Chapter 2
Automatic Speech Recognition Models, Tools, and Techniques: A Systematic Review

Puneet Mittal
Mangalore Institute of Technology and Engineering, Karnataka, India

Sukhwinder Sharma
Mangalore Institute of Technology and Engineering, Karnataka, India

ABSTRACT

Automatic speech recognition (ASR) has gained wide popularity in last decade. Various devices like mobile phones, computers, vehicles, and audio/video players are now being equipped with ASR technology. The increasing use and dependence on ASR technology leads to research enhancements and opportunities in this domain. This chapter provides a detailed review of various advancements in ASR systems development. It highlights history of speech recognition followed by detailed insight into recent advancements and industry leaders providing latest solutions. ASR framework has been discussed in detail which includes feature extraction techniques, acoustic modeling techniques, and language modeling techniques. The chapter also lists various popular data sets available and discusses generation of new data sets. This work will be helpful for the researchers who are new to this field and are exploring development of new speech recognition techniques.

INTRODUCTION

Since the beginning of human life, communication has been the most important aspect for humans. Speech communication is the easiest and widely popular way of interaction between human to human. But when we interact with a machine, it is often limited to some controls or buttons present on it. Automatic Speech Recognition (ASR) means machine will automatically recognize what a person speaks, i.e., it can convert speech into text, operate based on the spoken command and respond. Communication with machine using speech has given a new dimension for interaction with machines like mobile phones,

DOI: 10.4018/978-1-6684-6001-6.ch002

computers, vehicles and televisions, therefore, gained wide popularity in last few decades. It has made interactions with machine more natural and easier (Neustein, 2010; Kumar et al., 2011; Tan & Lindberg, 2008). Now, humans can talk to machines just like human-to-human communication takes place. The technology is helpful in conditions where a person is busy in some physical tasks like driving a vehicle and cannot operate the machine with hands. Eyes and hands-free communication is quite helpful in such situations. It is proving to be a boon for people with disabilities like blindness and handless, who are struggling to use devices like normal persons. With the advancements in speech technology, controlling various devices is becoming much easier. The execution of an appropriate action depends on the recognition accuracy of underlying ASR system. If the system is unable to clearly identify the spoken words, either no action or some inappropriate action will likely to be taken. Therefore, an accurate ASR system is required.

In latest times, the need to build an efficient interface that provides speech recognition is mounting swiftly due to the increasing use of mobile phones and similar devices. These devices are having a variety of applications such as voice calling, SMS, listening music, setting alarms and reminders, internet surfing, smart home automation, speech based dictation, and control of various devices.

Speech is basically a signal produced by human's speech organs and detected by human ears. Speech articulatory organs are vocal cords, pharyngeal wall, glottis, jaw, soft palate, hard palate, nasal cavity, tongue, alveolar ridge, teeth and lips. The sound is produced when air passes through these organs. Human auditory organs perceive sound in the form of vibrations, and this sound is transduced into nerve impulses, which is further perceived by brain. The brain processes the sound and extracts meaningful words from it or takes action or responds accordingly.

Speech Recognition by machines requires machines to perceive the speech and recognize it. For developing the ASR, various researchers have proposed algorithms, tools, techniques and models in various languages of the world. This paper highlights the contributions made by researchers in the field of ASR. The chapter has been structured in the following order: Section 2 illustrates history of ASR and pioneering work done in this field. Section 3 highlights the recent advancements being done in ASR, while Section 4 elaborates the ASR architecture showcasing various techniques for feature extraction, acoustic modeling, language modeling and pronunciation dictionary modeling. It also gives brief details about various toolkits and decoders available for ASR model generation. Section 5 presents the challenges being faced by ASR technology followed by conclusions in last section.

BACKGROUND

History of Speech Recognition

Researchers have been exploring the ASR field for the past seven decades (Kikel, nd). Research commenced in this field in the year 1952 when Davis et al. developed Aurdey at Bell Labs, which is the first known speech recognizer, which recognizes digits with an accuracy of 97-99%.

IBM shoebox technology (IBM ShoeBox) was introduced in the year 1960. This device recognized digits from 0 to 9 and commands like plus, minus and total. Total 16 words were there in the vocabulary.

In 1975, Baker proposed The Dragon System, which was going to be the basis of modern ASR research. The model was based on the Markov process probabilistic function which has revolutionized modern speech recognition.