



Language-Aware and Language-Agnostic Multilingual Speech Recognition with a Single Model

Karol Nowakowski¹ ^a and Michal Ptaszynski² ^b

¹*Tohoku University of Community Service and Science, Sakata, Yamagata, Japan*

²*Kitami Institute of Technology, Kitami, Hokkaido, Japan*

Keywords: Speech Recognition, Multilingual Learning, Adapters, Language Identifiers, Slavic Languages.

Abstract: In recent years, there has been increasing interest in multilingual speech recognition systems, where a single model can transcribe speech in multiple languages. Additional benefit of multilingual learning is that it allows for cross-lingual transfer, often leading to better performance, especially in low-resource languages. On the other hand, multilingual models suffer from errors caused by confusion between languages. This problem can be mitigated by providing the information about language identity as an additional input to the model. In this research, we carry out experiments using a modern state-of-the-art ASR system architecture based on a pretrained multilingual wav2vec 2.0 model and adapter modules trained for the downstream task, and confirm that multilingual supervised learning with language identifiers is a viable method for improving the system's overall performance. Furthermore, we find that training with language identifiers still yields a model with better average performance than the model trained without such information, even if language identity is unknown at inference time.

1 INTRODUCTION


Previous works on automatic speech recognition (Toshniwal et al., 2018; Conneau et al., 2021; Babu et al., 2021), as well as other language processing tasks (Johnson et al., 2017; Conneau et al., 2020), have shown that combining data in multiple languages to train a single model that can support all of them, instead of developing separate monolingual models, not only saves resources, but can also result in better performance, especially for low-resource languages. This is particularly true if the languages used in training are closely related or share common linguistic traits (Conneau et al., 2021; Nowakowski et al., 2023).


In the most basic approach where data in multiple languages is simply pooled together and no additional information is supplied, the model needs to learn to infer the input language identity. This requirement can be removed by incorporating language identifiers (LID) into the system, in the form of a language vector (Li et al., 2018; Toshniwal et al., 2018) or LID prefixes attached to each utterance in

the data (Nowakowski and Ptaszynski, 2023), resulting in improved performance. With the exception of (Nowakowski and Ptaszynski, 2023), previous studies seem to make an assumption – either implicitly or explicitly, as in (Zhou et al., 2022; Houston et al., 2024) – that language identifiers are only useful if used both in training and inference.

In recent years, state-of-the-art multilingual ASR systems are often being built by combining an initial speech recognition model or speech representation model – such as wav2vec 2.0 (Baevski et al., 2020) – pretrained on multilingual data, with language-specific fine-tuning (Kannan et al., 2019; Conneau et al., 2021). Recently, (Pratap et al., 2023) developed a speech recognition model with separate adapter modules (Rebuffi et al., 2017; Houlisby et al., 2019) for more than 1,000 languages. On the other hand, (Nowakowski et al., 2023) and (Nowakowski and Ptaszynski, 2023) demonstrated that multilingual supervised fine-tuning can obtain improved results in a low-resource setting.

In this research, we build a speech recognition model supporting seven Slavic languages by training a single adapter module and using language identifiers. We show that on average the proposed approach achieves lower error rates than monolingual adapters.

^a  <https://orcid.org/0000-0001-7435-4061>

^b  <https://orcid.org/0000-0002-1910-9183>

Furthermore, we investigate the performance of a model trained with LID information on test data without such information, and find that it yields similar or lower error rates than a multilingual model trained in a language-agnostic manner. This means that the same model can be utilized both in scenarios where the input language is explicitly specified and when it is unknown, without loss in quality of the system’s output compared to a setup using two separate models.

The remainder of this paper is organized as follows. Section 2 provides an overview of related research. In Section 3, we explain our research methodology. Section 4 describes the experimental setup. In Section 5, we present and analyze our experimental results. Section 6 discusses the limitations of our study. Finally, Section 7 offers concluding remarks and ideas for future improvements.

2 RELATED WORK

Multilingual neural speech recognition with language identifiers injected into model’s input has been studied, among others, by (Li et al., 2018; Toshniwal et al., 2018; Zhu et al., 2020; Nowakowski and Ptaszynski, 2023).

Alternative methods for utilizing language identity information include using a multi-task learning architecture for jointly recognizing speech and predicting language identity (Toshniwal et al., 2018; Zhang et al., 2022), implementing a language-specific gating mechanism (Kim and Seltzer, 2018), and language-specific attention heads (Zhu et al., 2020).

Recently, and in particular after the paradigm shift in language processing that took place with the advent of pretrained language representation models such as BERT (Devlin et al., 2019) and wav2vec 2.0 (Baevski et al., 2020), the development of multilingual ASR systems has often been performed as a two-stage process, where multilingual pretraining is followed by language-specific fine-tuning (Kannan et al., 2019; Conneau et al., 2021). On the other hand, studies by (Nowakowski et al., 2023; Nowakowski and Ptaszynski, 2023) demonstrated that fine-tuning jointly on multiple languages can lead to improved results in a low-resource setting. In order to avoid catastrophic forgetting and reduce computational cost, fine-tuning is in many cases only applied to a subset of the model layers (Liu et al., 2024) or utilizes adapter modules inserted into the pretrained network (Kannan et al., 2019; Pratap et al., 2023).

Using a combination of multilingual adapter fine-tuning and LID in multilingual ASR was previously

investigated by (Shen et al., 2023). Unlike in our study, they did not observe improvements in overall performance in comparison to monolingual baselines when using a large-sized (>300M parameters) pretrained model.

(Nowakowski and Ptaszynski, 2023) found the positive effects of using language identifiers in model training to persist, even if they are unavailable at inference time. Based on these results, they proposed a hypothesis that “the additional knowledge about the relationships and differences between the languages used in fine-tuning, learned by the agency of the language identifiers, can be to a large extent reused in inference regardless of their presence in the new data”. However, they only experimented with bilingual and trilingual models and performed all the evaluations on a single language.

3 PROPOSED METHOD

Our system is based on the MMS architecture (Pratap et al., 2023), which uses a pretrained wav2vec 2.0 model and language-specific adapter modules trained to transcribe speech. Compared to the original framework proposed in (Pratap et al., 2023), we introduce the following modifications: (i) training a single adapter module jointly on data in multiple languages, and (ii) adding language identity information to the model’s input.

3.1 Multilingual Adapter Training

We investigate the possibility of improving overall ASR performance in a multilingual setting by training a single adapter module for transcribing speech in multiple languages, rather than separate adapters for each language. To that end, we simply pool the labeled data of all languages together to form a single multilingual training set.

3.2 Language-Aware Training

With the aim of reducing the number of errors caused by confusion between languages, we provide the model with the information about language identity (LID). Specifically, we follow (Nowakowski and Ptaszynski, 2023) in including this information directly in the data, in the form of short, artificially generated audio clips (different for each language) prefixed to every audio file in the corpus. The LID prefixes are 25ms long, which corresponds to the receptive field of the wav2vec 2.0 encoder (Baevski et al., 2020).

Table 1: Statistics of the Common Voice data used in our experiments.

	ces	bul	slk	srp	mkd	slv	hsb	total
Data split	Train							
Utterances	20,144	4,849	3,258	1,879	1,686	1,388	808	34,012
Hours	26.6	7.0	3.5	1.5	2.0	1.3	1.5	43.4
Data split	Validation							
Utterances	9,009	2,766	2,588	1,583	1,289	1,232	172	18,639
Hours	11.5	4.3	3.1	1.1	1.5	1.3	0.3	23.1
Data split	Test							
Utterances	9,067	3,201	2,647	1,539	1,097	1,242	444	19,237
Hours	11.6	4.9	3.2	1.4	1.5	1.4	0.8	24.8

In addition to performing experiments with LID information available both in training and evaluation, we test the hypothesis – suggested by (Nowakowski and Ptaszynski, 2023) – that language-aware training can be helpful in multilingual ASR even in a scenario where the input language identity is unknown at inference time.

4 EXPERIMENT SETUP

4.1 Data

We use a subset of the Common Voice Corpus 17.0 (Ardila et al., 2020), obtained through the Hugging Face Datasets library¹ (Lhoest et al., 2021). Specifically, we fine-tune and test our models on the data in seven Slavic languages with varying resource levels: Czech (*ces*), Bulgarian (*bul*), Slovak (*slk*), Serbian (*srp*), Macedonian (*mkd*), Slovene (*slv*), and Upper Sorbian (*hsb*). Table 1 shows the data statistics per language. We use the official train, validation and test splits. We resample the audio data to 16 kHz. Transcriptions are preprocessed by removing punctuation and lowercased.

4.2 Training and Inference

We use a pretrained MMS model checkpoint² as the base for our ASR models and fine-tune it by re-initializing and training the adapter layers and the output layer only, while freezing the rest of the model parameters. We train all the models for 10 epochs, using a batch size of 32. The learning rate is warmed up for the first 5% of updates to a peak of 1e-3, and linearly decayed after that. We evaluate on the validation data every 100 steps and at the end of training,

¹https://huggingface.co/datasets/mozilla-foundation/common_voice_17_0

²<https://huggingface.co/facebook/mms-1b-11107>

and select the best checkpoint based on Character Error Rate. Concerning other hyperparameters, we follow (von Platen, 2023). Each training experiment is run three times with a different random seed in each execution. Inference is performed without a language model. We carry out all experiments using the Hugging Face Transformers library (ver. 4.45.1) (Wolf et al., 2020).

5 RESULTS AND DISCUSSION

Table 2 compares the results obtained by monolingual baseline models and multilingual models trained and tested with LID prefixes in the data. Language-aware multilingual fine-tuning results in lower error rates on 4 out of 7 languages and on average. Furthermore, it yields more stable performance: while the error rates for models fine-tuned on monolingual Serbian (*srp*) and Slovene (*slv*) data exhibit very high variance, for multilingual adapters standard deviation never exceeds ± 1.0 . On the other hand, the results on the two languages with the largest amount of training data – namely, Czech (*ces*) and Bulgarian (*bul*) – are worse, which might suggest that this approach is not beneficial for high-resource languages.

Next, we examine the performance of multilingual models fine-tuned with LID prefixes when applied to data without them, and compare it to baseline multilingual models trained without LID information. The results are shown in the upper two rows of Table 3 (# audio prefixes = 0 and # audio prefixes = 7, respectively). Although removing LID information from the test data leads to a relative increase by 44% in average Character Error Rate, the models trained with language identifiers still perform better on four languages and on average.

We are interested in finding out whether strong performance of the models trained in a language-aware manner on test data without LID prefixes is due to positive impact of providing language iden-

Table 2: Comparison of monolingual baseline models and multilingual models trained and tested on data with language identifiers. We report the mean and standard deviation of the Character Error Rates and Word Error Rates obtained in three executions with different random seeds. Bold font indicates the best results for each language.

Metric	# training languages	ces	bul	slk	srp	mkd	slv	hsb	avg
CER	1	2.2 ± .0	3.5 ± .0	5.0 ± .1	12.9 ± 15.9	3.3 ± .0	31.0 ± 48.2	7.1 ± .0	9.3
CER	7	2.3 ± .0	3.6 ± .1	4.7 ± .1	3.0 ± .5	3.3 ± .0	3.2 ± .1	6.9 ± .0	3.9
WER	1	11.0 ± .1	17.3 ± .1	23.2 ± .5	27.2 ± 24.5	17.1 ± .3	42.7 ± 49.0	33.8 ± .3	24.6
WER	7	11.3 ± .0	18.2 ± .2	21.5 ± .4	9.9 ± .9	17.5 ± .3	14.5 ± .2	33.5 ± .3	18.1

Table 3: Comparison of the results on test data without LID prefixes, obtained by using multilingual models trained with (i) the original data without audio prefixes, (ii) the data modified by prepending a language-specific (LID) audio prefix to each training sample, and (iii) the data modified by adding the same audio prefix to all training samples. We report the mean and standard deviation of the Character Error Rates and Word Error Rates obtained in three executions with different random seeds. Bold font indicates the best results for each language.

Metric	# audio prefixes	ces	bul	slk	srp	mkd	slv	hsb	avg
CER	0	2.4 ± .0	3.7 ± .1	6.1 ± .2	12.9 ± 1.9	3.8 ± .0	6.0 ± .5	7.0 ± .2	6.0
CER	7	2.5 ± .1	3.6 ± .0	6.0 ± .4	10.3 ± 4.1	4.0 ± .3	5.9 ± .8	7.1 ± .1	5.6
CER	1	2.4 ± .0	3.7 ± .1	6.0 ± .2	14.2 ± 2.6	3.9 ± .0	6.3 ± .3	7.7 ± .8	6.3
WER	0	11.7 ± .1	18.4 ± .3	24.5 ± .2	22.2 ± 2.4	18.9 ± .4	17.9 ± .8	34.1 ± .6	21.1
WER	7	12.0 ± .2	18.1 ± .1	24.1 ± .4	19.8 ± 4.9	19.3 ± .3	17.9 ± .8	34.8 ± .7	20.9
WER	1	11.9 ± .1	18.4 ± .4	24.6 ± .4	24.3 ± 3.3	19.0 ± .2	18.8 ± .2	34.9 ± .7	21.7

Table 4: The ratios of out-of-vocabulary character counts in test predictions to the number of test samples for each language. We report the mean and standard deviation obtained by three models trained with different random seeds. Bold font indicates the best results for each language.

# audio prefixes	ces	bul	slk	srp	mkd	slv	hsb	avg
0	.02 ± .00	.03 ± .00	.33 ± .07	1.39 ± .25	.15 ± .02	.79 ± .14	.01 ± .02	.39
7	.02 ± .02	.01 ± .01	.32 ± .13	0.99 ± .53	.24 ± .17	.80 ± .25	.00 ± .00	.34
1	.02 ± .00	.02 ± .00	.30 ± .04	1.58 ± .37	.21 ± .04	.88 ± .11	.32 ± .49	.47

tivity information. An alternative possible explanation is that the audio prefixes might be implicitly regularizing the model and preventing or reducing overfitting to low-data languages. In order to verify this, we carry out an additional experiment where a single audio prefix is used for all training samples, regardless of the language. The results are presented in the bottom row of Table 3 (# audio prefixes = 1). On all of the languages where the models trained with language-specific audio prefixes outperformed the baseline, they also outperform this approach, which indicates that it is indeed the language identity information conveyed by the prefixes that contributes to their solid performance.

Additionally, we verify whether incorporating LID information during training helps mitigate the problem of language confusion in language-agnostic inference. To this end, for each language, we count the number of out-of-vocabulary characters³ in the

test predictions and normalize this value by dividing it by the size of the evaluation set for that language. The results, shown in Table 4, indicate that the model trained with language-specific audio prefixes is less prone to language confusion than the other systems.

The above results seem to corroborate the hypothesis that using explicit language identity information during multilingual training can contribute to learning a better model of speech features with robust decision boundaries between languages, facilitating not only language-aware, but also language-agnostic inference. While, with the exception of Serbian, the differences in error rates in favor of the models trained with LID are not large, and on three out of seven languages they are slightly outperformed by the baseline method, the observation that both approaches offer comparable performance is already important, as it means that a single model – trained in a language-

³We define an *out-of-vocabulary character* as a char-

acter that does not occur in the training data for the target language.

aware manner – can be used both when the input language is explicitly specified, in which case it performs substantially better than the fully language-agnostic counterpart, as well as in a setting where language identity is unknown at inference time and needs to be inferred by the model – which it can do as well as or better than a model trained without LID prefixes. This approach can potentially allow for reductions in the cost of developing and maintaining multilingual ASR models.

6 LIMITATIONS

Since our system architecture is based on a pre-trained wav2vec 2.0 model, its performance and capacity for cross-lingual transfer may be influenced by the amount and characteristics of the pretraining data (particularly the pretraining data in languages being considered in our experiments). This aspect was not analyzed in the present study.

Although our results demonstrate strong performance for multilingual models, the experiments were conducted using very small training datasets. Whether our approach would match the performance of dedicated monolingual models in a high-resource scenario remains an open question.

7 CONCLUSIONS AND FUTURE WORK

We have demonstrated the effectiveness of fine-tuning a multilingual pretrained speech representation model for speech recognition by training a single adapter module jointly on labeled speech data in multiple languages (instead of adding separate adapters for each language) and providing the information about language identity, in the form of language-specific audio prefixes attached to the data. Furthermore, in experiments using evaluation data without language identifiers, the proposed approach yielded better overall performance than models trained without LID prefixes, suggesting that the benefits of language-aware multilingual training can persist even when language identity information is absent during inference.

Due to limited computational resources, in this research we only focused on languages with very small training datasets (namely, less than 50 hours of labeled data). In the future we will investigate whether the observations made in our experiments also hold in a setting with relatively large amounts of data available. Apart from that, we are planning to perform

experiments on a group of unrelated languages. We will also investigate fine-tuning the base model (or a subset of its layers) as an alternative to re-training the adapter layers only.

ACKNOWLEDGEMENTS

This work was supported by JSPS KAKENHI Grant Number JP22K17952.

REFERENCES

- Ardila, R., Branson, M., Davis, K., Henretty, M., Kohler, M., Meyer, J., Morais, R., Saunders, L., Tyers, F. M., and Weber, G. (2020). Common Voice: A Massively-Multilingual Speech Corpus. In *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, pages 4211–4215.
- Babu, A., Wang, C., Tjandra, A., Lakhotia, K., Xu, Q., Goyal, N., Singh, K., von Platen, P., Saraf, Y., Pino, J., Baevski, A., Conneau, A., and Auli, M. (2021). XLS-R: Self-supervised Cross-lingual Speech Representation Learning at Scale. *arXiv*, abs/2111.09296.
- Baevski, A., Zhou, H., Mohamed, A., and Auli, M. (2020). wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. *ArXiv*, abs/2006.11477.
- Conneau, A., Baevski, A., Collobert, R., Mohamed, A., and Auli, M. (2021). Unsupervised Cross-lingual Representation Learning for Speech Recognition. In *Inter-speech*.
- Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. (2020). Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.
- Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., De Laroussilhe, Q., Gesmundo, A., Attariyan, M., and Gelly, S. (2019). Parameter-efficient transfer learning for NLP. In Chaudhuri, K. and Salakhutdinov, R., editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2790–2799. PMLR.

- Houston, B., Sadjadi, O., Hou, Z., Vishnubhotla, S., and Han, K. (2024). Improving multilingual asr robustness to errors in language input. In *Interspeech 2024*.
- Johnson, M., Schuster, M., Le, Q. V., Krikun, M., Wu, Y., Chen, Z., Thorat, N., Viégas, F., Wattenberg, M., Corrado, G., Hughes, M., and Dean, J. (2017). Google’s multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.
- Kannan, A., Datta, A., Sainath, T. N., Weinstein, E., Ramabhadran, B., Wu, Y., Bapna, A., Chen, Z., and Lee, S. (2019). Large-scale multilingual speech recognition with a streaming end-to-end model. In *Interspeech 2019*, pages 2130–2134.
- Kim, S. and Seltzer, M. L. (2018). Towards language-universal end-to-end speech recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, page 4914–4918. IEEE Press.
- Lhoest, Q., Villanova del Moral, A., Jernite, Y., Thakur, A., von Platen, P., Patil, S., Chaumond, J., Drame, M., Plu, J., Tunstall, L., Davison, J., Šaško, M., Chhablani, G., Malik, B., Brandeis, S., Le Scao, T., Sanh, V., Xu, C., Patry, N., McMillan-Major, A., Schmid, P., Gugger, S., Delangue, C., Matussière, T., Debut, L., Bekman, S., Cistac, P., Goehringer, T., Mustar, V., Lagunas, F., Rush, A., and Wolf, T. (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.
- Li, B., Sainath, T. N., Sim, K. C., Bacchiani, M., Weinstein, E., Nguyen, P., Chen, Z., Wu, Y., and Rao, K. (2018). Multi-dialect speech recognition with a single sequence-to-sequence model. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, page 4749–4753. IEEE Press.
- Liu, Y., Yang, X., and Qu, D. (2024). Exploration of whisper fine-tuning strategies for low-resource asr. *EURASIP J. Audio Speech Music Process.*, 2024(1).
- Nowakowski, K. and Ptaszynski, M. (2023). Improving low-resource speech recognition through multilingual fine-tuning with language identifiers and self-training. In Wu, J.-L. and Su, M.-H., editors, *Proceedings of the 35th Conference on Computational Linguistics and Speech Processing (ROCLING 2023)*, pages 63–70, Taipei City, Taiwan. The Association for Computational Linguistics and Chinese Language Processing (ACLCLP).
- Nowakowski, K., Ptaszynski, M., Murasaki, K., and Nieuważny, J. (2023). Adapting multilingual speech representation model for a new, underresourced language through multilingual fine-tuning and continued pretraining. *Information Processing & Management*, 60(2):103148.
- Pratap, V., Tjandra, A., Shi, B., Tomasello, P., Babu, A., Kundu, S., Elkahky, A., Ni, Z., Vyas, A., Fazel-Zarandi, M., Baevski, A., Adi, Y., Zhang, X., Hsu, W.-N., Conneau, A., and Auli, M. (2023). Scaling speech technology to 1,000+ languages.
- Rebuffi, S.-A., Bilen, H., and Vedaldi, A. (2017). Learning multiple visual domains with residual adapters. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.
- Shen, Z., Guo, W., and Gu, B. (2023). Language-universal adapter learning with knowledge distillation for end-to-end multilingual speech recognition.
- Toshniwal, S., Sainath, T. N., Weiss, R. J., Li, B., Moreno, P., Weinstein, E., and Rao, K. (2018). Multilingual speech recognition with a single end-to-end model. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4904–4908.
- von Platen, P. (2023). Fine-tuning mms adapter models for multi-lingual asr. Online: https://huggingface.co/blog/mms_adapters.
- Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T. L., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. M. (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.
- Zhang, C., Li, B., Sainath, T., Strohmaier, T., Mavandadi, S., Chang, S.-Y., and Haghani, P. (2022). Streaming end-to-end multilingual speech recognition with joint language identification. In *Interspeech 2022*, pages 3223–3227.
- Zhou, L., Li, J., Sun, E., and Liu, S. (2022). A configurable multilingual model is all you need to recognize all languages. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6422–6426.
- Zhu, Y., Haghani, P., Tripathi, A., Ramabhadran, B., Farris, B., Xu, H., Lu, H., Sak, H., Leal, I., Gaur, N., Moreno, P. J., and Zhang, Q. (2020). Multilingual speech recognition with self-attention structured parameterization. In *Interspeech*.