INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

END-to-END Cross-Lingual Spoken Language Understanding Model with Multilingual Pretraining

Xianwei Zhang[1], *Liang He*[1,2*]
[1]Department of Electronic Engineering, Tsinghua University, China
[2]Xinjiang University, China
zhangxw019@163.com, heliang@tsinghua.edu.cn

Abstract

The spoken language understanding (SLU) plays an essential role in the field of human-computer interaction. Most of the current SLU systems are cascade systems of automatic speech recognition (ASR) and natural language understanding (NLU). Error propagation and scarcity of annotated speech data are two common difficulties for resource-poor languages. To solve them, we propose a simple but effective end-to-end cross-lingual spoken language understanding model based on XLSR-53, which is a pretrained model in 53 languages by the Facebook research team. The end-to-end approach avoids error propagation and the multilingual pretraining reduces data annotation requirements. Our proposed method achieves 99.71% on the Fluent Speech Commands (FSC) English database and 79.89% on the CATSLU-MAP Chinese database, in intent classification accuracy. To the best of our knowledge, the former is the reported best result on the FSC database.

Index Terms: end-to-end, spoken language understanding, multilingual pretraining, cross-lingual transfer learning

1. Introduction

Spoken language understanding (SLU) is an important part of Human-Computer dialog systems [1]. For example, I say to the machine *"Turn the kitchen lights on"*. This sentence can be parsed into {*action:"activate", object:"lights", location: "kitchen"*}. The machine uses these slot values to match instructions and make responses to people.

Machines are better at analyzing fixed commands and symbols. Compared with text, the speech signal is complex and variable. Hence current SLU systems generally convert speech to text through automatic speech recognition (ASR), and then use text-based natural language understanding (NLU) technology for intent detection and slot filling [2, 3]. The pipeline will pass the ASR recognition error into the following NLU subtask. And using text for intent detection will lose the speaker's tone and emotional information [4, 5].

In the real world, we often encounter the situation where it is difficult to collect enough speech for low-resource languages. As transcribing and annotating the speech is expensive and time-consuming, the lack of annotated data for the target language can't allow us to get an effective ASR model [6]. The cross-lingual transfer learning of SLU has been paid a lot of attention [7, 8, 9, 10]. The typical approach is to train a model on available high-resource languages and then fine-tune the model on the low-resource languages aiming at the SLU task [7, 8, 9, 10]. With cross-lingual pretraining has been verified effective, models can jointly learn the latent semantic information shared across languages [11, 12, 13]. These released cross-lingual pretraining models boost researches in low-resource speech understanding.

Recently, Conneau *et al.* [13] released model (XLSR-53) which learns cross-lingual speech representations in 53 languages. The model has achieved amazing results on the ASR task. In this paper, we propose a two-part model for the SLU task. One is XLSR-53 used to extract multilingual universal semantic information, and the other is the intent classification module, which classifies the extracted information for intents. This model achieves a accuracy of 99.71% on the English dataset FSC. To the best of our knowledge, this is the highest accuracy on FSC. Then the model is transferred to the Chinese SLU task and also get better performance than baseline. We explore several ways of fine-tuning XLSR-53 to enhance its performance on intent classification task and give some effective suggestions based on the experimental results for reference.

The main contributions of this paper are as follows:

- We conduct a wealth of experiments to investigate different of fine-tuning methods of XLSR-53 on intent classification.
- We propose a model that is effective on both Chinese and English SLU databases.

2. Related Work

Qian *et al.* [14] developed an ASR-free end-to-end SLU module for a dialog application. Serdyuk *et al.* [15] and Chen *et al.* [16] also discussed SLU tasks without ASR deeply. Achieving high accuracy with these models needs a large amount of labeled training data. To avoid the hassle of collecting data, Lugosch *et al.* [4] proposed a pretraining strategy that the model is trained to predict words and phonemes firstly, and then using this model to classify intention. Their system achieves high accuracy even using a 10% subset of the training data. And they released an open-source SLU dataset, Fluent Speech Commands. Lugosch *et al.* [5] proposed using speech synthesis to generate a large synthetic training dataset for relieving the influence of less data. Radfar *et al.* [17] introduced a neural transformer-based approach for SLU.

The bottleneck (BN) features are extracted from a narrow layer of the neural network and encoded the phonetic information in a nonlinearly compressed form. Fér *et al.* [11] used multilingual training to generate the latent universal features. Comparing with monolingual BN features, using multilingual BN features have better performance on the task of spoken language recognition. Silnova *et al.* [12] released the BUT/Phonexia bottleneck feature extractor. It is a Python toolkit that allowing to extract of BN features or phoneme classes posterior probabilities from a given audio signal. The package provides three neural networks. Two of them are trained on labeled Fisher English which contains approximately 2k hours speech of English. The third network *BabelMulti* is trained on 17 languages from

* corresponding author

http://dx.doi.org/10.21437/Interspeech.2021-818

the IARPA Babel program. The BUT/Phonexia bottleneck feature extractor helps researchers who do not have enough speech data with their phonetic studies.

With Bidirectional Encoder Representations from Transformers (BERT) [18] has been particularly successful for natural language processing, the framework named wav2vec 2.0, for self-supervised learning of speech representations is proposed by Baevski *et al.* [19]. The biggest difference between wav2vec 2.0 with the pretraining models mentioned above is that the self-supervised learning of wav2vec 2.0 needs unlabeled speech data. Unlabeled speech is much easier to collect than labeled. With the transformer's powerful parallelizable capability [20], wav2vec 2.0 can train a large amount of speech data simultaneously. Conneau *et al.* [13] tried to learn cross-lingual speech representations (XLSR) by extending wav2vec 2.0. Just like BERT, the waveform of the audio are fed to a cascade system of a multi-layer convolutional neural network (CNN) and a large Transformer network to build universal language representations. XLSR-53 has 7 CNN layers and 24 transformer blocks, model dimension 1024, inner dimension 4096 and 16 attention heads. It trained on $56k$ hours of speech data in 53 languages. In multilingual ASR tasks, the XLSR-53 system fine-tuning on labeled data greatly reduces the error rate and achieves state-of-the-art results. They released XLSR-53 publicly.

3. Proposed Method

The proposed method in this paper is using XLSR-53 to extract latent universal features across multiple languages and then using an intent module to classify intentions. As shown in Figure 1, input is speech signal. As a universal features extractor, XLSR-53 mainly consists of two parts, stacked convolutional neural networks (SCNNs) and stacked transformers. The SCNNs $f : \mathcal{X} \mapsto \mathcal{Z}$ map raw audio $X_{(B,L)}$ to latent speech representations $Z_{(B,T,F_1)}$, where B is batch size, L is lengths of a raw audio, T is the number of frames, F_1 is the dimension of the feature in latent speech representations space. The stacked transformers $g : \mathcal{Z} \mapsto \mathcal{C}$ map $Z_{(B,T,F_1)}$ to context representations $C_{(B,T,F_2)}$, where F_2 is dimension of the feature in context representations space. Our downstream task is to classify intentions. The (*action, object, location*) forms a triad. For example, (*"activate", "lights", "kitchen"*) is a single label. There are 31 and 15 distinct labels in FSC and CATSLU-MAP, respectively. The context representation $C_{(B,T,F_2)}$ are fed to intent module $h : \mathcal{C} \mapsto \mathcal{Y}$ to get the classification results $Y_{(B,N)}$, where N is number of labels. Since slot filling needs time information, we adopt a gated recurrent unit (GRU) layer. A linear layer is used to reduce dimensions and get intent labels. We make modifications based on this repository[1].

We focus on how to fine-tune XLSR-53 for accommodating the intent classification with a specific language (Chinese and English in this paper) and propose the following three methods.

1. Freezing vs. Unfreezing.

   As for the utilization of XLSR-53, it can be directly used as a universal language feature extractor like BUT bottleneck feature extractor but also can be trained with unfreezing its weights and fine-tuning them for the SLU task with backpropagation. The advantage of the former is that it saves computing resources and time. After all, XLSR-53 has about 320 million parameters.

2. Complementing tandem features.

   In the previous works [11, 21], we know that features extracted from the pretraining model combined with traditional acoustic features (MFCC, PLP, etc) can achieve better results on some phonetic tasks. We concatenate the features extracted from XLSR-53 and the tandem features of the log mel spectrogram encoded by CNN as the input of the intent module.

3. Further pretraining.

   The XLSR-53 model is pre-trained in 53 languages. For a specific language, such as Singapore Hokkien [22], its data distribution may be different from XLSR-53. We further pretrain XLSR-53 with its original self-supervised method in the target language. The XLSR-53 can get data distribution that is more relevant to the target language.

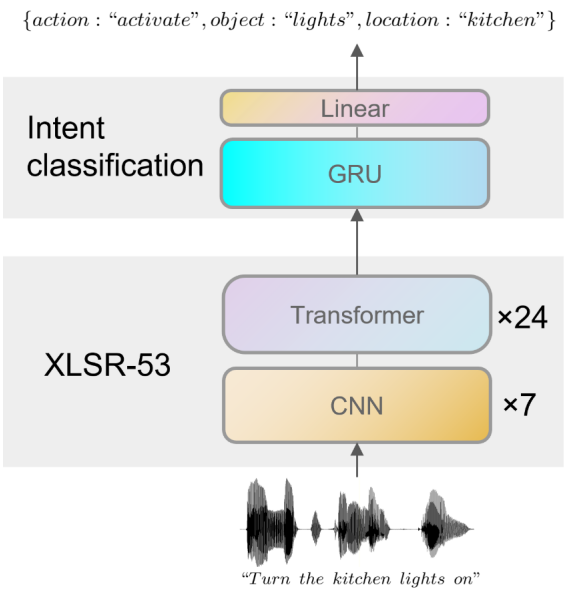

Figure 1: *Structures for our model: XLSR-53 extracts the semantic information from a waveform and the intent classification module generates the label.*

4. Experiments

4.1. Datasets

To test our method, we run a series of experiments on Fluent Speech Commands (FSC) and CATSLU-MAP. They are open-source SLU datasets. CATSLU-MAP can be downloaded here[2].

The FSC [4] is an English (en) SLU dataset that consists of 30043 audio files. These audios were recorded by 97 speakers and their format is 16 kHz single-channel .wav. Each audio contains an instruction to the smart assistant, like *"Turn the kitchen lights on"*. Its corresponding triple is {*action:"activate", object:"lights", location: "kitchen"*} and (*"activate", "lights", "kitchen"*) is the label. There are a total of 31 such distinct intents. It should be noted that for a piece of audio, such as "Switch the language", its label is (*"change language", "none", "none"*). The values of *object* and *location* can be *"none"*. Our SLU task is to do a single-label classification and

[1] https://github.com/lorenlugosch/end-to-end-SLU

[2] https://sites.google.com/view/catslu/home

does not consider the correlation between the triples. When classifying each audio, each slot value is filled, and the predicted three-slot values are consistent with the label, then it was judged to be correct. Our intent classification task is exactly the same as [4].

The CATSLU-MAP [23] is a Chinese (zh) SLU dataset. CATSLU means "the 1st Chinese Audio-Textual Spoken Language Understanding Challenge". Here we use Chinese to simulate the low-resource language. The utterances of CATSLU-MAP were collected from dialogues between users and a manageable spoken dialogue system in the map navigation domain. They are real-world commands, like "(*Chinese*) 导航云南省昆明市黄土坡(*Navigate to Huangtupo, Kunming City, Yunnan Province*)", "(*Chinese*) 西乡流塘 (*This is a Chinese place name*)", "(*Chinese*) 我在哪里(*Where am I now?*)". A total of 1788 users' voices. The office provides the audio information and manual transcripts, here we only use the audio information. Each utterance contains a command. In the original challenge tasks, not only had to match the slot values of *actions* and *objects* but also the *location* that are full of place names had to be identified. In our task, we only classify the labels composed of *action* and *object*. To keep it consistent with the English FSC task, we set all *locations* to *"none"*. So the three sentences mentioned above, corresponding labels are (*"inform", "operating", "none"*), (*"inform", "destination", "none"*) and (*"request", "position", "none"*). There are about $1/5$ of the utterances that have 2 or 3 labels, here we only take the first one. The 15 distinct labels can be used for the single-label classification task. The information about the number of users of the FSC and CATSLU-MAP, the number of utterances and labels can be found in Table 1.

Table 1: *Information about the FSC and CATSLU-MAP datasets.*

Dataset	#Uers	#Utterances			#Labels
		train	valid	test	
FSC (en)	97	23132	3118	3793	31
CATSLU -MAP (zh)	1788	5093	921	1576	15

4.2. Experimental settings

We use the XLSR-53 model with 24 transformer blocks, model dimension 1024, inner dimension 4096, and 16 attention heads, and combining XLSR-53 with intent module. In the freezing experiments on the FSC and CATSLU-MAP, we no longer update the parameters of the XLSR-53 model using backpropagation. We training the model with the batch size of 8, the learning rate of 1e-3. Audio signal $X_{(8,L)}$, where L is audio lens, passes through XLSR-53 to obtain the semantic information $C_{(8,T,1024)}$, where T is the number of frames. And $C_{(8,T,1024)}$ constructs time information $\tilde{C}_{(8,T,1024)}$ through GRU. $\tilde{C}_{(8,T,1024)}$ passes a linear layer and a max-pooling layer to generate $Y_{(8,N_s)}$, where N_s is the number of slots. In the unfreezing experiments on the FSC and CATSLU-MAP, we train the entire model with the batch size of 1, the learning rate of 6e-6.

We try to discuss the influence of traditional acoustic features on the model. Combining the multilingual representation information with the log mel spectrum. We use torchlibrosa [21] to extract log mel spectrum from audio with 32 mel filterbanks. The size of fast Fourier transform is 512, and the type of window is hanning. The shape of the log mel spectrum is (B, 1, T, M), where B is batch size, M is the number of mel bins of 32. The log mel spectrum passes through a batch normalization layer and a convolutional layer, its shape becomes (B, 64, T, 32). Reshape it to get $\hat{C}_{(B,T,1024)}$. We concatenate $\hat{C}_{(B,T,1024)}$ with the semantic representation feature $C_{(B,T,1024)}$ of dimension T to input the intent module. In the experiments of complementing the log mel spectrum, the batch size is 8 and the learning rate is 1e-4.

We collected the open source Mandarin speech corpus AISHELL1 [24] and AISHELL2 [25], a total of 1178 hours. We used the fairseq toolkit [26] to further pretrain XLSR-53 on 6 Nvidia Tesla v100 GPUs about 130 hours for 90 epochs, and the valid loss was reduced to 1.605. Then we got XLSR-53-ZH. In the Chinese further pretraining experiments, instead of XLSR-53, we will use XLSR-53-ZH as the pretrained model. We have released the pretrained model of XLSR-53-ZH[3].

The baseline of the experiments on FSC is the best result in [4]. For the Chinese baseline, we use the model of BUT/Phonexia bottleneck feature extractor [12] combined intent module with the stacked bottleneck features (SBN). The SBN extracted from the neural network of *BabelMulti* are used to intent classification.

4.3. Experimental results

Here we report results for three experiments on FSC and CATSLU-MAP: *Freezing* vs. *Unfreezing*, complementing Tandem features and further pretraining. Figure 2 and Figure 3 show the accuracy of these models over time on the FSC and CATSLU-MAP datasets, respectively.

Freezing is to freeze the XLSR-53 whose parameters are no longer updated by backpropagation. As a feature extractor, the frozen XLSR-53 processes the raw audio signal for intent classification. *Unfreezing* is to update the parameters of XLSR-53 during training. The experimental results of *Freezing* vs. *Unfreezing* are shown in Table 2. The accuracy in the table represents the highest accuracy over 30 epochs of the single-label classification. We can see that the *Freezing*'s accuracy is very low (green line in Figure 2 and 3). The features extracted from the frozen XLSR-53 that represents universal information across 53 languages. They are not suitable for the specific language SLU task. *Unfreezing* shows that our model outperforms the baseline models of 0.87% and 8.44%, respectively. It means that, in order to achieve excellent results on the SLU task, the model should be fine-tuned on the labeled data of the target language firstly.

Table 2: *Experiment results of Freezing vs. Unfreezing.*

Dataset	**Model**	**Accuracy**
FSC	Baseline[4]	98.84%
	Freezing(ours)	58.08%
	Unfreezing(ours)	**99.71%**
CATSLU -MAP	Baseline[12]	71.45%
	Freezing(ours)	40.67%
	Unfreezing(ours)	**79.89%**

Freezing+logmel represents the experiment that the features from the frozen XLSR-53 concatenating tandem features

[3]https://zenodo.org/record/4655324

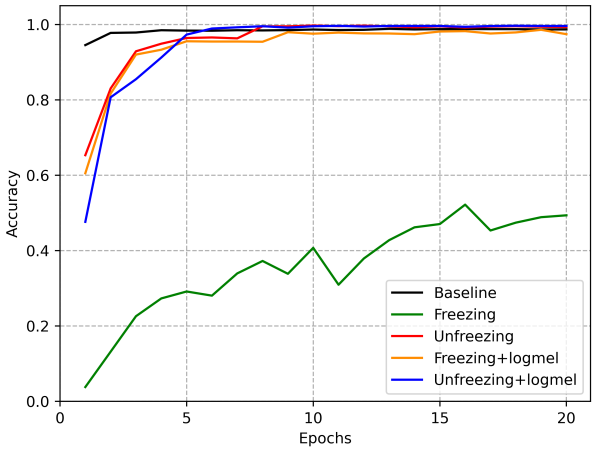

Figure 2: *Accuracy on the test set over time for models trained on FSC.*

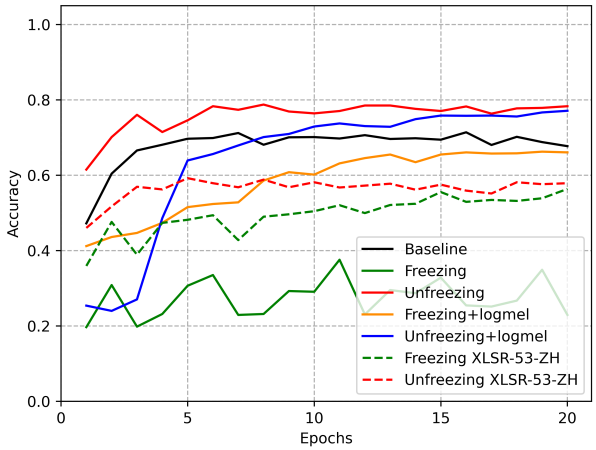

Figure 3: *Accuracy on the test set over time for models trained on CATSLU-MAP.*

Table 3: *Experiment results of complementing tandem features.*

Dataset	Model	Accuracy
FSC	*Freezing*	58.08%
	Freezing+logmel	**98.58%**
	Unfreezing	**99.71%**
	Unfreezing+logmel	99.66%
CATSLU -MAP	*Freezing*	40.67%
	Freezing+logmel	**66.88%**
	Unfreezing	**79.89%**
	Unfreezing+logmel	77.98%

of the log mel features encoded by the convolutional layer input into the intent module. *Unfreezing+logmel* means combining the log mel part, all parameters will be trained. Table 3 shows the effect of concatenating log mel spectrum. We can see that after combining the log mel spectrum, the *Freezing*'s performance of classification has been greatly improved (orange line in Figure 2 and 3). The accuracy has reached an acceptable value. From *Unfreezing* to *Freezing*, the training parameters of the model drop from 322 million to 5 million. Using *Freezing+logmel* can save computing resources and training time. But in *Unfreezing* experiments, concatenating log mel spectrum does not meet our expectations, and the accuracy even dropped. We guess that the log mel spectrum represents traditional acoustic features, and universal language information represents fine semantic features. Simply concatenating the two cannot effectively complement each other.

Freezing XLSR-53-ZH and *Unfreezing XLSR-53-ZH* (Figure 3, dashed line) just replaced the pretrained model XLSR-53 in *Freezing* and *Unfreezing* with XLSR-53-ZH. The experiment results of further pretraining are shown in Table 4. We want to use this experiment to demonstrate how much the large amount of unlabeled speech data in the target language plays a role in improving the performance of intent classification. XLSR-53-ZH has indeed learned Chinese grammar knowledge. It can be seen from the green dashed line in the Figure 3. The features from the frozen XLSR-53-ZH are more relevant to Chinese SLU tasks. But after unfreezing XLSR-53-ZH, the accuracy of the model is much lower than *Unfreezing*. This may be because the further pertraining of the Chinese dataset has destroyed the original universal data distribution of 53 languages. The parameters are more sensitive to Chinese grammar, but the huge amount of parameters leads to overfitting. How to perform the self-supervised training on XLSR-53 is still to be explored.

Table 4: *Experiment results of XLSR-53-ZH.*

Dataset	Model	Accuracy
CATSLU -MAP	*Freezing*	40.67%
	Freezing XLSR-53-ZH	**56.54%**
	Unfreezing	**79.89%**
	Unfreezing XLSR-53-ZH	59.33%

For the reason why the accuracy of this seemingly simple Chinese SLU task is lower than FSC, it is mainly because that there are about 1/5 of the utterances that have 2 or 3 labels in the test set, and we only chose the first one as their labels. This may cause a result that was predicted correctly by model to be judged as an error. And the utterances of CATSLU-MAP were collected from the real world. There are a lot of users. The sentences is flexible and mixed with dialects. The complex speech information adds challenges to the model.

5. Conclusion

In this paper, we propose an end-to-end model with XLSR-53 to challenge cross-lingual SLU tasks. And we conduct three types of experiments to investigate the different approaches to fine-tuning XLSR-53 for the specific language SLU tasks. We show that our model not only achieves state-of-the-art performances on the English dataset, but also can be transferred to the Chinese dataset even outperforms a competitive model. The experiments found that the universal language information from the frozen XLSR-53 should be combined with traditional acoustic features of target language for the SLU task. We try a strategy of further self-supervised pretraining. In the future, we will use this model to do experiments in another languages and consider how to compress the model.

6. References

[1] W. Zhang, Z. Chen, W. Che, G. Hu, and T. Liu, "The first evaluation of chinese human-computer dialogue technology," *ArXiv*, vol. abs/1709.10217, 2017.

[2] G. Mesnil, Y. Dauphin, K. Yao, Y. Bengio, L. Deng, D. Hakkani-Tur, X. He, L. Heck, G. Tur, D. Yu, and G. Zweig, "Using recurrent neural networks for slot filling in spoken language understanding," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, pp. 530–539, 2015.

[3] A. Coucke, A. Saade, A. Ball, T. Bluche, A. Caulier, D. Leroy, C. Doumouro, T. Gisselbrecht, F. Caltagirone, T. Lavril, M. Primet, and J. Dureau, "Snips voice platform: an embedded spoken language understanding system for private-by-design voice interfaces," *ArXiv*, vol. abs/1805.10190, 2018.

[4] L. Lugosch, M. Ravanelli, P. Ignoto, V. Tomar, and Y. Bengio, "Speech model pre-training for end-to-end spoken language understanding," *ArXiv*, vol. abs/1904.03670, 2019.

[5] L. Lugosch, B. Meyer, D. Nowrouzezahrai, and M. Ravanelli, "Using speech synthesis to train end-to-end spoken language understanding models," *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 8499–8503, 2020.

[6] Y. Huang, H. K. Kuo, S. Thomas, Z. Kons, and M. Picheny, "Leveraging unpaired text data for training end-to-end speech-to-intent systems," *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7984–7988, 2020.

[7] W. Chen, M. Hasegawa-Johnson, and N. F. Chen, "Topic and keyword identification for low-resourced speech using cross-language transfer learning," in *INTERSPEECH*, 2018.

[8] Q. Do and J. Gaspers, "Cross-lingual transfer learning for spoken language understanding," *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5956–5960, 2019.

[9] ——, "Cross-lingual transfer learning with data selection for large-scale spoken language understanding," in *EMNLP/IJCNLP*, 2019.

[10] X. Jia, J. Wang, Z. Zhang, N. Cheng, and J. Xiao, "Large-scale transfer learning for low-resource spoken language understanding," *ArXiv*, vol. abs/2008.05671, 2020.

[11] R. Fér, P. Matejka, F. Grézl, O. Plchot, K. Veselý, and J. H. Cernocký, "Multilingually trained bottleneck features in spoken language recognition," *Comput. Speech Lang.*, vol. 46, pp. 252–267, 2017.

[12] A. Silnova, P. Matejka, O. Glembek, O. Plchot, O. Novotný, F. Grézl, P. Schwarz, L. Burget, and J. Černocký, "But/phonexia bottleneck feature extractor," in *Odyssey*, 2018.

[13] A. Conneau, A. Baevski, R. Collobert, A. rahman Mohamed, and M. Auli, "Unsupervised cross-lingual representation learning for speech recognition," *ArXiv*, vol. abs/2006.13979, 2020.

[14] Y. Qian, R. Ubale, V. Ramanarayanan, P. L. Lange, D. Suendermann-Oeft, K. Evanini, and E. Tsuprun, "Exploring asr-free end-to-end modeling to improve spoken language understanding in a cloud-based dialog system," *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 569–576, 2017.

[15] D. Serdyuk, Y. Wang, C. Fuegen, A. Kumar, B. Liu, and Y. Bengio, "Towards end-to-end spoken language understanding," *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5754–5758, 2018.

[16] Y.-P. Chen, R. Price, and S. Bangalore, "Spoken language understanding without speech recognition," *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6189–6193, 2018.

[17] M. Radfar, A. Mouchtaris, and S. Kunzmann, "End-to-end neural transformer based spoken language understanding," in *INTERSPEECH*, 2020.

[18] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *NAACL-HLT*, 2019.

[19] A. Baevski, H. Zhou, A. rahman Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *ArXiv*, vol. abs/2006.11477, 2020.

[20] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *ArXiv*, vol. abs/1706.03762, 2017.

[21] Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, and M. D. Plumbley, "Panns: Large-scale pretrained audio neural networks for audio pattern recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2880–2894, 2020.

[22] R. Muhr, "Adrian tien: Chinese hokkien and its lexicon in singapore: evidence for an indigenised singapore culture," 2012.

[23] S. Zhu, Z. Zhao, T. Zhao, C. Zong, and K. Yu, "Catslu: The 1st chinese audio-textual spoken language understanding challenge," *2019 International Conference on Multimodal Interaction*, 2019.

[24] H. Bu, J. Du, X. Na, B. Wu, and H. Zheng, "Aishell-1: An open-source mandarin speech corpus and a speech recognition baseline," *2017 20th Conference of the Oriental Chapter of the International Coordinating Committee on Speech Databases and Speech I/O Systems and Assessment (O-COCOSDA)*, pp. 1–5, 2017.

[25] J. Du, X. Na, X. Liu, and H. Bu, "Aishell-2: Transforming mandarin asr research into industrial scale," *ArXiv*, vol. abs/1808.10583, 2018.

[26] M. Ott, S. Edunov, A. Baevski, A. Fan, S. Gross, N. Ng, D. Grangier, and M. Auli, "fairseq: A fast, extensible toolkit for sequence modeling," in *NAACL-HLT*, 2019.