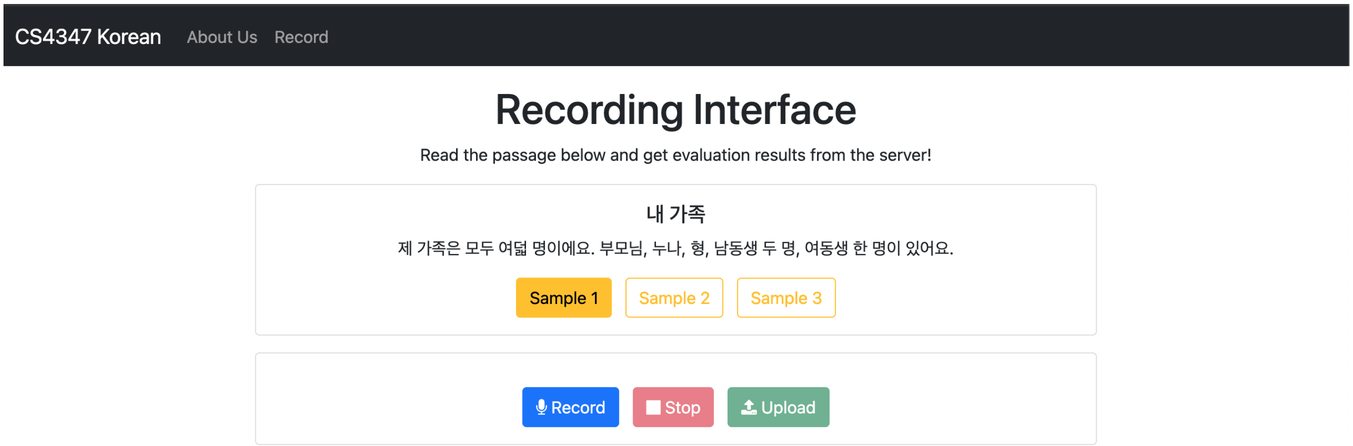
Automated Pronunciation Evaluation for Korean Language Learners

Dolly Agarwal[*]
e0674474@u.nus.edu
National University of Singapore
Singapore

Eugene Tan Yew Chin[*]
e0321481@u.nus.edu
National University of Singapore
Singapore

Ronald Santoso[*]
ronald.s@u.nus.edu
National University of Singapore
Singapore



Recording Interface

Read the passage below and get evaluation results from the server!

내 가족

제 가족은 모두 여덟 명이에요. 부모님, 누나, 형, 남동생 두 명, 여동생 한 명이 있어요.

Figure 1: A screenshot of the web application interface

ABSTRACT

The Korean Wave has successfully attracted many to the Korean culture, and this can be seen in the rise of number of Korean as a Second Language (L2) learners around the world. While providing individual corrective feedback on pronunciation is crucial for language learning, this process is particularly time-consuming for teachers due to the sheer number of students enrolled in the class. Additionally, if multiple examiners are involved in the assessment of students' pronunciation, it is hard to avoid subjectivity between examiners, leading to more time required for standardisation. This paper focuses on the development of an Automatic Speech Recognition (ASR) based system to evaluate the pronunciation of Korean L2 learners using Kaldi, an open-source speech recognition toolkit. Two different acoustic models had been developed: Hidden Markov Model - Gaussian Mixture Model (HMM-GMM) and Hidden Markov Model - Deep Neural Network (DNN), and the HMM-GMM model is used in the web application prototype. An experiment was also carried out to compare the performance between the HMM-GMM model and the Google Speech-to-Text (STT) Application Programming Interface (API). While Google STT outperforms our system in terms of accuracy, developing our own ASR system gives us flexibility in terms of obtaining a detailed evaluation result.

CCS CONCEPTS

• **Software and its engineering** → *Software prototyping*; • **Applied computing** → **E-learning**; **Sound and music computing**.

KEYWORDS

korean, e-learning, ASR, web development

ACM Reference Format:
Dolly Agarwal, Eugene Tan Yew Chin, and Ronald Santoso. 2018. Automated Pronunciation Evaluation for Korean Language Learners. In *Woodstock '18: ACM Symposium on Neural Gaze Detection, June 03–05, 2018, Woodstock, NY*. ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

In this globalised world, where people from various parts of the world are interacting on a regular basis, language does not only serve as a medium of communication, but also a gateway for cultural appreciation and mutual understanding. The acquisition of L2 allows individuals to interact with members of various communities and build more meaningful relationships with people of various backgrounds.

With the rise of the Korean Wave [1], the world is no stranger to the Korean culture: its songs, food, fashion, drama and many more. In fact, fans of Hallyu have surged to over 100 million with

[*] All authors contributed equally to this research.

Woodstock '18, June 03–05, 2018, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/18/06...$15.00
https://doi.org/10.1145/1122445.1122456

the happenings of stay-at-home, due to COVID-19 [2]. This has no doubt inspired a significant increase in Korean L2 learners, who may learn the language for a variety of reasons. For instance, one could be able to better appreciate Korean pop culture through television shows and songs, or that learners may be able to understand Korea better and interact with Korean locals on a vacation. Whatever the reason may be, learning a new language often involves a large amount of effort, from understanding the basic language structures to learning how to speak sentences.

A common problem of L2 learners is that they may be passionate about learning a new language, but it is difficult to tell if their progress is headed in the right direction. This is because they do not get timely corrective feedback on their pronunciation, which is essential for pronunciation learning. For students who learn under the guidance of language teachers, an issue of a possibly low teacher to student ratio often results in slower and possibly less fruitful feedback from these teachers because providing individual corrective feedback on pronunciation is particularly time-consuming for teachers. Additionally, if multiple examiners are involved in the assessment of students' pronunciation, it is hard to avoid individual differences (subjectivity) between examiners, leading to more time required for standardization.

Computer Assisted Language Learning (CALL) systems that make use of an ASR seem to offer an alternative for practicing pronunciation because they can offer specific feedback on individual errors and extra time for practicing at the learners' own pace. With this motivation, our team has decided to address the issues of manpower and grading standardization by automating the grading of speech samples, and we have chosen the increasingly popular Korean language to be the L2 language of focus.

Corrective feedback on pronunciation can be given on different aspects. In this paper, we focus on corrective feedback on both word and phoneme levels. For providing global level (word) feedback, pronunciation measures that are calculated over longer stretches of speech are used, while detailed feedback at the segmental level requires computing a score for each individual realization of a given phone.

We have designed two systems such that a comparison of the results can be carried out. The first is based on Kaldi, a state-of-the-art automatic speech recognition toolkit, and the second one is based on Google STT API[1]. Our Kaldi-based model is a GMM-HMM based acoustic model, which was trained on the Zeroth Korean dataset [2]. Given a learner's speech, we pass the audio to both systems, which returns us a forced alignment for our Kaldi-based system and a transcription for both our Kaldi-based system and the Google STT API-based system. We use this information to compute fluency and segmental accuracy. It can be noted that if we had labeled scores of recordings, these extracted features could be used to train a machine learning model to give us an overall proficiency score.

2 RELATED WORK

CALL and Computer Assisted Pronunciation Training (CAPT) have been widely researched topics for more than three decades [8]. One primary reason is the rising number of L2 learners, who expect to learn anytime and anywhere. This also led to a recent proliferation of web-based and mobile language learning apps. However, there have been two main concerns regarding the usefulness of such language learning systems: their ability to detect pronunciation errors and to provide appropriate feedback that can support the learning experience [9].

Most automated speech proficiency assessment systems contain three main components [3]: an ASR system that generates word hypotheses for a given speech sample along with other information, such as the duration of pauses between words, a set of modules based on digital signal processing and natural language processing (NLP) technologies that compute a number of features measuring various aspects of speech considered relevant by language assessment experts (e.g., fluency, pronunciation, segmental accuracy), and finally, a scoring model that maps features to a score using a supervised machine learning paradigm.

For such supervised learning systems, various features have been explored in the literature. ETS reported 29 candidate features to score non-native spontaneous speech in tests of spoken English, such as TOEFL. Most of the features that they reported are related to fluency, such as the number or duration of words or silences.

Another most widely used method for automatic mispronunciation detection is the Goodness Of Pronunciation (GOP) algorithm proposed by Witt [17, 18]. The GOP algorithm calculates the likelihood ratio that the realized phone corresponds to the phoneme that should have been spoken according to the canonical pronunciation. Thresholds, calculated beforehand, are used to decide which likelihood ratio scores corresponded to mispronounced sounds.

It is to be noted that most of the research for CAPT is done for English or European languages [4–7, 11–13]. Fewer studies have been conducted on automatic proficiency assessment of non-native Korean speech. Some research has focused on the analysis of pronunciation variabilities in non-native Korean speech. For instance, [10] examined variations of Korean segments produced by Japanese learners of Korean while [20] modeled pronunciation variations frequently produced by Chinese learners.

A study performed by [19] examined the segmental, phonological, accentual, and temporal correlations of native speakers' evaluation of L2 Korean proficiency produced by learners with various levels and nationalities. They reported that proficiency ratings by native speakers significantly correlate not only with the rate of speech but also with the segmental accuracy. Although phonological accuracy was expected to be highly correlated with the proficiency score, it was the least influential measure according to their study. Another new finding in their study was that the role of pitch and accent has been under-emphasized so far in the non-native Korean speech perception studies.

[16] proposed a method for automatic pronunciation assessment of Korean spoken by L2 learners. They experimented with various features to select the best feature set from a collection of the most well-known features in the literature. They found that by the result of selected features from the Best Subset Selection (BSS) model, most of the salient features correspond to speech rate and segments. Their results show RATE features have better performance than SILENCE, SEGMENT, and GOP features for automatic pronunciation assessment of Korean L2 learners.

[1] https://cloud.google.com/speech-to-text/
[2] https://github.com/goodatlas/zeroth

[14] presents an automatic proficiency assessment method for a non-native Korean read utterance using bidirectional long short–term memory (BLSTM)–based acoustic models (AMs) and speech data augmentation techniques. The study proposed methods for two scenarios: with and without prompted text. The proposed method with the prompted text performs (1) a speech feature extraction step, (2) a forced-alignment step using a native AM and non-native AM, and (3) a linear regression–based proficiency scoring step for the proficiency scores.

Among the spoken Korean CALL applications, this paper focuses on an ASR–based proficiency assessment for non-native Korean speech. We took inspiration from existing studies and focused on methods corresponding to prompted text while designing our system for automated proficiency assessment of Korean L2 learners. We extracted 2 types of features in this work: Fluency and Segmental Accuracy. We have not considered features like LM score, disfluency, repetition and the number of unique words, because these are more relevant in assessing spontaneous speech rather than read speech.

3 APPROACH

3.1 Dataset

Our project uses the Zeroth Korean dataset [15], which is a Korean Open-source Speech Corpus for Speech Recognition by Zeroth Project. The data set contains transcribed audio data for Korean. There are 51.6 hours transcribed Korean audio for training data (22,263 utterances, 105 people, 3000 sentences) and 1.2 hours transcribed Korean audio for testing data (457 utterances, 10 people). The details of the corpus is highlighted in Table 1. This corpus also contains a pre-trained/designed language model, lexicon and morpheme-based segmenter (morfessor).

	Duration (hours)	Utterances	Speakers
Training Data (41 Males, 64 Females)	51.6	22263	105
Test Data (4 Males, 6 Females)	1.2	457	10

Table 1: Distribution of samples in Zeroth Korean Dataset

The technical specifications of the audio recordings found in the Zeroth Korean dataset can be seen in the Table 2.

Audio Format	FLAC
Sampling Rate	16kHz
Resolution	16 bits
Channels	Mono

Table 2: The audio specification of recordings in the Zeroth Korean dataset

3.2 ASR Training

We worked on two systems for the proficiency assessment, one is based on the ASR trained on Kaldi and another is using Google STT API. To train our ASR, we used an NUS cluster as our server. After preparing the data and having set up all the requirements needed, such as obtaining computing resources, compiling Kaldi with GPU support and installing required libraries, we performed training for our Kaldi-based model. We have trained two different systems using Kaldi, one based on HMM-GMM and another based on HMM-DNN, which will be described in greater detail in the next two sub-sections.

3.2.1 HMM-GMM Training. We first trained a Hidden Markov Model-Gaussian Mixture Model system. We performed speech feature extraction to extract the Mel-frequency cepstrum coefficients (MFCC) acoustic features and compute the cepstral mean and variance normalization (CMVN) stats. We started by training a monophone model, which is an acoustic model that does not include any contextual information about the preceding or following phone. It is used as a building block for the triphone models, which make use of contextual information. It is to be noted that we train the initial models only on a subset of data mainly for efficiency.

The parameters of the acoustic model are then estimated in acoustic training steps. However, the process can be better optimized by cycling through training and alignment phases. By aligning the audio to the reference transcript with the most current acoustic model, additional training algorithms can then use this output to continually improve or refine model parameters. Therefore, each training step will be followed by an alignment step, where the audio and text are realigned.

To get a more refined model, we repeated the steps of realigning audio with the acoustic models and retraining with additional triphone training algorithms, such as delta+delta-delta training, Linear Discriminant Analysis – Maximum Likelihood Linear Transform (LDA-MLLT), and Speaker Adaptive Training (SAT). Delta+delta-delta training computes delta and double-delta features, or dynamic coefficients, to supplement the MFCC features. LDA-MLLT is used to build feature vectors in a reduced feature space and to derive a unique transformation for each speaker so that the differences amongst speakers can be minimised (speaker normalization). SAT also performs speaker and noise normalization by adapting to each specific speaker with a particular data transformation. After SAT, the acoustic model is no longer trained on the original features, but on speaker-normalized features.

Hence, for alignment, we remove the speaker identity from the features by estimating the speaker identity using the inverse of the fMLLR matrix, and then remove it from the model by multiplying the inverse matrix with the feature vector. We then use these quasi-speaker-independent acoustic models in the alignment process. Details of the final tri4 model trained can be seen in Table 3.

3.2.2 HMM-DNN Training. We next trained a Hidden Markov Model-Deep Neural Network system. We started DNN training from the labeled frames (phoneme-to-audio alignements), which were generated by the HMM-GMM system. This means that our HMM-DNN system's performance will be greatly affected by the quality of the HMM-GMM system we have trained previously.

Number of phones	194
Number of pdfs	3384
Number of transition-ids	31048
Number of transition-states	15484
Feature Dimensions	40
Number of Gaussians	40035

Table 3: Details of the trained HMM-GMM acoustic model

Firstly, we fed the audio feature frames into the input layer. The network will then assign a phoneme label to a frame, during which a comparison between what the neural net predicted and what the real phoneme was could be carried out. Using a loss function and back propagation, we iterated over all of our training frames to adjust the weights and biases of our network. Table 4 lists the parameters that we used to complete our nnet3 training. However, due to the limited GPU memory in the school server, we had to change a few parameters before training the network for fewer number of epochs so that it completes successfully. It is to be noted that our trained HMM-DNN system has a performance comparable to the author's model, albeit trained at a different configuration. This can be seen from the similar Word Error Rate (WER) shown in Table 4.

	Author's Version	Our Attempt
WER	10.55	11.87
Number of Epochs	10	2
Minibatch size (regular, minimum minibatch size) Note: the minimum minibatch size to be considered as a minibatch applies to the remainder of samples when there are less than the regular number of samples left.	128, 64	64, 32
Number of Jobs (Initial, Final) Note: use of the computing cluster/GPU machine our team adopted requires the maximum of only 1 job. This may affect the results, but at least gives a system that works.	2, 8	1, 1

Table 4: Details of the trained HMM-DNN acoustic model

Training of neural networks on a huge amount of data requires much computing resources. We faced many technical difficulties, including CUDA version issues and insufficient GPU memory. We spent a lot of time resolving these infrastructure related issues with the school's technical team. Although we successfully trained our HMM-DNN based acoustic model, we could not use it for our speech evaluation task because the GPU settings of the machine were somehow reverted by the school. Due to time constraint, we instead focused on computing assessment features using HMM-GMM acoustic model. Given more time and more control over the infrastructure, our system would very likely use a HMM-DNN model instead of the HMM-GMM tri4 model currently being used.

3.3 Assessment Modelling

In order to perform automatic pronunciation assessment for Korean L2 learners, we designed a pronunciation assessment modeling framework as shown in Figure 2.

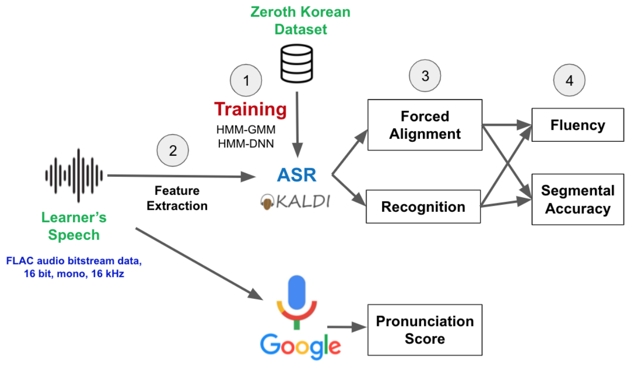


Figure 2: Assessment Modeling Framework

As illustrated in Figure 2, given a learner's speech, we first perform feature extraction to get MFCC features of the audio. It is then forced-aligned and recognized through the trained ASR. Using the results of forced alignment and recognition, we calculate the features for pronunciation assessment. Among the various features, we have selected and categorized features into fluency and segmental accuracy as fluency and segmental accuracy are highly correlated with L2 Korean proficiency [20].

3.3.1 Fluency. Many studies that measured fluency as the evaluation criteria agree that it is a useful measure that highly correlates with native listeners' perception of speech proficiency [4, 17, 18]. Fluency includes rate, which is related to the pace at which learners speak. Novice learners tend to speak slowly and pronounce each syllable separately, which is not observed in native speech. This would discount the fluency score. We rated fluency at the phoneme and word level.

Table 5 summarizes the features we computed to rate the speakers fluency. The definitions of these features based on [11] are as follows:

- **Rate of Speech (ROS)**: ROS is defined as the ratio between the number of speech phones and total duration. [11] reported that ROS had a correlation of 0.81 with the manual pronunciation scoring.
- **Articulation rate (AR)**: AR is defined as the ratio between the number of phones and duration of speech without internal pauses. [5, 12] demonstrated that AR had a correlation of 0.83 with manual ratings for read speech, while it had weak correlation for spontaneous one.
- **Phonation time ratio (PTR)**: It is defined as the ratio between the duration of speech without internal pauses and

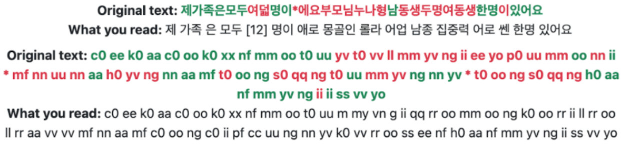
total duration. This feature also had a strong correlation with manual ratings for read speech, according to [12].

- **Words count per minute (WPM)**: To compute WPM, we divide the number of words spoken by the time (in minutes) it took to read them.
- **Words correct per minute (WCPM)**: WCPM is calculated by subtracting the total number of errors from the total number of words read, then dividing it by the time (in minute) it took to read them.

Type	Feature	Formula	Description
Phoneme level	Rate of Speech (ROS)	$\frac{N_{phones}}{T_{total}}$	N_{phones} and T_{total} denote the number of phones and the total duration respectively.
	Articulation Rate (AR)	$\frac{N_{phones}}{T_{nopause}}$	$T_{nopause}$ denotes the duration of speech without internal pauses.
	Phonation Time Ratio (PTR)	$\frac{T_{nopause}}{T_{total}}$	
Word level	Words Count Per Minute (WPM)	$\frac{N_{words}}{T_{total}}$	N_{words} and T_{total} denote the number of words and the total duration (in minutes) respectively.
	Words Correct Per Minute (WCPM)	$\frac{N_{correct}}{T_{total}}$	$N_{correct}$ denotes the number of words read correctly.

Table 5: Phoneme-level and word-level features for fluency

3.3.2 Segmental Accuracy. Studies have shown that proficiency ratings by native speakers significantly correlate not only with fluency, but also with segmental accuracy [13]. To compute segmental accuracy, we phonetically transcribed all segments and rated segmental accuracy by counting the number of mismatch between the canonical and realized pronunciations. We also compared the transcription obtained from ASR with the canonical text at word level. An example can be seen in Figure 3. Green marks correctly realized phonemes/words whereas red marks mispronounced phonemes/words.



Original text: 제가족은모두여덟명이*에요부모님누나형남동생두명여동생한명이있어요

What you read: 제 가족 은 모두 [12] 명이 애로 몽골인 롤라 어업 남종 집중력 어로 쎈 한명 있어요

Original text: c0 ee k0 aa c0 oo k0 xx nf mm oo t0 uu yv t0 vv ll mm yv ng ii ee yo p0 uu mm oo nn ii * mf nn uu nn aa h0 yv ng nn aa mf t0 oo ng s0 qq ng t0 uu mm yv ng nn yv * t0 oo ng s0 qq ng h0 aa nf mm yv ng ii ii ss vv yo

What you read: c0 ee k0 aa c0 oo k0 xx nf mm oo t0 uu m my vn g ii qq rr oo mm oo ng k0 oo rr ii ll rr oo ll rr aa vv vv mf nn aa mf c0 oo ng c0 ii pf cc uu ng nn yv k0 vv rr oo ss ee nf h0 aa nf mm yv ng ii ss vv yo

Figure 3: Segmental accuracy at word and phoneme level

3.4 Web Development

In order to showcase a prototype of the Automated Pronunciation Evaluation system, a simple web application is developed using Python Flask. This website is hosted on a virtual machine in the NUS School of Computing's server and a reverse proxy is required to allow any external parties to access the website hosted on the intranet. A working prototype of the website can be seen here: https://cs4347-korean.comp.nus.edu.sg/record. Refer to Figure 4 for the architecture diagram of the web application.

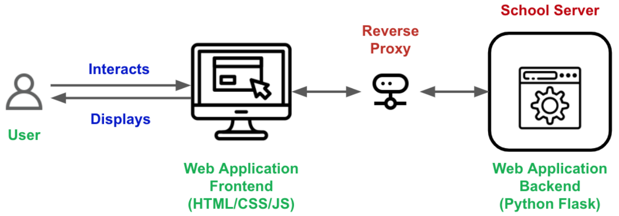

School Server
Reverse Proxy
Interacts
Displays
User
Web Application Frontend (HTML/CSS/JS)
Web Application Backend (Python Flask)

Figure 4: Architecture diagram of the web application

The website's recording interface consists of three basic functionalities at this point: a sentence selection interface, a recording interface and an evaluation interface. The sentence selection interface consists of three different Korean sentences catered towards learners of Korean 1 module in National University of Singapore. It is worth noting that these sentences are hard-coded in this prototype, but it is possible to implement a database storing all the different sentences that the teachers wish to test the students on. The recording interface is built on Recorder.js, a JavaScript library that allows audio recording from a browser. This library enables users to record their voice, listen to a playback of their voice and save this recording. Finally, the evaluation interface consists of two different parts: a Kaldi-based evaluation and a Google STT-based evaluation. Under Kaldi-based evaluation, segmental accuracy and fluency features at both the word and phoneme level will be displayed to the user. Under Google STT-based evaluation, only segmental accuracy at the word level will be displayed to the user.

The workflow of the website is as follows. A student will select the desired sentence to read aloud from the sentence selection interface. Using the recording interface, the student will record himself or herself reading the sentence aloud. When this student uploads the recording to the server, the recording is sent to both the Kaldi-based system and the Google STT-based system. Once both systems return the evaluation result, the evaluation interface will display the evaluation results to the user.

It is worth noting that while the performance of Google STT is better than our Kaldi system as it uses deep neural network, our Kaldi system is more flexible as it is able to produce acoustic features from the forced alignment that can be used as a metric to evaluate the fluency and accuracy of the students' pronunciation at a more fine-grained level.

4 RESULT AND DISCUSSION

In order to evaluate the performance of our pronunciation evaluation system, we have decided to compare the accuracy of our

Kaldi-based system transcription with the transcription generated from Google STT API.

In the first experiment, the following sentence below was read out and evaluated by both systems: 집에 부모님, 저, 동생이 있어요. 하지만 누나는 지금 일본에 있어요. When reading the sentence, the experimenter supplies a recording that is exactly the same as the canonical text. The result of the experiment can be seen below.

	Positive	Negative
Positive	26	0
Negative	0	0

Table 6: Confusion matrix for Google STT (Experiment 1)

	Positive	Negative
Positive	15	11
Negative	0	0

Table 7: Confusion matrix for Kaldi (Experiment 1)

As seen in Table 6 and 7, the Google STT transcription is exactly the same as the canonical test, leading to all 26 characters being transcribed correctly. However, the Kaldi-based system only managed to transcribe 15 characters correctly, and 11 other characters which were correctly pronounced but falsely classified as being incorrectly pronounced (false negatives).

In the second experiment, the following sentence was read out and evaluated by both systems: 제 가족은 모두 여덟 명이에요. 부모님, 누나, 형, 남동생 두 명, 여동생 한 명이 있어요. When reading the sentence, the experimenter deliberately reads a sentence that is slightly different from the canonical text to emulate the mistakes of an L2 learner. A total of 4 words were deliberately pronounced wrongly. The result of the experiment can be seen below.

	Positive	Negative
Positive	28	0
Negative	0	4

Table 8: Confusion matrix for Google STT (Experiment 2)

	Positive	Negative
Positive	16	12
Negative	0	4

Table 9: Confusion matrix for Kaldi (Experiment 2)

As seen from Table 8 and 9, the transcription by Google STT API is 100% accurate, similar to the result of the first experiment. The Kaldi system is able to catch all the deliberately wrong pronounced words, but also has classified 12 more words that are supposed to be correct as wrong (false negatives).

It is noteworthy that among the false negatives, most of the characters that have been transcribed wrongly do share a similar pronunciation to the character in the canonical text, suggesting that the acoustic model is probably trained well. On the other hand, these similar-sounding words do not make a lot of sense in the context of the sentence being formed. This shows that while the performance of our system might not be as optimum, the accuracy can be improved by tuning the language model of the Kaldi-based system. Additionally, as mentioned earlier, shifting to a HMM-DNN system in the future instead of using the current HMM-GMM system might also be able to improve the accuracy of our Kaldi-based system.

Besides comparing the accuracy of both systems, it is also worth noting that the evaluation time of both systems also differs. While Google STT API is able to return an evaluation result within 5 seconds, our Kaldi-based system takes much longer, returning an evaluation result after a minute on average. While this means that there are rooms for optimisation in our Kaldi-based system, it should be noted that our Kaldi-based system is running on the school server with minimal resources and Google STT is running on commercial-level engines that have been optimised for fast querying. It may not be realistic to expect our Kaldi-based system to run at the same speed as a commercial platform.

5 LIMITATIONS

As we trained our models on a remote computing cluster with GPU support, the team did not have full control over hardware specifics and drivers. Development of the HMM-DNN model (using the nnet3 provided by Kaldi Zeroth-Korean recipe) was therefore arduous and difficult due to the specific version that Kaldi required and many issues on the computing cluster and machines where our models were trained on. We strongly advise potential researchers and developers of the system to find a reliable computing cluster/GPU machine, where they have control over such technicalities.

6 CONCLUSION AND FUTURE WORKS

This paper proposes a method for automatic pronunciation assessment of Korean spoken by L2 learners using an ASR-based system. We surveyed features that have been used for similar tasks and extracted fluency and segmental accuracy in our work. We trained two acoustic models using Kaldi, one based on HMM-GMM and another based on HMM-DNN. After training our acoustic models, we designed two workflows: one based on our trained HMM-GMM model and another based on Google STT API. Although we have successfully trained our HMM-DNN acoustic model, we could not deploy it on the web application prototype due to technical challenges at the server side. We also compared the results from both workflows and found that Google STT outperforms our Kaldi system in terms of both accuracy and evaluation time. However, developing our own Kaldi-based system is still crucial as forced alignment allows us to generate various fluency features, which cannot be obtained from commercial tools such as Google STT.

Future works could focus on implementing the HMM-DNN acoustic model on the web application interface for a better evaluation accuracy. Additionally, the extracted fluency and segmental accuracy features could be used to train a machine learning model via supervised learning to give learners a numerical overall proficiency score. This requires Korean L2 teachers to label each student's recording whose features have been extracted by our Kaldi-based system.

7 STATEMENT OF CONTRIBUTIONS

Overall, the team displayed good teamwork with weekly meetings to find out about each others' progress and to assign new tasks to perform via our team's internal issue tracker. Whenever needed, we assisted one another and provided realistic timelines to perform our tasks, which has resulted in a successful deployment of our system. The contributions of each team member are as follows:

- **Dolly Agarwal**: Training of models, researching suitable Korean speech evaluation measures, evaluation scripts, fixing technical issues.
- **Eugene Tan Yew Chin**: Training of models, evaluation scripts, fixing technical issues.
- **Ronald Santoso**: Development of web interface, integration of Kaldi and web application interface, fixing technical issues.

REFERENCES

[1] [n.d.]. Korean Wave. https://en.wikipedia.org/wiki/Korean_wave. Accessed: 2021-04-20.

[2] [n.d.]. Korean's popularity rises despite COVID. http://www.koreatimes.co.kr/www/culture/2021/02/703_302463.htmle. Accessed: 2021-04-20.

[3] Lei Chen, Klaus Zechner, Su-Youn Yoon, Keelan Evanini, Xinhao Wang, Anastassia Loukina, Jidong Tao, Lawrence Davis, Chong Min Lee, Min Ma, Robert Mundkowsky, Chi Lu, Chee Wee Leong, and Binod Gyawali. 2018. Automated Scoring of Nonnative Speech Using the SpeechRaterSM v. 5.0 Engine. *ETS Research Report Series* 2018, 1 (2018), 1–31. https://doi.org/10.1002/ets2.12198 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/ets2.12198

[4] Catia Cucchiarini, Helmer Strik, and Lou Boves. 1997. Automatic evaluation of Dutch pronunciation by using speech recognition technology. In *1997 IEEE workshop on automatic speech recognition and understanding proceedings*. IEEE, 622–629.

[5] Catia Cucchiarini, Helmer Strik, and Lou Boves. 2000. Different aspects of expert pronunciation quality ratings and their relation to scores produced by speech recognition algorithms. *Speech Communication* 30, 2-3 (2000), 109–119.

[6] Catia Cucchiarini, Helmer Strik, and Lou Boves. 2000. Quantitative assessment of second language learners' fluency by means of automatic speech recognition technology. *The Journal of the Acoustical Society of America* 107, 2 (2000), 989–999.

[7] Catia Cucchiarini, Helmer Strik, and Lou Boves. 2002. Quantitative assessment of second language learners' fluency: Comparisons between read and spontaneous speech. *the Journal of the Acoustical Society of America* 111, 6 (2002), 2862–2873.

[8] Maxine Eskenazi. 2009. An Overview of Spoken Language Technology for Education. *Speech Commun.* 51, 10 (Oct. 2009), 832–844. https://doi.org/10.1016/j.specom.2009.04.005

[9] Gertraud Havranek. 2002. When is corrective feedback most likely to succeed? *International Journal of Educational Research* 37, 3-4 (2002), 255–270.

[10] Hyejin Hong, Sunhee Kim, and Minhwa Chung. 2013. A corpus-based analysis of Korean segments produced by Japanese learners. In *Speech and Language Technology in Education*.

[11] Guimin Huang, Jing Ye, Yan Shen, and Ya Zhou. 2017. A evaluating model of english pronunciation for Chinese students. In *2017 IEEE 9th International Conference on Communication Software and Networks (ICCSN)*. IEEE, 1062–1065.

[12] Sandra Kanters, Catia Cucchiarini, and Helmer Strik. 2009. The goodness of pronunciation algorithm: a detailed performance study. (2009).

[13] Leonardo Neumeyer, Horacio Franco, Vassilios Digalakis, and Mitchel Weintraub. 2000. Automatic scoring of pronunciation quality. *Speech communication* 30, 2-3 (2000), 83–93.

[14] Yoo Rhee Oh, Kiyoung Park, Hyung-Bae Jeon, and Jeon Gue Park. 2020. Automatic proficiency assessment of Korean speech read aloud by non-natives using bidirectional LSTM-based speech recognition. *ETRI Journal* 42, 5 (2020), 761–772.

[15] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, et al. 2011. The Kaldi speech recognition toolkit. In *IEEE 2011 workshop on automatic speech recognition and understanding*. IEEE Signal Processing Society.

[16] Hyuksu Ryu, Hyejin Hong, Sunhee Kim, and Minhwa Chung. 2016. Automatic pronunciation assessment of Korean spoken by L2 learners using best feature set selection. In *2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*. IEEE, 1–6.

[17] Silke Maren Witt. 1999. Use of speech recognition in computer-assisted language learning. (1999).

[18] Silke M Witt and Steve J Young. 2000. Phone-level pronunciation scoring and assessment for interactive language learning. *Speech communication* 30, 2-3 (2000), 95–108.

[19] Seung-Hee Yang and Minhwa Chung. 2017. Linguistic Factors Affecting Evaluation of L2 Korean Speech Proficiency.. In *SLaTE*. 53–58.

[20] Seung-Hee Yang, Minsoo Na, and Minhwa Chung. 2015. Modeling pronunciation variations for non-native speech recognition of Korean produced by Chinese learners.. In *SLaTE*. 95–99.