Literature Review

Mario Peng Lee
XCS224U - Natural Language Processing, Fall 2022

General problem

Translating natural language into phonological representations such as those in the International Phonetic Alphabet (IPA) might sound like a trivial task yet it is significantly understudied. The main reason might be due to the fact that this task is not necessary to obtain acoustic interpretation technology. Text-to-speech (TTS) models have been around for many years and more recently multilingual TTS and voice cloning models all joined the phonology research landscape, none of them requiring the encoding acoustic features into IPA. Although overlooked, this translation task would greatly contribute to the field of model interpretability. By translating natural language into IPA, we should be able to better understand what a model is learning in its representation using different methods of analysis, mainly by observing input/output pairs. One example, the wug test, which evaluates morphemic and phonological understanding, cannot be tested in textual contexts – which is the default case for most language models – thus, by having phonetic outputs, we might observe interesting results that differ from text-only language models. The five papers in this literature review converge in answering what algorithm or model would be the best for this machine translation task and why, each giving their specific reasons and building on top of each other. The first two papers explore the innovative bidirectional pretrained model BERT on NMT, and the last three papers explore different applications and explorations of IPA in machine translation and speech recognition. In an indirect way, these papers also narrow down the scope and expands the motivations of word-to-IPA translation.

Concise summaries of the articles

BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

Devlin et al., 2019

In BERT, the pretraining of bidirectional representations is explored as an alternative to traditional transformer representations by analyzing an unlabeled text through reading both left and right context in all layers. One of the highlights of pretrained BERT is that, at the time of fine tuning, it only needs one additional output layer to achieve state-of-the-art performance on a wide variety of tasks. (Pre-training of Deep Bidirectional Transformers for Language Understanding).

Incorporating BERT Into Neural Machine Translation

Zhu et al., 2020

Aggregating on BERT's original paper, neural machine translation (NMT) using BERT was explored. Due to the lack of previous work, the researchers had to start from the beggining. First they tried to initialize downstream models and fine-tuning them afterwards. This method turned out unsuccessful yielding no significant improving. The second method was to use BERT representations as context-aware embeddings for other downstream models. This lead to the exploration of the BERT-fused models, which were the main finding of their paper: BERT is used to obtain the representations of the input sequence, and these representations are fused with each layer of the NMT encoder-decoder transformer model through attention mechanisms.

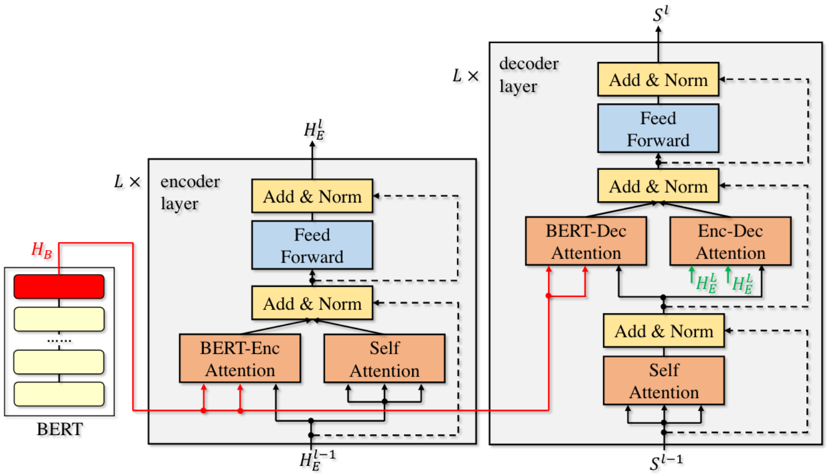

Illustration of BERT-fused model.

Multilingual and Crosslingual Speech Recognition Using Phonological-vector Based Phone Embeddings

Zhu et al., 2021

This paper brought JoinAP (Joining of Acoustics and Phonology) to the discussion, a new method that encodes each phone in the IPA (International Phonetic Alphabet) to its own phonological vector. This vector representation is the phone embedding that is used by the model for multilingual speech recognition, yielding outstanding results on the CommonVoice dataset – which includes German, French, Spanish, and Italian – and the AISHLL-1 dataset – which includes Mandarin Chinese. Through the experiments, it is clear that nonlinear phone embeddings perform superiorly over linear phone embeddings when using the JoinAP method.

Revisiting IPA-based Crosslingual Text-to-Speech

Zhang et al., 2021

IPA is a good source of information for phonological analysis, arguably the most complete phonetic alphabet that exists, however, there is not enough studies of it on cross-lingual text-to-speech (CL TTS). By using IPA characters as input, this paper explores building a a cross-lingual TTS model for cross-lingual voice cloning (CL VC). They found that if only one speaker was used per language, the TTS model would not achieve reliable results on CL VC. The main reason being due to language unique tone, stress, and phonetic inventory leaking speaker specific information.

Grapheme-to-Phoneme Models for (Almost) Any Language

Deri and Knight, 2016

Although CL VC with IPA has not been thoroughly explored, general cross lingual IPA interpretation using grapheme-to-phoneme (g2p) models have been around for decades. One method was to create phoneme and language distance metrics on high resource languages, to create models that would work for low-resource languages, which rarely have enough g2p training and evaluation dataset. In this project, the dataset Phoible was used, an online repository of CL phonological data. It contains phoneme inventories and unique phoneme feature vectors for 1674 languages; on top of that, it has named entity resources for 384 languages. This dataset could prove itself useful for any phonological model.

Compare and contrast

Although these papers differ greatly in the main tasks they are showcasing, they all contribute to the discussion of NMT from natural language to IPA. G2p models used to be the most reliable way of achieving NMT form a specific target language to IPA, however, transformers do a better job at machine translation. More specifically, bidirectional transformers take into account the context from both left and right sides, thus yielding better results in MT tasks. The same could be expected from English to IPA for example, since phonology is heavily dependent on phonotactic constraints – which requires to take into account both previous and preceding phonemes – bidirectional representaitons would in theory work better. Concerning cross lingual tasks, although encoding phone embeddings seems a viable option, IPA for any cross lingual task seems to not be the best approach.

The papers reveal that one of the main difficulties of cross lingual phoneme vectorization and acoustic-phonological translation is that different languages may have different interpretations of the same sound. Consider the sound ‘b’ in ‘bee’ and ‘p’ ‘pea’ in English, in this context, these two phonemes differ only on voice onset time, and to speakers of other languages such as Spanish, voice onset time thresholds for distinction are different. This already creates ambiguous references of the same acoustic representation to different phonemes, despite the IPA having unique feature vectors for each individual phoneme. Additionally, even speakers of the same language have different perceptions of phonetic thresholds of the same phonemes; for example, native English speakers from Boston perceive the ‘ɔ’ phoneme in wider ranges as compared to those from California. In synthesis, a language’s acoustic features are on a scale and gradual; labelling them to categorical phonemes inherently suffers from perceptual categorization, even within the same language. Thus, since crosslingual tasks do not benefit from the intermediary step of language-to-IPA, for interpretability work on word-to-IPA, using only one language is going to be more appropriate than incorporating multilingual capabilities in the model.

Future work

What is the performance of single language text-to-speech with BERT bidirecitonal embeddings?

Although crosslingual TTS was explored with a DNN (Deep Neural Network), how will BERT’s bidirectional representations change performance?

Will a BERT-fused-like model be able to translate word-to-IPA?

BERT-fused models seem promising for NMT, however, word-to-IPA is not exactly the same task as translating from one language to another, how will this extrapolate?

Additional questions

- Is bidirectionality important for phonetic encoding?
- Will an IPA intermediary step aid in TTS or Voice Cloning?
- Can there exist objective thresholds for each phoneme in existance?

References

1. Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. “BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding.” arXiv. https://doi.org/10.48550/arXiv.1810.04805.
2. Zhu, Jinhua, Yingce Xia, Lijun Wu, Di He, Tao Qin, Wengang Zhou, Houqiang Li, and Tie-Yan Liu. 2020. “Incorporating BERT into Neural Machine Translation.” arXiv. https://doi.org/10.48550/arXiv.2002.06823.
3. Zhu, Chengrui, Keyu An, Huahuan Zheng, and Zhijian Ou. 2021. “Multilingual and Crosslingual Speech Recognition Using Phonological-Vector Based Phone Embeddings.” arXiv. https://doi.org/10.48550/arXiv.2107.05038.
4. Zhang, Haitong, Haoyue Zhan, Yang Zhang, Xinyuan Yu, and Yue Lin. 2021. “Revisiting IPA-Based Cross-Lingual Text-to-Speech.” arXiv. https://doi.org/10.48550/arXiv.2110.07187.
5. Deri, Aliya, and Kevin Knight. 2016. “Grapheme-to-Phoneme Models for (Almost) Any Language.” In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 399–408. Berlin, Germany: Association for Computational Linguistics. https://doi.org/10.18653/v1/P16-1038.