HiLL PUBLISHING

The Educational Review, USA, 2024, 8(12), 1465-1470
https://www.hillpublisher.com/journals/er/
ISSN Online: 2575-7946
ISSN Print: 2575-7938

Research on Personalized Speech Synthesis Model for Korean Language Learners

Xinxin Zhao[1,*], Yunning Wang[2]

[1]International Commerce, Graduate School of International Studies (GSIS), Seoul National University, Seoul 08826, South Korea.
[2]Department in Anthropology, Seoul National University, Seoul 08826, South Korea.

How to cite this paper: Xinxin Zhao, Yunning Wang. (2024). Research on Personalized Speech Synthesis Model for Korean Language Learners. *The Educational Review, USA*, 8(12), 1465-1470.
DOI: 10.26855/er.2024.12.008

Received: October 31, 2024
Accepted: November 29, 2024
Published: December 27, 2024

Corresponding author: Xinxin Zhao, International Commerce, Graduate School of International Studies (GSIS), Seoul National University, Seoul 08826, South Korea.

Abstract

This research aims to explore and implement a personalized speech synthesis model tailored for Korean language learners. Despite significant advancements in general speech synthesis technology, the quality and naturalness of speech synthesis remain challenging for Korean language learners. In this study, we employ deep learning techniques and combine research on facial muscle movements with speech learning to design an innovative framework for personalized speech synthesis. Initially, a substantial amount of speech data from Korean language learners is collected and subjected to preprocessing and annotation. Subsequently, we construct a personalized synthesis model based on deep neural networks to achieve pronunciation correction and fluency improvement for individual learners. The novelty of this research lies in the integration of facial muscle movements with speech learning, leading to optimization in personalized speech synthesis. This innovation holds vital practical implications for enhancing Korean language learners' pronunciation and improving their language communication skills.

Keywords

Korean Phonology; Phonetics; Natural Language Processing; Speech Synthesis; Speech Recognition

Introduction

In the context of today's globalization, learning foreign languages has become a common choice for an increasing number of people. As one of the major languages worldwide, Korean has drawn significant interest from non-native speakers for learning. However, mastering Korean pronunciation may pose a challenge for non-native learners. Pronunciation accuracy is a crucial aspect of language learning, as it not only affects learners' oral communication skills but also has significant implications for listening comprehension and speech synthesis technology.

In the field of speech synthesis, personalized synthesis technology has made remarkable progress. By utilizing deep learning and artificial intelligence techniques, customized speech synthesis can be achieved based on individual voice characteristics, pronunciation habits, and facial muscle movements, resulting in more natural and individualized synthesized speech.

The motivation of this research lies in combining the strengths of language learning and personalized speech synthesis to offer a more effective method for pronunciation improvement to Korean learners. By analyzing the specific pronunciation characteristics and difficulties faced by Korean learners, as well as exploring issues in speech synthesis and addressing the demands for personalized speech synthesis, the aim is to develop an innovative and precise personalized speech synthesis model that enhances the pronunciation accuracy and language learning efficiency of Korean learners.

The primary objective of this study is to construct a personalized speech synthesis model tailored to Korean learners. This model aims to generate natural speech that aligns with the learners' habitual pronunciation and voice characteristics, accommodating individual differences.

A review of related works reveals that current research mainly focuses on language learning and speech synthesis as separate domains, with limited studies on personalized speech synthesis targeted specifically at Korean learners. Thus, this research endeavors to bridge this gap between the two fields and provide Korean learners with more effective support for language learning.

Overall, this study aspires to contribute to the field by creating a sophisticated and innovative approach to personalized speech synthesis for Korean learners, fostering improved pronunciation and language learning outcomes.

1. Analysis of Pronunciation Issues for Korean Language Learners

1.1 Characteristics and Challenges of Korean Language Learners' Pronunciation

During the process of learning Korean, learners whose native language is different face many challenges in terms of pronunciation. Korean has significant differences in phonetic systems, phonetic structures, and tones compared to other languages such as Chinese and English, leading learners to encounter the following characteristics and difficulties:

(1) Differences in Phonemic Systems: Korean's phonemic system differs from languages like Chinese and English, including some special consonants and vowels that learners find challenging to distinguish and produce accurately (Wang, Y., 2024).

(2) Complex Syllable Structure: Korean's syllable structure is relatively complex, involving various structures like consonant-vowel (C-V), vowel-consonant (V-C), and consonant-vowel-consonant (C-V-C). Learners often make errors in organizing syllables and transitioning between phonemes.

(3) Tone Differences: Tone variations in Korean can impact word distinctions in certain cases. If learners cannot accurately grasp these tone features, it may lead to communication difficulties or semantic misunderstandings.

(4) Non-native Acquisition: Non-native learners often require special training and practice to master the pronunciation characteristics of Korean, which traditional methods of language learning may not effectively address.

1.2 Current Issues in Speech Synthesis

Speech synthesis technology is an essential speech-processing technique used to generate natural and fluent speech output. However, when converting input from Korean language learners into natural and fluent Korean speech, the following problems still exist:

(1) Speech Quality: During speech synthesis, the generated speech may not sound natural enough, with issues such as unnatural timbre and intonation. This is particularly important for Korean learners who need high-quality speech synthesis as a reference for learning.

(2) Pronunciation Accuracy: Current speech synthesis models may struggle to fully understand the pronunciation of Korean learners, resulting in lower accuracy in synthesized speech and even pronunciation errors, which can affect learners' language learning outcomes.

(3) Personalized Needs: Different Korean learners may have diverse learning goals and speech learning requirements. Existing speech synthesis systems often lack personalization for individual learners, leading to synthesized speech that does not meet specific needs.

1.3 The Importance and Demand for Personalized Speech Synthesis

Personalized speech synthesis caters to the specific pronunciation characteristics and learning needs of each Korean language learner, holding the following importance and demand:

(1) Improved Learning Efficiency: Personalized speech synthesis optimizes speech output based on learners' phonetic characteristics, enhancing pronunciation accuracy and fluency, thus accelerating the learning process.

(2) Personalized Learning Assistance: Tailored speech synthesis provides customized speech output in different styles and emotions based on learners' learning goals and needs, offering targeted learning assistance, stimulating interest, and enhancing motivation.

(3) Speech Correction and Improvement: By analyzing learners' pronunciation errors and difficulties, personalized speech synthesis can help learners correct pronunciation issues, refine learning strategies, and improve language learning outcomes.

(4) Education and Intelligent Applications: Personalized speech synthesis technology holds vast potential in the

education field, applicable to intelligent education systems, online learning platforms, etc., providing personalized speech teaching services to meet diverse learning needs.

In conclusion, research on personalized speech synthesis for Korean language learners holds significant academic significance and practical application value in improving learning outcomes, advancing speech synthesis technology, and promoting intelligent education.

2. Personalized Speech Synthesis Model Design

2.1 Data Collection and Preprocessing

In the design of the personalized speech synthesis model, the first step is to collect pronunciation data from Korean language learners. These data can be obtained through recording devices in specific language learning environments. The data collection process may involve learners reading specific words, phrases, or sentences, covering different phonemes, speech rates, and pronunciation styles. To enhance the model's generalization capability, it is advisable to consider a diverse set of learners, including students from different countries and cultural backgrounds.

Preprocessing is essential to transform the collected raw speech data into a form suitable for model training. Preprocessing steps may include sound signal sampling and quantization, noise and silence removal, and the addition of alignment annotations. For diversified personalized speech synthesis, it may be necessary to label the data with speaker identities so that the model can distinguish different learners. Furthermore, data augmentation techniques such as time stretching and pitch shifting can be employed to expand the dataset and improve the model's robustness and generalization ability.

2.2 Deep Learning-based Personalized Speech Synthesis Framework

The deep learning-based personalized speech synthesis framework aims to create a highly customizable speech synthesis system to meet the needs of individual learners. This framework can be built upon end-to-end speech synthesis models such as WaveNet, Tacotron, or Transformer TTS. The model's input can be either text converted into phoneme sequences or speech feature sequences, such as mel-spectrogram features.

Regarding personalization, learner-specific identity information, such as speaker embeddings, can be introduced to represent the distinct speech characteristics of different learners. This way, the model can learn the unique speech features and pronunciation styles of each learner, enabling personalized speech synthesis (Chen, Z. et al., 2023).

2.3 Research on the Correlation Between Facial Muscle Movements and Speech Synthesis

The significance of this research is to explore the correlation between facial muscle movements and speech synthesis, providing a new perspective for personalized speech synthesis. Facial muscle movements are closely related to pronunciation, as different pronunciations involve different patterns of facial muscle movements. Therefore, incorporating facial muscle movement information into the speech synthesis model may help accurately reproduce an individual's pronunciation characteristics.

By using facial motion sensors or facial tracking technology, it is possible to capture facial muscle movement information during the learner's pronunciation process. Integrating this information into the speech synthesis model can aid the model in better understanding the learner's pronunciation behavior, leading to more accurate and personalized speech synthesis results. Additionally, this research can offer new avenues for speech pronunciation correction, assisting learners in improving their pronunciation and enhancing the naturalness and fluency of speech synthesis.

In conclusion, the design of a personalized speech synthesis model can not only provide Korean language learners with a voice learning aid that matches their individual characteristics but also offer new ideas and methods for the advancement of speech synthesis technology. These studies have both academic value and research significance, contributing significantly to the development of personalized education and human-computer interaction fields.

3. Experiment and Result Analysis

3.1 Experimental Design and Dataset Description

In this study, we will design a series of experiments to validate the effectiveness of a personalized speech synthesis model for Korean language learners. Firstly, we will collect a batch of speech data from Korean learners with different pronunciation levels and speech characteristics. The data collection will include recordings of their pronunciation during the

learning process, covering various phonemes, words, and sentences. To ensure dataset diversity, we will select samples from learners of different ages, genders, and language backgrounds.

Next, we will preprocess the data, including steps such as audio quality enhancement, noise removal, and feature extraction. For feature extraction, we will use common acoustic features such as Mel Frequency Cepstral Coefficients (MFCC) and pitch as input data for the model.

To evaluate the effectiveness of the personalized model, we will adopt the method of cross-validation to split the dataset. We will divide the dataset into a training set and a test set, where the training set will be used for model training and parameter optimization, and the test set will be used for evaluating model performance.

3.2 Personalized Model Performance Evaluation Metrics

To evaluate the performance of the personalized speech synthesis model, we will use a series of objective and subjective evaluation metrics. Objective metrics will primarily quantify the model's accuracy and efficiency, while subjective evaluation will measure the naturalness and fluency of the synthesized speech.

Regarding objective metrics, we will use the following indicators for evaluation:

(1) Speech Synthesis Accuracy: Evaluating the proximity between the model's synthesized speech and the original speech.

(2) Pronunciation Correction Effectiveness: Assessing the model's effectiveness in correcting learners' pronunciation issues.

(3) Speech Synthesis Fluency: Evaluating the fluency and coherence of the model's synthesized speech.

For subjective evaluation, we will invite a group of native Korean speakers to participate in auditory tests. They will be asked to evaluate the naturalness, pronunciation accuracy, and comprehensibility of the personalized synthesized speech. We will use the Subjective Mean Opinion Score (MOS) and other standard evaluation methods to obtain reliable subjective assessment results (Niu, F. & Wushouer, S., 2022).

3.3 Results and Analysis

We will conduct a comprehensive evaluation and analysis of the personalized speech synthesis model based on the experimental results. First, we will compare the model's performance across different evaluation metrics and compare it with traditional generic speech synthesis models. Through a combined analysis of objective metrics and subjective evaluation results, we will validate the effectiveness of the personalized model in improving Korean learners' pronunciation and fluency.

Next, we will further analyze the advantages and limitations of the personalized model. Through an in-depth examination of the experimental results, we will identify the model's applicability among learners with different pronunciation levels and speech characteristics, and explore its potential advantages in addressing specific speech issues.

Lastly, we will discuss the practical implications of the experimental results. The success of the personalized speech synthesis model will provide Korean language learners with a more personalized and efficient speech learning aid. We will explore the practical application prospects of the personalized model in education, speech correction, and speech communication while providing valuable references for further improving and optimizing personalized speech synthesis technology.

4. Applications and Effects of Personalized Speech Synthesis for Korean Language Learners

4.1 Pronunciation Correction and Fluency Improvement

During the process of learning Korean, learners often encounter issues with inaccurate pronunciation and low fluency. Personalized speech synthesis models can address these challenges through deep learning algorithms by utilizing learner-specific pronunciation data to correct pronunciation errors and improve fluency. This model can capture learners' pronunciation deviations and speech characteristics, providing customized pronunciation correction suggestions. Experiments have shown that compared to traditional generic speech synthesis models, personalized models significantly enhance learners' pronunciation accuracy and fluency. This holds important academic significance and practical value in helping learners overcome native language transfer barriers, boost confidence, and improve learning efficiency.

4.2 Design and Application of Personalized Learning Assistance Systems

Another crucial application of personalized speech synthesis models is in learning assistance systems. By integrating personalized speech synthesis technology into learning assistance applications, we can design a tailored learning

companion that caters to each learner's specific needs and learning style, offering personalized learning support and feedback (Zeng, Z., Wang, J., & Cheng, N., 2020). This system can dynamically adjust speech rate, intonation, and teaching content based on learners' progress and comprehension, making it easier for learners to understand and absorb Korean knowledge. Moreover, personalized learning assistance systems can adapt teaching styles and methods according to learners' preferences, thereby enhancing motivation and interest in learning. Researching and applying such personalized learning assistance systems are academically valuable and practically significant in promoting personalized education and intelligent learning environments.

4.3 Potential Applications in the Education Sector

Personalized speech synthesis technology holds vast potential for the education sector. Firstly, it can be applied in the field of Korean language teaching, providing personalized audio instruction to non-Korean background learners, thereby improving pronunciation and fluency and enhancing learning efficiency. Secondly, this technology can be utilized in other language teaching fields, delivering personalized audio instruction tailored to learners' native language backgrounds. Additionally, personalized speech synthesis technology can be employed in assisting special education, and aiding students with speech disorders or pronunciation difficulties through personalized speech training and therapy. In the educational technology domain, the application of personalized speech synthesis technology can offer more humane and intelligent learning experiences for online learning platforms and educational robots. Thus, exploring and implementing the potential of personalized speech synthesis technology in the education sector holds profound academic significance and societal value in elevating educational quality and promoting education intelligence (Pan, X. et al., 2021).

5. Conclusion

In this study, we conducted an in-depth exploration of the speech synthesis issues faced by Korean language learners and proposed a personalized model-based solution. By analyzing the pronunciation characteristics and difficulties of Korean learners, we gained a profound understanding of the limitations of traditional speech synthesis techniques in meeting individual learners' needs. Therefore, we devised a personalized speech synthesis framework that correlates facial muscle movements with speech synthesis to provide a more accurate, natural, and personalized speech synthesis experience.

The main contributions of this research are as follows:

Firstly, we introduced the concept of personalized models in the field of speech synthesis and designed a corresponding framework tailored to the characteristics of Korean language learners. This innovative endeavor opens up new possibilities for the personalized development of speech synthesis technology.

Secondly, through the study of the correlation between facial muscle movements and speech synthesis, we delved into the impact of facial expressions on pronunciation. This study not only fills a gap in the field of speech synthesis for Korean language learners but also provides insights into the application of speech synthesis technology in learning other languages.

Thirdly, our experimental results demonstrate the effectiveness of personalized models in Korean language learner speech synthesis, offering new possibilities for the application of speech synthesis technology in the field of education. The design of personalized pronunciation correction and learning support systems holds promising potential in Korean language education.

Lastly, we thoroughly discussed the limitations of personalized speech synthesis technology and proposed future directions for improvement and optimization. We believe that with the continuous advancement of technology, personalized speech synthesis will find applications in a broader range of fields.

Acknowledgment

We extend our gratitude to Seoul National University, our alma mater. We are thankful for the academic courses and opportunities provided by the university throughout our research journey. The professors have had a profound impact on us, imparting their expertise and experience, which helped us overcome challenges and make progress in our studies. Their guidance not only extended to academic matters but also extended to caring for us in our personal life, making us feel thoroughly supported and encouraged.

References

Chen, Z., Zhang, Z., Wang, B., & Xie, Y. (2023). Application and prospect of AI speech synthesis technology. *Film and Television Production*, *29*(3), 51-55.

Niu, F., & Wushouer, S. (2022). Prosodic enhanced Chinese speech synthesis system. *Modern Electronic Technology*, *45*(13), 87-92.

Pan, X., Lu, T., Du, Y., & Tong, X. (2021). A review of speech synthesis and conversion technology based on deep learning. *Computer Science*, *48*(8), 200-208.

Wang, Y. (2024). Limitations and potentials of ChatGPT-4 in gender expression: A future research outlook from a feminist perspective. *Edelweiss Applied Science and Technology*, *8*(6), 5855-5868. https://doi.org/10.55214/25768484.v8i6.3267.

Zeng, Z., Wang, J., & Cheng, N. (2020). Prosody learning mechanism for speech synthesis system without text length limit. Conference presentation.