Data Selection using Spoken Language Identification for Low-Resource and Zero-Resource Speech Recognition

Jianan Chen*, Chenhui Chu*, Sheng Li† and Tatsuya Kawahara*
* Graduate School of Informatics, Kyoto University, Kyoto, Japan
E-mail: {jichen, kawahara}@sap.ist.i.kyoto-u.ac.jp, chu@nlp.ist.i.kyoto-u.ac.jp
† National Institute of Information and Communications Technology (NICT), Kyoto, Japan
E-mail: sheng.li@nict.go.jp

***Abstract*—Large-scale pre-trained models have become common for Automatic Speech Recognition (ASR) tasks. They utilize large-scale, multilingual datasets to learn acoustic features and then are finetuned on downstream ASR tasks. However, their performance degrades when applied to low-resource and zero-resource languages lacking data. This paper introduces a new data selection method with Spoken Language Identification (SLI) models to bring non-target language speech data into training. With the help of phonetic labels as a universal intermediate representation to link low-resource languages to those with rich resources, we enhance the ASR system's performance on low-resource and even zero-resource languages. We conducted ASR experiments on Marathi, Assamese, and Panjabi with augmented non-target Hindi data in the CommonVoice corpora. The experimental results show that the proposed method can train ASR systems with little target language resource.**

I. Introduction

Multilingual pre-trained models have become common when dealing with multilingual Automatic Speech Recognition (ASR) tasks, because of their ability to transcribe speech into many languages with only one model. These models are usually pre-trained on large-scale datasets containing different languages. The pre-trained models can leverage shared acoustic features and improve downstream task performance [1]–[3]. However, these models still require a lot of annotated data for finetuning to perform well. It is still challenging to transfer them to low-resource or even zero-resource languages not included in the datasets used for pre-training [4].

Compared to the low-resource language, it is easier to get data from languages with rich resources. Several studies have investigated phonetic pre-training on various tasks, revealing that introducing non-target language data can improve performance across languages [5]–[8]. However, despite the success of introducing other languages to improve low-resource language performance, not all the data from other languages are useful for improving performance [9]. With low-quality data, the performance might even degrade. As a result, the challenge remains of how to select the useful non-target language and data to improve the target ASR systems.

Several studies have investigated the similarities between languages [10]–[12]. However, their studies did not include acoustic features, which are essential for training ASR systems. Even if one language is similar to another, the detailed pronunciation and articulation can vary significantly. Several studies used Many studies have investigated the semantic similarities between text sentences and sentences [13]–[15]. None of the studies focused on how similar one speech audio can be to another language.

Motivated by the fact that SLI is a task that classifies single speech audio into its languages, we propose to use SLI models to measure how similar one speech audio is to another language including acoustic features. This will help us to select the optimal non-target language data to assist the training of ASR systems for low-resource, even zero-resource languages. By identifying the language of each speech audio clip, even if different from the target languages, we can ensure the selected data reflects the acoustic features of the target languages to some extent.

Soky et al. (2023) [16] utilized SLI to enhance the performance of low-resource languages. That work involves using target language data from different domains and non-target language data from the same domains to supplement the original target language data, demonstrating that introducing non-target language data can improve the ASR performance of the target language. However, that work used all non-target language data, regardless of the quality of the data. Meanwhile, Ma et al. (2023) [17], and Wang et al. (2023) [18] used SLI to improve code-switch ASR. They used SLI as a router to select the neural network segments. They have shown that SLI can be used to code-switch ASR. However, they did not test the performance of SLI on multilingual ASR.

Samuel et al. (2016) [19] used SLI to cluster similar languages into groups from multilingual datasets. However, they also used all language data for training. Chuangsuwanich et al. (2016) [20] used SLI to select frames of non-target language audio clips that are similar to non-target languages. However, their work trains frame-level speech features for DNNs instead of end-to-end ASR. Since that work focuses on the frame-level features, they might also lose some structured features in languages.

To the best of our knowledge, this work is the first attempt

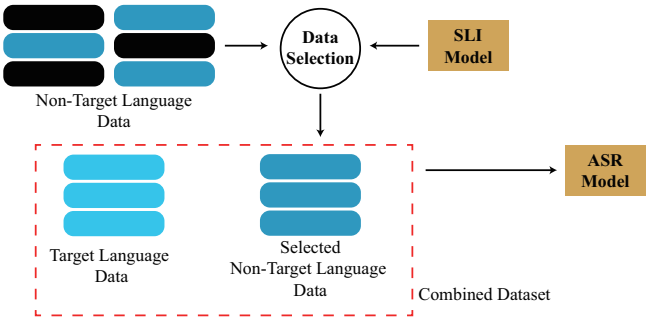


Fig. 1. Illustration of the pipeline to train ASR in this work. We used a data selection module to select non-target language data that sounds like the target language. Added to the original target language data, we obtained a combined dataset to train the ASR system for the target language.

to utilize SLI models to select non-target language data for augmentation in end-to-end ASR training.

The experimental results show that non-target speech data of other languages can benefit the ASR training for low-resource languages. Even though much of the data is filtered out by the proposed SLI selection method, the performance can be better than the baselines using all non-target language data. In addition, it is shown that even without any target language data, we can build ASR systems for zero-resource languages with the help of high-quality non-target language data.

II. Data Selection with Spoken Language Identification

Spoken Language Identification (SLI) is a task that involves calculating the probabilities of the language to which a speech audio belongs. It can be utilized to measure how similar a speech audio clip is to another language. The probabilities given by SLI models can serve as objective metrics to measure how similar a speech audio clip can sound in one specific language. After selecting audio clips similar to the target language, we can use them to supplement the dataset for a low-resource target language.

We propose to use an SLI model to select non-target language data from different languages for the ASR training of the low-resource languages. As shown in Fig. 1, part of the non-target language data are selected with the SLI model, and in addition to the target language data, we obtain an augmented dataset for the following ASR training. We denote this method as **SLI Selection**. The process of selecting data from the original datasets is as follows:

1) **Train SLI model:** Train an SLI model with a multilingual dataset, including the target and the non-target languages.
2) **Calculate the Probabilities of Each Language:** Use the trained SLI model to predict the language of speech audio clips in the multilingual datasets. Here, $\mathbf{x}_i$ is the i-th speech clip in the speech dataset $\mathcal{X}$ = [$\mathbf{x}_1$, $\mathbf{x}_2$ ... $\mathbf{x}_N$], N is the size of the dataset; the probability vector of each language $\mathbf{p}_i^{(langs)}$ is calculated on each $\mathbf{x}_i$.

$$\mathbf{p}_i^{(langs)} = \mathrm{SLI}(\mathbf{x}_i) \tag{1}$$

3) **Select If Target Language in Top-K** Sort the language candidates according to the probabilities. If the target language is in the top-k candidates, the SLI model considers this clip to sound similar to the target language. Thus, we select this clip for the new dataset.
We use the top-k selection because SLI models have high accuracies, causing the probability of the true language to be outstandingly higher than the rest of the probabilities and thus the probabilities of the other languages are not stable and reliable.

III. Experimental Setups

A. Dataset

Four Indian Languages of Hindi, Marathi, Assamese, and Panjabi from the CommonVoice corpora [21] are utilized for training and evaluations. CommonVoice is a large open-source voice dataset built by volunteers all over the world. The dataset is designed especially for the ASR task. It has many releases, and we use the 16.1 release for this work. The detailed information of the data used in this work is shown in Table I.

We fine-tune the XLSR-53 [22] on the downstream ASR task to transcribe Marathi, Assamese, and Panjabi. Hindi, one of India's most spoken languages and geologically near the regions where target languages are spoken, is used only in the training process as an extra language. We then implement the algorithms presented in Section II, selecting audio clips that are similar to the target languages from the Hindi dataset. The models trained with the proposed method are referred to as *top-k*, where k is the rank parameter used for data selection.

In the baseline setups, we build the ASR systems with 1) all the Hindi, 2) all the target language data, or 3) a combination of Hindi and the target language data. The baseline is referred to as *all*. In addition, to eliminate the effect of the quantity of data, we randomly select part of the data from the original Hindi dataset to build an additional baseline. This baseline is referred to as *random*.

We also conduct both few-shot and zero-shot experiments. For the few-shot experiments, we first combine the target language (Marathi Assamese or Panjabi) and the non-target language (Hindi) to obtain a combined training dataset. For the zero-shot experiments, we use the non-target language data and do not use the target language data in the training process.

B. Training SLI and data selection

We used the ECAPA-TDNN architecture [23] provided by SpeechBrain [24] to build the SLI model. The model was trained on VoxLingua107 [25], which is a dataset designed for SLI tasks. The sampling rate of the single-channel speech audio clips is 16kHz.

TABLE I
SPECIFICATIONS OF THE TRAINING DATASETS IN COMMON VOICE USED IN THE EXPERIMENTS.

Language	Audio clips	Duration (hours)
Hindi (hi)	4675	6.7
Marathi (mr)	2215	3.9
Assamese (as)	658	1.2
Panjabi (pa)	733	1.2

C. Training ASR

We utilize the Wav2Vec 2.0 XLS-R 53 as the ASR encoder. The encoder's inputs are 16kHz single-channel speech audio clips, and its outputs are feature embeddings extracted by the encoder. Then, we added a linear projection layer after the encoder and fine-tuned the model with data with a CTC loss. We use the Adam optimizer with a 5e-5 final learning rate and the linear scheduler, setting the batch size to 16 and accumulating the gradient for two steps. The learning rate is set to 5e-5, and we utilize the linear learning rate scheduler. We train the models on the data for at most 200 epochs on 2 48GB A6000 GPUs.

IV. RESULTS AND ANALYSIS

This section presents the results of the experiments. There are mainly three parts of the experiments:

1) **SLI Selection on Hindi:** We first conduct the top-k SLI selections on the Hindi datasets and filter the audio clips similar to the three target languages.
2) **Few-Shot Experiments:** We conduct few-shot end-to-end ASR experiments in this subsection. Both the target language data (Marathi, Assamese, or Panjabi) and the extra non-target language data (Hindi) are utilized as training data.
3) **Zero-Shot Experiments:** We conduct zero-shot end-to-end ASR experiments in this subsection. Only extra non-target language data (Hindi) is utilized as training data.

A. SLI Selection on Hindi

This subsection presents the results of applying SLI selection on the Hindi train dataset in the CommonVoice Dataset. Following is the analysis of the results:

1) Number of audio clips after Selection: We applied the SLI selection on the dataset with k from 1 to 10 to select Hindi subsets that are similar to Marathi, Assamese, and Panjabi pronunciations, respectively. As shown in Table I, the original Hindi train dataset consists of about 4675 audio clips, and the total duration of the dataset is about 6.7 hours.

The number of audio clips after selection is shown in Table II. It can be seen that when k is set at 1, the selected non-target Hindi dataset contains only a small number of audio clips, which is neglectable compared to the original target language dataset size. As a result, we choose to conduct Marathi and Panjabi experiments with k=2 and 5 and Assamese experiments with k=7 and 10 in the following few-shot and zero-shot experiments to make a comparable amount of training data.

2) SLI Selection Filters More Target-like Audio Clips: Fig. 2 presents the probabilities of 19 top languages after data selection. We applied selection to the Hindi dataset and got a new dataset for training Marathi ASR. We selected 15 audio clips and used the SLI model again to predict the possible language IDs of the clips. We omitted the true language ID hi (Hindi) in the probability heatmap because Hindi's probabilities are outstandingly higher than those of the other languages.

Figure (a) presents that the probabilities with random selection in ur (Urdu, an Indic language sounds similar to Hindi, but in a different writing system) are higher than the other languages. This indicates that the SLI model considers the Hindi speech clip to be very similar to Urdu, instead of the target language, Marathi. Meanwhile, Figures (b) and (c) show the probabilities of SLI selection with k equals 3 and 2, respectively. It is observed that with smaller parameter k=2, the proposed method can select more target-like data, showing higher probabilities in mr than ur.

B. Few-Shot Experiments

The few-shot experiments were conducted on three target languages: Marathi, Assamese, and Panjabi. The character error rate is calculated to measure the performance of phonetic label ASR systems. Table III presents the experimental results and following is the analysis:

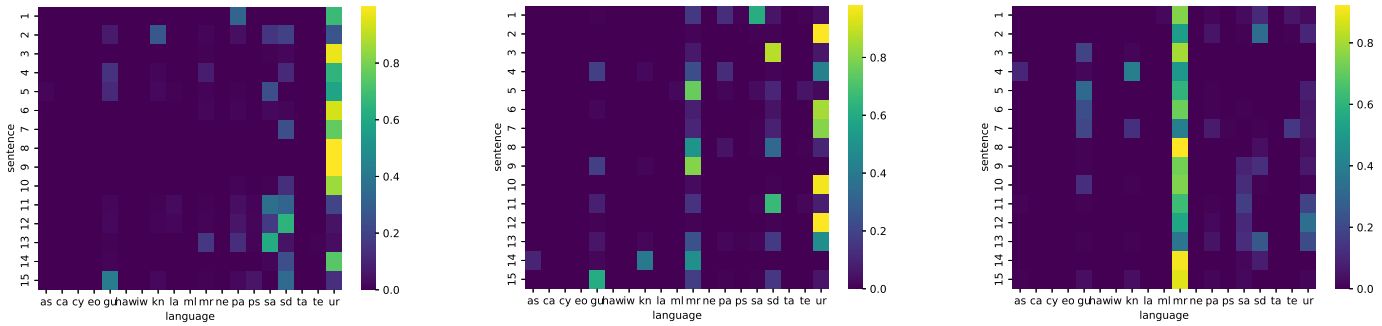
1) Ratio of Target and Non-Target Language: One of our findings is that the ratio of the target and the non-target language data is essential in few-shot experiments. By comparing experiment no. 1 (only target language data) and experiment no. 2 (both target language data and non-target language data are used), we can see that introducing all target language data decreased the performance of the ASR system in the target language. However, by comparing experiment no. 2 (all non-target extra data), experiment no. 4 to 6 (part of the extra data), it can be observed that using fewer extra data might not only perform better than using only target language data but also perform better than using all the non-target language data. In addition, no. 6 (top-2) obtained the best result.

Thus, the ratio of the target and non-target language data is crucial because the non-target language data should not be too many or too few compared to the target language data amount. The same phenomenon can be seen in all three target languages. However, the best ratio of these data remains a problem that needs to be investigated.

2) Quality of Non-Target Language Data: Another finding is that the quality of the extra non-target language data is important. In the experiments, we used two methods to select subsets of the non-target language data for training. One is the random selection, and the other is the SLI selection. Despite the same amount of data for training, the experimental results differ depending on how we selected the data.

TABLE II
NUMBER OF SELECTED AUDIO CLIPS AFTER APPLYING SLI SELECTION ON THE HINDI TRAIN DATASET IN COMMONVOICE. (BOLD FONTS ARE USED FOR ASR)

Top-k	k=1	2	3	4	5	6	7	8	9	10	all
Marathi (mr)	36	**497**	1154	1931	**2729**	3360	3743	4020	4198	4303	4675
Assamese (as)	3	21	67	123	215	319	**480**	692	950	**1224**	4675
Panjabi (pa)	83	**807**	2251	3168	**3670**	4000	4225	4366	4448	4493	4675



(a) random (b) top-3 (c) top-2

Fig. 2. Probabilities of the Hindi data with different data selection methods on 19 languages (as, ca, cy, eo, gu, haw, hi, iw, kn, la, ml, mr, ne, pa, ps, sa, sd, ta, te, ur). (a) is the probabilities for random selection; (b) is for SLI selection with k=3; (c) is with k=2.

In Marathi, the result of experiment no. 6 (top-2) is about 14% better than no. 4 (same amount, randomly selected); In Assamese, no. 11 (top-10) is about 1% better than experiment no. 9 (same amount, ramdonly selected); In Panjabi, no. 18 (top-2) is 5% better than no. 16 (same amount, randomly selected). It can be concluded that even with the same amount of final training data, the data selected with the proposed method is better than random selection in quality.

3) Effect of the Proposed Method: Finally, from Table III, the proposed method achieved the best CER of 11.61 for Marathi by a 2% improvement; the best CER of 19.21 for Assamese by a 3% improvement; and the best CER of 27.14 of Panjabi by a 16% improvement compared to the baselines using only target language data. In addition, the proposed method outperformed all the corresponding baselines with random selection.

These experimental results showed that the proposed data selection with SLI models could select better non-target language data for training the ASR systems, even though the amount of the selected data is much smaller (less than 1 hour).

One possible reason is that introducing low-quality non-target language data might harm the training of ASR systems for target languages. In random selection, not only high-quality data but also low-quality non-target language data are selected. The distribution of the data selected randomly should be similar to the original non-target language dataset. As mentioned before, introducing non-target language data directly might degrade the performance. Meanwhile, as shown in Fig. 2, the data selection method can select data more similar to the target language, helping the ASR models to learn more similar acoustic features.

TABLE III
RESULTS (CER%) OF THE ASR MODELS TRAINED WITH BOTH TARGET (TGT) AND NON-TARGET (EXTRA) SPEECH DATA (LOW-RESOURCE) AND PHONETIC LABEL SCRIPTS. THE BEST RESULTS ARE SHOWN IN BOLD.

Target	**No.**	**Method**	**Duration (Hours)**			**CER%**
			tgt	**extra**	**total**	
Marathi (mr)	1	all	3.9	-	3.9	11.81
	2	all	3.9	6.7	10.6	12.18
	3	random	3.9	3.9	7.8	13.80
	4	random	3.9	0.7	4.6	11.65
	5	top-5	3.9	3.9	7.8	11.90
	6	top-2	3.9	0.7	4.6	**11.61**
Assamese (as)	7	all	1.2	-	1.2	19.82
	8	all	1.2	6.7	7.9	22.00
	9	random	1.2	1.7	2.9	19.42
	10	random	1.2	0.7	1.9	19.80
	11	top-10	1.2	1.7	2.9	**19.21**
	12	top-7	1.2	0.7	1.9	19.71
Panjabi (pa)	13	all	1.2	-	1.2	32.48
	14	all	1.2	6.7	7.9	66.17
	15	random	1.2	5.3	6.5	28.86
	16	random	1.2	1.1	2.3	28.63
	17	top-5	1.2	5.3	6.5	27.28
	18	top-2	1.2	1.1	2.3	**27.14**

4) Limitation of SLI Slection: Although SLI selection with k=2 leads to the best results in Marathi and Panjabi, a larger k was used in Assamese. In Assamese, the best result comes at k=10. This might be because the amount of the similar Hindi data is smaller than the other 2 languages, because the SLI model takes Hindi and Assamese not as similar as the other

two languages. The current results suggest that the optimal amount of non-target language data will be between 1000 and 3000 clips. The best amount and the method to find it remain to be solved for future studies.

C. Zero-Shot Experiments

The zero-shot experiments were conducted on the three target languages. Hindi was used in the training process in the zero-shot experiments. The target languages are only used for the evaluations after the ASR systems are trained with the non-target language data. Table IV presents the experimental results. The table columns are similar to Table III, except that Duration is the non-target Hindi data hours. Following is the analysis for the results:

1) SLI Selection Shows Effectiveness in Zero-Shot:: From Table IV, it can be seen that the proposed method has effect on the zero-shot experiments. Although the amount of data is much less than those in few-shot experiments, the systems can gain improvements on the baselines. In Assamese, no. 9 (top-10) obtained around 60% improvements from no. 7 (same amount, randomly selected). In Panjabi, no. 15 (top-2) obtained around 56% improvements from no. 13 (same amount, randomly selected).

Comparing Table III and Table IV, in Assamese and Panjabi, the proposed method achieved comparable zero-shot results to the few-shot results. This indicates that the proposed method can be utilized to construct ASR systems on zero-resource languages.

2) Limitation of SLI Selection in Zero-Shot: Unfortunately, the proposed SLI selection method did not work consistently in the zero-shot experiments. In the zero-shot Marathi ASR experiments, the best result comes at no. 2 (randomly selected, 3.9 hours of non-target language data) and data selection does not show any effect. Moreover, there is a large gap between few-shot and zero-shot in Marathi. This suggests that the SLI selection is not stable. More stable data selection methods should be studied for future works.

TABLE IV
RESULTS (CER%) OF THE ASR MODELS TRAINED WITH ONLY NON-TARGET SPEECH DATA (ZERO-SHOT) AND PHONETIC LABEL SCRIPTS. THE BEST RESULTS ARE SHOWN IN BOLD.

Target	No.	Method	Duration (Hours)	CER%
Marathi (mr)	1	all	6.7	34.03
	2	random	3.9	**33.15**
	3	random	0.7	98.61
	4	top-5	3.9	35.04
	5	top-2	0.7	33.63
Assamese (as)	6	all	6.7	73.01
	7	random	1.7	51.75
	8	random	0.7	66.72
	9	top-10	1.7	**20.74**
	10	top-7	0.7	22.85
Panjabi (pa)	11	all	6.7	88.60
	12	random	5.3	79.63
	13	random	1.1	75.90
	14	top-5	5.3	36.45
	15	top-2	1.1	**33.13**

V. CONCLUSIONS

This work has proposed a new data selection method with SLI models to select the suitable non-target language data to train ASR systems for target languages to augment low-resource and zero-resource languages. We trained the ASR systems in Marathi, Assamese, and Panjabi with the augmentation of the non-target language, Hindi. We concluded that 1) introducing non-target language data can improve the ASR results for low-resource languages; 2) SLI selection is more effective than the random selection baselines, especially in the zero-shot experiments.

The proposed SLI selection approach can help enhance the usage of non-target languages in training ASR systems. This work paves the way for future studies, particularly in utilizing SLI to select non-target language data for auxiliary training on low-resource languages. Future works can dive into finding the optimal amount non-target language data, which might help improve the data selection process for better ASR performance. Moreover, future works on more stable data selection methods should be conducted.

ACKNOWLEDGMENT

This work was partly supported by JST BOOST, Grant Number JPMJBS2407.

REFERENCES

[1] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "Wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems*, vol. 33, pp. 12 449–12 460, 2020.

[2] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 29, pp. 3451–3460, 2021.

[3] N. Vaessen and D. A. Van Leeuwen, "Fine-tuning wav2vec2 for speaker recognition," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2022, pp. 7967–7971.

[4] C. Yi, J. Wang, N. Cheng, S. Zhou, and B. Xu, "Applying wav2vec2. 0 to speech recognition in various low-resource languages," *arXiv preprint arXiv:2012.12121*, 2020.

[5] C. S. Anoop and A. G. Ramakrishnan, "Exploring a unified asr for multiple south indian languages leveraging multilingual acoustic and language models," in *2022 IEEE Spoken Language Technology Workshop (SLT)*, 2023, pp. 830–837. DOI: 10.1109/SLT54892.2023.10022380.

[6] Y. Qian, K. Yu, and J. Liu, "Combination of data borrowing strategies for low-resource lvcsr," in *2013 IEEE workshop on automatic speech recognition and understanding*, IEEE, 2013, pp. 404–409.

[7] Y. Qian and Z. Zhou, "Optimizing data usage for low-resource speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 394–403, 2022.

[8] Q. Xu, A. Baevski, and M. Auli, "Simple and effective zero-shot cross-lingual phoneme recognition," *arXiv preprint arXiv:2109.11680*, 2021.

[9] Y. Zhang, E. Chuangsuwanich, and J. Glass, "Language id-based training of multilingual stacked bottleneck features," in *Proc. Interspeech*, Citeseer, 2014, pp. 1–5.

[10] V. Beaufils and J. Tomin, "Stochastic approach to world-wide language classification: The signals and the noise towards long-range exploration," 2020.

[11] K. Batsuren, G. Bella, and F. Giunchiglia, "A large and evolving cognate database," *Language Resources and Evaluation*, pp. 1–25, 2022.

[12] J. Eronen, M. Ptaszynski, and F. Masui, "Zero-shot cross-lingual transfer language selection using linguistic similarity," *Information Processing Management*, vol. 60, no. 3, p. 103 250, 2023, ISSN: 0306-4573. DOI: https://doi.org/10.1016/j.ipm.2022.103250. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S030645732200351X.

[13] L. Wang, Y. Ling, Z. Yuan, *et al.*, "Gensim: Generating robotic simulation tasks via large language models," in *Arxiv*, 2023.

[14] A. Joulin, E. Grave, P. Bojanowski, and T. Mikolov, "Bag of tricks for efficient text classification," in *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, Association for Computational Linguistics, Apr. 2017, pp. 427–431.

[15] C. Orašan, "Aggressive language identification using word embeddings and sentiment features," in *Proceedings of the first workshop on trolling, aggression and cyberbullying (TRAC-2018)*, 2018, pp. 113–119.

[16] K. Soky, S. Li, C. Chu, and T. Kawahara, "Domain and language adaptation using heterogeneous datasets for wav2vec2. 0-based speech recognition of low-resource language," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2023, pp. 1–5.

[17] G. Ma, W. Wang, Y. Li, Y. Yang, B. Du, and H. Fu, "Lae-st-moe: Boosted language-aware encoder using speech translation auxiliary task for e2e code-switching asr," in *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, IEEE, 2023, pp. 1–8.

[18] W. Wang, G. Ma, Y. Li, and B. Du, "Language-routing mixture of experts for multilingual and code-switching speech recognition," *arXiv preprint arXiv:2307.05956*, 2023.

[19] S. Thomas, K. Audhkhasi, J. Cui, B. Kingsbury, and B. Ramabhadran, "Multilingual data selection for low resource speech recognition.," in *Interspeech*, 2016, pp. 3853–3857.

[20] E. Chuangsuwanich, Y. Zhang, and J. Glass, "Multilingual data selection for training stacked bottleneck features," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2016, pp. 5410–5414.

[21] R. Ardila, M. Branson, K. Davis, *et al.*, "Common voice: A massively-multilingual speech corpus," English, in *Proceedings of the Twelfth Language Resources and Evaluation Conference*, N. Calzolari, F. Béchet, P. Blache, *et al.*, Eds., Marseille, France: European Language Resources Association, May 2020, pp. 4218–4222, ISBN: 979-10-95546-34-4. [Online]. Available: https://aclanthology.org/2020.lrec-1.520.

[22] A. Babu, C. Wang, A. Tjandra, *et al.*, "Xls-r: Self-supervised cross-lingual speech representation learning at scale," *arXiv preprint arXiv:2111.09296*, 2021.

[23] B. Desplanques, J. Thienpondt, and K. Demuynck, "Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification," *ArXiv*, vol. abs/2005.07143, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:218630075.

[24] M. Ravanelli, T. Parcollet, and P. Plantinga, *SpeechBrain: A general-purpose speech toolkit*, arXiv:2106.04624, 2021. arXiv: 2106.04624 [eess.AS].

[25] J. Valk and T. Alumäe, "VoxLingua107: A dataset for spoken language recognition," in *Proc. IEEE SLT Workshop*, 2021.