech signal through an algorithm implemented as a computer program to a series of words (i.e., spoken words to text) [9].The job is to get a machine to understand the language spoken. Through "comprehending" we mean properly responding and translating the input speech into another form, e.g. text. Therefore, voice recognition is sometimes called speech-to-text (STT).A speech recognition system consists of a microphone for the person to speak in; speech recognition software; a computer for taking and interpreting speech; an input and/or output soundcard of good quality; a correct and accurate pronunciation. Automatic speech recognition [10] is a device shown in Fig 1 that performs automatic transformation of an acoustic input speech signal into a transcription of a text. Many such systems were developed and worked on to improve the ease of communication access. The real aim of ASR research is to allow a machine to recognize real-time speech with 100% accuracy regardless of noisy scenario, vocabulary size, speaker characteristics, accents and channeling conditions. Automatic speech recognition due to variations such as setting, speaker and meaning is one of the difficult research areas. Speaker recognition is a software or hardware's ability to receive voice signal, identify the speaker present in the speech signal, and then recognize the speaker. 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).

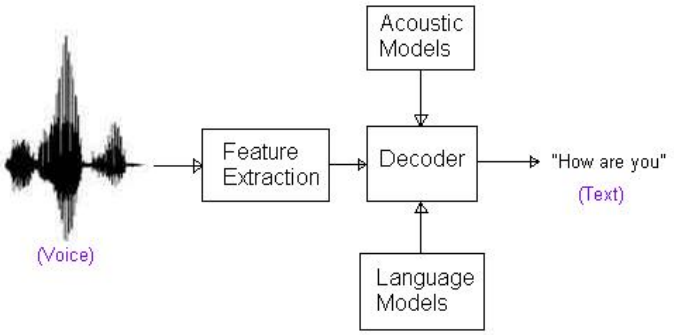

Fig. 1. Basic structure of Automatic Speech Recognition System
Source: https://www.rfwireless-world.com/Terminology/automatic-speech-recognition-system.html

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. These have shown progress over monolingual structures, and simplified teaching and serving by removing auditory, syntax, and language-specific models. An end-to-end (E2E) device can be trained as a single model, allowing for multilingual voice recognition in real time. A drastic improvement in the ASR output on several data-scarce languages using nine Indian languages, while still improving performance for the data-rich languages.

4 Research Process

The study intended to be a systematic review meets Shiva Kumar K M 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. Let us look at the various phases of research preparation as well as the steps involved in a 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 Fig.2.
1) Information Sources

A broad perspective is important for a wide and wide scope of writing. An selection of suitable repositories needs to be picked to increase the probability of very important posts. Using following online resources in this investigation to scan for important studies:

- Springer (www.springerlink.com)
- IEEE explore (http://ieeexplore.ieee.org)
- Science Direct (www.sciencedirect.com)

- ACM Digital Library (ww.acm.org)

Wiley Interscience (www3.interscience.wiley.org)

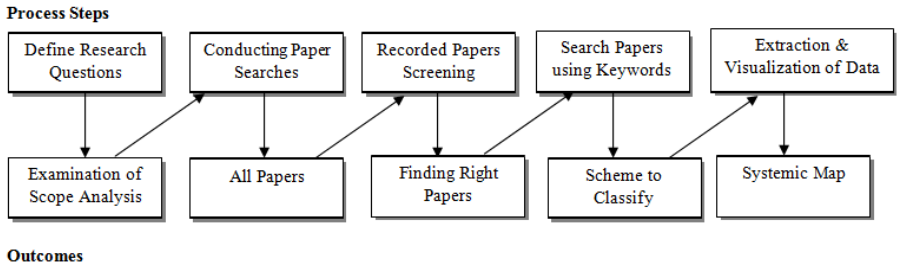

Fig. 2. Searching Process

2) Selection of sample

In the above electronic sources the search engines hit a number of research, papers. The quest is limited to keywords, title and description, so that the search target is guided and the inapplicable papers eliminated. We are inclined to try to extract the maximum amount of relevant literature to support the completeness of the analysis as possible. The outlines of the quest method and the range of results obtained are shown in table 1. The result shows the distribution of relevant papers over the years 2010 through 2017 analyzed as shown in Table 2.

3) Integration and Prohibition

Nevertheless, the studies that are applicable to Speech-to-text assisted metrics are included the studies that are not yet written, reports and theses. Almost in all searches, the keyword "Speech Recognition," "Speech to Text," "Tools and Techniques for Feature Extraction" is included in the abstract and attempted to retrieve all relevant studies.

Secondly unsuitable experiments on the concept of title are omitted from the returned research. If it was completely not clear from the abstract, otherwise unfitting experiments were omitted once the entire text of the papers was read. 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.

TABLE I: SEARCH SELECTION

S.No.	**E-resource**	**studies returned**	**Excluded**			**keyword used**
			based on title	**based on abstract**	**based on 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

TABLE II: DISTRIBUTION OF PAPERS OVER YEARS

Year / E-resource	2010	2011	2012	2013	2014	2015	2016	2017	2018
IEEE	1	-	-	2	-	-	-	2	3
ACM	-	2	-	-	2	-	-	-	2
Springer	1	-	4	-	-	-	1	-	2
Science Direct	1	-	-	2	1	-	-	1	1
Wiley	-	1	-	-	-	3	-	-	2

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 in Table III, IV, V and the comparison between every feature extraction technique is given in Table VI.

TABLE III: TOOL FACET

Category	Description
Hidden Markov Toolkit	Hidden Markov Model Toolkit (HTK) is a portable toolkit designed to build and manipulate hidden Markov models. HTK is primarily used for speech recognition research although it has been used for many other applications including speech synthesis testing, character recognition and DNA sequencing.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].
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]. It is an API used to interface with web servers including Apache.
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. This method platform is used to control and train Speech Process Program. It supports multiple languages, and is in essence complex. The components of a Speech Recognition System are Language Model, Dictionary Acoustic Model [11].
Praat	PRAAT is a computer program for discourse analysis, synthesis, and manipulation. It is established at the University of Amsterdam's Institute of Phonetic Sciences by Paul Boersma and David Weenink since 1992. This tool (Language Process Software) is used to process the voice files. A PRAAT utility is used to

	mechanically mark speech files with transcription files, and two folders are generated; clean and incorrect [14].
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.

TABLE IV: FEATURE EXTRACTION TECHNIQUES

Category	**Description**
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.
Cepstral Coefficients	This technique is purely based on Fast-Fourier Transformation (FFT) and is not much compatible with humans because of regular spaced filters.
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.
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 [15].

TABLE V: FEATURE CLASSIFICATION TECHNIQUES

Category	**Description**
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.
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.
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.
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].

TABLE VI: FEATURE EXTRACTION TECHNIQUES COMPARISON [21]

Category	**Type of Filter**	**Shape of filter**	**What is modeled**	**Speed of computation**	**Type of coefficient**	**Noise resistance**	**Sensitivity to quantization/additional noise**	**Reliability**	**Frequency captured**

Mel frequency cepstral coefficient (MFCC)	Mel	Triangular	Human Auditory System	High	Cepstral	Medium	Medium	High	Low
Linear prediction coefficient (LPC)	Linear Prediction	Linear	Human Vocal Tract	High	Autocorrelation Coefficient	High	High	High	Low
Linear prediction cepstral coefficient (LPCC)	Linear Prediction	Linear	Human Vocal Tract	Medium	Cepstral	High	High	Medium	Low & Medium
Line spectral frequencies (LSF)	Linear Prediction	Linear	Human Vocal Tract	Medium	Spectral	High	High	Medium	Low & Medium
Discrete wavelet trans	Lowpass&highpass	—	—	High	Wavelets	Medium	Medium	Medium	Low & High

form (DWT)									
Perceptual linear prediction (PLP)	Bark	Trapezoidal	Human Auditory System	Medium	Cepstral& Autocorrelation	Medium	Medium	Medium	Low & Medium

6 Conclusion and Future Work

Speech Recognition is a difficult problem that needs to be addressed. In this paper, we have attempted to provide a summary of how much progress this technology has made in past years. 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. Text synthesis speech is a vital field of research and implementation in the field of multimedia interfaces. Such speech systems have steadily enhanced their ability to ' hear ' and understand a wider variety of words, phrases, and accents due to the introduction of new technologies such as cloud-based computing as well as ongoing data collection initiatives. 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 and L. Deng, "An Overview of Modern Speech Recognition", Handbook of Natural Language Processing, Second Edition, Chapter 15, Chapman & Hall/CRC, **(2010)**, pp. 339-366.
2. B. Raghavendhar Reddy,E. Mahender*,"Speech to Text Conversion using Android Platform"*, International Journal of Engineering Research and Applications (IJERA),Vol. 3, Issue 1, January -February 2013, ISSN: 2248-9622.
3. Sanjib Das, *"Speech Recognition Technique: A Review"*, International Journal of Engineering Research and Applications (IJERA), Vol. 2, Issue 3, May-Jun 2012, ISSN: 2248-9622.

4. Y. Kumar and N. Singh, “An automatic speech recognition system for spontaneous Punjabi speech corpus,” *Int. J. Speech Technol.*, vol. 0, no. 0, p. 0, 2017.
5. A. H. Unnibhavi and D. S. Jangamshetti, "A survey of speech recognition on south Indian Languages," 2016 International Conference on Signal Processing, Communication, Power and Embedded System (SCOPES), Paralakhemundi, 2016, pp. 1122-1126.
6. C. Vimala, “A Review on Speech Recognition Challenges and Approaches,” vol. 2, no. 1, pp. 1–7, 2012.
7. B. W. Gawali, “A Review on Speech Recognition Technique,” vol. 10, no. 3, 2010.
8. A. Speech, “A REVIEW ON SPEECH TO TEXT CONVERSION,” vol. 4, no. 7, pp. 3067–3072, 2015.
9. Huang, Xuedong & Acero, Alex & Hon, Hsiao-Wuen. (2001). Spoken Language Processing: A Guide to Theory, Algorithm, and System Development.
10. Rabiner, L. Juang, B. H., Yegnanarayana, B. 2010. Fundamentals of speech recognition, Pearson publishers.
11. K. M. Shivakumar, V. V Jain, and K. P. P, “A study on impact of Language Model in improving the accuracy of Speech to Text Conversion System,” pp. 1148–1151, 2017.
12. S. Mittal and R. Kaur, "Implementation of phonetic level speech recognition system for Punjabi language," 2016 1st India International Conference on Information Processing (IICIP), Delhi, 2016, pp. 1-6.
13. S. Sultana, M. A. H. Akhand, P. K. Das and M. M. Hafizur Rahman, "Bangla Speech-to-Text conversion using SAPI," 2012 International Conference on Computer and Communication Engineering (ICCCE), Kuala Lumpur, 2012, pp. 385-390.
14. S. Rauf, A. Hameed, T. Habib and S. Hussain, "District names speech corpus for Pakistani Languages," 2015 International Conference Oriental COCOSDA held jointly with 2015 Conference on Asian Spoken Language Research and Evaluation (O-COCOSDA/CASLRE), Shanghai, 2015, pp. 207-211.
15. Ravinder Singh, Dr. Anand Sharma, "Speech Recognition Method Applied for Different Punjabi Language Accents", vol. 6, no. 1, pp. 2319-8354, 2017.
16. Santosh, K.Gaikwad & Bharti, W.Gawali & Yannawar, Pravin (2010)," A Review on Speech Recognition Technique International Journal of Computer Applications", 10. 10.5120/1462-1976.
17. Shaheena Sultana, M. A. H. Akhand, Prodip Kumer Das, M. M. Hafizur Rahman, "Bangla Speech -to-Text Conversion using SAPI", 3-5 July 2012
18. P. Heracleous, H. Ishiguro and N. Hagita, "Visual-speech to text conversion applicable to telephone communication for deaf individuals," 2011 18th International Conference on Telecommunications, Ayia Napa, 2011, pp. 130-133
19. S. Tripathy, N. Baranwal and G. C. Nandi, "A MFCC based Hindi speech recognition technique using HTK Toolkit," 2013 IEEE Second International Conference on Image Information Processing (ICIIP-2013), Shimla, 2013, pp. 539-544.
20. Miss Prachi Khilari, Prof. Bhope V.P., "A Review on Speech to Text Conversion on Methods", Volume A, Issue 7, July 2015.
21. Alim, Sabur & Alang Md Rashid, Nahrul Khair. (2018), "Some Commonly Used Speech Feature Extraction Algorithms", 10.5772/intechopen.80419.
22. Yanhua Long, Yijie Li, Qiaozheng Zhang, Shuang Wei, Hong Ye, Jichen Yang,"Acoustic data augmentation for Mandarin-English code-switching speech recognition", Applied Acoustics,Volume 161,2020,107175, ISSN 0003-682X.
23. Patil, Sagar & Phonde, Mayuri & Prajapati, Saranga & Lahane, Anita. (2016). Multilingual Speech and Text Recognition and Translation using Image. International Journal of Engineering Research and. V5. 10.17577/IJERTV5IS040053.