



Ant Multilingual Recognition System for OLR 2021 Challenge

Anqi Lyu, Zhiming Wang, Huijia Zhu

Tiansuan Lab, Security BG, Ant Group
Shanghai, China

{lyuanqi.laq, zhiming.wang, huijia.zhj}@antgroup.com

Abstract

This paper presents a comprehensive description of the Ant multilingual recognition system for the 6th Oriental Language Recognition(OLR 2021) Challenge. Inspired by the transfer learning scheme, the encoder components of language identification(LID) model is initialized from pretrained automatic speech recognition(ASR) networks for integrating the lexical phonetic information into language identification. The ASR model is encoder-decoder networks based on U2++ architecture [1]; then inheriting the shared conformer encoder [2] from pretrained ASR model which is effective at global information capturing and local invariance modeling, the LID model, with an attentive statistical pooling layer and a following linear projection layer added on the encoder, is further finetuned until its optimum. Furthermore, data augmentation, score normalization and model ensemble are good strategies to improve performance indicators, which are investigated and analysed in detail within our paper. In the OLR 2021 Challenge, our submitted systems ranked the top in both tasks 1 and 2 with primary metrics of 0.0025 and 0.0039 respectively, less than 1/3 of the second place¹, which fully illustrates that our methodologies for multilingual identification are effectual and competitive in real-life scenarios.

Index Terms: multilingual identification; language recognition; OLR 2021; conformer; transfer learning

1. Introduction

As globalization is becoming a trend and multilingual communication is common in real-life scenarios, multilingual speech technologies are increasingly important and gain a lot of attention from research and industrial communities. Among them, language identification is one of the crucial tasks that aims to predict the language category of the given utterance. Sometimes, language identification is a prepositive module of many commercial ASR decoders, which choose appropriate multilingual ASR services in the light of the predicted language category for the inputting audio stream.

The Oriental Language Recognition(OLR) Challenge is organized annually to encourage and promote in-depth study in the field of multilingual speech. The challenge in 2021 [3] includes four tasks: (1) constrained LID on close domain data, only the data provided by the organizer can be used with the exception of non-speech data; (2) unconstrained LID on wild data, any accessible data (except evaluation data) is allowed; (3) constrained multilingual ASR; (4) unconstrained multilingual ASR. Our team has participated in tasks 1 and 2, which both involve multilingual identification².

¹http://csit.riit.tsinghua.edu.cn/mediawiki/index.php/OLR_Challenge_2021, here is OLR 2021 Challenge official website.

²Our team name is X-Voice, which means to explore the unknown voice world.

There have been intensive studies for language recognition. Some state-of-the-art researches adopt the paradigm of speaker verification that incorporates i-vector [4] or deep neural networks (DNN) based x-vector [5] frameworks. For example, E-TDNN [6] was proposed to interleave dense and TDNN layers, and ECAPA-TDNN [7] was to integrate Res2Net and Squeeze-and-Excitation blocks, both resulting in enhanced performance. Apart from that, the methods of transfer learning and unsupervised learning are investigated a lot for LID task and proved to be effective: for the former, [8] initialized the LID network model by pretraining as one ASR task; for the latter, [9] utilized unsupervised pretrained models, by way of CPC [10], APC [11] or wav2vec series [12], to provide informative speech representations for the downstream LID recognition task.

Inspired by the transfer learning scheme, we believe that ASR pretraining can integrate the lexical phonetic information that helps to strengthen language identification performance. To be specific, we adopt encoder-decoder network model based on U2++ architecture [1] at the ASR pretraining stage; then based on the shared conformer encoder [2] which is inherited from pretrained ASR model and verified to be effective at global information capture and local invariance modeling, the LID model is further finetuned until its optimum; back-end scoring methods and model fusion strategies are fully explored to gain improved performance. Coupled with diversified data augmentation, our proposed systems ultimately won the first place in both tasks 1 and 2, solidly confirming their superiority at cross domain and wild scenarios.

This paper presents a comprehensive description of the Ant multilingual recognition system for the OLR 2021 Challenge. The subsequent parts are organized as follows: in Section 2, we describe the data information for the first two tasks, especially for the unconstrained task; Section 3 illustrates our multilingual identification system detailedly; in Section 4 are experimental settings and results; conclusion is given in Section 5.

2. Data Preparation

2.1. Data Profile

The OLR 2021 Challenge [3] provides the following datasets for model training: OLR16-OL7, OLR17-OL3, OLR17-dev, OLR17-test, OLR18-test, OLR19-dev, OLR19-test, OLR20-dialect, and OLR20-test, which involve up to 280 hours of recordings in 17 different languages.

For the 1st task, 13 target languages(i.e., Indonesian, Japanese, Russian, Korean, Vietnamese, Mandarin, Cantonese, Sichuanese, Shanghainese, Hokkien, Tibetan, Kazakh and Uyghur) are considered, and the corresponding audio samples are picked from the OLR datasets mentioned above.

For task 2, all languages except for Cantonese in task 1 are included, and there are five additional languages of interest: Thai, Malay, Telugu, Hindi, English. Since there is no data

constraint for this task, we bring in extra open-source datasets to strengthen diversities, which include VoxLingua107 [13], OpenSLR³, CommonVoice⁴, Librispeech [14], WenetSpeech [15]; full details are given in Appendix A.1. Some of them, such as WenetSpeech and Librispeech, are of large size for a single language; out of consideration of data balance, only part of their corresponding data are sampled at random and employed in our experiments.

For both tasks, data are randomly divided into training and development sets. With the help of given speaker information, there is no speaker overlap in the split for a better cross-validation indicator of model’s performance [16]. In addition, data with transcriptions in the two separate datasets are further selected to train and evaluate the ASR system.

2.2. Data Augmentation

To improve model’s robustness, four strategies of data augmentation are performed:

- Speed perturbation, with speed factors of 0.9, 1.0, 1.1;
- Mixture of noise from MUSAN(that is SLR17³), and following the challenge’s data protocol, only non-speech noises are used in task 1;
- Reverberation injection, using simulated room impulse responses from SLR28³;
- SpecAugment [17], with one frequency mask in [0, 10], and one time mask in [0, 5].

Except for the differences in data constitution, both tasks of 1 and 2 share the same system framework as discussed in Section 3.

3. System Description

3.1. Input Features

Audios are sampled at 16,000 Hz. 80-dimensional logarithm Mel filter banks are generated within a 25ms sliding window using a hop step size of 10ms; and cepstral mean normalization(CMN) is performed within a 3-second sliding window. No voice activity detection(VAD) is used.

3.2. Backbone Model

Similarly as proposed in [8], the training procedure is divided into two stages: first, an end-to-end multilingual encoder-decoder ASR network model is pretrained to integrate lexical phonetic information; then inheriting the shared encoder components from the pretrained ASR network model, the LID classifier is further finetuned to achieve the optimal.

3.2.1. Pretrain encoder-decoder ASR network model

The encoder-decoder ASR network model based on U2++ architecture [1] is shown in the right part of Figure 1, which consist of a shared encoder, and two attention decoders including *left-to-right* and *right-to-left* ones respectively with the forward and backward text sequence information. In comparison with U2 architecture [18] using only one *left-to-right* decoder, U2++ architecture demonstrates better decoding performance, even in CTC prefix beam search mode. The configurations of the ASR network model are as follows: the shared encoder is comprised

of 4-factor subsampling layers and 12 conformer blocks [2], and each attention decoder is of 3 transformer blocks [19]. The ASR model is trained in an end-to-end way with joint CTC and attention loss using the Wenet toolkit [20], as in Eq.(1),

$$L = \lambda L_{ctc} + (1 - \lambda)(1 - \alpha)L_{att-l2r} + (1 - \lambda)\alpha L_{att-r2l}, \quad (1)$$

where CTC weight $\lambda = 0.3$ and *right-to-left* attention weight $\alpha = 0.3$ are empirically tuned.

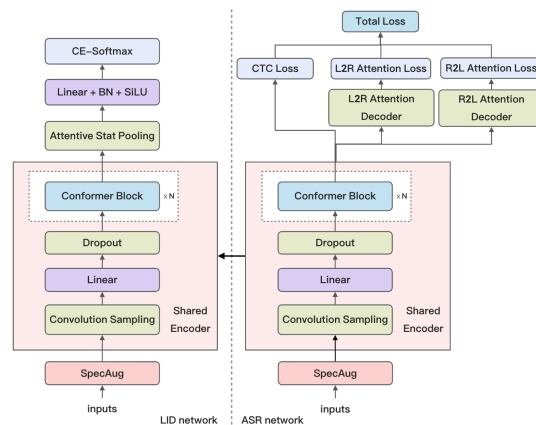


Figure 1: Transfer learning scheme. The right part is the encoder-decoder ASR network model based on U2++ architecture [1]; the left part is the LID system, with shared encoder from the pretrained ASR network model.

For each language, a sentencepiece [21] model is trained to define respective modeling units; then all of them are combined together to constitute an overall dictionary. For most of languages, the number of tokens is less than 2000, while for some languages such as Mandarin, more than 2000 tokens is allowed due to their large and various character space.

3.2.2. Finetune LID classifier

At the finetuning stage, on top of the shared conformer encoder from the pretrained ASR network model, an attentive statistical pooling layer [22] is used to map frame-level representations into segment-level vector representation, which is then projected by a linear layer with batch normalization and non-linear activation to the utterance-level embedding that serves as multilingual identification feature. Following that is a language classifier. The LID model is finetuned with cross entropy(CE) softmax loss; other loss function like AAM-Softmax [23] is an alternative. The LID system is depicted in the left of Figure 1.

3.3. Scoring Methods

The utterance embeddings of original training data and augmented audio samples by the front three strategies in 2.2 are computed for enrollment sets. For more discriminative embedding features, the dimension of the embeddings is reduced to a fixed size with Linear Discriminative Analysis(LDA) method; we find the number of language categories minus 1 is the best choice for LDA dimensional size, may attributed to the orthogonal low dimensional subspace maximizing inter-class separability. Dimension reduced features are averaged into one enrollment embedding vector for each language.

After LDA projection and averaging, cosine similarity was computed as the score of each trial. Logistic regression(LR)

³<http://www.openslr.org>.

⁴<https://commonvoice.mozilla.org/zh-CN>.

was also considered which takes dimension reduced features as input, and similar performances were observed. Since cosine similarity is much more simpler than LR prediction, the former is first chosen to compute scores. As for score normalization, simple min-max normalization as in Eq.(2) is performed independently for each score vector x , which comprises scores for one test utterance with all enrollment languages.

$$\hat{x} = \frac{x - \min_i(x_i)}{\max_i(x_i) - \min_i(x_i)}. \quad (2)$$

4. Experimental Settings and Results

4.1. Experimental Settings

For the shared 12-block conformer encoder of ASR and LID models, output dimension of each block is 256, linear dimension is 2048, and the number of attention heads is 4. For LID model, the attentive pooling layer is 1536-dimensional multi-layer perceptron(MLP), and the following projection layer has 400 hidden units.

Both ASR and LID models are trained with ADAM optimizer. The learning rate schedule follows that from transformer [19] with 25,000 warmup steps, and the peak learning rate 1e-3 for ASR pretraining, 1e-4 for LID finetuning. For the LID training task, as in [24], randomly selected T frames of audio segments are taken as a mini-batch at each iteration; to be specific, $T \sim Uniform(200, 400)$ in general unless otherwise specified.

4.2. Main Results

As in [3], we adopt average cost function(C_{avg}) as primary metric, and equal error rate(EER) as auxiliary one; smaller values of them correspond to better performances.

4.2.1. Performances of a single model

For fair comparison, the baseline E-TDNN x-vector model released by the organizer⁵ and our proposed conformer-based LID model are both trained from scratch, and their performances are reported in Table 1. As is shown here, a conformer-based LID model outperforms the baseline by a large margin, which means conformer architecture is competitive in language recognition task. It is believed that conformer block is good at capturing global information by attention mechanism, and also benefits from local invariance modeling by convolution.

As indicated in Table 1, when pretrained as an ASR task and then finetuned, the performance is further improved, which confirms the assumption that the phonetic information integrated by ASR task helps to enhance language identification. As C_{avg} and EER decrease to 0.003, 0.3015% on the development set, the proposed conformer-based LID model with ASR pretraining might already surpasses many ensemble systems, in reference to the overall performance in OLR 2020 Challenge [25]⁶.

Table 1: Results of conformer-based LID model w or w/o ASR pretraining in Task 1

Models	Dev	
	C_{avg}	EER(%)
Baseline	0.0608	5.8030
Conformer model from scratch	0.0165	1.6620
Conformer model with ASR pretraining	0.0030	0.3015

4.2.2. Fusion and ensemble

For a single system, the weights of the top K (i.e., $K = 5$) checkpoints with lower validation losses on the development datasets are averaged into those of one model, which avoids local fluctuation and brings in better generalization. For multi-systems, the output scores from different networks are linearly weighted into a regression value, and the optimal weights are tuned on the development datasets with grid search method. We observe that, with some techniques, ensemble model always leads to the better performance for mutual complementation, which is also our final submission to the challenge.

To be specific, for task 1, four top performing conformer-based LID models are selected: A. trained with cross entropy softmax loss; B. trained with AAM-Softmax [23] loss; C. based on A, with additive babble noises from training datasets; D. based on A, but $T \sim Uniform(300, 400)$. Their independent evaluations and fusion performances are shown in Table 2. Our final submission to the evaluation set achieved C_{avg} 0.0025 and EER 0.2708 %, ranking the top.

Table 2: Model’s independent and fusion evaluations in Task 1

Models	Weights	Dev		Eval	
		C_{avg}	EER(%)	C_{avg}	EER(%)
A	0.10	0.0026	0.2546	-	-
B	0.45	0.0024	0.2479	-	-
C	0.15	0.0021	0.2211	-	-
D	0.30	0.0021	0.2144	-	-
Fusion	-	0.0018	0.1809	0.0025	0.2708
Baseline	-	-	-	0.0817	8.9770

For task 2, two different conformer-based LID models are chosen based on the above A and B methods respectively. Their independent and fusion evaluations are reported in Table 3. In the leader-board for the evaluation set, our submitted fusion system achieved C_{avg} 0.0039 and EER 0.4212 %, ranking the first as well, which confirms the robustness of our proposed system in real-life scenarios⁶.

Table 3: Model’s independent and fusion evaluations in Task 2

Models	Weights	Dev		Eval	
		C_{avg}	EER(%)	C_{avg}	EER(%)
A	0.4	0.0058	0.8298	-	-
B	0.6	0.0058	0.7587	-	-
Fusion	-	0.0053	0.7208	0.0039	0.4212

4.3. Ablation Analysis

4.3.1. Impact of scoring methods

There are two mainstream approaches to compute confidence scores for each trial: simply take regression logits of LID model as the final result, or adopt the paradigm of speaker recognition which is inclusive of extracting utterance-level embedding, dimension reduction, calculating score, etc. As listed in Table 4,

⁵<https://github.com/Snowdar/asv-subtools/tree/master/recipe/olr2021-baseline>.

⁶According to the official website¹, in both tasks 1 and 2, the primary metric C_{avg} of our submitted system is less than 1/3 of the second place. Additionally, referring to Table 2 and 3 where fused models have minor improvement over single ones, we conclude that our proposed conformer-based single model is superior to other ensemble systems.

the latter is more flexible and achieves better results with great potentials.

To be specific, we find that LDA dimensional size has great impacts on model performance; for example, when LDA projection dimensional size is 50 or 100, worse performances are observed. Unlike the speaker recognition system that needs to deal with uncertain number of speakers, the number of language categories minus 1 is optimal for LDA projection in the close-set LID task. Furthermore, augmented enrollment datasets and min-max score normalization could also bring in performance improvement to varying degrees. In addition, there is no big difference in performances between cosine similarity and logistic regression, also as demonstrated in Table 4.

Table 4: Performances of different scoring methods in Task 1

Scoring Settings	Dev	
	C_{avg}	EER(%)
LID output logits only	0.0036	0.3552
LID embeddings with postprocessing		
LDA dim = 100, Cosine	0.0111	1.3940
LDA dim = 50, Cosine	0.0042	0.4289
LDA dim = 12, LR	0.0029	0.2881
LDA dim = 12, Cosine	0.0028	0.2814
+ augmented enrollment datasets	0.0026	0.2680
+ min-max normalization	0.0026	0.2546

4.3.2. Impact of data augmentation

According to Section 2.2, there are four different strategies of data augmentation in our experiments. As reported in Table 5, each of them respectively contributes to the boosted performances; especially, the mixture of noise and reverberation injection have huge impacts on the evaluation set, which may be due to their compensation for different recording environments of the evaluation datasets. Their aggregation achieves the best generalization performance in the end, though not optimal on the development set.

Table 5: Performances of data augmentation in Task 1

Augmentation	Dev		Eval	
	C_{avg}	EER(%)	C_{avg}	EER(%)
No Aug	0.0040	0.3954	0.0272	3.5050
Speed	0.0024	0.2546	0.0223	2.6630
Noise	0.0035	0.3552	0.0066	0.6664
Reverb	0.0026	0.2680	0.0099	1.0500
SpecAug	0.0026	0.2680	0.0204	2.4620
All	0.0028	0.2814	0.0043	0.4534

4.3.3. Relationship between LID and ASR performances

To investigate the influence of ASR performance on the LID task, three ASR models with the same model structure but different training approaches are chosen; e.g., A. employing all samples from four augmentation strategies; B. based on A, removing audios from babble noise; C. using augmented data just from speed perturbation and SpecAugment. For the sake of fairness, the corresponding LID models are further finetuned under the same training condition. Table 6 shows that pretrained ASR model with lower character error rate(CER) leads to better LID performances; however, with performance progress of the ASR model which focuses more on recognizing harder characters, the marginal gain of the LID task would be reduced.

Table 6: Influence of ASR performance on LID task 1

Models	ASR-Dev	LID-Dev	
	CER(%)	C_{avg}	EER(%)
A	28.3	0.0046	0.4758
B	25.3	0.0034	0.3351
C	24.5	0.0028	0.2814

4.3.4. Confusion matrix

Based on the released evaluation dataset of Task 1, we draw the confusion matrix of our submitted system as in Figure 2, where languages with fairly less errors are removed. It is seen that misidentified cases are mainly dominated by Chinese dialects, especially between Hokkien and Cantonese, due to their highly similarity of pronunciation and shared vocabulary. This indicates that dialect identification is still a challenging task worthy of more investigations, e.g., a promising research orientation is non-overlapping token units among different dialects.

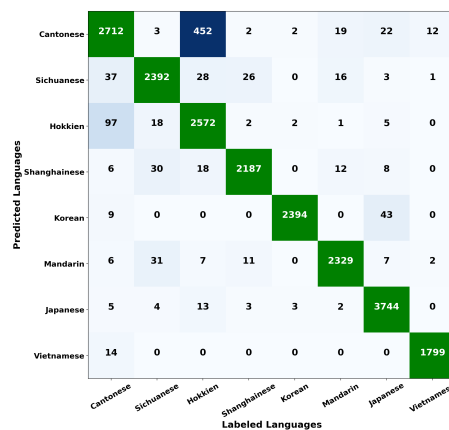


Figure 2: Confusion matrix of our submitted system on the evaluation dataset of Task 1, with the counted number of recognized cases in the box.

5. Conclusion

In this paper, we describe the Ant multilingual recognition system for the OLR 2021 challenge. A conformer-block based LID model is introduced, and its encoder components are initialized from pretrained ASR networks by transfer learning scheme. We verify that conformer architecture is powerful and effective at global information capturing and local invariance modeling, in both ASR and LID tasks; and the pretrained ASR shared encoder would integrate lexical phonetic information, greatly improving performance of language identification. Additionally, scoring methods are useful tricks for better performance: we bring in simple min-max score normalization for the first time in LID task; empirically tuned LDA dimensional size is the number of language categories minor 1, otherwise leading to notable performance degradation. Data augmentation and model ensemble are also conducted in our experiments for improved performance indicators. Ultimately in the challenge, our systems surpassed other opponents a lot and won the first in both tasks 1 and 2. By means of analysing the confusion matrix, we reveal that dialect identification is still a challenging task which deserves more research in the future.

6. References

- [1] D. Wu, B. Zhang, C. Yang, Z. Peng, W. Xia, X. Chen, and X. Lei, “U2++: unified two-pass bidirectional end-to-end model for speech recognition,” *CoRR*, vol. abs/2106.05642, 2021. [Online]. Available: <https://arxiv.org/abs/2106.05642>
- [2] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu *et al.*, “Conformer: Convolution-augmented transformer for speech recognition,” *arXiv preprint arXiv:2005.08100*, 2020.
- [3] B. Wang, W. Hu, J. Li, Y. Zhi, Z. Li, Q. Hong, L. Li, D. Wang, L. Song, and C. Yang, “Olr 2021 challenge: Datasets, rules and baselines,” *arXiv preprint arXiv:2107.11113*, 2021.
- [4] N. Dehak, P. A. Torres-Carrasquillo, D. Reynolds, and R. Dehak, “Language recognition via i-vectors and dimensionality reduction,” in *Proc. Interspeech 2011*, 2011, pp. 857–860.
- [5] D. Snyder, D. Garcia-Romero, A. McCree, G. Sell, D. Povey, and S. Khudanpur, “Spoken Language Recognition using X-vectors,” in *Proc. The Speaker and Language Recognition Workshop (Odyssey 2018)*, 2018, pp. 105–111.
- [6] D. Snyder, D. Garcia-Romero, G. Sell, A. McCree, D. Povey, and S. Khudanpur, “Speaker recognition for multi-speaker conversations using x-vectors,” in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 5796–5800.
- [7] B. Desplanques, J. Thienpondt, and K. Demuynck, “ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification,” in *Proc. Interspeech 2020*, 2020, pp. 3830–3834.
- [8] D. Wang, S. Ye, X. Hu, S. Li, and X. Xu, “An End-to-End Dialect Identification System with Transfer Learning from a Multilingual Automatic Speech Recognition Model,” in *Proc. Interspeech 2021*, 2021, pp. 3266–3270.
- [9] H. Yu, J. Zhao, S. Yang, Z. Wu, Y. Nie, and W.-Q. Zhang, “Language Recognition Based on Unsupervised Pretrained Models,” in *Proc. Interspeech 2021*, 2021, pp. 3271–3275.
- [10] A. van den Oord, Y. Li, and O. Vinyals, “Representation learning with contrastive predictive coding,” *CoRR*, vol. abs/1807.03748, 2018. [Online]. Available: <http://arxiv.org/abs/1807.03748>
- [11] Y.-A. Chung, W.-N. Hsu, H. Tang, and J. Glass, “An Unsupervised Autoregressive Model for Speech Representation Learning,” in *Proc. Interspeech 2019*, 2019, pp. 146–150.
- [12] S. Schneider, A. Baevski, R. Collobert, and M. Auli, “wav2vec: Unsupervised Pre-Training for Speech Recognition,” in *Proc. Interspeech 2019*, 2019, pp. 3465–3469.
- [13] J. Valk and T. Alumäe, “Voxlingua107: a dataset for spoken language recognition,” in *2021 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2021, pp. 652–658.
- [14] 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)*. IEEE, 2015, pp. 5206–5210.
- [15] B. Zhang, H. Lv, P. Guo, Q. Shao, C. Yang, L. Xie, X. Xu, H. Bu, X. Chen, C. Zeng *et al.*, “Wenetspeech: A 10000+ hours multi-domain mandarin corpus for speech recognition,” *arXiv preprint arXiv:2110.03370*, 2021.
- [16] R. Durosele, M. Sahidullah, D. Jovet, and I. Illina, “Language Recognition on Unknown Conditions: The LORIA-Inria-MULTISPEECH System for AP20-OLR Challenge,” in *Proc. Interspeech 2021*, 2021, pp. 3256–3260.
- [17] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, “SpecAugment: A simple data augmentation method for automatic speech recognition,” *arXiv preprint arXiv:1904.08779*, 2019.
- [18] B. Zhang, D. Wu, Z. Yao, X. Wang, F. Yu, C. Yang, L. Guo, Y. Hu, L. Xie, and X. Lei, “Unified streaming and non-streaming two-pass end-to-end model for speech recognition,” *CoRR*, vol. abs/2012.05481, 2020. [Online]. Available: <https://arxiv.org/abs/2012.05481>
- [19] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention is all you need,” in *Advances in neural information processing systems*, 2017, pp. 5998–6008.
- [20] B. Zhang, D. Wu, C. Yang, X. Chen, Z. Peng, X. Wang, Z. Yao, X. Wang, F. Yu, L. Xie *et al.*, “Wenet: Production first and production ready end-to-end speech recognition toolkit,” *arXiv preprint arXiv:2102.01547*, 2021.
- [21] T. Kudo and J. Richardson, “Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing,” *arXiv preprint arXiv:1808.06226*, 2018.
- [22] K. Okabe, T. Koshinaka, and K. Shinoda, “Attentive statistics pooling for deep speaker embedding,” *arXiv preprint arXiv:1803.10963*, 2018.
- [23] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, “Arcface: Additive angular margin loss for deep face recognition,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4690–4699.
- [24] Z. Wang, F. Xu, K. Yao, Y. Cheng, T. Xiong, and H. Zhu, “Antvoice neural speaker embedding system for ffsvc 2020,” *Proc. Interspeech 2021*, pp. 1069–1073, 2021.
- [25] J. Li, B. Wang, Y. Zhi, Z. Li, L. Li, Q. Hong, and D. Wang, “Oriental Language Recognition (OLR) 2020: Summary and Analysis,” in *Proc. Interspeech 2021*, 2021, pp. 3251–3255.

A. Appendix

A.1. Extra datasets used in Task 2

Full details of third-party open-source datasets used in Task 2 are summarized in Table 7.

Table 7: Extra datasets used in Task 2

Data Source	Included Languages	Hours
VoxLingua107	13 languages: Thai, Malay, Indonesian, Japanese, Russian, Korean, Vietnamese, Telugu, Hindi, English, Kazakh, Tibetan, Mandarin	884
OpenSLR	6 languages: Uyghur, Korean, Malay, Telugu, Hindi and Kazakh (that is, SLR22, SLR40, SLR63, SLR66, SLR97, SLR102 and SLR103 are covered)	382
CommonVoice	7 languages: Indonesian, Hindi, Japanese, Russian, Thai, Uyghur and Vietnamese	497
Librispeech	English	960
WenetSpeech	Mandarin	10005