INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

PronScribe: Highly Accurate Multimodal Phonemic Transcription From Speech and Text

Yang Yu[⋆], *Matthew Perez*[†1], *Ankur Bapna*[⋆], *Fadi Haik*[‡], *Siamak Tazari*[⋆], *Yu Zhang*[⋆]

Google USA[⋆], University of Michigan[†], Google Israel[‡]

yangyuai@google.com, mkperez@umich.edu, ankurbpn@google.com, fadih@google.com, staz@google.com, ngyuzh@google.com

Abstract

We present *PronScribe*, a novel method for phonemic transcription from speech and text input based on careful fine-tuning and adaptation of a massive, multilingual, multimodal speech-text pretrained model. We show that our model is capable of phonemically transcribing pronunciations of full utterances with accurate word boundaries in a variety of languages covering diverse phonological phenomena, achieving phoneme error rates in the vicinity of 1-2% which is comparable to human transcribers.

We show that PronScribe can effectively learn this task from relatively little training data, making it attractive even in low-resource settings. It learns from text and speech simultaneously in a coherent way, and is better than previous models using speech, text or both. Additionally, the model's good transfer learning characteristics in multilingual settings can effectively boost performance for lower-resourced languages.

1. Introduction

Phonemic transcriptions build the foundation of most text-to-speech synthesis (TTS) and automatic speech recognition (ASR) systems of today [1, 2]. TTS voices are typically built using phoneme level transcription and alignment of the recording script, and a pronunciation layer is a necessary component of the text normalization front-end of any controllable TTS system [3, 4]. Similarly, despite trends toward E2E modeling, many ASR systems rely on pronunciations to connect their acoustic and language models, and for controllability and biasing.

Traditionally pronunciation lexicons are curated through manual annotation which is both expensive and time-consuming, often becoming a critical bottleneck for scaling these systems to wider language bases. This is especially challenging for low-resource languages, which lack the data and/or experts for labeling. Besides the lexicon, transcription of *full utterances* presents additional challenges: resolution of homographs (which includes diacritics recovery in languages like Arabic, Hebrew), reduced variant pronunciations (e.g. for function words), and sandhi effects where the pronunciation of a word may be affected by its surroundings (e.g. liaison in French, 3rd tone sandhi in Mandarin, pitch accent in Japanese) are some examples. The task is so nuanced that even human experts exhibit a level of error in their transcriptions which we will examine in more depth later, as the transcription of phoneme sequences needs to encapsulate all detailed speech complexities such as syllabification, stress, tone, pitch accent, etc. With this in mind, having a neural model learned from unaligned, speech and text samples to generate the proper phoneme sequence would provide immense value in helping automate the creation of pronunciation transcriptions and lexicons. Besides, this neural model has further possible applications in alignment, language learning, grapheme-to-phoneme (G2P) development.

In this work, we present *PronScribe*, a multi-modal model framework for transcribing pronunciations from text and speech by utilizing both inputs simultaneously in a coherent way, improving significantly over baselines with different inputs:

- Speech-only (e.g. classic ASR/phoneme recognition [5])
- Text-only (e.g. classic TTS frontend such as Kestrel [4])
- Speech+Text (e.g. PronLearning [6])

We show phoneme and word error rates on a data set of rich utterances (see Section 4) for a diverse set of languages, and establish that its quality is close to human annotators, often correcting mistakes in the "ground-truth" annotations. We demonstrate the PronScribe model's effectiveness at generating accurate phoneme sequences including word and syllable boundaries, stress, tone, etc., while resolving the aforementioned issues such as homographs, variants, and sandhi effects. Additionally, we show that small amount of training data can lead to an acceptable baseline quality, and quickly improve thereafter, developing a recipe for forecasting the model's performance based on the amount of available data, and showing the suitability of our method in low-resource settings. Finally, we study the transfer learning characteristics of the model with multiregional and multilingual training showing favorable results.

2. Related Work

A previous pronunciation learning system [6] used a classic ASR acoustic model and a graph G2P FST as its language model. This required the existence of a strong G2P and thus was not suitable for low-resource languages. Its focus was on learning pronunciations for single new words in a high-resource language rather than transcribing entire utterances. We use this model as a comparative baseline in our work.

Some pronunciation literature focuses on analysis for language learning purposes [7, 8]. Although related, these applications are typically limited in scope by focusing on identifying a limited set of phonemes. These models are tuned to detect nuanced mispronunciations rather than learn patterns to generate accurate phoneme transcriptions at scale.

A closely related problem is grapheme-to-phoneme conversion (G2P). G2P only uses text and is typically applied to single words. It is richly studied in the literature, e.g. [9, 10, 11, 12, 13]. By utilizing audio in addition to the text, our model can surpass the quality of traditional G2P models, can do so in context, and can produce training data for obtaining better G2P

[1] The work is done while the author was an intern at Google.

10.21437/Interspeech.2023-1097

models. Route et. al. [14] investigate a multimodal approach for learning phonemic sequences utilizing both text and speech, and show that multilingual training leads to large performance improvements for low-resource languages. Route et. al. use the audio signal as an auxiliary output by trying to recreate MFCCs from text input with the final goal being a text-to-text G2P. This is contrast to our work using the audio as an additional input signal rather than an output.

The full-sentence text-to-phoneme transcription problem *without audio* is in essence the TTS text normalization and pronunciation problem. Modular rule- and lexicon-based systems such as Kestrel [4] are still widely used which we utilize as an additional comparative baseline for our work. Neural models for text normalization (excluding pronunciation) [15, 16, 17] and fully E2E text-to-phoneme models [18] generally could achieve very high accuracy, if there is a preexisting text normalization system or there are vast amounts of training data, which are very difficult to obtain. We provide the first scalable way of generating high-quality data for such models from speech+text sources, not relying on a preexisting working system.

3. Model

Our PronScribe framework consists of 2 stages: Pretraining a massive multilingual, multimodal speech-text model; and careful finetuning and adaptation to the pronunciation task.

3.1. Multilingual and Multimodal Pretraining

In the pretraining stage, we require a model that encodes both speech and text and learns a joint representation of both modalities. For our experiments, we used mSLAM [19] which has demonstrated a strong ability to learn cross-lingual, cross-modal representations of speech and text. It combines pretraining on over 800k hours of unlabeled speech data spanning over 51 languages, vast amounts of unlabeled text from 101 languages, as well as some paired speech with transcripts. Note that the pretraining does not involve pronunciations in any form.

The separate encoders for text and speech are followed by a multimodal encoder consisting of a deep stack of Conformer layers [20]. There are multiple pretraining objectives ensuring that in addition to learning each of the modalities individually, the correlation and a joint representation between them are effectively learned using the comparatively smaller amounts of paired speech and text data (see [19] for details). Ultimately, it has been shown to be broadly successful on a diverse set of downstream tasks including multilingual speech translation, speech classification, speech recognition, and text classification, making it a promising candidate for pronunciation transcription.

3.2. Pronunciation Finetuning

For finetuning shown in Fig. 1, the encoder side essentially reuses and continues to finetune the encoders from mSLAM. We removed self-supervision losses and the masking used in pretraining. We always concatenate both encoded modalities before feeding them into the Conformer layers of the multimodal encoder, as the task usually uses both inputs. An RNN-T decoder [21] is added after the encoder, which is optimized over a connectionist temporal classification (CTC) [22] loss function. We also compared RNN-T to an attention-based decoder [23], and observed that while those decoders exhibit a somewhat higher overall accuracy, they occasionally miss some parts of the output causing problems for downstream tasks.

The final model is finetuned to the task of phoneme sequence generation. Specifically, the output that we train on is the phonemic transcription of the full utterance with word and syllable boundaries including stress and tone as relevant (see Table 3). There are a few additional important task-specific adaptations:

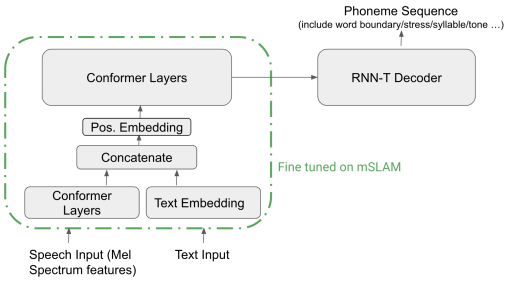

Figure 1: *Model architecture of PronScribe finetuning*

- We used separate learning rates for encoders and decoders.
- We restricted the maximum length of the output phoneme sequence to 1024 so as to cover about 99% of the available data while balancing model capacity, training efficiency, inference speed, and final quality.
- We augmented the data with speech-only input.

The latter point about data augmentation was a particularly crucial step for the success of the model. We initially observed that the task of jointly learning from multiple modalities is still very challenging during finetuning, and that the model was relying mostly on the text as it is shorter and more discretized. However, for the phonemic transcription task, we prefer the model to regard the speech as the source of truth more so than the text. We designed two data augmentation ways to help the model utilize the speech signal more strongly. The first is to mask some portion of the text for each input utterance. We call this method *masking*. The other is to completely remove the text for certain random utterances and feed them as additional speech-only training examples besides the original examples with paired input. We call this method *mixing*. As discussed in Section 5.2, the mixing method produced better results, and incidentally made the model also suitable for inference from speech-only data.

4. Data Sets

We used 3 data sets in our work:

(A) LibriSpeech: Widely-used and openly available.

(B) Single Words: With associated audio and pronunciation strings in English (proprietary).

(C) Rich Utterances: Recorded in studio conditions with text and phonemic transcripts in multiple languages (proprietary).

LibriSpeech is noisy both in terms of audio quality and text accuracy [24], and lacks phonemic transcriptions. To overcome the latter, we used an internal version of Kestrel [4] (see Section 2) to obtain reference pronunciations which are high quality and comparable to state-of-the-art methods (but not perfect).

Some statistics on the Rich Utterances set are available in Table 2. The phonemic transcriptions in this set were partially human-curated and are expected to match the audio very closely; nevertheless, they are not truly golden. We observed that finetuning only on set (C) gave us the best results and this is assumed in all experiments below.

Dataset	Base PER	Base WER	PER	WER
LibriSpeech	63	71.4	5.4	15.3
Single Words	7.1	14.5	0.7	2.2
Rich Utterances	24.9	42.6	1.9	4.9

Table 1: *Comparison to the baseline [6].*

Language	pretraining speech (hrs)	finetune utt.	finetune speakers	PER	WER
en-US	$> 70,000$	$> 200K$	25-30	1.9	4.9
fr-FR	$> 20,000$	$> 100K$	5-10	1.9	6.5
de-DE	$> 20,000$	$> 40K$	5-10	1.6	5.3
cmn-CN	< 50	$> 40K$	5-10	1.8	5.8
he-IL	0	$> 20K$	5-10	0.8	4.5
ar-XA	< 10	$> 10K$	5-10	1.9	10.4
am-ET	0	$> 10K$	4	2.2	11.5
fa-IR	≈ 200	$\approx 1.5K$	1	2.3	10.9

Table 2: *Results on the Rich Utterances set. Based on a single pretrained 51-language mSLAM model with the amounts of included speech hours indicated for the relevant ones.*

Text (EN)	Take a deep breath as I read today's mindfulness tip
Ref.	"t eI k | @ | . . . | "aI | "r\E d | t @ . "d eI z | . . .
Model	"t eI k | @ | . . . | "aI | **"r\i: d** | t @ . "d eI z | . . .
Why	The model reflects the homograph pronunciation of "read".
Text (EN)	Yank says what you doing johnny?
Ref.	"j { N k | . . . | "j u: | "d u: . @ N | "dZ A: . %n i:
Model	"j { N k | . . . | @` | **j** @ | "d u: . @ N | "dZ A: . %n i:
Why	Speaker said "are y' ", "are" and reduced "you" are added.
Text (FR)	Va prendre l'air.
Ref.	v a | p R A~d R | l E R
Model	v a | p R A~**. d R** @ | l E R
Why	The model resolves the correct sandhi effect on "prendre".
Text (FA)	در سراسر جهان شناخته شده بود
Ref.	d { 4 | s { . 4 Q: . s { 4 | dZ { . h Q: n | . . .
Model	d { 4 | s { 4 . **?** Q: . s { . **4 e** | dZ { . h Q: n | . . .
Why	All vowels are recovered without diacritics in the input including the vowel /e/ between the second and third word which is added due to a grammatical sandhi effect. A minor error is the introduction of the glottal stop /?/.

Table 3: *A few examples show why the model is different from reference. Transcriptions are in X-SAMPA notation.*

5. Experiments and Results

We primarily use phoneme error rate (PER) and word error rate (WER) as the two main metrics to evaluate and compare performance. PER is defined as the Levenshtein distance between the reference phoneme sequence and the predicted one, including syllable boundary and stress/tone symbols, compared to the total number of phonemes. Since predicting word boundaries is part of our task, we can define WER in an analogous way, where a word is considered wrong if any one symbol (phoneme/stress/. . .) inside it is wrong.

5.1. Comparison to Baselines

Most directly comparable is the baseline introduced in [6]. Recall that it combines an off-the-shelf ASR AM and a G2P FST, and has no trainable components of its own (see Sec. 2). Table 1 shows significant improvements over this system on all three US English data sets. Note that [6] was designed mainly for single-word pronunciation learning, and we found that it often fails at predicting anything on long or noisy speech, e.g. as found in LibriSpeech. In contrast, our model successfully transcribes long utterances and is robust to noise; and significantly improves over the baseline, even on single words.

Less comparable are text-only models, such as the latest transformer-based G2P systems [10, 12], reporting WER well over 20% on English single words versus our 2.2% WER; and a full text-to-phonemes system such as Kestrel [4] with access to a lexicon achieved 7% WER on the Rich Utterances set versus our 4.9% in Table 1 – showing that PronScribe utilizes the added speech effectively.

5.2. Effect of Multiple Modalities

We further studied the performance of the model when only speech modality is presented (text input is empty). We observed that the WER on LibriSpeech goes up to 16.8%, an absolute increase of 1.5% (compared to Table 1 that uses both modalities), which proves that the model using both modalities is better.

As mentioned in Section 3.2, we experimented with 2 methods for data augmentation in order to make the model follow the speech more closely: *masking* and *mixing*. Without any of these augmentations, the WER on LibriSpeech with **speech-only** input was 56.4%. We experimented with augmentation ratios from 10% to 80%, and found that 50% was the sweet spot. The best result with the *masking* technique delivered 26.6% WER, whereas the best result with the *mixing* technique (i.e. adding 50% speech-only utterances to the training) achieved the 16.8% reported above (and using speech+text at inference time delivers the 15.3% in Table 1).

5.3. Covering Diverse Phonological Phenomena

In Table 2, we examined the model on different languages with diverse linguistic and phonological phenomena such as diacritics recovery (ar-XA, he-IL, fa-IR), stress (en-US, de-DE), tone (cmn-CN), and sandhi (fr-FR, cmn-CN, fa-IR). We can see that our model performs consistently well in all cases even when very small amounts of data are available either during pretraining or finetuning.

5.4. Qualitative Analysis

As discussed in Section 4, the transcriptions of the Rich Utterances are of high quality but not truly golden. They were curated in an iterative way using human annotators and automatic tools and lexicons. To further understand the model's *true* accuracy, we selected a set of utterances that were known to have had human corrections applied on. We believe this set to contain considerably fewer errors and to be closer to a true golden set. We evaluated the system and obtained *true* PER 0.9% and WER 2.1% in en-US (cf. the 1.9% PER and 4.9% WER in Table 2). Even there we observed numerous examples where PronScribe improved over the human annotation.

While in other languages we did not have such a bespoke corrections set, in our analysis of wins and losses we would very often observe that the model was right and the reference annotation was wrong. Some examples in English and other languages are given in Table 3. Overall this leads us to believe that across languages:

(i) The true PER and WER are likely better than in Table 2; and

(ii) Humans make a comparable number of errors when facing the daunting task of full phonemic transcription.

5.5. Low-Resource Training

We are interested in seeing the performance of PronScribe in resource-restricted settings. We simulate low-resource training by segmenting the training set of two high-resource languages

(US English and German) into varying smaller sub-partitions. The results are shown in Fig. 2. We observe that with 2000 training examples, a PER of $< 5\%$ can be achieved offering an acceptable baseline. We observe a standard power-law relationship between the performance and the amount of training data (R^2>0.996), which highlights PronScribe's ability to perform relatively well with limited amounts of training data available.

We then average the power-law functions of US English and German to approximate a function that depicts the relationship between PronScribe performance and available training data in Figure 3. This forecast function is particularly useful for approximating performance given the amount of labeled data for new languages. We find that the efficacy of this forecast function is highlighted by estimating similar error rates to most of those shown in Table 2 for low- and high-resource languages.

5.6. Significance of Pretraining

Our model effectively learns from large amounts of unpaired speech and text data and some paired data during pretraining and forms inherent unnamed representations for all phones (cf.[25]). The finetuning essentially teaches the model to map those phone representations to specific phoneme symbols of the language. Therefore, without pretraining, it would likely be more difficult for the model to efficiently learn this mapping. We confirmed this with experiments in several different languages: Compared to the results on Arabic and Persian (10.4% and 10.9% WER) in Table 2, without pretraining, it can only achieve 12.3% and 33.6%, respectively.

5.7. Multilingual and Multiregional Transfer Learning

To understand the transfer learning characteristics of our model, we first started investigating its ability to leverage different regional data from the same language. We selected several regional English data (i.e. American, British, Australian, Indian, Nigerian, Singaporean), and compared the monolingual performance with 2 multi-region versions: One with and one without lang-id. In experiments with lang-id, we prepend the text input with two tokens indicating language and region.

The results are shown in Fig. 4. We observed performance improvements across all English regions for multi-region models with lang-id, and significant improvements for low-resource languages such as en-NG and en-SG. Even for en-US, we found that multilingual training improved performance over the monolingual baseline. This suggests that PronScribe can effectively leverage data from multiple different regions to improve performance, and provides a practical alternative to simply collecting more data for a single language.

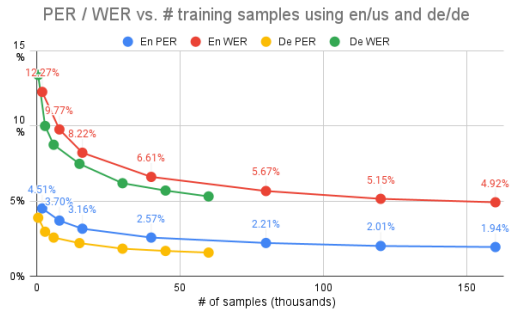

Figure 2: *Resource-restricted training for en-US and de-DE.*

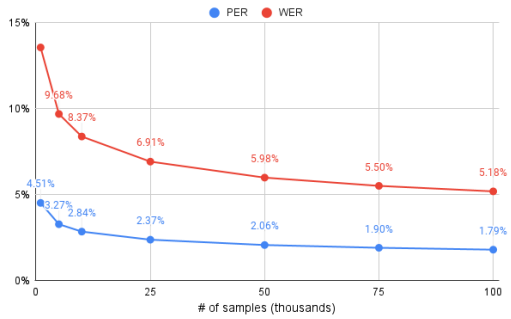

Figure 3: *Forecast performance trend for training PronScribe.*

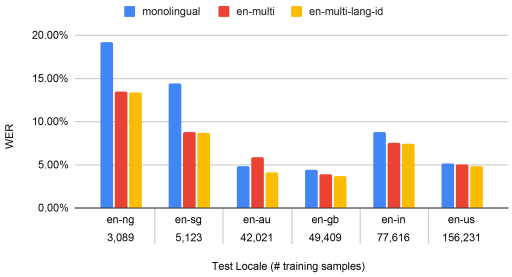

Figure 4: *WER of multi-region training on different EN locales.*

Although explicitly providing lang-id delivered consistently better results, an interesting finding is that contrary to G2P experiments [11], in most cases, multilingual training without lang-id still improves performance over the monolingual baseline. This suggests that the speech input allows the model to infer similar characteristics to an explicitly defined lang-id which ultimately facilitates multilingual training.

An experiment with Romance languages, including Spanish, French, French Canadian, Italian, and Romanian indicated similar results where particularly the lower-resourced regions benefited significantly. However, Romanian improved only slightly suggesting that the closeness of the languages matters.

6. Conclusion

We presented *PronScribe*, a novel method for phonemic transcription from speech and text input based on careful finetuning and adaptation of a massive, multimodal, multilingual pretrained model. We showed that our model is capable of phonemically transcribing full utterances in a variety of languages covering diverse phonological phenomena, significantly improving over previous methods, whether they used only text, only speech, or both. We argue that the accuracy we achieve is reasonably close to human transcribers, and that it presents a scalable way for generating large amounts of data for text-to-phoneme models.

Furthermore, we studied the model and its properties in depth, developing a recipe for predicting the accuracy of the model based on the available amount of training data, and particularly showing its suitability in low-resource settings. Finally, we observed good transfer learning capabilities, and that lower-resourced regions benefit significantly from joint training with higher-resourced ones. With the system, we have generated over 3 million high quality text-to-phoneme data samples in multiple languages.

7. References

[1] E. Trentin and M. Gori, "A survey of hybrid ann/hmm models for automatic speech recognition," *Neurocomputing*, vol. 37, no. 1-4, pp. 91–126, 2001.

[2] S. R. Mache, M. R. Baheti, and C. N. Mahender, "Review on text-to-speech synthesizer," *International Journal of Advanced Research in Computer and Communication Engineering*, vol. 4, no. 8, pp. 54–59, 2015.

[3] H. Zhang, R. Sproat, A. H. Ng, F. Stahlberg, X. Peng, K. Gorman, and B. Roark, "Neural Models of Text Normalization for Speech Applications," *Computational Linguistics*, vol. 45, no. 2, pp. 293–337, 06 2019. [Online]. Available: https://doi.org/10.1162/coli_a_00349

[4] P. Ebden and R. Sproat, "The kestrel TTS text normalization system," *Nat. Lang. Eng.*, vol. 21, no. 3, pp. 333–353, 2015. [Online]. Available: https://doi.org/10.1017/S1351324914000175

[5] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao, J. Wu, L. Zhou, S. Ren, Y. Qian, Y. Qian, J. Wu, M. Zeng, and F. Wei, "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," *CoRR*, 2021.

[6] A. Bruguier, D. Gnanapragasam, L. M. Johnson, K. Rao, and F. Beaufays, "Pronunciation learning with rnn-transducers," in *INTERSPEECH*, 2017.

[7] A. Asif, H. Mukhtar, F. Alqadheeb, H. F. Ahmad, and A. Alhumam, "An approach for pronunciation classification of classical arabic phonemes using deep learning," *Applied Sciences*, vol. 12, no. 1, p. 238, 2021.

[8] Y. Bi, C. Li, Y. Benezeth, and F. Yang, "Impacts of multicollinearity on capt modalities: An heterogeneous machine learning framework for computer-assisted french phoneme pronunciation training," *Plos one*, vol. 16, no. 10, p. e0257901, 2021.

[9] K. Rao, F. Peng, H. Sak, and F. Beaufays, "Grapheme-to-phoneme conversion using long short-term memory recurrent neural networks," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 4225–4229.

[10] S. Yolchuyeva, G. Németh, and B. Gyires-Tóth, "Transformer Based Grapheme-to-Phoneme Conversion," in *Proc. Interspeech 2019*, 2019, pp. 2095–2099.

[11] A. Sokolov, T. Rohlin, and A. Rastrow, "Neural machine translation for multilingual grapheme-to-phoneme conversion," in *Interspeech 2019*, 2019.

[12] M. Yu, H. D. Nguyen, A. Sokolov, J. Lepird, K. M. Sathyendra, S. Choudhary, A. Mouchtaris, and S. Kunzmann, "Multilingual grapheme-to-phoneme conversion with byte representation," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 8234–8238.

[13] K. Vesik, M. Abdul-Mageed, and M. Silfverberg, "One model to pronounce them all: Multilingual grapheme-to-phoneme conversion with a transformer ensemble," in *Proceedings of the 17th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*. Online: Association for Computational Linguistics, Jul. 2020, pp. 146–152. [Online]. Available: https://aclanthology.org/2020.sigmorphon-1.16

[14] J. Route, S. Hillis, I. Czeresnia Etinger, H. Zhang, and A. W. Black, "Multimodal, multilingual grapheme-to-phoneme conversion for low-resource languages," in *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019)*. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 192–201. [Online]. Available: https://aclanthology.org/D19-6121

[15] H. Zhang, R. Sproat, A. H. Ng, F. Stahlberg, X. Peng, K. Gorman, and B. Roark, "Neural models of text normalization for speech applications," *Comput. Linguistics*, vol. 45, no. 2, pp. 293–337, 2019. [Online]. Available: https://doi.org/10.1162/coli_a_00349

[16] C. Mansfield, M. Sun, Y. Liu, A. Gandhe, and B. Hoffmeister, "Neural text normalization with subword units," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Industry Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 190–196. [Online]. Available: https://aclanthology.org/N19-2024

[17] S. Tyagi, A. Bonafonte, J. Lorenzo-Trueba, and J. Latorre, "Proteno: Text normalization with limited data for fast deployment in text to speech systems," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Industry Papers*. Online: Association for Computational Linguistics, Jun. 2021, pp. 72–79. [Online]. Available: https://aclanthology.org/2021.naacl-industry.10

[18] A. Conkie and A. Finch, "Scalable multilingual frontend for tts," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 6684–6688.

[19] A. Bapna, C. Cherry, Y. Zhang, Y. Jia, M. Johnson, Y. Cheng, S. Khanuja, J. Riesa, and A. Conneau, "mslam: Massively multilingual joint pre-training for speech and text," *arXiv preprint arXiv:2202.01374*, 2022.

[20] A. Gulati, J. Qin, C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, and R. Pang, "Conformer: Convolution-augmented transformer for speech recognition," in *Interspeech*. ISCA, 2020, pp. 5036–5040. [Online]. Available: https://doi.org/10.21437/Interspeech.2020-3015

[21] A. Graves, "Sequence transduction with recurrent neural networks," *CoRR*, vol. abs/1211.3711, 2012. [Online]. Available: http://arxiv.org/abs/1211.3711

[22] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks," in *Proceedings of the 23rd International Conference on Machine Learning*, ser. ICML '06. New York, NY, USA: Association for Computing Machinery, 2006, p. 369–376. [Online]. Available: https://doi.org/10.1145/1143844.1143891

[23] W. Chan, N. Jaitly, Q. Le, and O. Vinyals, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 4960–4964.

[24] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An asr corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 5206–5210.

[25] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., vol. 33. Curran Associates, Inc., 2020, pp. 12 449–12 460. [Online]. Available: https://proceedings.neurips.cc/paper/2020/file/92d1e1eb1cd6f9fba3227870bb6d7f07-Paper.pdf