Abstract

The educational services market is currently undergoing dynamic and diverse changes, mostly due to the advancement of the Internet and convenient communication methods. E-learning is a whole new style of learning that involves the use of electronic media, most notably the Internet. Furthermore, it allows students to learn from people from many cultures, countries, and geographic zones. In such situations, a discourse (i.e. lesson) between several individuals of different nationalities is possible. It is obvious that e-learning is only one of multiple examples of situations where participants represent many different nationalities and linguistic backgrounds. Others might include work and business related conference calls between employees of international companies, or any other interactions where the commonly used language is not the first language for the participating speakers.

Such circumstances need the use of a common language that is not the native tongue of at least one person to the conversation. Communication between two or more persons using a common language necessitates that all participants be proficient in that language. As a result, it may be useful for all parties involved in the encounter to employ a variety of accessible tools to improve communication and mutual understanding. Automatic speech recognition (ASR) techniques, for example, can be used to provide real-time closed captioning for the interaction. It is relatively simple to develop a support software for an e-learning platform that could use any existing cloud-based ASR system. However, the effectiveness of cloud ASR systems varies significantly depending on whether the sound samples represent the speech of a native or non-native speaker. In essence, the non-native speech recognition problem relates to the situation where the speaker's mother tongue is different from the language he speaks at the time of the speech recognition process. For example, we might consider a Japanese person speaking English and using the English language ASR system.

Supervised learning techniques are commonly used in traditional approaches to developing speech recognition classifiers. While this approach is ideal for recognizing speech in the majority of the world's languages, it will not produce good classifiers for non-native speakers.

The fundamental reason for this is a lack of large labeled datasets of non-native speech to be used as a training set in a supervised learning system. ASR technologies are becoming more accurate and depending on the benchmark used, can achieve up to 90-95 percent accuracy. However, such high levels of accuracy can only be achieved when the system is used to recognize native speakers' speech (e.g. English language for North American people).

The rationale for the ASR's lower accuracy is the presence of patterns connected to the speaker's

mother tongue that can impact the grammar and accent traits of the second language. In order to solve the aforementioned issues it is possible to focus on the problem either from the perspective of optimizing the process of creating ASR models adapted for non-native speech or approach the problem by trying to modify (convert) the speech (accent) which is problematic for the existing ASR systems to process correctly. The current rate of globalization requires the effective recognition of non-native speakers, who today account for the vast majority of users. The research presented in this dissertation focuses on automatic speech recognition for non-native speakers. This thesis represents the research studying the aforementioned issues and is organized into five chapters.

Chapter 1 describes in detail the problem of non-native speech recognition, especially in the context of non-native speakers. It describes in detail the main reasons behind the differences between speech recognition systems utilized by native speakers and the same process involving non-native speakers. In this chapter we also underline the several dimensions at which the native and non-native speech might differ and explain the rationale behind this logic. Furthermore we provide the motivation for conducting this research. We also state the purpose of this dissertation and represent it as a problem in the machine learning domain, and as such we approached it using developed algorithms. We outline the dissertation's layout and describe the content of each chapter.

Chapter 2 provides a summary of extensive research conducted on the problem of speech recognition for both native and non-native speakers. There have been multiple approaches evaluated by researchers that come from a wide variety of national and linguistic backgrounds. The research done by them involves multiple languages and nationalities of the non-native speakers. In this chapter we explain that while the research that had been done in the past yields an incremental progress, the scalable solution to the non-native speech recognition problem is still to be found and developed.

Chapter 3 describes a research conducted in order to address the problem of non-native speech data scarcity from the point of view of creating an ASR adapted for non-nartive speakers, from the ground up. The methodology designed and evaluated within the scope of this chapter represents an approach called dual supervised learning. The idea presents a setup of slightly modified actor-critic model adapted for training speech recognition models. We designed a custom feedback loop including a language model and speech model acting as critics, and speech-to-text with speech synthesis models becoming actors. The method focuses on training and updating the state of actor models while preserving the state of the critic models. Critic models had been trained before using a set of unlabelled non-native speech samples, that usually come in large amounts. Actor models are trained using the aforementioned feedback loop setup. The experiments conducted within the scope of this

chapter prove that it is possible to leverage the much greater availability of unlabelled datasets in order to create an ASR system adapted for non-native speech. In this chapter we also described a method for creating a conversion methodology between sentences produced by non-native speakers of any particular language in his or her second language and corresponding sentences as if they were formed by native speakers of the language. The methodology plays a supplementary role to the dual supervised learning method described above. The approach uses a sequence-to-sequence encoder-decoder setup which realizes the conversion between sentence pairs. It plays the role of a language model which has been widely utilized in ASR techniques and as such, is typically employed right after the ASR output was received for a non-native speech sample. The language model becomes a support mechanism which is adapted and trained for examples of sentences produced by non-native speakers of a particular language. We evaluated the idea using an experimental setup designed to compare the model trained using dual supervised learning to the baseline model trained using a traditional supervised approach. The experiment yielded a 3 % increase in the model accuracy when trained with a dual supervised approach compared to the model trained with the traditional way.

Chapter 4 represents research conducted on methodologies for real-time modification of accent for non-native speech samples. The chapter studies the performance of already existing ASR solutions when applied to non-native speech. We presented the problems that often affect the speech-to-text systems upon being used by a non-native speaker. To overcome the problems we designed the method to modify the non-native speech samples on-the-fly, as a data adaptation technique with the purpose of increasing the ASR performance with respect to non-native speech. The chapter presents the methodology which is based on the neural style transfer approach used in the context of speech and sound, instead of graphics and image. The approach involves creating an algorithm based on convolutional networks to extract certain information related to style (accent) and content (phonemes, words). The style transfer idea allows us to extract such information from non-native and native speech samples as well as a sample created by our speech style converter. Next step involves designing the loss function that minimizes the style difference between these samples, thus allowing the converted sample to resemble the native speech style to a higher extent. The audio style transfer solution was evaluated with the experimental setup that involved a Google Cloud Speech to Text service as a baseline ASR model. We evaluated its accuracy using the dataset of Japanese students speaking English. It was compared to the experiment which involved executing style transfer right before the converted samples were fed into the Google Cloud ASR. The comparison yielded a 40 %

relative improvement. The idea presented in the chapter provides the possibility of using already productionalized ASR systems for recognition of non-native speech not constrained by nationality and mother tongue.

Chapter 5 provides a brief description of all the research done within the range of this dissertation as well as the main contributions. In this chapter we provide brief information related to research problems presented in the thesis. We describe the significance of the research method for creating ASR models adapted for non-native speech, as well as the method of real-time speech conversion, in the context of increasing the ASR accuracy. It also highlights the issues that are still yet to be solved and discusses directions for future research on the topic.