様式8

論文内容要旨

報告番号	甲 先 第 431 号	氏 名	Zolzaya Byambadorj
学位論文題目	A study on Mongolian text-to-speech system based on deep neural network (ディープニューラルネットワークに基づくモンゴル語のテキスト音声合成システムに関する研究)		

内容要旨： There are about 7,000 languages spoken today in the world. However, most natural language processing and speech processing studies have been conducted for high resource languages such as English, Japanese and Mandarin. Preparing large amounts of training data is expensive and time-consuming, which creates a significant hurdle when developing some systems for the world's many, less widely spoken languages. We proposed to build a text-to-speech system (TTS, also called speech synthesis) for the low resource Mongolian language. We present two studies within this TTS system, “*text normalization*” and “*speech synthesis*,” on the Mongolian language with limited training data. TTS system converts written text into machine-generated synthetic speech. One of the biggest challenges to developing a TTS system for a new language is converting transcripts into a real “spoken” form, the exact words that the speaker said. This is an important preprocessing for TTS systems known as text normalization. In other words, text normalization is transforming text into a standard form and is an essential part of the speech synthesis system. The followings are brief descriptions of each part.

Text normalization: The huge increase in social media use in recent years has resulted in new forms of social interaction, changing our daily lives. Social media websites are a rich source of text data, but the processing and analysis of social media text is a challenging task because written social media messages are usually informal and ‘noisy’. Due to increasing contact between people from different cultures as a result of globalization, there has also been an increase in the use of the Latin alphabet, and as a result a large amount of transliterated text is being used on social media. Although there is a standard for the use of Latin letters in the language, the public does not generally observe it when writing on social media. Therefore, social media text also contains many noisy, transliterated words. In this thesis, our first goal is to convert noisy, transliterated text into formal writing in a different alphabet. Therefore, it poses more challenges in the text normalization task. We propose a variety of character level sequence-to-sequence (seq2seq) models for normalizing noisy, transliterated text written in Latin script into Mongolian Cyrillic script, for scenarios in which there is a limited amount of training data available. When there is a limited amount of training data, and the rules for writing noisy, transliterated text are not limited, we encounter a difficult challenge when attempting to normalize out-of-vocabulary (OOV) words. Therefore, we applied performance enhancement methods, which included various beam search strategies, N-gram-based context adoption, edit distance-based correction and dictionary-based checking, in novel ways to two basic seq2seq models. We experimentally evaluated these two basic models as well as fourteen enhanced seq2seq models and compared their noisy text normalization performance with that of a transliteration model and a conventional statistical machine translation (SMT) model. The proposed seq2seq models improved the robustness of the basic seq2seq models for normalizing OOV words, and most of our models achieved higher normalization performance than the conventional method.

Speech synthesis: Deep learning techniques are currently being applied in automated TTS systems, resulting in significant improvements in performance. These methods require large amounts of text-speech pair data for model training however, and collecting this data is costly. Tacotron 2 we used, a state-of-the-art end-to-end speech synthesis system, requires more than 10 hours of training data to produce good, synthesized speech. Therefore, our second goal is to build a single-speaker TTS system containing both a spectrogram prediction network and a neural vocoder for the target Mongolian language, using only 30 minutes of target Mongolian language text-speech paired data for training. We evaluate three methods for training the spectrogram prediction models of our TTS system, which produce mel-spectrograms from the input phoneme

sequence; (1) cross-lingual transfer learning, (2) data augmentation, and (3) a combination of the previous two methods. In the cross-lingual transfer learning method, we used two high-resource language datasets, English (24 hours) and Japanese (10 hours). We also used 30 minutes of target language data for training in all three methods, and for generating the augmented data used for training in methods (2) and (3) mentioned above. We found that using both cross-lingual transfer learning and augmented data during training resulted in the most natural synthesized target speech output. We also compare single-speaker and multi-speaker training methods, using sequential and simultaneous training, respectively. The multi-speaker models were found to be more effective for constructing a single-speaker, low-resource TTS model. In addition, we trained two Parallel WaveGAN (PWG) neural vocoders, one using 13 hours of our augmented data with 30 minutes of target language data and one using the entire 12 hours of the original target language dataset. Our subjective AB preference test indicated that the neural vocoder trained with augmented data achieved almost the same perceived speech quality as the vocoder trained with the entire target language dataset. We found that our proposed TTS system consisting of a spectrogram prediction network and a PWG neural vocoder was able to achieve reasonable performance using only 30 minutes of target language training data.