PHONETIC RNN-TRANSDUCER FOR MISPRONUNCIATION DIAGNOSIS

Daniel Yue Zhang, Soumya Saha, Sarah Campbell

Amazon Alexa AI
{dyz, soumyasa, srh}@amazon.com

ABSTRACT

Non-autoregressive models and, in particular, Connectionist Temporal Classification (CTC) models have been the most popular approaches towards mispronunciation detection and diagnosis (MDD) task. In this paper, we identify two important knowledge gaps in MDD that have not been well studied in existing MDD research. First, CTC-based MDD models often assume conditional independence in the predicted phonemes, and therefore prominent mispronunciation patterns are underutilized. Second, existing MDD approaches are constrained to use training data from a language-specific phoneme set, and therefore cannot distinguish accented sounds that are not present in the predefined phoneme set. In this paper, we propose a set of autoregressive phonetic Recurrent Neural Network Transducer (RNN-T) MDD models that are capable of capturing temporal dependence of mispronunciation patterns. We further devise an extended phoneme set and weakly supervised training strategy to allow the model to distinguish similar sounding phonemes from different languages. We evaluate the proposed method on the public L2-ARCTIC dataset. Results have shown the proposed phonetic RNN-T model achieves significant improvements in false acceptance rate compared to state-of-the-art methods.

1. INTRODUCTION

Recent history has seen a massive growth in computer assisted pronunciation training (CAPT) with the ever increasing popularity of language learning. A key enabling technology for CAPT is mispronunciation detection and diagnosis (MDD) which identifies pronunciation errors and provides corrective feedback to guide non-native (L2) language learners [1]. Various models have been proposed to address the MDD task. Roughly, these models can fall into three categories: 1) force-alignment methods that align a reference transcript against the actual pronunciation, followed by a heuristic scoring function such as Goodness of Pronunciation (GOP) [2, 3]; 2) phoneme recognition methods that first convert input L2 speech into recognized phonemes. Then mispronunciation feedback is provided by aligning recognized phonemes with canonical phonemes converted from reference text [4, 5, 6]; and 3) end-to-end (E2E) methods that directly predict the pronunciation errors (e.g., insertion, deletion, substitution) without the intermediate phoneme output [7]. In this study, we focus on the phoneme recognition-based MDD methods, which can provide fine-grained phonetic-level feedback and have been shown to achieve state-of-the-art (SOTA) performance.

We found two important knowledge gaps that have not been well studied in existing MDD approaches. We refer to the first knowledge gap as the *Curse of conditional dependence*. In particular, we found that a vast majority of phonetic-level MDD models are non-autoregressive models trained on Connectionist Temporal Classification (CTC) loss. CTC-MASK [8], CNN-RNN-CTC [4], and more recent Wav2Vec2.0+CTC based MDD models [6, 9] are typical examples. One of the main advantages that led to the prevalence of these CTC models is conditional independence - the prediction of the phoneme is irrelevant to its precedent output. This is often believed to be a desirable feature because conditional dependency in the MDD case can lead to smoothing out mispronunciation and causing false acceptance of mispronunciation [8]. Consider an example where a language learner pronounces "mor<u>n</u>ing" as "mor<u>t</u>ing", an autoregressive model would likely to favor predicting the "n" sound rather than "t" based on previous phonemes, thus failing to capture the mispronunciation. However, non-autoregressive CTC approaches, without the constraints of a "language model", can also suffer from a lack of fidelity in the predicted outcome [10], such as predicting consecutive identical phonemes, or predicting unpronounceable phoneme sequences. We found the trade-offs in choosing different levels of conditional dependence for phoneme recognition in MDD have not been well studied in the existing literature.

The second knowledge gap lies in dealing with *cross-lingual phoneme disambiguation*. In particular, when L2 speakers learn a new language, they often carry over the phonemes from their native languages while pronouncing a foreign word [11]. However, most existing MDD approaches only use the phoneme set of the target language to be learnt. For example, for English language learning, most MDD models are trained with datasets annotated with English phoneme set such as the CMU Dictionary [12] or TIMIT [13]. We found such constrained phoneme set cannot accurately capture pronunciation variances from the native language or

accents of the language learner. In particular, these MDD models either map the language learner's pronunciation (L2) to the most similar English (L1) phoneme. For example, recognizing "rrabit" with a Spanish "rr" sound as "rabit" in English, will lead to inaccurate phoneme recognition and thus false acceptance of mispronunciation.

In this paper, we jointly address the above knowledge gaps by proposing phonetic RNN-T modeling for phonetic-level MDD task. A series of techniques are devised to relax the innate autoregressiveness of RNN-T, so that it achieves a good balance of false rejection and false acceptance rates of the resulting MDD performance. A novel weakly supervised data augmentation strategy with an extended phoneme set is proposed for the disambiguation of L1 and L2 phonemes. We benchmarked our proposed model as well as a comprehensive list of candidate approaches on the public L2-ARCTIC [14] dataset. Results have shown that proposed phonetic RNN-T based MDD model achieves comparable or even better performance with SOTA CTC models, and achieves 16.8% improvements in false accept rate for Spanish language learners.

2. METHOD

2.1. Phonetic RNN-T Model for MDD

The overview of Phonetic RNN-T is presented in Figure 1. Following [15], the proposed RNN-T model consists of an encoder ($\mathbf{E}$), a prediction network ($\mathbf{D}$) and a joint network ($\mathbf{J}$). The encoder receives raw acoustic feature vectors $\mathbf{x} = x_1, x_2, ..., x_T$ from input audio with frame length of T, and converts them into a sequence of hidden states $h^E t = f^E(x_t)$, where t is the time/frame index. The prediction network acts as a "language model" that takes previous non-blank subword label prediction $\tilde{y_{u-1}}$ as input, and produces hidden representation. Formally, $h^D u = f^D(y_{u-1})$, where u is the label index. Note that this formulation is different from CTC where conditional independence is assumed, i.e., $y_u \perp\!\!\!\perp y_j$ where $j < u$. RNN-T removes this independence assumption by instead conditioning on the full history of previous non-blank labels.

The joint network is a feed-forward network that takes each combination of encoder output and prediction network output and computes output logits $z^J_{u,t} = f^J(h^E t, h^D u)$. Then a softmax layer is applied on top of the logits to produce a final posterior for the next output token. The RNN-T model is trained by minimizing the RNN-T loss:

$$\mathcal{L}_{\text{RNN-T}} = -logP(y|x) \tag{1}$$

During inference, the N-best list of the utterance is generated using beam search decoding [15]. We propose two variants of the phonetic RNN-T model (Figure 1(b)) that capture different degrees of conditional dependence.

Monophone RNN-T. The training data for the phonetic RNN-T is formatted as a pair of audio and its corresponding monophone sequence.

Syllable RNN-T. The training data is the audio and its syllable sequence. To convert the phonemes into syllable, we apply Maximum Onset Principle [16]. The intuition for designing the syllable RNN-T is to explore the effect of further enforcing conditional dependency among phonemes.

The proposed monophone and syllable phonetic RNN-T (denoted as **rnnt-phn** and **rnnt-syl**) models both consist of a six-layer LSTM encoder with 1,024 units each and a two-layer LSTM prediction network with the same number of units. We use Adam optimizer with a warm-up, hold, and exponential learning rate decay policy and a mini-batch size of 768. We extract 64-dimensional Log-Mel-Frequency features with a filter bank size of 64 every 10ms from the input speech signal. Three consecutive frames are stacked, resulting in a 192-dimensional feature vector. We used 4,000 wordpieces for input embedding.

2.2. Taming Autoregressiveness of RNN-T

Unlike non-autoregressive models that completely remove the conditional dependence of predicted phoneme sequences, we took an alternative approach by taming the autoregressiveness of RNN-T. The intuition is to allow the model to both 1) leverage the conditional dependence (i.e., phoneme history) to capture the mispronunciation patterns of language learners; and 2) avoid over-smoothing of the prediction output, which leads the model to be insensitive to mispronunciation. In particular, we argue that one of the main reasons that leads to undesirable conditional dependence is when the training data lacks diversity and misses certain mispronunciation patterns. In the previous "morning" vs. "morting" example, the training data has seen dominantly "morning", and thus fail to capture the mispronounced "t" sound.

To tame the conditional dependence of phonetic RNN-T, we first propose to enrich the training data with synthetic L2 pronunciation with diversified mispronunciation patterns. We adopt the L2-GEN approach [5], which takes a small amount of seed L2 training data and uses a neural phoneme paraphraser and a neural TTS module to convert L1 phoneme sequences into diversified accented L2 speech. We further break the conditional dependence among words by injecting a special word break token $\langle wb \rangle$ in the training data. Finally, we modify the beam search for RNN-T inference to allow the top-N output phonemes to be diversified. Formally, we adopt the diversified beam search technique [17]. We define G diversified beam groups, each of which has a beam width of $K' = |K/G|$. During the decoding step for group $g \in G$, for each candidate monophone p, *diversity penalty* (Δ) is added to the original beam search loss:

$$\Delta(p_{[t]}, \bar{g}) = \lambda \cdot \sum_{h \in \bar{g}} \sum_{p \in T^h_{[t]}} dist(p, p_{[t]}) \tag{2}$$

where λ is a normalization scalar that maps $\Delta(p_{[t]}, \bar{g})$ to [0,1], $dist(\cdot)$ is the edit distance between two phoneme sequences.

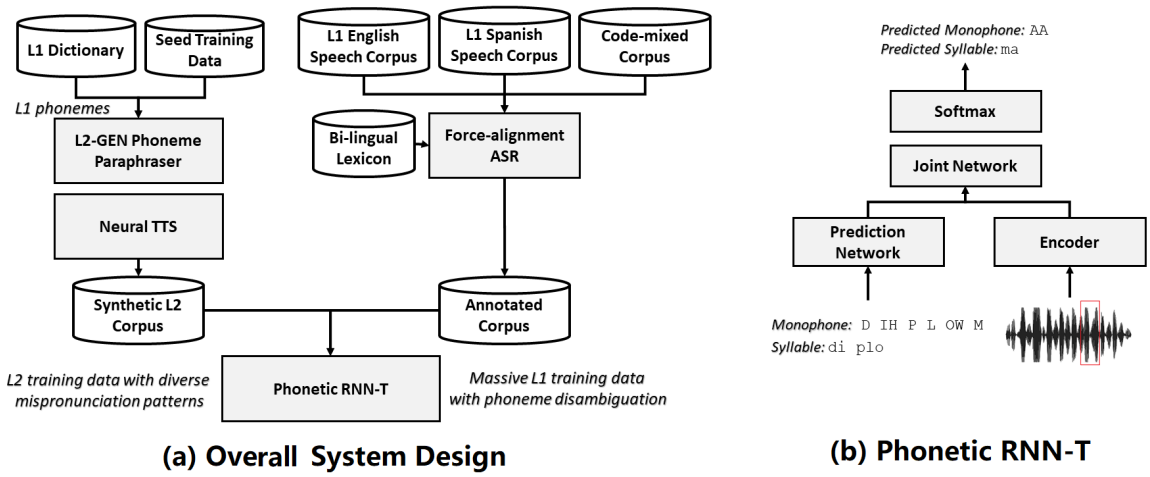

Fig. 1. Proposed Phonetic RNN-T Framework

2.3. Cross-Lingual Phoneme Disambiguation

Next, we address the cross-lingual phoneme disambiguation by building an extended phoneme set, and apply weakly-supervised learning to improve the MDD model's recognition on non-native phonemes. Consider the MDD use case for Spanish speakers learning English. Intuitively, to distinguish Spanish accents (Spanish phonemes when speaking English) from canonical reference pronunciation, we can train the MDD model using audio clips collected from L2 Spanish speakers speaking English and the audio clips need to be annotated with a universal phoneme set such as full IPA. However, such L2 speech is hard to obtain, and performing cross-lingual phonetic transcription takes tremendous manual effort and cost.

In this work, we propose a novel weakly supervised data augmentation approach that does not require non-native (L2) speech nor human annotation. This technique allows our RNN-T model to learn and disambiguate native and non-native phonemes using a massive speech corpus. To this end, we first create a bi-lingual lexicon by merging the L1 and L2 lexicons. During this process, we merge the phonemes that are (almost) indistinguishable such as "ll" in Spanish vs. "y" in English, while keeping unique L2 phonemes such as the hard "r" sound in Spanish. Then in the bi-lingual lexicon, each word has multiple valid pronunciations (reference phonemes) from both L1 and L2 language.

In the next step, we collect massive amount of non-transcribed speech corpora including 1) native Spanish speech; 2) native English speech, and 3) code-mixed Spanish/English speech. Then we apply a teacher force-alignment model [18] to generate pseudo phonetic transcriptions using the bi-lingual lexicon. The phonetic transcribed mixture of L1 and L2 speech are then used to train the phonetic RNN-T model. We illustrate the framework in Figure 1(a).

3. EXPERIMENTS

3.1. Experiment Setup

We use the popular L2-ARCTIC [14] dataset to benchmark the proposed method. Additional TIMIT corpus is also used for training MMD models. All speech data was downsampled to 16 kHz. The dataset statistics are summarized in Table 1. We chose False Rejection Rate (FRR), False Acceptance Rate (FAR), Precision, Recall, F1-score, and phoneme error rate (PER) as evaluation metrics for phonetic-level MDD performance.

	TIMIT	L2-ARCTIC		
	Train	Train	Dev	Test
Speakers	630	12	6	6
Hours	4.5	1.84	0.94	0.88
Utterances	6300	1800	897	900
L1 Language	English	Spanish, Hindi, Korean, Mandarin, Arabic,Vietnamese		

Table 1. Dataset Summary.

We use state-of-the-art baseline models for comparison, including *CNN-RNN-CTC* [4], and Wav2Vec2.0+CTC based MDD models [6, 9] referred to as *base-ctc* and *xlsr-ctc*. The wav2vec2.0 base model was pretrained on the Librispeech dataset with no additional finetuning. The multilingual XLSR model was pretrained on 56,000 hours of speech data covering 53 languages. We finetuned all baselines using the L2-ARCTIC + TIMIT data. We used Adam optimizer with learning rate $1e^{-8}$ and CTC criterion. All candidates were trained on a NVIDIA V100 GPU.

3.2. Results

We first present the overall MDD performance for all compared candidates in Table 2. For a fair comparison, proposed

phonetic RNN-T models were not trained on any corpora other than TIMIT and ARCTIC. We can observe that monophone RNN-T model closes the gap with the XLSR CTC model in terms of F1-score, FRR, and PER, and achieves better FAR. We attribute this performance gain to the L2-GEN data augmentation technique defined in 2.2 together with the conditional dependence of RNN-T, which allows the RNN-T candidates to capture more mispronunciation patterns. It is also observed that monophone RNN-T performs better than syllable RNN-T. We suspect that the strong conditional dependence constraints of syllable RNN-T prevent the model from distinguishing individual phonemes. These findings highlight the trade-offs of choosing the right level of autoregressiveness.

Model	FAR	FRR	Pre	Rec	F1	PER
rnnt-phn	**0.375**	0.072	0.601	**0.572**	0.586	15.73
rnnt-syl	0.412	0.095	0.593	0.568	0.580	16.92
cnn-rnn-ctc	0.435	0.132	0.571	0.520	0.544	17.99
base-ctc	0.446	0.058	0.614	0.554	0.583	15.68
xlsr-ctc	0.447	**0.053**	**0.634**	0.553	**0.591**	**15.47**

Table 2. Overall MDD Performance

Next, we took a closer look at the top performing RNN-T and CTC candidates and evaluate the effect of cross-lingual disambiguation. In this experiment, we adopted additional weakly supervised data augmentation techniques in 2.3 to train the phonetic RNN-T candidate. We focus on a scenario of native Spanish speakers learning English. The English and Spanish speech corpus were collected from multilingual Librispeech (MLS) dataset [19] as well as additional code-mixed corpus from an anonymous and de-identified commercial voice assistant traffic. The candidates were benchmarked using only the subset of L2-ARCTIC where the testers are native Spanish speakers.

The results are summarized in Table 3. We observed that with cross-lingual phoneme disambiguation, Monophone RNN-T achieves slightly better F1-score and PER compared to XLSR CTC and 16.8% reduction in FAR. Compared to the same model without the disambiguation module, it also significantly improved FAR and slightly degraded FRR. The results showcase that the proposed cross-lingual phoneme disambiguation mechanism allows the phonetic RNN-T model to better distinguish Spanish-accented phonemes and therefore detect more pronunciation errors (e.g., heavy accents). We further observe worse FRR compared to the CTC model. However, the phonetic RNN-T model achieves a better balance between FAR and FRR.

Candidate	FAR	FRR	Pre	Rec	F1	PER
rnnt-phn	**0.376**	0.071	0.642	**0.594**	**0.617**	**15.41**
w/o disambiguation	0.420	0.068	0.648	0.577	0. 610	15.79
xlsr-ctc	0.452	**0.049**	**0.673**	0.562	0.612	15.55

Table 3. MDD Performance on Native Spanish Speakers

We then summarize the common phoneme-level errors of *rnnt-phn* and *xlsr-ctc* in Table 3.2. In the top half of the table, we summarize the top three hypotheses phonemes for the L1 (English) phoneme set. We can observe that the phoneme errors for both candidates are similar-sounding phonemes, highlighting the effectiveness of both approaches. In the bottom half of the table, we summarize the top hypothesis phonemes for three most frequently mis-recognized Spanish phonemes: "nn" ($\tilde{n}$ sound as in "$niña$"), "gg" (g sound as in "$gente$"), and "rr" (r sound sound as in "$Pero$"). Two additional columns of "r" and "n" sounds from the L1 English phoneme set are also added to study how L1 sounds are mapped to L2 Spanish phonemes. We found the proposed phonetic RNN-T is able to distinguish similar sounding phonemes across L1 and L2 phoneme sets (e.g., "rr" vs "r"). In contrast, the candidate CTC approach either maps the accented Spanish phoneme to a similar English phoneme, causing false acceptance; or fails to recognize it at all (denoted as "-" for deletion error). The results further highlight the effectiveness of the proposed cross-lingual disambiguation mechanism.

Top L1 Phonemes Errors					
Candidates	**Reference Phonemes**				
	aa	z	er	ih	ow
xlsr-ctc	aa (80%)	z (79%)	er (89%)	ih (91%)	ow (93%)
	ax (12%)	s (17%)	ax (10%)	iy (7%)	aa (4%)
	ow (7%)	d (4%)	aa (1%)	ax (2%)	ax (3%)
rnnt-phn	aa (78%)	z (80%)	er (92%)	ih (84%)	ow (91%)
	ax (14%)	s (12%)	ax (5%)	iy (11%)	aa (6%)
	ow (8%)	d (8%)	eh (3%)	aa (5%)	ax (3%)
Top Cross-Lingual Phonemes Errors					
Candidates	**Reference Phonemes**				
	rr	r	nn	n	gg
xlsr-ctc	r (75%)	r (96%)	y (46%)	n (96%)	g (66%)
	aa (18%)	l (3%)	- (29%)	l (3%)	k (28%)
	-* (7%)	e (1%)	n (25%)	ih (1%)	y (6%)
rnnt-phn	**rr (65%)**	r (94%)	**nn (61%)**	n (95%)	**gg (48%)**
	r (22%)	l (5%)	n (28%)	l (3%)	g (37%)
	h (13%)	rr (1%)	y (11%)	y (2%)	k (15%)

* - stands for deletion error (i.e., no phoneme recognition).

Table 4. Comparing Top Errors for English (L1) and Cross-Lingual Phonemes

4. CONCLUSION

This paper presents novel phonetic RNN-T based models for mispronunciation detection and diagnosis. Compared to existing CTC approaches that ignore previous history during phoneme recognition, our autoregressive method can learn and leverage mispronunciation patterns from L2 phoneme sequences. With an extended phoneme set and weakly supervised training corpus composed of massive L1, L2, and code-mixed corpora. Empirical experiments have demonstrated the proposed phonetic RNN-T approach significantly reduces false acceptance rate, while beating CTC-based models in F1-score and PER for native Spanish language learners.

5. REFERENCES

[1] Chesta Agarwal and Pinaki Chakraborty, "A review of tools and techniques for computer aided pronunciation training (capt) in english," *Education and Information Technologies*, vol. 24, no. 6, pp. 3731–3743, 2019.

[2] Sweekar Sudhakara, Manoj Kumar Ramanathi, Chiranjeevi Yarra, and Prasanta Kumar Ghosh, "An improved goodness of pronunciation (gop) measure for pronunciation evaluation with dnn-hmm system considering hmm transition probabilities.," in *INTERSPEECH*, 2019, pp. 954–958.

[3] Vincent Laborde, Thomas Pellegrini, Lionel Fontan, Julie Mauclair, Halima Sahraoui, and Jérôme Farinas, "Pronunciation assessment of japanese learners of french with gop scores and phonetic information," in *Annual conference Interspeech (INTERSPEECH 2016)*, 2016, pp. 2686–2690.

[4] Wai-Kim Leung, Xunying Liu, and Helen Meng, "Cnn-rnn-ctc based end-to-end mispronunciation detection and diagnosis," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 8132–8136.

[5] Daniel Zhang, Ashwinkumar Ganesan, Sarah Campbell, Daniel Korzekwa, and Amazon Alexa AI, "L2-gen: A neural phoneme paraphrasing approach to l2 speech synthesis for mispronunciation diagnosis," *Proc. Interspeech 2022*, pp. 4317–4321, 2022.

[6] Minglin Wu, Kun Li, Wai-Kim Leung, and Helen Meng, "Transformer based end-to-end mispronunciation detection and diagnosis," *Proc. Interspeech 2021*, pp. 3954–3958, 2021.

[7] Daniel Korzekwa, Jaime Lorenzo-Trueba, Thomas Drugman, Shira Calamaro, and Bozena Kostek, "Weakly-supervised word-level pronunciation error detection in non-native english speech," in *Interspeech 2021*, Hynek Hermansky, Honza Cernocký, Lukás Burget, Lori Lamel, Odette Scharenborg, and Petr Motlícek, Eds. 2021, pp. 4408–4412, ISCA.

[8] Hsin-Wei Wang, Bi-Cheng Yan, Hsuan-Sheng Chiu, Yung-Chang Hsu, and Berlin Chen, "Exploring non-autoregressive end-to-end neural modeling for english mispronunciation detection and diagnosis," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 6817–6821.

[9] Xiaoshuo Xu, Yueteng Kang, Songjun Cao, Binghuai Lin, and Long Ma, "Explore wav2vec 2.0 for mispronunciation detection.," in *Interspeech*, 2021, pp. 4428–4432.

[10] William Chan, Navdeep Jaitly, Quoc Le, and Oriol 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)*. IEEE, 2016, pp. 4960–4964.

[11] Marta Nowacka, "Questionnaire-based pronunciation studies: Italian, spanish and polish students' views on their english pronunciation," *Research in Language*, vol. 10, no. 1, pp. 43–61, 2012.

[12] Robert Weide et al., "The carnegie mellon pronouncing dictionary," *release 0.6, www. cs. cmu. edu*, 1998.

[13] John S Garofolo, Lori F Lamel, William M Fisher, Jonathan G Fiscus, and David S Pallett, "Darpa timit acoustic-phonetic continous speech corpus cd-rom. nist speech disc 1-1.1," *NASA STI/Recon technical report n*, vol. 93, pp. 27403, 1993.

[14] Guanlong Zhao, Sinem Sonsaat, Alif Silpachai, Ivana Lucic, Evgeny Chukharev-Hudilainen, John Levis, and Ricardo Gutierrez-Osuna, "L2-arctic: A non-native english speech corpus.," in *INTERSPEECH*, 2018, pp. 2783–2787.

[15] Alex Graves, "Sequence transduction with recurrent neural networks," *arXiv preprint arXiv:1211.3711*, 2012.

[16] San Duanmu, "The revised max onset: Syllabification and stress in english," in *Prosodic Studies*, pp. 61–79. Routledge, 2019.

[17] Ashwin K Vijayakumar, Michael Cogswell, Ramprasath R Selvaraju, Qing Sun, Stefan Lee, David Crandall, and Dhruv Batra, "Diverse beam search: Decoding diverse solutions from neural sequence models," *arXiv preprint arXiv:1610.02424*, 2016.

[18] Jing Liu, Rupak Vignesh Swaminathan, Sree Hari Krishnan Parthasarathi, Chunchuan Lyu, Athanasios Mouchtaris, and Siegfried Kunzmann, "Exploiting large-scale teacher-student training for on-device acoustic models," in *International Conference on Text, Speech, and Dialogue*. Springer, 2021, pp. 413–424.

[19] Vineel Pratap, Qiantong Xu, Anuroop Sriram, Gabriel Synnaeve, and Ronan Collobert, "MLS: A large-scale multilingual dataset for speech research," in *Interspeech 2020*, Helen Meng, Bo Xu, and Thomas Fang Zheng, Eds. 2020, pp. 2757–2761, ISCA.