To see what is heard: A framework for adapting Speech-To-Text for Pronunciation Monitoring

Abhishek Purushothama
University of Colorado Boulder
abhishek.purushothama@colorado.edu

Abstract

People may speak the same language differently, due to various historic, social and personal reasons. When learning to *speak* a second language, learners may adapt parts of accents of their peers and teachers. Accents can also become obstacles in verbal communication including teaching. There are many established pedagogical methods for helping understand their speech, pronunciation and if desired adapt accents. In this paper we propose the task of pronunciation monitoring, a technical framework for building a pronunciation monitoring system. We additionally demonstrate how to build such a system using state-of-art of Automatic Speech Recognition system.

1 Introduction

Pronunciation is foundational to speech in spoken language. *Accents* or Dialects of languages are characterized by the speech patterns, of which variation in pronunciation is a major component (Wolfram and Schilling, 2015).

When learning to speak in a second language, based on various social, geographical and political reason. The speakers may also acquire an accent in the language, that may differ from other first language and second language speaker accents(Yanilis Romero, 2017).

The differences in accents may become an obstacle even if the speakers speak the *same language* despite having the same vocabulary and diction. This can be be true in pedagogical aspects in many specialized domains(Wolfram and Schilling, 2015). The characterization of accents may vary due to various factors. Accents are mainly variations in speech pattern, pronunciation being a significant part of it.

Active observation of the speakers pronunciation by themselves, peers or teachers in order to gain awareness of ones speech patterns and accents is a popular method for learning to pronounce with an accent. This can additionally help learners adapt, acquire and understand ability to speak and comprehend various accents of the same language. We refer this method as "Pronunciation Monitoring".

Pronunciation monitoring is generally done in a controlled environment such as classrooms settings, were a *monitor* is actively listening to the *speaker* with special attention to the pronunciation of the speaker. This will always involve the *active* involvement of the *monitor*, may additionally involve meta-monitoring by the *speaker* themselves (Derwing and Munro, 2005).

This requirement actively limits the amount of time, money and effort that can be put into pronunciation monitoring for learners. Providing *monitors* and *speakers* with tools to ease this process can allow speakers to accelerate the process of learning. Add citations

Automatic Speech Recognition (ASR) refers to the processes and techniques by which *speech* (in any language) can be *recognized automatically* by machines. The recognition can come in the form of transcription (Speeech-to-text) or even in the form of understanding(intent or emotion)(Jurafsky and Martin). Both forms have been available with varied success over the years. The latter has pervaded daily life in the form of voice assistants on our smartphones or even dedicated devices.

Neural methods for ASR have shown tremendous success and can be considered an important reason for the availability of ASR in many languages and accents (Kumar et al., 2018). Transformer architectures, in combination with pre-training with self-supervision learning objectives have provided avenues for easy adaptation of ASR for specific purposes. Although self-supervision training objectives allow for training with unlabelled corpus, adaption (fine-tuning) require datasets labelled for the purpose(Baevski et al., 2020).

In this paper we propose the task of pronunciation monitoring, a technical framework for building a pronunciation monitoring system. We additionally demonstrate how to build such a system using state-of-art of Automatic Speech Recognition system.

2 Pronunciation Monitoring

Most scripts do not explicitly encode pronunciation into the text. This means that script of the language is not sufficient to represent elements of pronunciation. International Phonetic Alphabet (IPA) was specifically crafted for that purpose. A Singular English text based on accent can be pronounced differently and hence would have differential IPA representation.

The task of pronunciation monitoring can be divided into two parts. The first part is the identification each pronunciation element and the second would be to analyse it for errors.

IPA acts as the best choice for representing pronunciation visually. So, effectively the sub-tasks can be considered as Speech-to-IPA and IPA Analysis.

3 Speech to IPA

Speech to Text(STT) is the specific subdomain of ASR that deals with converting Speech into Text, predominantly to a single script. STT hence specifically deals with converting *sounds* of speech into elements of a textual script or *transcription*.

Since the first sub-task of Pronunciation monitoring is IPA Transcription, we can formulate it as a conditional variant of Speech to Text, where the target script is always IPA.

Given this formulation of Speech to Text, the adaptation of STT methods for STI is straightforward. A system for STI would be an adaption where the target transcript is in IPA alphabet/script.

4 Speech to IPA dataset

In order to use STT methodologies to build STI, we have similar data requirements. For neural STT systems, the basic requirement would be parallel speech-text data. In our case, since the target script is IPA we would require speech-IPA data.

There are no explicit speech-IPA datasets available that can be used for the purpose.

Since IPA is a pronunciation representation, IPA text can be produced from standard text, given a specific rules of pronunciation. An accent such as American English has one such set of rules.

Split	Hours of Speech Data
Train	10
Validation	1
Test	4

Table 1: Librispeech Data Selection

CMU Pronunciation Dictionary is a popular resource for getting American English IPA representation of english words. It is also computationally oriented, meant for usage in language processing systems.

For a given accent, with a tool such as a pronunciation dictionary available, we can attempt to convert the text to IPA.

For our prototype with American English we select 15 hours of parallel data from the Librispeech dataset(Panayotov et al., 2015) with the splits shown in Table 1. We do not perform explicit verification of the speech being in American English Accent since we use the *train.100* split from the original dataset, which the authors had separated as American English using unsupervised methods.

We use huggingface *datasets* library (Lhoest et al., 2021) as our dataset manager and use *eng-to-ipa* as our tool for conversion of English text to IPA. The package provides a simple python interface for english to IPA utilizing the CMU Pronunciation dictionary.

We refer to this transformed dataset as *the dataset* in the rest of the paper.

A sample entry from the train split is shown in **??**

5 Wav2Vec2 for STI

Wav2Vec 2.0 (Wav2Vec2 from now on) in addition to being a framework is also a set of pre-trained models for speech representation trained with self-supervised learning objective. It is based on the Transformer architecture and provides ease for adaptation with fine-tuning similar pre-trained models in the language representation. It has also shown great performance on the TIMIT dataset meant for phoneme recognition, making it an excellent choice for adaptation to STI task, since it involves being able to distinguish each pronunciation character.

Wav2Vec has shown good performance for En-

Duration	15.625 seconds
Sampling Rate	16KHz
Number of Values	250000
English Text	’FOR HER BY HIS PROMPT RECOGNITION OF HER RARITY BY PRECEDING HER IN A FRIENDLY SPIRIT AS HE HAD THE EAR OF SOCIETY WITH A SHARP FLASHLIGHT OR TWO HE MET POOR DENSHER THESE ENQUIRIES AS HE COULD’
Converted IPA Text	’fr hr ba hz prmpt rkgnn v hr rrti ba prisid■ hr n frndli sprt z hi hæd ð r v sosati w rp flælat r tu hi mt pur densher* ðiz enquiries* z hi kd ’

Table 2: Caption

glish ASR with less than 15 hours of parallel speech-text data.

We design a prototype model for STI for American Spoken English (ASE) using Wav2Vec2. We fine-tune the pretrained model using the huggingface transformers library with the standard process of using a final linear layer with the output vocabulary of 47 IPA characters(Appendix 1).

6 Adaptation

6.1 Training

We trained(fine-tuned) our prototype on the training split of the dataset using the constructs of the *transformers*(Wolf et al., 2020) library.

We replicated the huggingface guide for using Wav2Vec2 for ASR.[1]. There are some differences and are discussed below.

6.1.1 Target Vocabulary

When fine-tuning for ASR in English, the target output is limited to the 26 english alphabetical characters, it may include punctuation and numerical characters for a more robust or original system.

Our target vocabulary is IPA and hence includes a slightly larger set of characters hence the vocabulary needs to be prepared the same way.

The hyper parameters of the training are listed in the table3.

Our efforts to fine-tune Wav2Vec2 for our training data failed with various changes sets of hyperparameters. The loss stagnates in less than two epochs, with the model only predicting the ’[PAD]’ token. This is very a likely an error in the implementation of the code. In training neural models, the models explicit care needs to taken to neglect the ’[PAD]’ tokens when calculating loss so that the training objective doesn’t get trained optimize to just predict ’[PAD]’.

Hyperparameter	Value
Number of Epochs	10
Learning Rate	1E-4
Loss Function	CTCLoss

Table 3: Hyperparameters for the training

When this is not done correctly, the model starts to predict ’[PAD]’ since it is the optimal output given the context.

6.2 Results

Since the the Wav2Vec2 training has not succeeded, our results for the prototype are unavailable.

7 IPA Analysis

We had divided the task of Pronunciation monitoring into Speech-To-IPA and IPA Analysis. The first part as shown, can be easily achieved to significant success by adapting ASR neural systems. The second part is the functional element of the monitoring.

IPA Analysis in current pedagogical environment is done by the meta-knowledge of the learners peers and teachers. This would be harder element to automate since we are trying to replicate skilled actions of humans.

Below we propose various ways to perform IPA Analysis for pronunciation monitoring, mainly inspired by English as Second language pedagogical techniques.

[1] https://huggingface.co/blog/fine-tune-wav2vec2-english

7.0.1 Text Transcript Comparison

In many cases, the monitoring of pronunciation is done in comparison to prepared content. The textual content can be converted to IPA and compared to the *transcripted IPA* and the differences highlighted.

7.1 Reverse Translation

In a automated manner, we can attempt to *transcript IPA* into text reversing the Text-To-IPA process. Failures to transcript snippets would be highlighted to the learner who can actively attempt corrections for the sections and improve learning.

7.2 Color Vowels

Color Vowel Chart [2] is a pedagogical pronunciation tool that inspired authors’ attempt into pronunciation monitoring. Fifteen American English vowel sounds are associated with colors and things, e.g. ’i:’ sound is associated with color Green which and item Tea both which contain the specific sound. This provides the learners with a vowel sound - color - item relation to help remember and practice. Learners can practice to color all the vowel sounds they need to use and attempt to hit the sounds. Since each vowel sound is associated with specific IPA characters, the *transcripted IPA* can be *colored* for easy reference of users.

8 Bias and Fairness

Technologies can have strong impact on society and neural methods have been shown to replicate and amplify biases and lead to unfair implementations in social, civil and political context. With this in mind we analyse (without experiments) bias and fairness that could be embedded in our prototypes and plans to mitigate the same.

Our framework is meant to build system for facilitating learning of spoken languages. It is meant for individual usage, but can be deployed and used in physical and virtual classroom settings. This necessitates that *fairness* would need to be analyzed with more emphasis than bias.

The framework does not include bias mitigation and fairness analysis. Neural methods used can especially inherit biases from data and data selection.

8.1 Bias

The framework does not require any specific method or data to be used and hence the biases would be specific to the dataset used and data selection.

Our prototype uses two main elements that should be analysed for biases.

8.1.1 Librispeech and Selection

The Librispeech dataset is a dataset created from audiobooks and their textual parallels. Audiobooks are generally *read* by trained professionals and would encode social, political, geographical and demographical biases of the speakers. Additionally, the model is trained selected American English Speech, since the goal is for pronunciation monitoring of ASE. But this may also mean that it encodes pronunciations of the selected individuals, who are professionally trained speakers of ASE and would not be expected to perform equally well on non-represented speech.

8.1.2 Wav2Vec2

The Wav2Vec2 model itself is trained with Librispeech dataset and has the bias problems discussed in the previous sections apply, at scale.

8.2 Fairness

8.2.1 Is IPA Fair?

There are many ways to adjudicate and quantify fairness, but we won’t be attempting to do the same of IPA. Rather we will discuss some of the salient and known critiques of IPA and how that affects the Fairness of our system.

Although IPA is maintained and published by a private organization,it is based on open discussion and consensus among experts in various languages. IPA does try and represent a variety of sounds. It contains characters to represent clicks such as [|, !, ’f, II] which were used by phoneticians working on Nguni and Khoisan languages. (Ladefoged, 1990). There is no objective evaluation of IPA’s fairness. Since IPA is the central substrate, IPA’s fairness gets inherited by our framework.

8.2.2 Robustness of the Framework

The framework lacks elements of robustness, it considers presence of parallel speech and text(IPA) sufficient to build a system for pronunciation monitoring. It fails to account for problems in the data/dataset itself across gender, age, race and other boundaries. Pronunciation sounds even in ASE is

[2] https://americanenglish.state.gov/resources/color-vowel-chart

likely differ across these boundaries, in human perceptible and imperciptible ways. Neural methods have been known to be fickle and pattern abusing, and are likely to be fail significant fairness tests. Librispeech dataset which is used in our prototype has been shown (Liu et al., 2021) to have performance differences across groups of the attributes.

The framework should account for robustness and fairness, and we discuss further on some of the options in Drawbacks section.

9 Drawbacks

There a few drawbacks to the current framework that the authors have identified and discuss options to mitigate the same.

9.1 Robustness

The framework does not enforce conditions on the *requirements* the dataset has to satisfy for it to be usable for Pronunciation monitoring. For the system to be robust across varied voice patterns, speech rates the initial dataset would need to satisfy variety and consistency constraints. Although there has been work around fairness of the data, the progress is insufficient to guarantee robustness by itself. The framework would need to account for methodologies that can be applied to bring in robustness. The framework would need to include a pre-processing step which provides guidelines to encourage if not enforce robustness.

10 Conclusions

In this paper, we proposed a framework for adapting ASR for pronunciation monitoring. We intuitively demonstrate how pronunciation monitoring can be broken into the ASR task of Speech-to-IPA and the pedagogical task of IPA Analysis. We additionally analyse some of the weaknesses of the framework with respect to Fairness and Robustness. We also provide design for a prototype for Pronunciation Monitoring of American Spoken English.

11 Future Work

We have discussed some of the drawbacks of the framework so we put forth the following avenues for improvement.

11.1 Robustness and Fairness

The framework should be expanded to bring elements to encourage if not enforce fairness and robustness in the system being built.

11.2 Applicability Analysis

The applicability of framework should be verified with multiple diverse languages, accents. This requires search, identification and experimentation with various datasets.

11.3 Formalization of IPA Analysis

The sub-task of IPA Analysis needs to be formalized and comparable techniques and their applicability needs to be verified with the help of pedagogical experts.

12 Resources

1. CMU Pronunciation Dictionary link
2. eng-to-ipa package link
3. Colab Notebook for Fine-tuning Wav2Vec2 with librispeech-IPA dataset link.
4. Librispeech Dataset link
5. Drive Folder Scripts for the project utilizing *eng-to-ipa* and *datasets* to prepare Custom IPA Dataset

   link
6. An unrefined WandB Report with WER and Losses across some of the latest monitored attempts at Wav2Vec2 link

Acknowledgements

We would like to thank Matthew Morley, who is an Instructor in the International English Center at CU Boulder, for his insights into pronunciation, and American Spoken Englush. We would like to acknowledge and express gratitude to Rafael Frongillo and Enrique Nueve for their support, feedback and guidance throughout the CSCI 5802 course at CU Boulder.

This work utilized the Summit supercomputer, which is supported by the National Science Foundation (awards ACI-1532235 and ACI-1532236), the University of Colorado Boulder, and Colorado State University. The Summit supercomputer is a joint effort of the University of Colorado Boulder and Colorado State University.

This work utilized the Blanca condo computing resource at the University of Colorado Boulder. Blanca is jointly funded by computing users and the University of Colorado Boulder.

References

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems*, volume 33, pages 12449–12460. Curran Associates, Inc.

Tracey M. Derwing and Murray J. Munro. 2005. Second language accent and pronunciation teaching: A research-based approach. *TESOL Quarterly*, 39(3):379–397.

Daniel Jurafsky and James H. Martin. https://web.stanford.edu/ jurafsky/slp3/26.pdf.

Akshi Kumar, Sukriti Verma, and Himanshu Mangla. 2018. A survey of deep learning techniques in speech recognition. *2018 International Conference on Advances in Computing, Communication Control and Networking (ICACCCN)*, pages 179–185.

Peter Ladefoged. 1990. Some reflections on the ipa. *Journal of Phonetics*, 18(3):335–346. Phonetic Representation.

Quentin Lhoest, Albert Villanova del Moral, Yacine Jernite, Abhishek Thakur, Patrick von Platen, Suraj Patil, Julien Chaumond, Mariama Drame, Julien Plu, Lewis Tunstall, Joe Davison, Mario Šaško, Gunjan Chhablani, Bhavitvya Malik, Simon Brandeis, Teven Le Scao, Victor Sanh, Canwen Xu, Nicolas Patry, Angelina McMillan-Major, Philipp Schmid, Sylvain Gugger, Clément Delangue, Théo Matussière, Lysandre Debut, Stas Bekman, Pierric Cistac, Thibault Goehringer, Victor Mustar, François Lagunas, Alexander Rush, and Thomas Wolf. 2021. Datasets: A community library for natural language processing. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 175–184, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Chunxi Liu, Michael Picheny, Leda Sarı, Pooja Chitkara, Alex Xiao, Xiaohui Zhang, Mark Chou, Andres Alvarado, Caner Hazirbas, and Yatharth Saraf. 2021. Towards measuring fairness in speech recognition: Casual conversations dataset transcriptions.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: An asr corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Walt Wolfram and Natalie Schilling. 2015. *American English : Dialects and Variation*. John Wiley & Sons, Incorporated, Hoboken, UNITED STATES.

Milton Pájaro Manjarres Yanilis Romero. 2017. *How Does the First Language Have an Influence on Language Learning? A Case Study in an English ESL Classroom*, volume 10 of *English Second Language*, chapter 10. Canadian Center of Science and Education.

A ASE IPA Vocabulary for Wav2Vec2

The following is the vocabulary used to fine-tune Wav2Vec2 for STI for ASE.

```
"": 0,
"": 1,
"p": 2,
"f": 3,
"t": 4,
"": 5,
"n": 6,
"*": 7,
"l": 8,
"æ": 9,
"r": 10,
"u": 11,
"a": 12,
"": 13,
"w": 14,
"": 15,
"b": 16,
"e": 17,
"g": 18,
"ð": 19,
"i": 20,
"": 21,
"v": 22,
"": 23,
"h": 24,
"m": 25,
"′": 26,
"■": 27, (velar nasal)
"c": 28,
"z": 29,
"x": 30,
"o": 31,
"": 32,
"k": 33,
"": 34,
"s": 35,
    "q": 36,
    "y": 37,
    "j": 38,
    "": 39,
    "": 40,
    "": 41,
    "": 42,
    "d": 43,
    "|": 44,
    "[UNK]": 45,
    "[PAD]": 46
```