IMPROVING MULTILINGUAL SPEECH RECOGNITION SYSTEMS

BY

HETING GAO

THESIS

Submitted in partial fulfillment of the requirements
for the degree of Master of Science in Electrical and Computer Engineering
in the Graduate College of the
University of Illinois Urbana-Champaign, 2021

Urbana, Illinois

Adviser:

Professor Mark Hasegawa-Johnson

ABSTRACT

End-to-end trainable deep neural networks have become the state-of-the-art architecture for automatic speech recognition (ASR), provided that the network is trained with a sufficiently large dataset. However, many existing languages are too sparsely resourced for deep learning networks to achieve as high accuracy as their resource-abundant counterparts.

Multilingual recognition systems mitigate data sparsity issues by training models on data from multiple language resources to learn a speech-to-text or speech-to-phone model universal to all languages. The resulting multilingual ASR models usually have better recognition accuracy than the models trained on the individual dataset.

In this work, we propose that two limitations exist for multilingual systems, and resolving the two limitations could result in improved recognition accuracy: (1) existing corpora are of the considerably varied form (spontaneous or read speech), corpus size, noise level, and phoneme distribution and the ASR models trained on the joint multilingual dataset have large performance disparities over different languages. We present an optimizable loss function, equal accuracy ratio (EAR), that measures the sequence-level performance disparity between different user groups and we show that explicitly optimizing this objective reduces the performance gap and improves the multilingual recognition accuracy. (2) While having good accuracy on the seen training language, the multilingual systems do not generalize well to unseen testing languages, which we refer to as cross-lingual recognition accuracy. We introduce language embedding using external linguistic typologies and show that such embedding can significantly increase both multilingual and cross-lingual accuracy. We illustrate the effectiveness of the proposed methods with experiments on multilingual and multi-user and multi-dialect corpora.

To my beloved father and mother.

ACKNOWLEDGMENTS

I gratefully acknowledge funding for my education and research from the Institute for Information & Communication Technology Promotion (IITP), IBM-UIUC Center for Cognitive Computing Systems Research (C3SR), and the ECE Department of University of Illinois at Urbana-Champaign.

I thank my adviser, Prof. Mark Hasegawa-Johnson, for his vision and guidance of this work, and my colleagues Junrui Ni and John Harvill, for their assistance and advice in the experiments.

I am grateful to my father and mother, Kaike Gao and Zhijun Luo, who raised me and educated me, and my family, who love me and support me.

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

Modern end-to-end neural network based speech recognition systems (ASR) have achieved great success in resource-rich languages such as English and Mandarin. These networks can reduce the word error rate to as low as 2% even in an open-vocabulary task [1, 2] when the networks are trained with a sufficiently large amount of data.

However, there are thousands of languages with billions of speakers that are resource-deficient [3]. These languages do not have large corpora of digitized text and recorded speech, making it hard for neural networks to achieve similar accuracy [4]. Multilingual speech recognition systems partially solve the resource-deficient problem by combining data from multiple languages so that an acoustic model universal to all languages is learned on the joint data pool. The resulting models usually outperform the models trained on an individual language dataset [5].

There have been a number of studies to further improve the performance of the ASR systems trained on the joint dataset, including enlarging the datasets for low-resource dialects [6], conducting language-dependent training [7] and language-dependent adaptation [8], and adding language/dialect related information as additional features [9]. While improving the ASR performance, these methods have addressed the performance disparity as a provisional resource problem, rather than treating the disparity as a problem equal in importance with the attainment of a low average word error rate.

On the other hand, research in data mining has demonstrated that artificial intelligence (AI) trained in an unfair environment will learn the unfairness of its teachers unless specifically instructed not to do so [10], [11], [12]. Inspired by methods in the AI fairness literature such as demographic parity, equal odds, and equal opportunity, that have been used to minimize discriminatory predictions based on race or gender, we seek to design multilingual ASR that works well for all languages and all users: a goal that has been described as

"inclusive speech technology" [13].

We attempt to reduce the performance disparity across different user groups. For this purpose, we propose a new measure derived from published measures of algorithmic fairness, which we call the equal accuracy ratio (EAR), and we integrate this measure into a standard neural speech recognizer to reduce the performance disparity in ASR systems. Experiments on multilingual, multi-dialect, and multi-user datasets show the proposed EAR measure reduces the performance disparity between different user groups without sacrificing the recognition accuracy. In fact, the experiments show EAR helps the network achieve higher average accuracy compared to the base model.

It is also desired that an ASR system trained on one set of languages not only performs well on the seen language set but also generalizes well to another set of unseen languages. For this purpose, we convert the language-specific transcriptions to language-universal phone transcriptions, i.e., International Phonetic Alphabet (IPA), and train a phone transcription model instead of a text transcription model. The phonetic recognition model is expected to capture acoustic information universal to all the languages and therefore have a decent recognition accuracy when tested on unseen languages. We refer to this setting as cross-lingual speech recognition.

While existing multilingual systems have a good performance on the seen training languages [14, 15, 16], they, unfortunately, do not have equally good accuracy when testing on 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 that are easily obtainable in the real world – phylogenetic similarity and phone inventory. For phylogenetic similarity, we extract phylogenetic information from Glottolog [17], 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 [18] 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 [19]. 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) without any transcribed data from the unseen test set. Ablation study shows that both the phylogenetic and phone inventory information are crucial for performance improvement.

CHAPTER 2

RELATED WORK

2.1 Speech Recognition

End-to-end neural network based speech recognition systems can achieve very high performance given sufficient training data. State-of-the-art deep neural architectures for speech recognition combine acoustic and linguistic encoders with a mechanism for reducing the length of the sequence, from a larger number of input frames to a smaller number of output symbols [20]. There is currently active research comparing the capabilities of connectionist temporal classification (CTC) [21], attention-based encoder-decoder structures [1], and hidden Markov models [20] for modifying the sequence length. CTC modifies the sequence length by ignoring repeated or blank phone symbols, thereby focusing the training procedure on a small number of conditionally independent frame classifications. Baidu's Deepspeech [22] is a recurrent neural network (RNN)-CTC model that achieves high performance in both English and Mandarin. Recent successful CTC-based models include LipNet [23] and the DeepMind RNN-CTC model [24].

Denote the spectral features of the i^{th} utterance as a set of frames $x^{(i)} = [x_1^{(i)}, x_2^{(i)}, \ldots, x_T^{(i)}]$ where T is the number of frames. Denote the reference transcription as $y^{(i)} = [y_1^{(i)}, y_2^{(i)}, \ldots, y_{S_i}^{(i)}] \in \mathcal{Y}^+$, and the ASR output hypothesis as $\hat{y}^{(i)} = [\hat{y}_1^{(i)}, \hat{y}_2^{(i)}, \ldots, \hat{y}_{\hat{S}_i}^{(i)}] \in \mathcal{Y}^+$, where S_i and $\hat{S}_i$ are the lengths of the reference and hypothesis transcriptions of i^{th} sample and $\mathcal{Y}$ is the set of all transcription characters. The true conditional probability distribution $p_{Y|X}(y|x)$ is unknown; the ASR computes an estimated distribution $p_{\hat{Y}|X}(y|x)$ in order to minimize the cross-entropy of the training corpus,

$$\mathcal{L}_{CE} = -\sum_{i=1}^{|\mathcal{S}|} \ln p_{\hat{Y}|X}(y^{(i)}|x^{(i)}), \tag{2.1}$$

where $\mathcal{S} = \left\{(x^{(1)}, y^{(1)}), \ldots, (x^{(|\mathcal{S}|)}, y^{(|\mathcal{S}|)})\right\}$ is a training corpus containing utterances with known transcriptions.

In connectionist temporal classification (CTC [21]) framework, the network performs time-scale modification by positing an alignment sequence, $\Pi^{(i)} = [\Pi_1^{(i)}, \ldots, \Pi_T^{(i)}]$ whose instance value is $\pi^{(i)} = [\pi_1^{(i)}, \ldots, \pi_T^{(i)}]$. Each time-aligned character $\pi_t^{(i)}$ is either one of the transcription characters ($\pi_t = y_s$ for some s), or $\pi_t = \varnothing$ where $\varnothing$ is a special "blank" character. For example, suppose we have a five-character text "hello" ($S = 5$) encoded in a 14-frame speech waveform ($T = 14$); the transcription and alignment might be

$$y = [h, e, l, l, o], \qquad \pi = [h, h, e, e, e, \varnothing, \varnothing, l, l, l, \varnothing, l, \varnothing, o].$$

Training data are often provided with only the transcriptions, and the alignment information is not given. If the alignments are known, it would be easier to estimate the cross-entropy given in Eq. (2.1) by taking the sum of the log probabilities of the correct alignment at each frame.

Since alignment is not known, CTC computes the cross-entropy by marginalizing over all the possible alignments that can be mapped to the true transcription using a surjective time-compression function defined as:

$$\mathcal{B} : (\mathcal{Y} \cup \{\varnothing\})^+ \to \mathcal{Y}^+.$$

A commonly used $\mathcal{B}$ first removes repeated labels and then removes all "blank" characters. For any valid alignment π, $\mathcal{B}(\pi)$ is a unique y. For any valid y , $\mathcal{B}^{-1}(y)$ is the set $\{\pi : \mathcal{B}(\pi) = y\}$. The log-probability of a transcription $y^{(i)}$ given the input frames $x^{(i)}$ can therefore be computed as

$$\begin{aligned}
\mathcal{L}_{CTC}^{(i)} &= -\ln p_{\hat{Y}|X}(y^{(i)}|x^{(i)}), \\
&= -\ln \sum_{\pi \in \mathcal{B}^{-1}(y)} \prod_{t=1}^{T} \exp(e_t(\pi_t)), \\
&= -\operatorname*{logsumexp}_{\pi \in \mathcal{B}^{-1}(y)} \sum_{t=1}^{T} e_t(\pi_t),
\end{aligned}$$

where $e_t(\pi_t)$ is the log output of a softmax layer predicting the transcription label at time t. The input of this softmax layer can be a bidirectional LSTM, Transformer, or other neural network parameterized by θ and having access

to the whole sequence x.

In the attention-based transducer framework [25], the log-probability is estimated by modeling each character output $y_s^{(i)}$ as a conditional distribution given the previous characters $y_{1:s-1}^{(i)}$ and the input signal $x^{(i)}$. Using the chain rule, the negative log-probability $\mathcal{L}_{ATT}$ is computed as

$$\begin{aligned}\mathcal{L}_{ATT}(x^{(i)}, y^{(i)}) &= -\ln p_{\hat{Y}|X}(y^{(i)}|x^{(i)}), \\ &= -\sum_{s=1}^{S} \ln p(y_s^{(i)}|x^{(i)}, y_{1:s-1}^{(i)}). \end{aligned} \tag{2.2}$$

The two frameworks can be used individually [22, 25] or hybridly [26]. When the two framework are used together, the training loss is the weighted summation over CTC loss $\mathcal{L}_{CTC}$ and attention loss $\mathcal{L}_{ATT}$ of each speech sample as

$$\mathcal{L}_{CE} = \sum_{i=1}^{|\mathcal{S}|} \alpha \mathcal{L}_{CTC}^{(i)} + (1-\alpha)\mathcal{L}_{ATT}^{(i)},$$

where $\alpha \in [0, 1]$ is a hyperparamter to be tuned; $alpha = 1$ means only CTC loss and $alpha = 0$ means only attention loss.

One concern with modern deep neural ASR models is that the model does not have equal performances for different user groups, potentially creating disparity of opportunity over region, age, gender, race, educational status, disability, class, etc. Experiments on the Japanese Newspaper Article Sentences corpus [27] show 10% higher word error rate for older voices than for younger voices [28]. A study examining YouTube's automatic captions reports lower accuracy for female speakers [29]. Experiments on a neural ASR system trained using seven different dialects of English from America, India, Britain, South Africa, Australia, Nigeria & Ghana, and Kenya report large disparities in word error rate ranging from 10.6% for American English to 33.4% for Ghana & Kenya English in dialect-dependent training [9]. Recognition systems trained on different Arabic dialects (Egyptian, Gulf, Levantine and Maghrebi) suffer similar error rate disparities [7], ranging from 26.3% for Maghrebi Arabic to 34.0% for Egyptian Arabic. Recently, a study on state-of-the-art ASR systems from Amazon, Apple, Google, IBM, and Microsoft reports that all these systems have obvious racial disparities [6]. The average

word error rate for the black speakers is twice as large as that of the white speakers.

Methods proposed in improving these models include adding group-specific features in the training [9], fine-tuning the model on data from each group of users, switching models based on group information [30], etc. These methods improve the accuracy for each user group and thus the overall accuracy of the model. However, they do not emphasize the inclusiveness of the ASR model. Our proposed method, the equal accuracy ratio, estimates the inclusiveness of the ASR during the training of a CTC-based sequence-to-sequence transcription model, and explicitly balances the relative importance of inclusiveness against the average error rate over a given set of training corpora.

2.2 Fairness Measures

A classifier trained on a corpus can be unfair toward certain groups of users due to historical bias or insufficient minority group training data. In one of the earliest studies of AI fairness, credit prediction models decided whether or not to accept a loan application [10]. Trained models were reported to be age-discriminatory according to the criterion of demographic parity, which requires that there be no difference between the average outcomes for different user groups:

$$|p_{\hat{Y}|A}(1|0) - p_{\hat{Y}|A}(1|1)| = 0, \tag{2.3}$$

where we define $p_{\hat{Y}|A}(y|a)$ to be the probability that the hypothesis result, $\hat{Y}$, takes value y, given that the protected attribute (e.g., age) has a value $A = a$. The demographic parity gap can be reduced by massaging dataset labels or giving more weight to samples from disadvantaged groups [11]. A recent paper [31] proposes that demographic parity can select data to create a fair training set, instead of modifying the training labels in an existing dataset.

Demographic parity is less useful when there is a desired or ground truth result, Y, which is known, and which is correlated with the protected attribute. If the desired result is correlated with A, then imposing Eq. (2.3) reduces accuracy.

When the ground truth is known and desirable, the equal odds and equal

opportunity criteria [12] are more desirable than demographic parity. The equal odds criterion requires conditional independence between hypothesis and attributes given ground truth, i.e.,

$$|p_{\hat{Y}|A,Y}(\hat{y}|0,y) - p_{\hat{Y}|A,Y}(\hat{y}|1,y)| = 0 \quad \forall y, \hat{y} \in \{0,1\}\,, \tag{2.4}$$

where Y is the ground truth and $\hat{Y}$ is the hypothesis. Equal odds requires that any particular mistake is made with equal probability, regardless of the setting of the protected attribute. "Equal opportunity" is a relaxation of equal odds, which focuses only on the error rate of the classifier: Eq. (2.4) is enforced only when the hypotheses match the ground truths ($\hat{y} = y$).

Predictive rate parity [32] takes a different perspective, arguing the prediction should reflect the real performance of the group. The ground truth should be conditionally independent of group attributes given the predictions:

$$|p_{Y|A,\hat{Y}}(y|0,\hat{y}) - p_{Y|A,\hat{Y}}(y|1,\hat{y})| = 0 \quad \forall y, \hat{y} \in \{0,1\}\,. \tag{2.5}$$

These fairness requirements can not all be simultaneously satisfied [33]; it is necessary for the users of a particular AI technology to decide which of these fairness criteria are the most desirable for their technology. Given such a fairness specification, a model can be trained to be fairer by optimizing the gap between predicted probabilities for each pair of groups. However, in their original published forms (as shown above), all of these criteria assume binary outcome variables ($\hat{Y}$). Extensions to real-valued and real-vector outcome variables have been published, but no previous study has published an extension to variable-length sequential outcome variables.

2.3 Language Embeddings

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 [16]

or language-specific decoders with private phone [15, 34, 35] or character inventories [36, 37, 38]. Multilingual ASR can benefit from the use of self-supervised pretraining algorithms such as contrastive predictive coding [39, 40, 41], 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, 16, 15, 34, 35] as they benefit from increased amount and diversity of data.

Language or dialect embedding has been shown to improve multilingual ASR systems [42, 43, 44, 45]. The embedding can be a one-hot vector specifying language ID [42, 44] or a vector learned from acoustic data under a standard multilingual model [43, 45] and can be used as additional input features to the network [42, 44], as adapter modules for language-specific adjustments [44] or as interpolation weights for the encoder [43]. 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 [46, 47] or a few hours [48] of transcribed data in the target language. Perhaps the prior work most similar to the proposed language embedding is a set of experiments using the Phoible [19] phoneme inventory of a language to define an untrained, knowledge-based linear output layer called the "signature matrix" [49, 16]; our phone token masking strategy is a simplification of the signature matrix, and our proposed language encoding is an enrichment of the same.

CHAPTER 3

ALGORITHMS

3.1 Equal Accuracy Ratio

The proposed equal accuracy ratio is an adaptation, to sequence-learning models, of the equal opportunity training criterion. Equal opportunity was defined in [12] as:

$$|p_{\hat{Y}|A,Y}(y|0,y) - p_{\hat{Y}|A,Y}(y|1,y)| = 0 \quad \forall y \in \mathcal{Y}. \tag{3.1}$$

There are three candidate definitions we can use for the purpose of adapting the equal opportunity criterion to ASR:

1. Matched frames: $p_{\hat{Y}|A,Y}(y|a,y)$ could be measured using sets of frames, with different values of the protected attribute $A = a$, for which the recognizer should output character y. However, matched frames would need a ground truth alignment, which are not required for CTC training, and are rare in practice.

2. Matched transcription: $p_{\hat{Y}|A,Y}(y|a,y)$ could be measured using sets of waveforms, with different values of the protected attribute, that has exactly the same transcription. Corpora that provide identical texts spoken by members of different groups exist (e.g., UASPEECH [50] and TIMIT [51]), but are rare and small.

3. Matched accuracy: The sentence accuracy of an ASR, for user group a, is given by

$$p_{\hat{Y}|A}(Y|a) = \sum_{y} p_{Y|A}(y|a) p_{\hat{Y}|A,Y}(y|a,y). \tag{3.2}$$

   Inclusiveness of an ASR might be reasonably defined to mean that

accuracy is the same for different demographic groups, even if they do not say exactly the same things. The equal-accuracy marginalization of Eq. (3.1) is

$$|p_{\hat{Y}|A}(Y|0) - p_{\hat{Y}|A}(Y|1)| = 0. \tag{3.3}$$

We will use definition 3 since it codifies the criterion that matters most to users (the accuracy of the speech recognizer), without requiring any additional constraints.

Extending equal opportunity in Eq. (3.3) to the multi-group case, the ASR provides equal opportunity if and only if

$$|p_{\hat{Y}|A}(Y|a) - p_{\hat{Y}|A}(Y|a')| = 0 \quad \forall a, a', \tag{3.4}$$

$$|\ln p_{\hat{Y}|A}(Y|a) - \ln p_{\hat{Y}|A}(Y|a')| = 0 \quad \forall a, a'. \tag{3.5}$$

Taking the logarithm on both sides of Eq. (3.5) does not alter the equality, but provides computational benefits as we will show shortly. After the manipulation, minimizing the gap on the left-hand side is actually forcing the ratio of accuracies to be one. Therefore we call the objective "**equal accuracy ratio**."

In practice, equal accuracy is rarely achieved. An ASR can be explicitly trained to minimize violations of equal accuracy, however, by training it to minimize the equal accuracy ratio, $\mathcal{L}_{EAR}$, defined as the total absolute difference between the cross-entropy rates of groups a and a', summed over all pairs of different groups:

$$\mathcal{L}_{EAR} = \frac{1}{2} \sum_{a,a'} \left| \ln p_{\hat{Y}|A}(Y|a) - \ln p_{\hat{Y}|A}(Y|a') \right| + C(\theta, a, a'), \tag{3.6}$$

where $C(\theta, a, a')$ is an offset term to be described shortly.

We do not have $p_{\hat{Y}|A}(Y|a)$, but we can estimate it based on the portion of the training data spoken by people from group a, thus

$$\ln p_{\hat{Y}|A}(Y|a) \approx \frac{1}{|S_a|} \sum_{x^{(i)}, y^{(i)} \in S_a} \ln p_{\hat{Y}|X}(y^{(i)}|x^{(i)}), \tag{3.7}$$

where $|S_a|$ is the number of training utterances available from group a. The

log-probability $p_{\hat{Y}|X}(y^{(i)}|x^{(i)})$ in Eq. (3.7) is no more than the negative cross-entropy loss of the training pair $(x^{(i)}, y^{(i)})$.

There are two possibilities to optimize $\mathcal{L}_{EAR}$, either increasing the performance of the group with lower accuracy or decreasing the performance of the one with higher accuracy. Apparently, the latter situation is not desirable. In order to avoid the latter, we can modify the equal accuracy ratio by adding an offset term, equal to the average of the two group-dependent cross entropies:

$$C(\theta, a, a') = -\frac{\ln p_{\hat{Y}|A}(Y|a) + \ln p_{\hat{Y}|A}(Y|a')}{2}. \tag{3.8}$$

With the offset constant defined in Eq. 3.8, the equal accuracy ratio becomes a weighted average of the per-group cross-entropy losses:

$$\begin{aligned}\mathcal{L}_{EAR} &= \sum_{a,a'} \max\big\{ -\ln p_{\hat{Y}|A}(Y|a), -\ln p_{\hat{Y}|A}(Y|a')\big\}, \\ &= -\sum_{a} N_{\leq a} \ln p_{\hat{Y}|A}(Y|a),\end{aligned} \tag{3.9}$$

where $N_{\leq a}$ is the number of other groups that have lower cross-entropy loss than group a. The resulting $\mathcal{L}_{EAR}$ as a measure of inclusiveness is intuitive. It is a weighted sum of cross-entropy loss over each dialect where the dialects with larger loss are given larger weights during training.

The loss function $\mathcal{L}_{CE}$ penalizes high average error rates, but ignores high inter-group error rate disparities; $\mathcal{L}_{EAR}$ penalizes high inter-group disparities, but ignores the error rate of the best-performing system. Multi-task training seeks to balance these two objectives by minimizing

$$\mathcal{L}_{MT} = \mathcal{L}_{CE} + \lambda \mathcal{L}_{EAR}, \tag{3.10}$$

where λ is a hyperparameter that can be tuned.

3.2 External Language Embedding

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

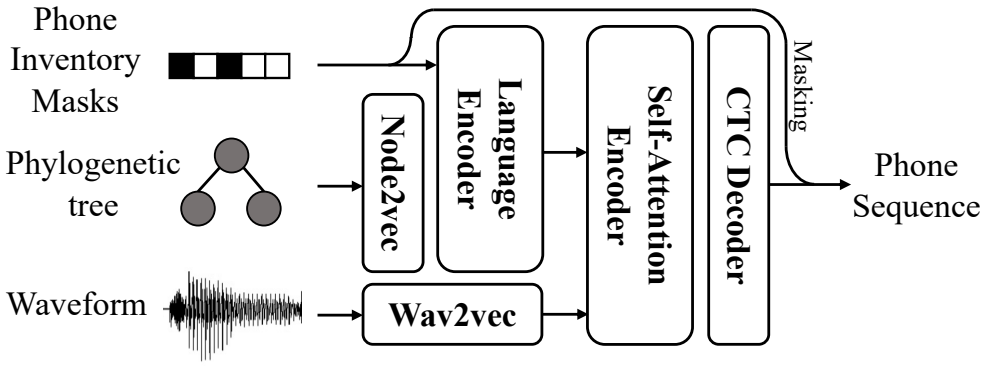


Figure 3.1: Architecture overview

testing language. Figure 3.1 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 [40], (2) phoneme inventory masking similar to [16], and (3) the proposed typology-based language encoder.

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 [18] 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 [16], we also include phone inventory information from Phoible [19], 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 [41]) from wav2vec2.0 that is pretrained on 960 hours of LibriSpeech [52].[1]

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

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.

CHAPTER 4

EXPERIMENTS

4.1 Datasets

4.1.1 Multi-Dialect Dataset

In order to better study the proposed EAR measure, a new benchmark dataset was created by combining existing datasets that cover seven racially, ethnically, and geographically diverse dialects of English as listed in Table 4.1. Most are retrieved from publicly available sources. Names for the American dialects (Standard American and African American) are based on the discussion in [53]. The Afrikaans English and Xhosa English corpora are named as in their source distribution [54]; the Latin American, UK Broadcast News and Indian English corpora are each named for the country or countries in which they were recorded.

The African American corpus is part of the Corpus of Regional African American Language (CORAAL) [55], which provides recorded conversational speech data from people who self-identify as African American, including audio recordings, time-aligned orthographic transcription and speaker information. We use the DCA version which focuses on African American Language in the Washington DC region. We omit speech by the interviewers and retain only speech segments by the self-identified speakers of African American Language.

The Standard American English corpus is collected from the LibriSpeech ASR corpus [52] of audiobook readings. Data from this corpus is not dialect-homogeneous: all regional dialects of the United States are represented, as are samples of dialects from outside the United States. Nevertheless, empirical results reported later in this work suggest that this corpus is more dialect-homogeneous than any of our other purportedly single-dialect cor-

Table 4.1: Sources of data used in our multi-dialect dataset "Abbr" column is the abbreviated dialect name used in performance tables. "#Utts" column shows the number of utterances in the training set. "Len" column shows the total duration of all utterances, in minutes.

Dialect	Abbr	Corpus	#Utts	Len
African American	AA	CORAAL	13908	491
Afrikaans Eng	AF	AST Afrikaans	3799	133
Standard American	AM	Librispeech	28533	6035
UK Broadcast News	BR	LDC95S24	10980	1221
Latin American	LA	LDC2014S05	281	28
Indian Eng	IN	MaheshChandra	358	16
Xhosa Eng	XH	AST Black	3323	116

pora, possibly because speakers modulate their speech, somewhat, to match a standard audiobook reading style. The data contains audio waveforms and their associated transcriptions. Librispeech is a very large corpus; we use only the "train-clean-100" partition.

The Latin American English corpus is extracted from Hispanic-English Database (LDC2014S05) [56] that contains a mixture of read speech and conversational speech along with their transcriptions. Participants were adult native speakers of Spanish as spoken in Central America and South America who resided in the Palo Alto, California area, had lived in the United States for at least one year and demonstrated a basic ability to understand, read and speak English. We only include the read speech part of the corpus.

UK Broadcast News is extracted from WSJCAM0 Cambridge Read News (LDC95S24) corpus [57]. The subjects in WSJCAM0 were native speakers of British English, reading in a standardized dialect. The corpus provides standard orthographic transcripts as well as time alignment between waveform and both word and phonetic transcriptions. The audio is originally in NIST SPHERE format and we convert it to wav format for fast data loading.

The AST Afrikaans English corpus and AST Xhosa English [54] are collected and published by African Speech Technology. AST Afrikaans English corpus contains a mixture of spontaneous and read speech by native speakers of Afrikaans, a language primarily spoken by white South Africans. AST Xhosa English is collected in a similar form but is spoken by native speakers of isiXhosa, a language primarily spoken by black South Africans. Both corpora are distributed in 8 kHz alaw format, and were converted to 16-bit

16 kHz wav files using `FFmpeg`.

The Indian English corpus is composed of three small corpora posted to Voxforge by Mahesh Chandra. Unlike other dialects, this corpus contains the speech of only one speaker, reading short sentences.

These corpora vary considerably in difficulty. The Indian English and Latin American corpora are difficult to recognize because they are small, and because the sampled dialects are quite different from the others in the list. The African American corpus is difficult because it is composed exclusively of spontaneous speech. The Afrikaans English and Xhosa English corpora each contain a mixture of spontaneous and read speech; the Standard American and UK Broadcast News corpora each contain exclusively read speech.

Transcriptions are cleaned by removing special characters and punctuation except apostrophe. We retain audio files with a duration longer than 1 second. Short-time Fourier transform (STFT) is computed using a 16 kHz sampling rate, and a Hamming window with a window size of 0.02 s and a window stride of 0.01 s. ASR features consist of the natural logarithm of one plus the magnitude of STFT, normalized by subtracting the mean and dividing by standard deviation.

4.1.2 CORAAL Dataset

The CORAAL dataset is part of the multi-dialect dataset, and has been described in Sec. 4.1.1 [55]. It contains complete and detailed information about both interviewers and interviewees. We extract interviewee information and identify four attributes, namely `age`, `work`, `education`, and `gender`, that are complete and meaningful to be used as sensitive attributes to partition the corpus. Details of the attributes are provided in Table 4.2. Additional experiments are performed on this dataset to verify the effectiveness of the equal accuracy ratio.

Table 4.2: Partition of CORAAL dataset. "Abbr" column shows the abbreviated group name used in performance tables. "#Utts" column shows the number of utterance in the training set. "Len" column shows the total duration in minutes of the utterances.

Attr	Group	Abbr	#Utts	Len
Age	-19		7320	250
	20-29		2776	104
	30-50		2590	99
	51+		1122	37
Work	Lower Working Class	LW	3516	125
	Upper Working Class	UW	4359	146
	Lower Middle Class	LM	3647	131
	Upper Middle Class	UM	1159	46
	Upper Class	U	824	28
	Unknown.	Unk	403	13
Edu	Elementary School	ES	169	6
	Student in Middle School	StMS	3190	107
	Student in High School	StHS	3510	118
	Some High School.	SHS	1206	41
	High School	HS	3156	108
	Student in College	StCO	192	7
	Some College	SCO	1485	63
	College	CO	847	32
	Graduate School	GS	153	5
Gender	Male	M	9155	317
	Female	F	4753	174

4.1.3 Multilingual Dataset

The dataset used for multilingual and cross-lingual experiments is a corpus that consists of 20 languages: 8 from IARPA Babel project corpora, 1 from CGN (Spoken Dutch Corpus) [58] and 11 from Globalphone [59] (GP) as summarized in Table 4.3. We only use the read speech part of CGN. 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.

Table 4.3: 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. "EAR" column shows if the language is used in equal accuracy ratio experiment. "LE" column shows if the language is used in external language embedding experiment.

Language	**Abbr**	**Corpus**	**Type**	**Family**	**Len**	**EAR**	**LE**
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	✗	✓

4.2 Equal Accuracy Ratio Experiment Settings

4.2.1 Multi-Dialect Dataset Setttings

The multi-dialect dataset was split into train, dev, and test sub-corpora with a ratio of 8:1:1. The training subcorpus was used to train a `deepspeech` network [22].[1] The deepspeech model has two convolutional layers, each with batch normalization and tanh activation. The convolution kernel sizes are 41×11 and 21×11 respectively. The convolution output is passed to five batch-normalized bidirectional LSTM layers, whose output is fed into a fully

[1] `https://github.com/SeanNaren/deepspeech.pytorch`

connected layer. The output is softmaxed to predict a label distribution at each frame, which is then used to compute CTC loss $\mathcal{L}_{CTC}$ and EAR loss $\mathcal{L}_{EAR}$ in Eq. (3.10).

Ideally, the average CTC loss for each dialect should be calculated within a large batch containing all utterances. Due to GPU memory limitations, our model can only be trained with a batch size of 16. The average CTC loss of each dialect is calculated cumulatively with new incoming batches within an epoch, so as the training proceeds, the estimated average CTC loss and equal accuracy ratio become increasingly accurate. The model is optimized using Adam optimizer with a learning rate of 0.001. Each model is trained for 30 epochs and the model with the least validation cross-entropy loss is used to calculate performance scores. Models are evaluated on both the development dataset and test dataset with character error rate (CER) as evaluation metrics.

4.2.2 CORAAL Dataset Setting

The CORAAL corpus is too small to train a complete deepspeech network; initial experiments overfit the training dataset, producing unstable performance on the development dataset. Since the equal accuracy ratio regularization does not depend on any specific neural architecture, we turn to a much simpler RNN-CTC architecture for a more stable performance measure. The model is composed of four layers of bidirectional LSTM with 128-dimensional hidden states, a batch normalization layer, and four fully connected layers, each with `tanh` activation. The output of the last layer is softmaxed and is used to compute CTC loss. Due to the simple RNN-CTC architecture, we are able to increase the batch size to 32 in training. Other settings are the same as those of dialect experiments.

The CORAAL corpus contains metadata about each interviewee, including `age`, `work`, `education`, and `gender` (Table 4.2). We ran four sets of experiments using this corpus, each of which used one of the metadata variables as an attribute protected by the equal accuracy ratio training criterion.

4.2.3 Multilingual Dataset Setting

We evaluate the EAR model on a subset of 14 languages (Bengali, Vietnamese, Zulu, Amharic, Javanese, Georgian, Dutch, Czech, French, Mandarin, Thai, Cantonese, Lao and Spanish) from the 20-language multilingual corpus as shown in Table 4.3. We use ESPnet as our ASR framework [60], which offers a complete ASR pipeline including data preprocessing, Transformer network implementation, network training and decoding. Due to the sampling rate difference between different corpora, we first upsample all audio signals to 16 kHz. Using Kaldi [61], we then extract 80-dim log Mel spectral coefficients with 25 ms frame size and 10 ms shift between frames, and augment the frame vectors with three extra dimensions for pitch features. The transcriptions are converted to IPA symbols using LanguageNet grapheme-to-phone (G2P) [62] models and the unique IPA symbols, including base phones, diacritics and suprasegmentals, in all 14 languages are collected as the shared phonetic token inventory. The resulting inventory size is 102.

The encoder of our model architecture is similar to the transformer architecture in [26], which starts with two 2D convolutional layers with a subsampling factor of 4, followed by 12 self-attention encoder layers, each having four heads, an attention dimension of 256 and a 2048-dim position-wise feed-forward layer. The decoder has six layers, each having a four-head, 256-dim self-attention layer to encode masked transcriptions, a four-head, 256-dim attention layer to align the encoded spectral feature with encoded transcriptions, and a 2048-dim position-wise feed-forward layer. The CTC loss weight is set to $\alpha = 0.3$.

The transformer model is trained on two GeForce RTX 2070 Graphics Card with half precision using `apex`.[2] We use dynamic batch size such that each batch contains at most 4200k frames during training.

[2] `https://github.com/NVIDIA/apex`

4.3 External Language Embedding Experiment Settings

4.3.1 Multilingual Dataset Setting

The data processing of the multilingual dataset for external language embedding experiments is almost the same as that for equal accuracy ratio experiment in Sec 4.2.3. Since the external language embedding experiment has a cross-lingual setting, 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 therefore 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/.

We experiment with two types of transformations to generate the language embedding, a three-layer fully connected transformation and a three-layer graph-convolutional transformation[3] on the language representations extracted from Glottolog [17] and Phoible [19]. 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.

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 [41] model that directly takes the 16 kHz waveform as input. We fix the weights of the wav2vec feature extractor during training. The encoder of our model architecture is similar to the one introduced in Sec 4.2.3. 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 [48]. Therefore, we discard the self-attention decoder in [26] and apply a dense layer to the encoder output to

[3] `https://github.com/tkipf/gcn`

compute the frame-wise phoneme posteriors and the CTC loss.

CHAPTER 5

RESULTS

5.1 Equal Accuracy Ratio Results

5.1.1 Multi-Dialect Results

Multi-dialect recognition experiments are tested with different λ values ranging from 0 to 1 where $\lambda = 0$ is the baseline, and $\lambda = 1$ gives equal weight to the cross-entropy and worst-cross-entropy loss terms. Resulting character error rates (CER: percent) are listed in Table 5.1. Mean CER is the average per-dialect character error rate, averaged uniformly over the seven dialect groups. The standard deviation of CER over all dialect groups is a measure of fairness. In general, we find that increased emphasis on the equal accuracy ratio (higher λ) results in a lower standard deviation of the CER (increased fairness).

Table 5.1: Multi-dialect experiments. Refer to Table 4.1 for the meanings of the abbreviations.

Dialect	λ=0	λ=0.001	λ=0.01	λ=0.1	λ=1	λ=10
AA	43.08	**39.07**	42.99	44.28	45.72	46.36
AF	20.88	**18.18**	23.70	22.26	24.81	20.98
AM	14.19	**10.94**	13.73	14.50	18.21	16.12
BR	14.56	**12.21**	17.36	17.09	19.23	16.98
IN	52.80	51.38	**50.95**	51.36	53.67	52.80
LA	38.41	**30.00**	41.70	36.28	32.14	36.46
XH	26.60	**22.11**	29.29	27.58	28.26	26.43
Mean	30.07	**26.27**	31.39	30.48	31.72	30.87
Std	14.97	14.85	14.11	13.97	**13.39**	14.61

5.1.2 CORAAL Results

Table 5.2: Character error rate (CER: percent), measured as a function of dialect and of the regularization weight λ, in experiments using CORAAL. Refer to Table 4.2 for the meanings of the group abbreviations.

		Multitask Regularization Weight λ					
		0	0.001	0.01	0.1	1	10
Age	-19	55.59	56.60	**53.96**	56.23	55.94	56.72
	20-30	55.56	55.99	**53.73**	55.82	56.60	57.13
	30-50	56.31	56.99	**54.94**	56.24	56.61	57.04
	50+	59.31	59.97	58.59	**58.53**	59.33	59.79
	Mean	56.69	57.39	**55.30**	56.70	57.12	57.67
	Std	1.78	1.77	2.25	**1.23**	1.50	1.42
Work	LM	56.16	**54.97**	58.03	55.64	57.05	56.90
	LW	55.30	**54.30**	57.44	55.06	56.76	55.60
	UW	56.03	**54.68**	58.32	55.55	56.96	56.81
	UM	58.01	**55.62**	58.27	55.69	58.15	57.78
	U	58.76	**57.25**	59.06	57.33	59.31	57.99
	Unk	56.86	**54.71**	57.41	57.46	56.71	56.36
	Mean	56.85	**55.26**	58.09	56.12	57.49	56.91
	Std	1.31	1.07	**0.62**	1.01	1.04	0.89
Edu	ES	61.94	61.00	61.54	62.35	**59.24**	60.19
	StMS	55.54	**54.86**	55.95	57.03	57.28	56.93
	StHS	55.40	**54.55**	56.48	57.31	56.71	55.83
	SHS	**55.20**	55.25	56.70	57.73	56.87	55.57
	HS	57.27	**56.04**	58.63	59.13	58.06	56.69
	StCO	**51.95**	53.25	55.03	59.17	54.79	57.28
	SCO	56.12	**55.54**	57.27	57.99	57.48	56.65
	CO	54.18	**53.79**	55.70	55.62	55.28	55.04
	GS	**54.42**	54.97	54.83	57.04	56.22	55.39
	Mean	55.78	**55.47**	56.90	58.15	56.88	56.62
	Std	2.74	2.24	2.09	1.92	**1.36**	1.54
Gender	M	55.74	55.28	55.55	57.32	58.07	**55.21**
	F	55.93	56.41	55.56	57.57	57.46	**55.44**
	Mean	55.84	55.85	55.55	57.45	57.76	**55.32**
	Std	0.13	0.80	**0.01**	0.17	0.43	0.16

Table 5.2 provides experimental results from the CORAAL corpus. Four sets of experiments were conducted using this dataset, each treating a different metavariable as the protected attribute: `age`, `work`, `education`, and `gender`. The figures in the table are character error rates (CER: percent). Optimal

Table 5.3: Multilingual Experiment Results. The numbers presented in the table are PTER in percentage.

Language	λ_{ctc}=0 λ_{att}=0	λ_{ctc}=0.1 λ_{att}=0	λ_{ctc}=0.1 λ_{att}=0.1
Amharic	43.5	41.7	**41.0**
Bengali	38.7	38.2	**37.9**
Cantonese	37.6	35.8	**35.1**
Javanese	**49.1**	51.6	52.2
Vietnamese	51.2	48.5	**47.0**
Zulu	45.2	41.8	**40.7**
Georgian	37.3	36.3	**35.4**
Lao	44.8	42.3	**41.3**
Dutch	**18.2**	18.4	18.4
Czech	10.9	10.8	**10.6**
French	**13.1**	13.4	13.3
Mandarin	**21.8**	**21.8**	22.3
Spanish	**11.3**	11.4	11.6
Thai	22.5	**21.1**	**21.1**
Mean	31.84	30.94	**30.56**
Std	14.32	13.72	**13.42**

inclusiveness (minimum inter-group standard deviation) is achieved for different values of λ, depending on the metavariable being protected. According to the empirical results, variation across different educational groups is minimized with $\lambda = 1$, variation across age groups is minimized with $\lambda = 0.1$, and variation across work or gender is minimized with $\lambda = 0.01$. Similarly, optimal average accuracy (minimal inter-group average error rate) is achieved for different values of λ, ranging from $\lambda = 0.001$ (`age` and `education`) to $\lambda = 10$ (`gender`).

5.1.3 Multilingual Results

The phone token error rate is summarized in Table 5.3. We conduct three sets of experiments with different settings of λ_{ctc} and λ_{att}. The leftmost result is the baseline where $\lambda_{ctc} = 0$ and $\lambda_{att} = 0$ denote no regularization. The middle column applies regularization only on CTC loss and the rightmost column applies regularization on both CTC and attention loss. λ values are set according to preliminary experiments on multi-dialect and multi-user

corpora. EAR stands for "equal accuracy ratio" and is calculated using the formula in Eq. (3.9).

From the table, we observe that fairness-loss regularization helps in increasing the overall accuracy and fairness of the Transformer model. With regularization on both CTC and attention loss, the model achieves the lowest phonetic token error rate (PTER) and the lowest equal accuracy ratio gap at the same time, with 1.25% absolute reduction in average PTER and 2.15% absolute reduction in EAR gap.

The equal-accuracy ratio regularization formulated in Eq. (3.9) is a weighted sum of losses over all languages with more weights assigned to the languages having higher losses. Such regularization is expected to increase the recognition accuracy for difficult languages, while sacrificing the accuracy in the easier languages. This is exactly what we observe from Table 5.3. The baseline model achieves slightly better accuracy on the languages that have around 10% to 20% PTER, i.e., the languages with read speech data (Dutch, French, Mandarin and Spanish). On the other hand, the regularized model achieves better accuracy on the languages that have over 30% PTER, i.e., the languages with spontaneous speech data (Amharic, Bengali, Cantonese, Vietnamese, Zulu, Geogian and Lao).

The PTER reduction on difficult languages brought by equal accuracy ratio regularization is larger than the PTER increase on easy languages. Therefore we obtain a model with better overall performance and fairness.

5.2 External Language Embedding Results

5.2.1 Multilingual and Cross-lingual Phonetic Recognition

We train and test on our 20-language dataset with seven different models, namely, "*Base*", "*W2V*", "*W2VM*", "*W2VL*", "*W2VLM*", "*W2VG*", "*W2VGM*". 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 "*L*" and "*G*" labels have an additional linear or graph-convolutional transformation network to compute the language embeddings respectively. Models with "*M*" apply phone inventory

Table 5.4: 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" is the average PTER over the 15 seen languages and the column "AvgU" is the average PTER over the 5 unseen languages.

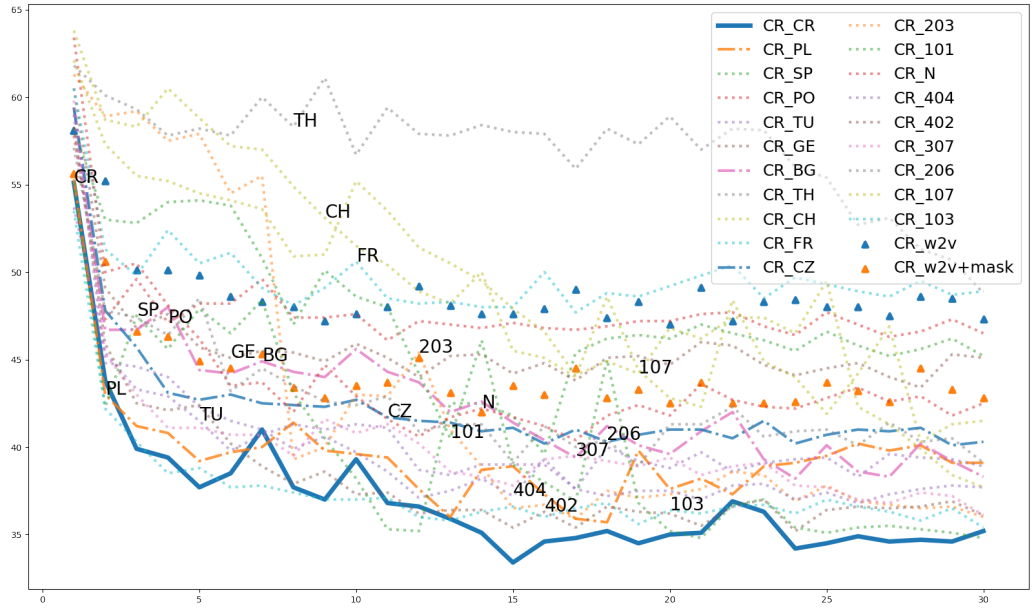
Exp	*Base*	*W2V*	*W2VM*	*W2VL*	*W2VLM*	*W2VG*	*W2VGM*
103	40.2	41.3	41.1	39	39	**38.2**	**38.2**
107	52.3	36.6	36.6	32.6	32.6	**32.0**	**32.0**
206	42.4	39	38.8	35.9	35.9	**35.2**	**35.2**
307	44.7	43.1	43.1	39.1	39.1	**38.0**	**38.0**
402	47	48.9	48.4	44.9	44.9	**44.2**	**44.2**
404	**38.0**	42.2	41.7	39.1	39.1	38.6	38.6
N	21.3	15.3	15.3	14	14	**13.2**	**13.2**
CZ	11	10.5	10.5	9.1	9.1	**8.5**	**8.5**
FR	13.7	14.8	14.8	12.9	12.9	**12.1**	**12.1**
CH	30	17.2	17.2	15.9	15.9	**15.5**	**15.5**
TH	26.1	22.2	22.2	19.9	19.9	**18.9**	**18.9**
GE	26.1	25.1	25.1	23.2	23.2	**22.3**	**22.3**
PO	18.4	18.7	18.7	16.3	16.3	**16.0**	**16.0**
TU	21.3	21	21	19.3	19.3	**18.4**	**18.4**
BG	27	30.2	30.2	28.2	28.2	**26.9**	**26.9**
101	77	77.9	76.5	74.6	**73.1**	76.1	**73.1**
203	78.2	79.3	76.8	76.3	72.8	72.4	**69.3**
CR	47.8	47.3	42.8	41.3	**35.2**	50.8	39.6
SP	38.1	39	36.8	37.3	**34.4**	37.5	35.3
PL	62.5	66.7	61.2	59.8	**54.0**	61.9	56.3
AvgS	30.6	28.4	28.3	26	26	**25.2**	**25.2**
AvgU	60.7	62	58.8	57.9	**53.9**	59.7	54.7

masking to the softmax output layer of the decoder.

The performance is shown in Table 5.4, where both proposed models ("*W2VLM*" and "*W2VGM*") outperform the "*Base*" model; "*W2VGM*" model achieves the lowest multilingual error rate, while "*W2VLM*" 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.

Figure 5.1: PTER of "*W2VLM*" model tested on Croatian with correct and fake language labels.

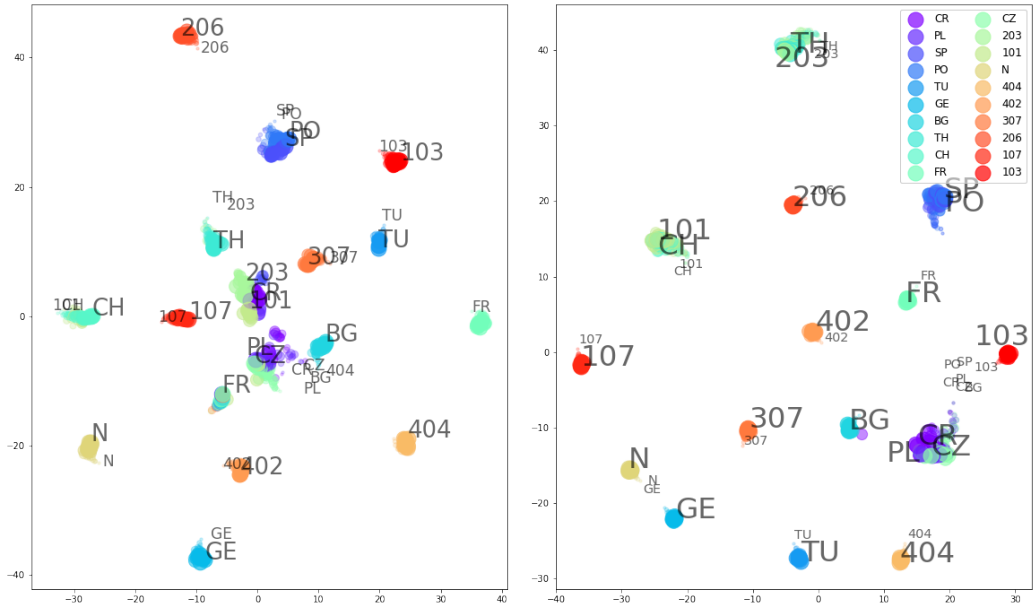


By comparing the average test PTER (AvgU) of "*W2V*", "*W2VL*" and "*W2VG*" with that of "*W2VM*", "*W2VLM*" and "*W2VGM*", we see that masking out the non-existing phonetic tokens in the test language greatly improves the recognition accuracy, possibly due to the reduced prediction space. The "*W2VGM*" model, which places the most emphasis on language-family structure, gains the largest improvement from phone masking, but still does not outperform the "*W2VLM*" model, suggesting that applying the graph constraint a second time (GCN on top of node2vec embeddings) provides no extra reduction of error rates.

5.2.2 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 5.1 shows the PTER of "*W2VLM*" model tested on Croatian. The blue and orange triangle points are PTERs of the "*W2V*" and "*W2VM*" models respectively. The blue solid line labeled "CR_CR" is the PTER curve with correct Croatian embedding and the dash-

Figure 5.2: t-SNE plot of language embedding. The left side is the plot of the embeddings from "*W2VL*" and the right side is from "*W2VG*".



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 ("*W2VM*"). 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.

5.2.3 Visualization of Language Embedding

We visualize the language embeddings of "*W2VL*" and "*W2VG*" using t-SNE [63] in Figure 5.2. 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

Table 5.5: Phonetic token error rates (PTER) Ablation Study.

W2VLM	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

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) toward 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 5.4, this flexibility reduces the cross-lingual error rate.

5.2.4 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 "*W2VL*" model with only Glottolog vector, with only Phoible vector, and with both. The results are shown in Table 5.5. First, providing external information reduces error: all three settings ("glotto", "phoible", "glotto+phoible") beat the "*W2VM*" 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 "*W2VGM*" 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; "*W2VLM*" 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.

CHAPTER 6

DISCUSSION

There are over 7000 languages [64] in the world and most of them lack a sufficient amount of digitized speech-text data for the training of ASR. By training multilingual ASR systems on data from different languages, we seek to improve ASR's performance on these low-resources languages. This idea has been shown promising by several previous works on multilingual speech recognition systems [5, 65].

However, simply putting together data from different languages without any consideration of the varied recording quality, speech type (read vs. spontaneous) and the inherent difficulty of the languages can end up in an ASR system having large performance disparity across different languages, as suggested by our multilingual experiment result in Table 5.3. Such observations apply not only to the multilingual dataset, but also to the multi-dialect dataset in Table 5.1. We, therefore, develop the EAR loss that quantifies the sequence-level performance disparity and show that explicitly optimizing the EAR loss as a regularization term together with the ASR loss can reduce the performance gap across different linguistic and dialectical user groups. We further test the EAR loss on the CORAAL dataset, a multi-user-group dataset defined by age, work, education and gender, and obtain similar results in Table 5.2.

Furthermore, the EAR experiments on the three datasets all indicate that it is possible to reduce the performance gap while improving the average ASR accuracy over all user groups provided we carefully tune the regularization factor. The improvement in the overall ASR accuracy might be explained by the fact we are assigning different weights to different user groups in the datasets defined by languages, dialects and other user attributes, which implicitly incorporate the group information into the model's training.

The cross-lingual experiments in Sec 5.2.1 explicitly incorporate the group information through external language embeddings. These external embed-

dings are extracted from open linguistic resources, Glottolog and Phoible. Node2vec and binary encoding ensure that phylogenetically and phonetically similar languages are embedded closely in the latent space. Comparing the PTERs of the models without language embedding ("*W2V*") with that of the models with language embedding ("*W2VL*" and "*W2VG*"), we observe that incorporating group (language) information improves the multilingual recognition accuracy on the seen training languages.

The "*W2VL*" and the "*W2VG*" models outperform the "*W2V*" not only on the seen training language set but also on the unseen testing language set. This result indicates that the language embeddings provide consistent and meaningful group information for both seen and unseen language sets and that the group information about the languages can improve the model's out-domain generalization even if the model is not trained on the out-domain data.

The proposed language embedding using Glottolog [17] and Phoible [19] works for thousands of existing languages as Glottolog tree contains around 8500 dialect, languages and languages families and Phoible contains phoneme inventory for around 2000 languages. A large multilingual ASR system with language embeddings could be trained using as many as the languages recorded in the two databases. Considering the varied difficulty in the training languages, EAR could be used to further improve the multilingual ASR system.

CHAPTER 7

CONCLUSION

In this work, we improve the multilingual speech recognition systems by proposing a novel inclusiveness loss, equal accuracy ratio, that can be seamlessly integrated into ASR neural architecture a regularizer to explicitly reduce the performance disparity of the system and by incorporating external linguistic typological knowledge to guide the neural model during training and testing.

We illustrate the effectiveness of the proposed EAR loss with experiments on a multi-dialect corpus, a multi-user corpus, and a multilingual corpus. Our results demonstrate that the models with fairness regularization generally have a smaller performance gap among user groups, without increasing the overall average error rate, as compared to the baseline.

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.

REFERENCES

[1] J. Li, V. Lavrukhin, B. Ginsburg, R. Leary, O. Kuchaiev, J. M. Cohen, H. Nguyen, and R. T. Gadde, "Jasper: An end-to-end convolutional neural acoustic model," 2019, pp. 71–75.

[2] 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.

[3] E. Bender, "The benderrule: On naming the languages we study and why it matters," *The Gradient*, 2019.

[4] S. Zhou, S. Xu, and B. Xu, "Multilingual end-to-end speech recognition with a single transformer on low-resource languages," *arXiv preprint arXiv:1806.05059*, 2018.

[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] A. Koenecke, A. Nam, E. Lake, J. Nudell, M. Quartey, Z. Mengesha, C. Toups, J. R. Rickford, D. Jurafsky, and S. Goel, "Racial disparities in automated speech recognition," *Proceedings of the National Academy of Sciences*, vol. 117, no. 14, pp. 7684–7689, 2020.

[7] M. G. Elfeky, P. Moreno, and V. Soto, "Multi-dialectical languages effect on speech recognition: Too much choice can hurt," *Procedia Computer Science*, vol. 128, pp. 1–8, 2018.

[8] G. I. Winata, S. Cahyawijaya, Z. Liu, Z. Lin, A. Madotto, P. Xu, and P. Fung, "Learning fast adaptation on cross-accented speech recognition," *arXiv preprint arXiv:2003.01901*, 2020.

[9] B. Li, T. N. Sainath, K. C. 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," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 4749–4753.

[10] F. Kamiran and T. Calders, "Classifying without discriminating," in *2009 2nd International Conference on Computer, Control and Communication*, 2009, pp. 1–6.

[11] T. Calders, F. Kamiran, and M. Pechenizkiy, "Building classifiers with independency constraints," in *2009 IEEE International Conference on Data Mining Workshops*. IEEE, 2009, pp. 13–18.

[12] M. Hardt, E. Price, and N. Srebro, "Equality of opportunity in supervised learning," in *Advances in Neural Information Processing Systems*, 2016, pp. 3315–3323.

[13] O. Scharenborg, "Inclusive speech technology: Developing automatic speech recognition for everyone," June 2021, webinar delivered to the TU Delft Safety and Security Institute and Campus, The Hague, The Netherlands.

[14] 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.

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

[16] 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 *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 8249–8253.

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

[18] 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.

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

[20] R. Schlüter, "Survey talk: Modeling in automatic speech recognition: Beyond hidden markov models," 2019, survey talk presented at Interspeech 2019.

[21] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks," in *Internat. Conf. Machine Learning (ICML)*, 2006, pp. 369–376.

[22] A. Hannun, C. Case, J. Casper, B. Catanzaro, G. Diamos, E. Elsen, R. Prenger, S. Satheesh, S. Sengupta, A. Coates et al., "Deep speech: Scaling up end-to-end speech recognition," *arXiv preprint arXiv:1412.5567*, 2014.

[23] Y. M. Assael, B. Shillingford, S. Whiteson, and N. de Freitas, "Lipnet: Sentence-level lipreading," *arXiv preprint arXiv:1611.01599*, vol. 2, no. 8, 2016.

[24] B. Shillingford, Y. M. Assael, M. W. Hoffman, T. Paine, C. Hughes, U. Prabhu, H. Liao, H. Sak, K. Rao, L. Bennett, M. Mulville, B. Coppin, B. Laurie, A. W. Senior, and N. de Freitas, "Large-scale visual speech recognition," *CoRR*, vol. abs/1807.05162, 2018. [Online]. Available: http://arxiv.org/abs/1807.05162

[25] W. Chan, N. Jaitly, Q. Le, and O. Vinyals, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 4960–4964.

[26] 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.

[27] K. Itou, M. Yamamoto, K. Takeda, T. Takezawa, T. Matsuoka, T. Kobayashi, K. Shikano, and S. Itahashi, "JNAS: Japanese speech corpus for large vocabulary continuous speech recognition research," *Journal of the Acoustical Society of Japan (E)*, vol. 20, no. 3, pp. 199–206, 1999.

[28] R. Vipperla, "Automatic speech recognition for ageing voices," Ph.D. dissertation, The University of Edinburgh, 2011.

[29] R. Tatman, "Gender and dialect bias in YouTube's automatic captions," in *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, 2017, pp. 53–59.

[30] X. Yang, K. Audhkhasi, A. Rosenberg, S. Thomas, B. Ramabhadran, and M. Hasegawa-Johnson, "Joint modeling of accents and acoustics for multi-accent speech recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 1–5.

[31] H. Anahideh and A. Asudeh, "Fair active learning," *arXiv preprint arXiv:2001.01796*, 2020.

[32] M. B. Zafar, I. Valera, M. Gomez Rodriguez, and K. P. Gummadi, "Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment," in *Proceedings of the 26th International Conference on World Wide Web*, 2017, pp. 1171–1180.

[33] J. Kleinberg, S. Mullainathan, and M. Raghavan, "Inherent trade-offs in the fair determination of risk scores," in *8th Innovations in Theoretical Computer Science Conference (ITCS 2017)*, ser. Leibniz International Proceedings in Informatics (LIPIcs), C. H. Papadimitriou, Ed., vol. 67. Dagstuhl, Germany: Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, 2017. [Online]. Available: http://drops.dagstuhl.de/opus/volltexte/2017/8156 pp. 43:1–43:23.

[34] 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.

[35] 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.

[36] 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.

[37] 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.

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

[39] 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.

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

[41] 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.

[42] 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.

[43] 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.

[44] 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. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2019-2858 pp. 2130–2134.

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

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

[47] 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.

[48] 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.

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

[50] H. Kim, M. Hasegawa-Johnson, A. Perlman, J. Gunderson, T. Huang, K. Watkin, and S. Frame, "Dysarthric speech database for universal access research," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 09 2008, pp. 1741–1744.

[51] L. F. Lamel, R. H. Kassel, and S. Seneff, "Speech database development: Design and analysis of the acoustic-phonetic corpus," in *Proc. of the DARPA Speech Recognition Workshop*, February 1986, pp. 100–109.

[52] 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.

[53] W. Wolfram and N. Schilling, *American English: Dialects and Variation*. New York: Jon Wiley & Sons, 2015.

[54] J. C. Roux, P. H. Louw, and T. Niesler, "The African speech technology project: An assessment." in *Language Resources and Evaluation Conference LREC*. European Language Resources Association ELRA, 2004, pp. 93–96.

[55] T. Kendall and C. Farrington, "The corpus of regional African American language," *Version*, vol. 6, p. 1, 2018.

[56] W. Byrne et al., "Hispanic–English database ldc2014s05," 2014.

[57] T. Robinson et al., "WSJCAM0 Cambridge Read News (LDC95S24). Linguistic Data Consortium, Philadelphia, PA." 1995.

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

[59] 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.

[60] 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 2018*, pp. 2207–2211, 2018.

[61] 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 2011 workshop on automatic speech recognition and understanding*, no. CONF. IEEE Signal Processing Society, 2011.

[62] 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.

[63] L. Van der Maaten and G. Hinton, "Visualizing data using t-SNE." *Journal of Machine Learning Research*, vol. 9, no. 11, 2008.

[64] D. M. Eberhard, G. F. Simons, and C. D. Fennig, "Ethnologue: Languages of the world," 2021, twenty-fourth edition. Dallas, Texas: SIL International. [Online]. Available: http://www.ethnologue.com.

[65] V. Pratap, A. Sriram, P. Tomasello, A. Hannun, V. Liptchinsky, G. Synnaeve, and R. Collobert, "Massively multilingual ASR: 50 languages, 1 model, 1 billion parameters," *arXiv preprint arXiv:2007.03001*, 2020.