Chapter 1.7 Introduction to Speech Recognition

Sergio Suarez-Guerra
National Polytechnic Institute of Mexico, Mexico

Jose Oropeza-Rodriguez
National Polytechnic Institute of Mexico, Mexico

ABSTRACT

This chapter presents the state-of-the-art automatic speech recognition (ASR) technology, which is a very successful technology in the computer science field, related to multiple disciplines such as the signal processing and analysis, mathematical statistics, applied artificial intelligence and linguistics, and so forth. The unit of essential information used to characterize the speech signal in the most widely used ASR systems is the phoneme. However, recently several researchers have questioned this representation and demonstrated the limitations of the phonemes, suggesting that ASR with better performance can be developed replacing the phoneme by triphones and syllables as the unit of essential information used to characterize the speech signal. This chapter presents an overview of the most successful techniques used in ASR systems together with some recently proposed ASR systems that intend to improve the characteristics of conventional ASR systems.

INTRODUCTION

Automatic speech recognition (ASR) has been one of the most successful technologies allowing the man-machine communications to request some information from it or to request to carry out some given task using the natural oral communication. The artificial intelligence field has contributed in a remarkable way to the development of ASR algorithms.

The more widely used paradigm in ASR systems has been the phonetic content of the speech signal, which varies from language to language, but there are no more than 30 different phonemes without some variations, such as accentuation, duration, and the concatenation. The last one includes the co-articulation such as demisyllables and triphones. Considering all variations, the number of phonetic units will be increased considerably. Recently some researchers have considered the use of syllables instead of phonemes as an alternative for development of the ASR systems, because in general, the natural way to understand the language by the human brain is to store and recognize syllables not phonemes.

The automatic speech recognition is a very complex task due to the large amount of variations involved in it, such as intonation, voice level, health condition and fatigue, and so forth (Suárez, 2005). Therefore, in the automatic speech recognition system, for specific or general tasks, there is an immense amount of aspects to be taken into account. This fact has contributed to increase the interest in this field, and as a consequence, several ASR algorithms have been proposed during the last 60 years.

A brief review of ASR systems can be summarized as follows. At the beginning of 1950s, the Bell Laboratories developed an ASR system that was able to recognize isolated digits. The RCA Laboratories developed a single-speaker ASR for recognition of 10 syllables. The University College in England developed a phonetic recognizer, and in the MIT Lincoln Laboratory a speaker independent vowel recognizer was developed. During the 1960s, some basic tools for ASR systems were developed. The dynamic time warping is developed by the NEC Laboratories and Vintsyuk of the Soviet Union. In the Carnegie Mellon University, the automatic continuous speech recognition system with small vocabularies HAL 9000 was developed. During the 1970s, several isolated word recognition systems were developed, such as a large vocabulary ASR system by IBM. During this time also there was an important increase on the investment to develop ASR systems, such as the DARPA and HARPY project in the U.S. During the decade of the 1980s, the first algorithms for continuous speech recognition system with large vocabularies appeared. Also during this time the hidden Markov model (HMM) and neural networks were introduced in (development of) the ASR systems; one of these types of systems is a SPHINX system. The ASR systems have appeared as commercial systems during the decade of the 1990s thanks to the development of fast and cheap personal computers that allow the implementation of dictation systems and the integration between speech recognition and natural language processing. Finally, during recent years, it was possible to use the voice recognition systems in the operating systems, telephone communication system, and Internet sites where Internet management using voice recognition, Voice Web Browsers, as well as Voice XML standards are developed.

STATE-OF-THE-ART

During the last few decades, the study of the syllables as a base of language model has produced several beneficial results (Hu, Schalkwyk, Barnard, & Cole, 1996). Hu et al. (1996) realized an experiment where syllables belonging to the name of months in English were recognized. They created a corpus with a total of 29 syllabic units, and 84.4% efficiency was achieved in their system. Boulard (1996) realized similar works using the syllables in German. Hauenstein (1996) developed a hybrid ASR system: HMM-NN (hidden Markov models-neural networks) using syllables and phonemes as basic units for the model. He realized a performance comparison between the system using only syllables and another one using only phonemes, and he concluded that the system that combined both units (syllables and phonemes) presents higher performance than the system using syllables or phonemes separately. Wu, Shire, Greenberg, and Morgan (1997) proposed an integration of information at syllables level within the automatic speech recognizer to improve their performance and robustness (Wu, 1998; Wu et al., 1997), taking 10% of the recognition error for digits voices of OGI (Oregon Graduate Institute) corpus. In a work by Wu (1998), 6.8% of recognition error rate for digits data using corpus of digits from telephone conversation was reported; here, the RSA system was a phoneme-syllable