Automatic Speech Recognition-enabled Pronunciation Learning System Development for Second Language Speakers of Korean

Seung Hee Yang

Spoken Language Processing Laboratory, Interdisciplinary Program in Cognitive Science,
Seoul National University, Republic of Korea

sy2358@snu.ac.kr

1. Motivation

One-on-one tutoring is ideal in second language learning, especially during speaking practice. However, this can be costly for the learners, and technology can be a practical way of simulating realistic interactions in a private and stress-free environment, and thereby providing the beneficial effects of one-on-one tutoring. These systems offer pronunciation scores and feedback on oral proficiency on the word and utterance levels. There has been an increasing interest in developing such systems through the use of ASR (Automatic Speech Recognition) technology.

However, pronunciation variations in non-native speech are far more diverse than those observed in native speech. This poses a difficulty for Computer-Assisted Pronunciation Training (CAPT) systems to automatically recognize and assess learners' speech, detect mispronunciations, and provide corrective feedbacks [1]. Despite the growing popularity in learning Korean as a foreign language and the rapid growth in language learning applications [2], the existing Korean language learning systems do not utilize linguistic characteristics of L2 Korean speech. For an effective CAPT system, it is essential to identify frequent variation patterns, identify what actually matters in native speakers' intuitive judgements of accentedness, and evaluate the system against these language-dependent findings.

For these reasons, the goal of this thesis is to investigate how ASR, assessment, and feedback generation models can be successfully be designed, developed, and tested in a CAPT application for Korean. The present thesis describes and conducts three-fold experiments towards this end goal of automatic speech recognition-enabled language learning system development, which include automatic recognition and assessment of the learners' speech, and corrective feedback generation.

2. Key Issues: Identification and Solutions

A development of a CAPT system requires speech recognition and the functionalities necessary for pronunciation assessment and feedback generation. The three subsections below identify the key issues in each component, design solving approaches, conduct experiments, and present their findings.

2.1. Pronunciation Error Pattern Discovery and ASR for Non-native Speech

The key issue in employing ASR systems in this task is that they are generally not suited for CAPT applications, since they lack the flexibility to recognizing speech of low-proficient non-native speakers. As CAPT applications have to cope with non-native speech that is challenging, it is necessary to develop a dedicated ASR technology.

In order to conduct a speech recognition experiment, non-native Korean pronunciation error patterns were first identified. A corpus-based analysis indicates that lenition, non-realization of phonological rules, and coda deletions and insertions are characteristic of non-native Korean speech, and that optimizing the pronunciation model based on the analysis results improved speech recognition performance, reducing the relative error rate by 12.21% [3], as shown in Table 1.

Table 1: *Recognition results for pronunciation models for speaker levels (in WER). The best WER is achieved for all levels when phonological, substitution, and insertion variation patterns are modelled in the ASR system.*

Pronunciation Model	Beginner	Interme-diate	Advanced	Aver-age
Baseline	23.79	13.65	8.6	15.49
Deletion	23.85	13.65	8.37	15.43
Phonology	23.85	13.49	8.31	15.37
Substitution	22.08	11.72	7.31	13.87
Insertion	20.44	11.72	7.13	13.21
Subs. & Ins.	19.85	10.1	6.54	12.34
Phon. Subs. & Ins.	**19.67**	**10.02**	**6.42**	**12.21**
Del. & Phon. & Subs. & Ins.	19.91	**10.02**	6.48	12.32

2.2. Automatic Assessment of Non-native Speech

The development of CAPT systems involves an interplay between several individual components. As these are interconnected and interdependent, a full appreciation of the complexity observed in developing these systems cannot be obtained by considering it from a single perspective. This is the reason why, besides the research related to the individual components, the current thesis proposes a Generative Adversarial Network (GAN)-based approach [5]. The proposed method, thanks to the adversarial nature of GANs, has the potential to connect assessment and corrective feedback in a single model, reducing the difficulties in integrating independent modules into a single architecture.

The second experiment was conducted for the development of automatic speech assessment model using GAN. While traditional automatic speech assessment techniques use hand-crafted features to evaluate the goodness of second language learners' speech [4], the present thesis proposes a novel speech assessment method that works in an unsupervised way. The generator (G) performs mapping and generates fake samples that imitate the real data distribution. G learns this by adversarial training, where the discriminator D classifies whether an input is a fake sample generated by G or a real sample. The discriminator backpropagates and pass-

Table 2: *Classification results of the discriminator in GAN for Automatic Scoring*

Class \ Classified	Non-native	Native	Total
Non-native	80	23	103
Native	0	29	29

es information to G on what is real and what is fake, and in turn, G tries to generate better imitations by adapting its parameters towards the real distribution. This adversarial learning process is formulated as a minimax game between G and D:

$$\min_G \max_D V(D,G) = \mathbb{E}_{x\sim P_{data}(x)}[\log D(\mathrm{X})] + \mathbb{E}_{z\sim P_{z(z)}}[\log(1\text{-}D(G(\mathrm{Z}))]. \quad (1)$$

where $P_{data}(x)$ is the real data distribution, and $P_{z(z)}$ is the prior distribution. For a given x, D(X) specifies the probability x is drawn from $P_{data}(x)$ and D(G(Z)) specifies the probability that the generated distribution is drawn from $P_{z(z)}$.

The proposed model is trained on L2KSC (L2 as Korean Speech Corpus) [6]. There are 217 non-native speakers with 27 mother tongue backgrounds, and 107 native speakers of 54 females and 53 males. Each speaker read 300 short utterances, which are in average one second in length. Experiments confirm the discriminator's potential in performing assessment task [7]. Although the segmental accuracy is not strongly correlated with the assessment score (r=0.370, p<0.001), F1 classification score achieves 0.870, as shown in Table 2, which is comparable to the reported performances in the related works [8,9].

2.3. Self-Imitating Corrective Feedback Generation

In self-imitating feedback, the characteristics in native utterances are extracted and transplanted onto the learner's speech. Listening to the manipulated speech enables students to understand the differences between the accented utterances and the native counterparts, and to produce native-like utterances by self-imitation. The third experiment was conducted with the generator of a GAN, which can generate corrective feedback based on the learner's own voice, in which characteristics in native utterances are learned and transplanted onto learner's own speech input, and are played back to the learner as a corrective feedback.

Conditional GANs (cGANs) learn a conditional generative model where we condition on the input and generate objective corresponding output [10]. G tries to minimize the below objective against an adversarial D that tries to maximize it.

$$L_{cGAN}(\mathrm{G},\mathrm{D}) = \mathbb{E}_{x,y}[\log D(x,y)] + \mathbb{E}_{x,y}[\log(1\text{-}D(x,G(x,z))]. \quad (2)$$

[10] demonstrated that cGANs can solve a wide variety of problems by testing the method on nine different graphics and vision tasks, such as style transfer and product photo generation. By interpreting speech correction task as a spectrogram translation problem, we explore the generality of conditional GANs.

[11] introduced cycle consistency loss to further reduce the space of possible mapping functions. Cycle consistency loss was implemented in the present thesis in order to encourage the output to preserve the global structure, which is shared by native and non-native utterances. This is incentivized by the idea that the learned mapping should be cycle-consistent, which is trained by the forward and backward cycle-consistency losses:

$$L_{cyc}(G,F) = \mathbb{E}_{y\sim P_{data}(x)}[\|F(G(x) - x\|_1] + \mathbb{E}_{y\sim P_{data}(y)}[\|G(F(y)\text{-}y|_1]. \quad (3)$$

To train the corrective model, 97,200 utterances of native and non-native Korean speech were used. In this framework, each utterance is transformed into spectrograms. The generator transforms non-native to native speech spectrograms, which is then converted back to speech. Perceptual evaluation of nativeness and auditory transcription of the generated utterances on a held-out test set show that the newly proposed CycleGAN-based speech feedback method is able to correct common segmental errors in non-native Korean speech better than the traditional PSOLA (Pitch-Synchronous Overlap and Add) algorithm [12], as shown in human MOS evaluation scores in Table 3 [13].

3. Contribution Points and Future Works

This study lays the groundwork for speech recognition-enabled language learning software development for Korean as a foreign language. This is a first attempt, to the best of my knowledge, at using GAN for language learning application. Its unsupervised nature allowed many contribution points, such as independence from feature extraction and annotation efforts, simpler integration process of individual CAPT components, better or equivalent performances as the traditional methods, and possibility of easier language expansion. Moreover, another major contribution in this improvement is that the current method is able to correct both suprasegmental and segmental mispronunciations, which is significantly better than the traditional methods that are limited to suprasegmental correction. Since the segmental accuracy plays an important role in non-native communication in Korean [14], it is meaningful that the current method overcomes the limitations of the traditional approach, which has been limited to the prosodic level only.

According to the results, I consider the four following future works. First, the loss during Griffin-Lim inversion may be causing lower scores in generated sound qualities, and an improved inversion method seems desirable. Moreover, there is a room for improvement in CycleGAN's imitability score, which may be due to the diversity in reference styles. Future work can investigate better speaker voice imitation methods. Third, the feasibility of a real-time interactive response generation needs to be tested, including, but not limited to parallelization techniques for GAN algorithms. Last, I plan to apply and test the proposed method in an integrated CAPT system that is used by real language learners.

Table 3: *MOS values of perceptual test by four human experts on self-imitation feedback generation (SQ: Sound Quality)*

Model	Corrective Ability			Imit-ability	SQ	Avg.
	Holis-tic	Seg-mental	Supra-segmental			
PSOLA	3.118	3.029	3.324	4.029	2.794	3.259
Pix2Pix	1.970	2.485	2.152	2.697	1.636	2.188
Cycle-GAN	4.000	4.333	4.364	3.515	2.667	3.776

4. Acknowledgements

I express sincere gratitude to my advisor, Professor Minhwa Chung. I thank Kyuwhan Lee for support and valuable technical discussions.

5. References

[1] Eskenazi, M., 2009. “An overview of spoken language technology for education,” Speech Communication, 51, p. 832-844.

[2] Shin, Y., 2016. “Number of Foreign Students in Korea Reaches over 100,000,” Yonhap News. (In Korean)

[3] S.H. Yang, M. Na & M. Chung (2015). Modeling Pronunciation Variations for Non-native Speech Recognition of Korean Produced by Chinese Learners. Proceedings of SLaTE 2015, 95-99.

[4] K. Zechner, D. Higgins, X. Xi, and D. M. Williamson, “Automatic scoring of non-native spontaneous speech in tests of spoken English,” Speech Communication, vol. 51, pp. 883–895, 2009.

[5] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. WardeFarley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial nets,” *Advances in Neural Information Processing Systems*, 2014, pp. 2672–2680.

[6] S. Lee, and J. Chang, “Design and Construction of Speech Corpus for Korean as a Foreign Language (L2KSC),” The Journal of Chinese Language and Literature, vol. 33, 2005, pp. 35-53.

[7] S.H. Yang and M. Chung, "Speech Assessment using Generative Adversarial Network," Proceedings of Machine Learning in Speech and Language Processing Workshop, 2018

[8] A. Metallinou and J. Cheng, “Using deep neural networks to improve proficiency assessment for children English language learners,” in INTERSPEECH 2014 – 94th Annual Conference of the International Speech Communication Association, September 14-18, Singapore, Proceedings, 2014, pp. 1468–1472.

[9] F. Hönig, B. Anton, W. Karl, and N. Elmar, "Automatic assessment of non-native prosody for english as l2." In Speech Prosody 2010-Fifth International Conference. 2010.

[10] P. Isola, J.Y. Zhu, T. Zhou, A. Efros, "Image-to-Image Translation with Conditional Adversarial Networks," *Proceedings of CVPR*, 2017.

[11] J.Y. Zhu, T. Park, P. Isola, A. Efros, "Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks," *Proceedings of ICCV,* 2017.

[12] F. Charpentier and E. Moulines, “Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones,” *Proceedings of the First European Conference on Speech Communication and Technology, Eurospeech*, 1989, pp. 2013-2019.

[13] S.H. Yang and M. Chung, “Self-imitating Feedback Generation Using GAN for Computer-Assisted Pronunciation Training,” https://arxiv.org/abs/1904.09407. (To be published in Proceedings of INTERSPEECH 2019, Graz, Austria).

[14] S.H. Yang and M. Chung, "Linguistic Factors Affecting Evaluation of L2 Korean Speech Proficiency," *Proceedings of SLaTE 2017*, *Stockholm, Sweden,* 2017, pp. 53-58.