Pronunciation Assessment and Feedback Tool

Jessica Kuleshov
UNI: jjk2235
Columbia University in the City of New York

***Abstract*—With the globalization of English comes a rise of many people learning English as a second language. It can be difficult and tedious to evaluate pronunciation by hand. This paper proposes a pronunciation assessment and feedback tool that could be used by learners of English to improve their mutual intelligibility. Using a pre-trained TED-LIUM model force-aligned on LibriSpeech data, the tool takes as input speech and transcripts from the Speech Accent Archive and, judged on phone-level confidence scores, outputs a color-coded version of the transcript indicating the level of pronunciation.**

***Keywords*—Speech Recognition, ASR, GOP, phone, pronunciation, pronunciation assessment, TED-LIUM, LibriSpeech, Speech Accent Archive, Kaldi, CALL, CAPT.**

I. Background

Pronunciation is a key part of computer-assisted language learning (CALL) and often is the most difficult to assess and quantify in a helpful manner. More specifically, the field of computer-assisted pronunciation teaching (CAPT) has developed a multitude of methods for quantifying the "goodness of pronunciation" of speakers, based on the number and severity of the errors present [1]. These errors can be broken up into two main categories of phonemic (often "more severe") errors dealing with sound changes, insertions, and deletions; and prosodic errors dealing often with pitch and intonation, shown in Figure 1.

Feature Category	Feature Name
Phonemic	Phone-level log-likelihood scores, GOP
	Vowel durations, duration trigrams
	Phoneme pair classifiers
	spectral features (formants)
	Articulatory-acoustic features
Prosodic (Intonation, Stress, fluency)	distances between stressed and unstressed syllables
	Mean, max, min power per word (energy)
	F0 contours (slope and maximum)
	rate of speech (words per second/minute)
	Trigram models to model phoneme duration in context
	Phonation/time ratio, mean phoneme duration
	Articulation Rate (phonemes/sec)
	Mean and standard deviation of long silence duration
	Silences per second
	Frequency of disfluencies (pauses, fillers etc)
	Total and mean pause time (i.e. duration of interword pauses)

Fig. 1. List of features commonly used in pronunciation assessment [1].

Mastering these features means that you have mastered the "native" accent of a language; however, it is very difficult to assess these metrics by hand [2], and so the help of technology is greatly needed in this department. Pronunciation is a key part of language understanding and should not be overlooked in language learning, and providing a proper CALL tool for language learners to improve their pronunciation and intelligibility in their target language, without the intervention of hand-analyzed scoring is crucial to their success in becoming more fluent in that target language.

II. Previous Work

There have been several projects done working on pronunciation assessment by various parties. One of the earlier models is PLASER (Pronunciation Learning via Automatic Speech Recognition) [3]. This project was a pronunciation feedback tool that employed hidden Markov Models to represent position-dependent English phonemes and used the American English TIMIT corpus. Using elicitation, the model established a ground-truth transcription that it compared the user's speech to and then evaluated based on phone confidence. This model was the first of its kind to allow for non-native accents and was tested by 900 students in Hong Kong, who recommended it be used more often at the time. Another similar tool was Japañol, made for native Japanese speakers learning Spanish, where they rated pronunciation on a 1-10 scale based off of correct utterances out of total utterances [4].

More recent studies having been using DNNs for acoustic modeling, applying only acoustic features and not requiring native acoustic models. On larger data sets, DNN-HMM models report a 20% relative decrease in WER [5]. One study focused on L2 speakers of Japanese that also explored a DNN-based method for pronunciation assessment [6]. However, despite promising results, these results were not consistent across all sizes of data set and makeup of data, so this project will stay with traditional GMM-HMM models.

Phone-level confidence scores are a common metric for pronunciation assessment. In a paper from Qin et al. (2019), they used phone-level confidence scores and posteriorgrams from two sources to detect speech from people with aphasia. In that experiment, they used a CNN-based classifier in order to determine from a strong and weak recognizer the severity of aphasia in patients. [7] Other methods have used the GOP scoring method directly, but since it is considered by some to be fairly outdated, there have been other methods proposed. [9] One interesting method that has been proposed but was not

Data Set	Usage	Contents for this experiment	Source	Sampling Rate
TED-LIUM Release 3	Training set	TED Talks by English speakers, native and nonnative, 118 hours. Transcripts directly match recordings	openslr.org/51	16kHz
LibriSpeech	Forced alignment	Audiobook corpus of native English speech, 100 hours. Transcripts directly match recordings	openslr.org/12	16kHz
Speech Accent Archive	Test set	2137 recordings of nonnative English speakers reading the same passage. Transcripts do not fully align with spoken text	kaggle.com/rtatman/speechaccentarchive	16kHz (downsampled from various sampling rates)

Fig. 2. The data sets used in this experiment, with some background.

implemented in this paper is one form Sudhakara et al. (2019) discussing an improved GOP method that used a DNN-HMM system that incorporated HMM transition probabilities into the final scoring. [8]

There are also multiple commercial services that focus on pronunciation assessment, such as the SpeechRater tool used by TOEFL and AMEnglish.com, Versant from Pearson, EyeSpeak, and CarnegieSpeechAssessment. Each of these tools have different focuses - for example, the CSA tool provides feedback at the phone and sentence level, whereas SpeechRater provides feedback on stress, intonation, pause length, and other features. In general, the models that these websites use are not advertised and they are difficult to access without payment and as such it is difficult to compare across models.

III. Approach

The prevalence of assessment tools that are difficult to access and use, as well as various research showing the effectiveness and usefulness of pronunciation assessment as a standalone process in the improvement of people's language learning, are an onus for creating a more user-friendly and accurate pronunciation assessment tool that allows the pressure to be taken off of hand-scorers and analyzers. Due to its continued success and relatively simple implementation, a GMM-HMM model is used for training a model that can serve as the back end to scoring pronunciation. Using a forced alignment of a pre-trained model on a native speaker data set and comparing the differences in phones and phone confidence values, this results in the correctness of the speaker's pronunciation.

The text will first be converted into possible phonetic transcriptions and then phone boundaries will be determined in the recording based off of these transcriptions and the provided pre-trained HMM from semi-native speech [10]. Once the model is tested against the Speech Accent Archive data set, the model should be worked into a tool for L2 speakers of English to use to assess and improve their own pronunciation. The main feature of the tool will be asking the user to repeat a passage, either a word or a sentence or two. Then, depending on how "correct" the pronunciation ended up being, the word or phone will be given a color-associated "correctness" level (green for correct pronunciation, yellow for fair pronunciation, red for incorrect pronunciation).

IV. Experiment

A. Data Sets

The data for the training set was the TED-LIUM data set, as found on openslr.org [11]. This data set contains native and slightly-accented English speakers from various TED talks, some of which have accented speech. It contains about 100 hours of speech. The presence of slightly accented speech is very good as well because with this initial training data there are still some phones left over post-forced-alignment that, though they may not appear in the native English new alignment, still would be present in the lexicon and may more accurately reflect what the speakers are saying.

For the forced alignment, LibriSpeech was used, also found on openslr.org [12]. This dataset consisted of 100 hours of clean recordings and transcripts of native English speech reading audio books. This was selected because it would establish an excellent baseline for native English speech to compare to as the recordings are intended for a wide range of

English audiences and are typically done by large companies that focus on quality and intelligibility.

The Speech Accent Archive was used for the test set, and can be found on Kaggle.com [13]. This data set includes 2140 speech samples of speakers from 177 different countries who natively speak 240 languages. All of the speakers are reading the same set of sentences as well, which allows for ease of computation with the usage of the same basic text. This wide range of accents and demographics is an ultimate test for a speech recognition model, as the average age of onset is 8.8 with variance 71.4 and the average age of speaker is 33.1 with variance 208.9. The text being read for this assignment was as follows:

"Please call Stella. Ask her to bring these things with her from the store: Six spoons of fresh snow peas, five thick slabs of blue cheese, and maybe a snack for her brother Bob. We also need a small plastic snake and a big toy frog for the kids. She can scoop these things into three red bags, and we will go meet her Wednesday at the train station."

This set of sentences was constructed in order to account for all common sound combinations that English may have. Due to the various demographics and first languages of the speakers in the Speech Accent Archive, they typically have difficulty with different sections of the text, allowing for fairly varied data.

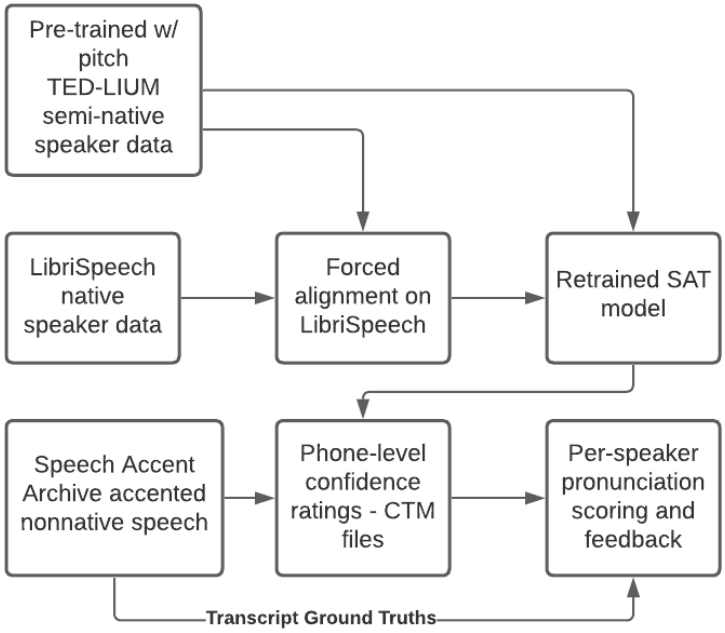

Fig. 3. General procedure for training, decoding, and scoring in this experiment.

B. Method

The procedure for the main script of this was based off of the TED-LIUM Release 3 example found in the Kaldi egs folder, but has been heavily modified in order to fit the needs of this project. The first step was to format the data properly. In order to get the LibriSpeech and Speech Accent Archive data sets into a format the is workable with Kaldi, the first step was to run a data preparation step to create the files that Kaldi needs, allowing for the wav.scp file to also convert the FLAC files to WAV. However, Speech Accent Archive required a bit more formatting, as it does not have its own data preparation script, so a Python script was written to reformat it, changing the FLAC files to WAV and downsampling to 16kHZ to match the previous two sets. Unlike the LibriSpeech dataset, because the utterances in the Speech Accent Archive were fairly short and each speaker only had one recording, utterances were not split into any portions and the results of each speaker totaled their one utterance. After that, make_mfcc_pitch.sh was run on both the LibriSpeech and Speech Accent Archive data in order to extract MFCC pitch features, and Cepstral Mean and Variance (CMVN) statistics were computed for each speaker in the data set.

After this initial data cleaning, the next step was to get the Speaker-Adapted Training (SAT) TED-LIUM model adjusted for the baseline of native English (a SAT model was employed, as it was hoped that this would normalize slightly better across features and be more accurate in its representation of phones.) The forced alignment was performed, reassessing the phone and pitch alignments in the pre-trained TED-LIUM model with the LibriSpeech data set. With a forced alignment, all phone-level confidence scores are set by default to 1, as the alignment is done using the ground truth, so this creates something substantial to which to compare later confidence levels. Using this new alignment, the TED-LIUM model was retrained in order to yield another SAT model with the established baseline. With the new TED-LIUM model, the Speech Accent Archive data was decoded using fMLLR characteristics that could be compared with the previously-trained Speaker-Adapted Model and the outputs of the word-level lattices were converted into CTM-format files with per-phone confidence levels.

A script was also created in order to score these confidence scores, which included aligning the phones with the lexicon-version transcripts and then comparing the most likely phones and their respective confidence levels. The four confidence levels used were Poor, OK, Good, and Excellent, color-coded Red, Orange, Yellow, and Green respectively. Certain statistics were also added in in order to give the user more information about their sentence.

V. Results and Discussion

The output of the model is shown in Figure 4. It can be very difficult to evaluate output from a pronunciation scoring metric, but generally the results look promising. The first speaker used was a native Turkmen speaker, whose recording sounded mostly mutually intelligible but the speaker definitely had a couple of difficulties. In the end, generally yellow or green pronunciations were scored, so this indicates an alignment with intuition. The words to work on are the group of OK and Poor ratings. In a hypothetical tool that returns this

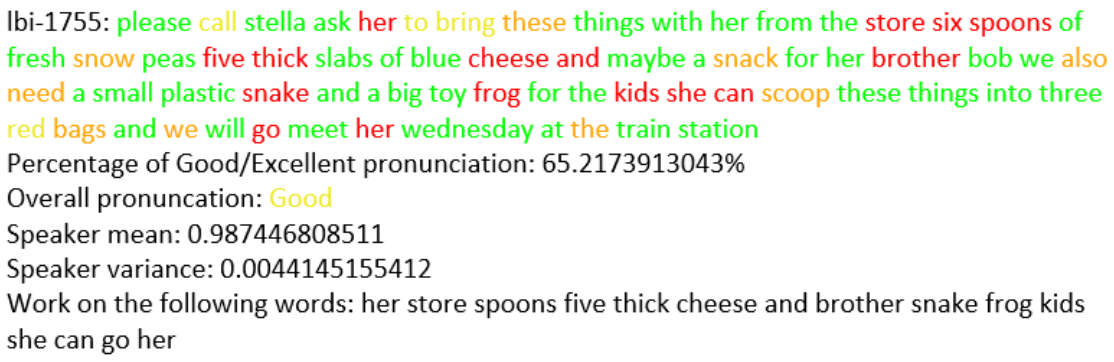

Fig. 4. Output of one speaker's assessed pronunciation. The colors/level correspondence is Poor-Red, OK-Orange, Good-Yellow, and Excellent-Green.

kind of input, it would be useful to save such lists for each user, such that they receive some positive feedback when the word comes up in practice again.

The statistics that were able to be gained from this, however, left much to be desired. The desire to quantify the pronunciation by phone-level confidences alone is great, but there is a much larger underlying issue - these phone confidence levels are only perfectly comparable when the phone predicted for the speaker and from the lexicon are the same. This results in much more difficulty assessing what the actual phone is, because from the output of native Kaldi phone-level confidence algorithms only show the best pronunciation, not the pronunciation for the expected phone. Often times, this phone confidence level ended up being 1, so when mean and variance calculations are shown in the above output, the mean is often very high, with a very low variance.

In order to account for this, the phones themselves were compared as best as possible, using the phone-level confidences solely as a supplement when the phones matched or did not match. When the phones matched exactly, the pronunciation would be ranked as Excellent if the phone confidence level was greater than a threshold (close to 1) and Good if lower. If the phone was incorrect and happened to align, then the lower confidence score in that phone was interpreted as better, as it was taken to mean that though this sound was the most confident,

This modification in evaluation resulted in more difficulties. One of the main issues with the Speech Accent Archive's format - that being, providing only the transcript that all of the speakers read - is that this format of transcript does not account for all of the mistakes in words themselves that the speakers tended to make. This would include pauses, going back on their words, saying entirely incorrect words, and other issues. This made the word-level alignments fairly difficult, and so a mildly manual approach was employed to check surrounding words' beginning and ending phones to see, when the words initially do not seem to line up, whether the two sets - of the speaker's phones and of the target phones - are missing a word somewhere.

Another problem that may have arose from this assessment which it could not evaluate for is the inevitable fact that some speakers enunciate words very clearly whereas others rush through speech. Typically very slow, disjointed enunciation is a sign that a person does not understand a language (as many native American English speakers replace vowels with schwa's when speaking quickly in typical conversational English); however, there are definitely speakers who speak quite slowly and pronounce every word as if it were being used individually, yet learned the language natively. As such, this system cannot discriminate between a speaker who does no understand pronunciation changes within a sentence versus a native speaker with slow and deliberate pronunciation.

VI. Plans for Evaluation

Unlike a typical machine-learning-based experiment or other speech recognition-based assignment, it is relatively difficult to assess the quality of pronunciation assessment without measuring its effectiveness when used with nonnative speakers over a duration of time. Effectiveness of pronunciation assessment can also be well-represented by precision and recall metrics looking at the effectiveness of classifying the mispronunciations. [14] Unfortunately, for the case of evaluating longer-term effectiveness, there was not enough time to complete this before the needed conclusion of the project, and for the case of assessing precision and recall, there was no ground truth in the form of per-speaker phonetic transcriptions available to base those assessments off of (in terms of "actual" mispronunciations.)

Instead, a proposed plan is outlined here for what would be required in order to properly evaluate the effectiveness of the model in the future. A proper assessment would require some amount of given phonetic transcriptions (done preferably by hand) for the given data set, as mentioned previously. This would give the proper baseline necessary for judging the model and make sure that whatever values are being outputted for word-level pronunciations here is more accurate, past the faults that the Kaldi lattices have with phone-level confidence scoring to begin with. For proper pronunciation assessment effectiveness as a learning tool (separate from the problem of phonetic transcriptions, as that is a model-quality rating issue), at least two groups would have to be involved in such a project: One group using a standard tool for pronunciation assessment

such as Duolingo or SpeechRater (TOEFL) that rate word-level pronunciations in a similar manner as described here and one group using a tool created based off of the method in this paper. Typically, pronunciation improvements can be noticeably observed within 3-4 weeks from the beginning of conscious attempts for pronunciation improvement or accent reduction. [15] With this experiment, it would benefit from having continuous evaluation for at least 4 weeks or longer, starting with an initial placement assessment for both groups and then tracking the rate of words/phone patterns to work on being added and removed from the user's information. An end assessment would then be completed, and with that amount of information, the average progress of a speaker can be gleaned, average effectiveness, and areas where the assessment and feedback tool were most useful in improving the speaker's speech. Though it could not be completed currently, this is an excellent place to start work in the future when creating a more developed and usable tool that can be an invaluable resource to second language learners of English.

VII. Conclusion and Further Work

Pronunciation evaluation is not as exact of a science as would be desired. Pronunciation scoring, especially with an inaccurate transcript on the per-speaker level, can be difficult to accomplish and requires more careful scoring. The reality of natural language is that even if attempting to elicit a certain sentence, the program is not guaranteed to get an audio that has that exact sentence as an input, causing issues with alignment and the trustworthiness of the output. With that in mind, however, using Kaldi toolkit and the method described in the paper, given enough time and tuning, shows promise in being useful in speaker evaluation. Having cleaned data and a better alignment process pre-scoring would be especially helpful in avoiding some more manual calculation required in order to get phones to align and be properly compared.

There are plenty of directions where this work may go. In addition to accounting for the changes suggested above, there are several possibilities as to how this could be implemented as a helpful pronunciation assessment tool past the proof-of-concept shown in this paper. One possibility would be to implement this in an applet, where users can log in, read from several pre-set transcripts testing common English constructions (such as the one provided in the Speech Accent Archive) and then give words and constructions to work on, almost as a Duolingo-like approach to language-learning but solely for pronunciation purposes. A beneficial addition to this would be to have the pronunciation levels over time become relative to the user's individual baseline, allowing them initially to set a baseline for their speech and then set the levels of pronunciation relative to their typical average confidence level. That way, it would encourage users to increase their average confidence over time and push subconsciously for better pronunciation.

VIII. Acknowledgements

The author would like to thank Professor Beigi for his invaluable assistance in making this project a reality, and Riddhima Narravula for providing the class with a pre-trained with pitch TED-LIUM model that was used in this experiment.

References

[1] S. Witt, "Automatic Error Detection in Pronunciation Training: Where we are and where we need to go," 2012.
[2] PM. Rogerson-Revell, "Computer-Assisted Pronunciation Training (CAPT): Current Issues and Future Directions," RELC Journal, 2021. pp.189–205.
[3] B. Mak, M. Siu, M. Ng, Y.C. Tam, Y.C. Chan, K.W. Chan, K.Y. Leung, S. Ho, F.H. Chong, J. Wong, and J. Lo, "PLASER: Pronunciation Learning via Automatic Speech Recognition." 2004.
[4] C. Tejedor-García, V. Cardeñoso-Payo, en D. Escudero-Mancebo, "Automatic Speech Recognition (ASR) Systems Applied to Pronunciation Assessment of L2 Spanish for Japanese Speakers", Applied Sciences, vol 11, no 15, 2021.
[5] A. Becerra, J. I. de la Rosa and E. González, "A case study of speech recognition in Spanish: From conventional to deep approach," 2016 IEEE ANDESCON, 2016. pp.1-4.
[6] K. Takai , P. Heracleous, K. Yasuda , and A. Yoneyama, "Deep Learning-Based Automatic Pronunciation Assessment for Second Language Learners," [Stephanidis C., Antona M. (eds) HCI International 2020 - Posters. HCII 2020. Communications in Computer and Information Science, vol 1225. Springer, Cham. 2020.
[7] Y. Qin, T. Lee and A. P. Hin Kong, "Combining Phone Posteriorgrams from Strong and Weak Recognizers for Automatic Speech Assessment of People with Aphasia," ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2019, pp. 6420-6424, doi: 10.1109/ICASSP.2019.8683835.
[8] S. Sudhakara, M. K. Ramanathi, C. Yarra, en P. K. Ghosh, "An Improved Goodness of Pronunciation (GoP) Measure for Pronunciation Evaluation with DNN-HMM System Considering HMM Transition Probabilities", in INTERSPEECH, 2019.
[9] S. Kanters, C. Cucchiarini, and H. Strik, "The Goodness of Pronunciation algorithm: a detailed performance study," Speech Communication, 2009.
[10] J. Yuan, W. Lai, C. Cieri, and M.Y. Liberman, "Using Forced Alignment for Phonetics Research," 2018.
[11] A. Rousseau, P. Deléglise, and Y. Estève, "TED-LIUM: an automatic speech recognition dedicated corpus," in Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12), May 2012.
[12] V. Panayotov, G. Chen, D. Povey and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2015, pp. 5206-5210, doi: 10.1109/ICASSP.2015.7178964.
[13] S. Weinberger, "Speech accent archive," George Mason University, 2013.
[14] M. Peabody, "Methods for Pronunciation Assessment in Computer Aided Language Learning", 01 2012.
[15] T. Raja. "How long does it take to reduce your accent?", 2021.