International Journal of Knowledge and Systems Science
Volume 15 • Issue 1

A Review on Speech Recognition for Under-Resourced Languages:
A Case Study of Vietnamese

Trung-Nghia Phung, Thai Nguyen University of Information and Communication Technology, Vietnam*
Duc-Binh Nguyen, Thai Nguyen University of Information and Communication Technology, Vietnam
Ngoc-Phuong Pham, Thai Nguyen University, Vietnam

ABSTRACT

Fundamental speech recognition technologies for high-resourced languages are currently successful to build high-quality applications with the use of deep learning models. However, the problem of "borrowing" these speech recognition technologies for under-resourced languages like Vietnamese still has challenges. This study reviews fundamental studies on speech recognition in general as well as speech recognition in Vietnamese, an under-resourced language in particular. Then, it specifies the urgent issues that need current research attention to build Vietnamese speech recognition applications in practice, especially the need to build an open large sentence-labeled speech corpus and open platform for related research, which mostly benefits small individuals/organizations who do not have enough resources.

KEYWORDS

Deep Learning, DNN-HMM, End to End, Machine Learning, Speech Corpus, Speech Recognition, Under-Resourced Languages, Vietnamese Speech Recognition

INTRODUCTION

Speech recognition is a type of problem in the field of pattern recognition, so there are difficulties similar to other recognition problems. There are also other problems with the random change nature of speech signals. So far, there are five classic and major problems affecting the accuracy and performance of a speech recognition system (Tebelskis, 1995; Duc, 2003; Jurafsky, 2008; Lei, 2006; Yu & Deng, 2016) which include the speaker-dependent problem, co-articulation problem, vocabulary (dictionary) size problem, noise problem, language-dependent problem.

Each speaker has a different structure of the sound articulators, so the characteristics of the voice that are emitted are greatly influenced by the speaker. Even when a speaker pronounces the same sentence, the voice speaking out can be different due to the amount of air escaping from the

DOI: 10.4018/IJKSS.332869 *Corresponding Author

lungs, emotional status, health status, etc. In terms of speaker-dependent characteristics, speech recognition can be divided into two types. The first type is speaker-dependent speech recognition, which is specifically developed to work with the voices of only one or a few people. The second type is speaker-independent recognition; that is a recognition system built to recognize the voice of any person. Normally, the speech recognition error rate of a speaker-independent system is usually higher than the speaker-dependent speech recognition system.

In a continuous pronunciation sequence, each sound is often greatly influenced by the sounds preceding and following it. Therefore, words with discrete pronunciation in recognition will have higher accuracy than words in a continuous pronunciation sequence. Since the recognition quality for a continuous pronunciation sequence is also dependent on detecting boundaries and silences between two words. When the speaker pronounces at a high speed, the silences and boundary between words will be narrowed, leading to each word segment being confused or overlapping, affecting the accuracy of word recognition.

The dictionary size is the number of all the different words that a particular recognition system is capable of recognizing. The larger the size of the dictionary is, the higher the complexity of the recognition system will be. The error rate of the recognition system is always proportional to the size of the dictionary. The speech recognition systems applied in practice today mostly require a large dictionary that covers all phonetic units to be able to recognize any sentence or word. These recognition systems require the training speech corpus to be large enough and cover all phonetic units such as phonemes in different contexts. While high-resourced languages such as English already had many labeled and widely used large speech corpora with hundreds of Terabytes, the large speech corpus problem for under-resourced languages including Vietnamese is still an unsolved problem.

In practice, the speech signal is often affected by noise from the outside environment such as traffic, animals' sounds, or the voices of one or more other people speaking at the same time. For humans, distinguishing and focusing on a person who is speaking to understand and distinguish semantics is quite simple, but for computers such cases cause special difficulties for identification because microphones pick up every type of audio signal in the frequency band in which it operates. Currently, even when applying optimal preprocessing methods on the received signal, and at the same time filtering out the signal of the speaker that wants to be identified, the recognition quality for these cases is still very low.

Each language has its own set of characters and phonemes. Researching and finding a set of standard phonemes for a language will improve recognition accuracy. For each language, the grammatical problem of pronunciation also greatly affects the quality of recognition. Pronunciations that follow a clear and complete syntactic structure are more accurately recognized than free pronunciations; that is words in pronunciation without specific grammatical constraints.

In this study, the authors review the results of fundamental research on speech recognition in general to handle the above 5 existing problems of speech recognition in Section 2; review research results on Vietnamese speech recognition - an under-resourced language in Section 3; and make recommendations for further studies on Vietnamese speech recognition in Section 4.

FUNDAMENTAL RESEARCH RESULTS ON SPEECH RECOGNITION

Currently, there have been several scientific publications on many different aspects to contribute to improving the quality of speech recognition. There are many ways of classifying studies on speech recognition. This paper reviews current studies based on four main components of a recognition system (Yu & Deng, 2016) including: Feature extraction; Acoustic model; Language model; The decoder, as well as the review of the latest research trends on the outstanding role of statistical learning methods and large corpus in speech recognition that aims to solve these five existing problems of speech recognition mentioned in section Introduction.

Related Content

Property Valuation Using Linear Regression and Random Forest Algorithm
Sam Goundar, Kunal Maharaj, Anirudh Kumarand Akashdeep Bhardwaj (2021). *International Journal of System Dynamics Applications (pp. 1-16).*
www.irma-international.org/article/property-valuation-using-linear-regression-and-random-forest-algorithm/273122

A Comparative Study of Neural Network and Fuzzy Logic Control Based Active Shunt Power Filter for 400 Hz Aircraft Electric Power System
Saifullah Khalid (2017). *International Journal of Applied Evolutionary Computation (pp. 1-12).*
www.irma-international.org/article/a-comparative-study-of-neural-network-and-fuzzy-logic-control-based-active-shunt-power-filter-for-400-hz-aircraft-electric-power-system/188709

A Collective-Intelligence View on the Linux Kernel Developer Community
Haoxiang Xia (2012). *Systems Approaches to Knowledge Management, Transfer, and Resource Development (pp. 188-200).*
www.irma-international.org/chapter/collective-intelligence-view-linux-kernel/68218

Fuzzy Time Series Model Based on Intuitionistic Fuzzy Sets for Empirical Research in Stock Market
Bhagawati P. Joshiand Sanjay Kumar (2012). *International Journal of Applied Evolutionary Computation (pp. 71-84).*
www.irma-international.org/article/fuzzy-time-series-model-based/74854

Evolution of IC Science and Beyond
Leif Edvinsson (2012). *Systems Approaches to Knowledge Management, Transfer, and Resource Development (pp. 15-27).*
www.irma-international.org/chapter/evolution-science-beyond/68208