Zero-shot Cross-Lingual Phonetic Recognition with External Language Embedding

Heting Gao[1], *Junrui Ni*[1], *Yang Zhang*[2], *Kaizhi Qian*[2], *Shiyu Chang*[2], *Mark Hasegawa-Johnson*[1]

[1]University of Illinois at Urbana-Champaign
[2]MIT-IBM Watson AI Lab

{hgao17,junruin2,jhasegaw}@illinois.edu, {Yang.Zhang2,kqian,Shiyu.Chang}@ibm.com

Abstract

Many existing languages are too sparsely resourced for monolingual deep learning networks to achieve high accuracy. Multilingual phonetic recognition systems mitigate data sparsity issues by training models on data from multiple languages and learning a speech-to-phone or speech-to-text model universal to all languages. However, despite their good performance on the seen training languages, multilingual systems have poor performance on unseen languages. This paper argues that in the real world, even an unseen language has metadata: linguists can tell us the language name, its language family and, usually, its phoneme inventory. Even with no transcribed speech, it is possible to train a language embedding using only data from language typologies (phylogenetic node and phoneme inventory) that reduces ASR error rates. Experiments on a 20-language corpus show that our methods achieve phonetic token error rate (PTER) reduction on all the unseen test languages. An ablation study shows that using the wrong language embedding usually harms PTER if the two languages are from different language families. However, even the wrong language embedding often improves PTER if the language embedding belongs to another member of the same language family.

Index Terms: speech recognition, phonetic recognition, external linguistic knowledge

1. Introduction

Modern end-to-end neural network based speech recognition systems (ASR) have achieved great success on resource-rich languages such as English and Mandarin [1]. However, most existing languages are resource-deficient, making it hard for neural networks to achieve similar accuracy.

Multilingual and Cross-lingual phonetic recognition attempt to partially solve the low-resource problem by building a universal phone recognizer that transcribes speech from different languages into corresponding phone sequences, under the assumption that there exists a universal acoustic model shared by all languages. If this assumption holds, an ideal recognizer should have low error rates on not only the languages it is trained on, *i.e. multilingual error rates*, but also the unseen languages, *i.e. cross-lingual error rates*, in a zero-shot setting.

However, although multilingual training is shown to improve the performance on seen languages [2, 3, 4], it does not greatly benefit zero-shot generalization to unseen languages [5]. This implies that acoustic models implicitly captured in these multilingual systems are language-specific, and thus would not generalize to unseen languages unless additional information about the unseen languages is supplied.

Motivated by this, we propose to improve the zero-shot cross-lingual recognition accuracy by incorporating a language embedding that captures two types of external knowledge – *phylogenetic similarity* and *phone inventory*. For phylogenetic similarity, we extract phylogenetic information from Glottolog [6], which is a large graph specifying the belonging relations between nodes of dialects, languages and language families. Assuming the closeness of the two languages in the graph captures the phylogenetic similarities between the languages, we use node2vec [7] to extract vector representations for each node. For the phone inventory information, we extract a binary vector to represent the phoneme inventory for each language from Phoible [8]. The two vectors are combined and fed into a language encoder and produce the language embedding, on which the multilingual phoneme classifier is conditioned. The phone inventory information is also imposed by masking on the output logits with the binary vector.

The experiments show that the proposed algorithm with language embedding and masking improves the performance over the baselines on the unseen languages in the zero-shot setting by a large margin (4%-8% absolute). Ablation study shows that both the phylogenetic and phone inventory information are crucial for performance improvement.

2. Related Work

There has been active research on multilingual recognition. A large number of languages do not have enough parallel speech and text data and deep learning models trained on these languages usually have high error rates [5]. Multilingual speech recognition mitigates the data sparsity by training the network on a combined dataset from several languages. The network usually has a common encoder that extracts acoustic information from audio features and can either have a common decoder with a shared phoneme inventory [4] or language-specific decoders with private phone [3, 9, 10] or character inventories [11, 12, 13]. Multilingual ASR can benefit from the use of self-supervised pretraining algorithms such as contrastive predictive coding [14, 15, 16], which pretrains a model on large amounts of unlabeled raw audio data to predict neighboring frame representations given the center frame. Multilingual models generally have better accuracy and robustness compared to monolingual models [5, 4, 3, 9, 10] as they benefit from increased amount and diversity of data.

Language or dialect embedding has been shown to improve multilingual ASR systems [17, 18, 19, 20]. The embedding can be a one-hot vector specifying language ID [17, 19] or a vector learned from acoustic data under a standard multilingual model [18, 20] and can be used as additional input features to the network [17, 19], as adapter modules for language-specific adjustments [19] or as interpolation weights for the

This work is supported by IBM-UIUC Center for Cognitive Computing Systems Research (C3SR).

encoder [18]. However, the embeddings in all these previous works depend on the test language being either one of the training languages (in the case of a one-hot embedding) or recorded in a fashion that makes its acoustic embedding vector a useful predictor of its phoneme-to-sound acoustic models.

Studies have found that multilingual models do not generalize well to unseen languages [5], without adapting to parallel data from that language. While multilingual training can yield error rates 10–20% below monolingual training, the leave-one-out cross-lingual error rate when applying the multilingual model to an unseen language can be 70–90%. Because of the high error rates of zero-shot cross-lingual ASR, most researchers studying cross-lingual ASR have chosen pragmatically to define that term to mean few-shot rather than zero-shot recognition, e.g., by fine-tuning using one hour [21, 22] or a few hours [23] of transcribed data in the target language. Perhaps the prior work most similar to the work in this paper is a set of experiments using the Phoible [8] phoneme inventory of a language to define an untrained, knowledge-based linear output layer called the "signature matrix" [24, 4]; our phone token masking strategy is a simplification of the signature matrix, and our proposed language encoding is an enrichment of the same.

3. Methods

Previous works have shown that it is hard to achieve good performance on zero-shot cross-lingual recognition without any knowledge about the testing language. We therefore consider incorporating extra information about the testing language. Figure 1 shows the overview of the proposed architecture. The proposed system is a CTC+Attention system based on [5], with three additions: (1) wav2vec-based feature extraction based on [15], (2) phoneme inventory masking similar to [4], and (3) the proposed typology-based language encoder.

The language encoder The language encoder includes two sets of information about the test language. The first is the language phylogenetic information extracted from Glottolog, which is a graph containing dialects, languages, language families as nodes and the belonging relationships as edges. We use node2vec [7] to embed the nodes so that the languages that are close in the graph have larger cosine similarities.

Similar to the multilingual allophone system in [4], we also include phone inventory information from Phoible [8], a cross-linguistic phonological inventory database for over 2000 distinct languages. We combine inventories for all the languages to create a shared phoneme inventory and use a binary vector to represent the phoneme set of each language.

The language node embedding and the binary phoneme inventory vector are concatenated, forming a general representation applicable to at least 2,000 languages. The vector is then fed into the language encoder, producing a language embedding as an additional input to the phoneme classifier.

Wav2vec Feature Extraction Considering the remarkable performance boost brought by pretrained unsupervised acoustic representation, we experiment on the feature extractor (referred to as feature encoder in [16]) from wav2vec2.0 [1] that is pretrained on 1000 hours of LibriSpeech [25].

Phone Inventory Masking In addition to feeding the phone inventory asks as an input to the language encoder, we also directly use it to mask out the non-existing phonetic tokens in the output layer, which has been shown to be effective in reducing the error rate, especially for unseen languages.

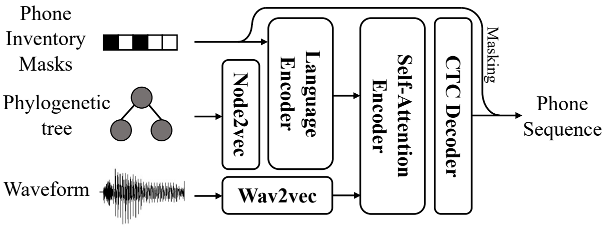

Figure 1: *Architecture overview*

Table 1: *Sources of data used in our cross-lingual experiment. The upper part is the training languages and the lower part is the testing languages. "Type" column denotes whether the corpus contains spontaneous (Sp.) or read speech. "Len" column shows the total duration of all utterances in hours. "Family" column shows the language family.*

Language	Abbr	Corpus	Type	Family	Len
Bengali	103	Babel	Sp.	Indo-Aryan	215
Vietnamese	107	Babel	Sp.	Vietic	215
Zulu	206	Babel	Sp.	Bantu	211
Amharic	307	Babel	Sp.	Ethiopic	204
Javanese	402	Babel	Sp.	Austronesian	204
Georgian	404	Babel	Sp.	Kartvelian	190
Dutch	N	CGN	Read	Germanic	64
Czech	CZ	GP	Read	West Slavic	29
French	FR	GP	Read	Romance	25
Mandarin	CH	GP	Read	Sinitic	31
Thai	TH	GP	Read	Tai	22
German	GE	GP	Read	Germanic	18
Portuguese	PO	GP	Read	Romance	26
Turkish	TU	GP	Read	Turkic	17
Bulgarian	BG	GP	Read	South Slavic	21
Cantonese	101	Babel	Sp.	Sinitic	215
Lao	203	Babel	Sp.	Tai	207
Croatian	CR	GP	Read	South Slavic	16
Spanish	SP	GP	Read	Romance	22
Polish	PL	GP	Read	West Slavic	24

4. Experiment Setup

4.1. Dataset

The performance of our model is evaluated on a corpus that consists of 20 languages, 8 from IARPA Babel project corpora, 1 from CGN (Spoken Dutch Corpus) [26] and 11 from Globalphone [27] (GP) as summarized in Table 1. We only use the read speech part of CGN corpus. We use the default 8:1:1 train-dev-test partition provided by Babel corpora and split CGN and Globalphone corpora into 8:1:1 partitions with non-overlapping speakers. Since our task is cross-lingual phonetic token recognition, the train and dev partitions of the testing languages are not used. We select 5 languages, namely Cantonese, Lao, Croatian, Spanish and Polish as the testing language set and use the remaining 15 languages as a training language set. Each testing language is selected to have a similar language belonging to the same language family in the training set.

[1] https://github.com/pytorch/fairseq/tree/master/examples/wav2vec

Table 2: *Phonetic token error rates (PTER) in percentage. The columns "103" to "BG" are PTER's evaluated on the 15 seen languages and the columns from "101" to "PL" are PTER's evaluated on the 5 unseen languages. The column "AvgS" the is the average PTER over the 15 seen languages and the column "AvgU" are the average PTER over the 5 unseen languages.*

Exp	103	107	206	307	402	404	N	CZ	FR	CH	TH	GE	PO	TU	BG	101	203	CR	SP	PL	AvgS	AvgU
base	40.2	52.3	42.4	44.7	47.0	**38.0**	21.3	11.0	13.7	30.0	26.1	26.1	18.4	21.3	27.0	77.0	78.2	47.8	38.1	62.5	30.6	60.7
w2v	41.3	36.6	39.0	43.1	48.9	42.2	15.3	10.5	14.8	17.2	22.2	25.1	18.7	21.0	30.2	77.9	79.3	47.3	39.0	66.7	28.4	62.0
w2v+mask	41.1	36.6	38.8	43.1	48.4	41.7	15.3	10.5	14.8	17.2	22.2	25.1	18.7	21.0	30.2	76.5	76.8	42.8	36.8	61.2	28.3	58.8
w2v+linear	39.0	32.6	35.9	39.1	44.9	39.1	14.0	9.1	12.9	15.9	19.9	23.2	16.3	19.3	28.2	74.6	76.3	41.3	37.3	59.8	26.0	57.9
w2v+linear+mask	39.0	32.6	35.9	39.1	44.9	39.1	14.0	9.1	12.9	15.9	19.9	23.2	16.3	19.3	28.2	**73.1**	72.8	**35.2**	**34.4**	**54.0**	26.0	**53.9**
w2v+gcn	**38.2**	**32.0**	**35.2**	**38.0**	**44.2**	38.6	**13.2**	**8.5**	**12.1**	**15.5**	**18.9**	**22.3**	**16.0**	**18.4**	**26.9**	76.1	72.4	50.8	37.5	61.9	**25.2**	59.7
w2v+gcn+mask	**38.2**	**32.0**	**35.2**	**38.0**	**44.2**	38.6	**13.2**	**8.5**	**12.1**	**15.5**	**18.9**	**22.3**	**16.0**	**18.4**	**26.9**	**73.1**	**69.3**	39.6	35.3	56.3	**25.2**	54.7

4.2. Data Preprocessing

We use ESPnet as our ASR framework [28] since ESPnet offers a complete ASR pipeline including data preprocessing, Transformer implementation, network training and decoding.

Due to the sampling rate difference between different corpora, we upsample all audio signals to 16kHz. Using Kaldi [29], we then extract 80-dim log Mel spectral coefficients with 25ms frame size and 10ms shift between frames, and augment the frame vectors with 3 extra dimensions for pitch features.

The transcriptions are converted to IPA symbols using LanguageNet grapheme-to-phone (G2P) [30] models and the unique IPA symbols, including base phones, diacritics and suprasegmentals, in all 15 training languages are collected as the shared phonetic token inventory. The resulting inventory size is 95. The test languages contain phones that are not present in any training languages, which causes an out-of-vocabulary (OOV) problem as our network cannot predict a phone it has never seen. We map each OOV phone to its closest in-vocabulary phone according to its articulatory features defined by IPA. For example, /β/ in Spanish is mapped to /v/.

4.3. Language Embedding

We experiment with two types of transformations to generate the language embedding, a 3-layer fully-connected transformation and a 3-layer graph-convolutional transformation[2] on the language representations extracted from Glottolog [6] and Phoible [8]. Each transformation layer is followed by a ReLU activation and a dropout layer with a dropout rate of 10%. The output of the transformation networks is used as language embedding and as input to the self-attention based ASR network.

4.4. Model

We experiment with two audio embedding modules. One consists of two 2D convolutional layers (randomly initialized) with a subsampling factor of 4 that takes the extracted 83-dim audio features as input, and the other is the feature extractor of a pretrained wav2vec2.0 [16] model that directly takes the 16kHz waveform as input. We fix the weights of the wav2vec feature extractor during training.

The encoder of our model architecture is similar to the transformer architecture in [31]. The audio embeddings are fed into 12 self-attention encoder layers, each having 4 heads, an attention dimension of 256 and a 2048-dim position-wise feedforward layer. The only difference is that input to each encoder layer is additionally concatenated with the correct language embedding to provide language information to the transformer.

Our preliminary experiments indicate that the self-attention decoder framework does not outperform a simple CTC decoder in cross-lingual recognition, which is consistent with the findings in [23]. Therefore, we discard the self-attention decoder in [31] and apply a dense layer to the encoder output to compute the frame-wise phoneme posteriors and the CTC loss.

4.5. Evaluation

We use phonetic token error rate (PTER) [5] to evaluate our models. It is calculated the same way as character error rate except that the model predicts a set of language-universal IPA tokens instead of normal orthographic characters. It treats diacritics (such as aspiration /ʰ/), suprasegmentals (such as long vowels /ː/ and primary stress symbol /ˈ/), and tones (such as high tone /˥/ and low tone /˩/) as separate tokens. It also splits diphthongs and affricates into individual symbols. For example /ˈtaː/ would be viewed as 4 tokens. Therefore, our PTER metric slightly differs from the phone error rate (PER) calculated in other multilingual literature such as [4].

5. Results

5.1. Multilingual and Cross-lingual Phonetic Recognition

We train and test on our 20-language dataset with 7 different models: "base", "w2v", "w2v+mask", "w2v+linear", "w2v+linear+mask", "w2v+gcn", "w2v+gcn+mask". All the models have a self-attention encoder and a CTC decoder. "base" model uses a randomly initialized 2D convolutional feature extractor and the models with "w2v" label instead use a pretrained wav2vec feature extractor. The models with "linear" and "gcn" labels have an additional linear or graph-convolutional transformation network to compute the language embeddings. Models with "mask" apply phone inventory masking to the softmax output layer of the decoder.

The performance is shown in Table 2, where both proposed models ("w2v+linear+mask" and "w2v+gcn+mask") outperform the "base" model; "w2v+gcn+mask" model achieves the lowest multilingual error rate, while "w2v+linear+mask" model achieves lowest cross-lingual error rate.

By comparing "base" and "w2v", we see that a pretrained wav2vec feature extractor reduces the average multilingual recognition error rate. In particular, the reduction is 15.7% on Vietnamese (107), 6% on Dutch (N) and 12.8% on Mandarin (CH). Although it slightly increases the cross-lingual error rate, we decide to build on "w2v" model instead of "base" model.

Comparing the average test PTER (AvgU) of "w2v", "w2v+linear" and "w2v+gcn" with that of "w2v+mask", "w2v+linear+mask" and "w2v+gcn+mask", we see that masking out the non-existing phonetic tokens in the test language greatly improves the recognition accuracy, possibly due to the reduced prediction space. The"w2v+gcn+mask" model, which places the most emphasis on language-family structure, gains

[2] https://github.com/tkipf/gcn

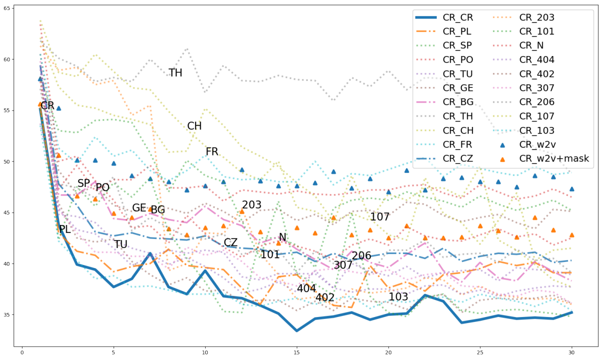

Figure 2: *PTER of “w2v+linear+mask” model tested on Croatian with correct and fake language labels.*

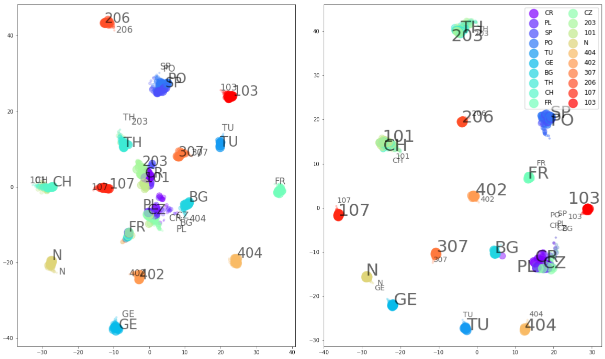

Figure 3: *t-SNE plot of language embedding. The left plot is the embedding from “w2v+linear” and the right plot is the embedding from “w2v+gcn”.*

Table 3: *Phonetic token error rates (PTER) Ablation Study.*

w2v+linear+mask	101	203	CR	SP	PL	Avg
glotto+phoible	73.1	72.8	35.2	**34.4**	54.0	53.9
glotto	**69.5**	73.4	**35.1**	34.8	55.7	**53.7**
phoible	76.0	**71.9**	36.6	38.8	**53.4**	55.3

the largest improvement from phone masking, but still does not outperform the “wav+linear+mask” model, suggesting that applying the graph constraint a second time (GCN on top of node2vec embeddings) provides no extra reduction of error rates.

5.2. Cross-lingual Phonetic Recognition with Fake Language Labels

To better understand how language embedding affects the model’s performance, we feed both true and fake language embeddings to the model and plot the test PTERs across epochs. Figure 2 shows the PTER of “w2v+linear+mask” model tested on Croatian. The blue and orange triangle points are PTERs of the “w2v” and “w2v+mask” models respectively. The blue solid line labeled “CR_CR” is the PTER curve with correct Croatian embedding and the dash-dotted lines or dotted lines are PTER’s of the model when provided with fake language embeddings.

We observe that when provided with correct language embedding (CR_CR), the model outperforms the masked wav2vec baseline (w2v+mask). The PTER of the model when provided with fake embedding varies from 35% to 80%. In particular, when provided with fake embeddings of languages from the same language family, Slavic family in this example, the model generally has a lower PTER compared to others, as shown by the curves of Polish (CR_PL), Bulgarian (CR_BG) and Czech (CR_CZ). This indicates that our model is able to leverage the phylogenetic and phonetic similarities for better accuracy.

5.3. Visualization of Language Embedding

We visualize the language embeddings of “w2v+linear” and “w2v+gcn” using t-SNE [32] in Figure 3. The small and light circles are the embeddings from earlier epochs and large and solid circles are from later epochs. We use small and light text to label the embeddings’ initial-epoch position and large and solid text to label the final-epoch position. In the right plot, we observe that graph convolutional transformation on language vectors largely preserves the phylogenetic information; the languages that are close in the initial epoch remain close in the final epoch. In contrast, the left plot shows that linear transformation preserves the phylogenetic information only partially. For example, while the Sinitic-language embeddings (CH and 101) are close initially, Cantonese (101) moves away from Mandarin (CH) towards the Slavic-languge embeddings (CR, CZ, PL and BG) as the training epoch increases. This observation indicates the linear transformation has larger flexibility to learn its embeddings; as shown in Table 2, this flexibility reduces the cross-lingual error rate.

5.4. Ablation Study on Language Representation

We conduct an ablation study to see the role of the Glottolog vector and Phoible vector in error rate reduction by training “w2v+linear” model with only Glottolog vector, with only Phoible vector and with both. The results are shown in Table 3. First, providing external information reduces error: all three settings (“glotto”, “phoible”, “glotto+phoible”) beat the “w2v+mask” baseline. Second, using only Glottolog vectors reduces the Cantonese (101) error rate to 69.5% but raises the Lao (203) error rate to 73.4%, which is close to the performance of the “w2v+gcn+mask” model, while using only Phoible vectors does the reverse, raising the Cantonese error rate but reducing the Lao error rate. These results show both vectors improve the performance in different ways; “w2v+linear+mask” finds a good trade-off between relying on phylogenetic information and phonetic information. Finally, we notice that using only Glottolog vectors (“glotto”) has nearly the same performance as both vectors (“glotto+phoible”). We hypothesize that phoneme masking is functioning as a substitution, reducing the necessity of the phoible vector.

6. Conclusions

In this work, we propose to use external phylogenetic and phonetic knowledge from language typologies to improve the cross-lingual phoneme recognizer. We study the performance of learning language embeddings using a linear transformation network and a graph convolutional network and show that both models outperform the baseline. In particular, we show both phylogenetic and phonetic knowledge are necessary for good cross-lingual accuracy and that a linear transformation network can flexibly leverage both types of information to learn a better phonetic model compared to a graph convolutional network.

7. References

[1] A. Zeyer, P. Bahar, K. Irie, R. Schlüter, and H. Ney, "A comparison of transformer and LSTM encoder decoder models for ASR," in *ICASSP*, 2019, pp. 8–15.

[2] J.-T. Huang, J. Li, D. Yu, L. Deng, and Y. Gong, "Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013, pp. 7304–7308.

[3] S. Dalmia, R. Sanabria, F. Metze, and A. W. Black, "Sequence-based multi-lingual low resource speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 4909–4913.

[4] X. Li, S. Dalmia, J. Li, M. Lee, P. Littell, J. Yao, A. Anastasopoulos, D. R. Mortensen, G. Neubig, A. W. Black *et al.*, "Universal phone recognition with a multilingual allophone system," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 8249–8253.

[5] P. Żelasko, L. Moro-Velázquez, M. Hasegawa-Johnson, O. Scharenborg, and N. Dehak, "That sounds familiar: an analysis of phonetic representations transfer across languages," in *Interspeech*, 2020, pp. 3705–3709.

[6] H. Hammarström, R. Forkel, M. Haspelmath, and S. Bank, *Glottolog 4.3*, Jena, 2020. [Online]. Available: https://glottolog.org/accessed2021-03-30

[7] A. Grover and J. Leskovec, "node2vec: Scalable feature learning for networks," in *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, 2016, pp. 855–864.

[8] S. Moran, D. McCloy, and R. Wright, "Phoible online," 2014.

[9] X. Li, S. Dalmia, D. Mortensen, J. Li, A. Black, and F. Metze, "Towards zero-shot learning for automatic phonemic transcription," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 05, 2020, pp. 8261–8268.

[10] G. I. Winata, G. Wang, C. Xiong, and S. Hoi, "Adapt-and-adjust: Overcoming the long-tail problem of multilingual speech recognition," *arXiv preprint arXiv:2012.01687*, 2020.

[11] S. Watanabe, T. Hori, and J. R. Hershey, "Language independent end-to-end architecture for joint language and speech recognition," in *IEEE Proceedings on Automatic Speech Recognition and Understanding*, 2017, pp. 265–271.

[12] J. Cho, M. K. Baskar, R. Li, M. Wiesner, S. H. Mallidi, N. Yalta, M. Karafiát, S. Watanabe, and T. Hori, "Multilingual sequence-to-sequence speech recognition: Architecture, transfer learning, and language modeling," in *Proc. IEEE Workshop on Spoken Language Technology (SLT)*, 2018, pp. 521–527.

[13] O. Adams, M. Wiesner, S. Watanabe, and D. Yarowsky, "Massive multilingual adversarial speech recognition," in *Proc. NAACL*, 2019.

[14] M. Rivière, A. Joulin, P.-E. Mazaré, and E. Dupoux, "Unsupervised pretraining transfers well across languages," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7414–7418.

[15] S. Schneider, A. Baevski, R. Collobert, and M. Auli, "wav2vec: Unsupervised pre-training for speech recognition," in *Interspeech*, 2019, pp. 3465–3469.

[16] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *34th Conference on Neural Information Processing Systems (NeurIPS)*, 2020.

[17] B. Li, T. Sainath, K. Sim, M. Bacchiani, E. Weinstein, P. Nguyen, Z. Chen, Y. Wu, and K. Rao, "Multi-dialect speech recognition with a single sequence-to-sequence model," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4749–4753, 2018.

[18] S. Toshniwal, T. N. Sainath, R. J. Weiss, B. Li, P. Moreno, E. Weinstein, and K. Rao, "Multilingual speech recognition with a single end-to-end model," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 4904–4908.

[19] A. Kannan, A. Datta, T. N. Sainath, E. Weinstein, B. Ramabhadran, Y. Wu, A. Bapna, Z. Chen, and S. Lee, "Large-Scale Multilingual Speech Recognition with a Streaming End-to-End Model," in *Proc. Interspeech*, 2019, pp. 2130–2134. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2019-2858

[20] X. Li, S. Dalmia, A. Black, and F. Metze, "Multilingual speech recognition with corpus relatedness sampling," in *INTERSPEECH*, 2019.

[21] J. Li and M. Hasegawa-Johnson, "Autosegmental neural nets: Should phones and tones be synchronous or asynchronous?" in *Interspeech*, 2020.

[22] M. Hasegawa-Johnson, P. Jyothi, D. McCloy, M. Mirbagheri, G. di Liberto, A. Das, B. Ekin, C. Liu, V. Manohar, H. Tang, E. C. Lalor, N. Chen, P. Hager, T. Kekona, R. Sloan, , and A. K. Lee, "Asr for under-resourced languages from probabilistic transcription," *IEEE/ACM Trans. Audio, Speech and Language*, vol. 25, no. 1, pp. 46–59, 2017.

[23] 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.

[24] X. Li, S. Dalmia, D. R. Mortensen, F. Metze, and A. W. Black, "Zero-shot learning for speech recognition with universal phonetic model," 2018.

[25] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an ASR corpus based on public domain audio books," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.

[26] N. Oostdijk, "The spoken dutch corpus. overview and first evaluation." in *LREC*. Athens, Greece, 2000, pp. 887–894.

[27] T. Schultz, N. T. Vu, and T. Schlippe, "Globalphone: A multilingual text & speech database in 20 languages," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013, pp. 8126–8130.

[28] S. Watanabe, T. Hori, S. Karita, T. Hayashi, J. Nishitoba, Y. Unno, N.-E. Y. Soplin, J. Heymann, M. Wiesner, N. Chen *et al.*, "Espnet: End-to-end speech processing toolkit," *Proc. Interspeech*, pp. 2207–2211, 2018.

[29] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The kaldi speech recognition toolkit," in *IEEE workshop on automatic speech recognition and understanding*. IEEE Signal Processing Society, 2011.

[30] M. Hasegawa-Johnson, L. Rolston, C. Goudeseune, G.-A. Levow, and K. Kirchhoff, "Grapheme-to-phoneme transduction for cross-language asr," in *International Conference on Statistical Language and Speech Processing*. Springer, 2020, pp. 3–19.

[31] S. Karita, N. Chen, T. Hayashi, T. Hori, H. Inaguma, Z. Jiang, M. Someki, N. E. Y. Soplin, R. Yamamoto, X. Wang *et al.*, "A comparative study on transformer vs RNN in speech applications," in *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2019, pp. 449–456.

[32] L. Van der Maaten and G. Hinton, "Visualizing data using t-SNE." *Journal of machine learning research*, vol. 9, no. 11, 2008.