Differentiable Allophone Graphs for Language-Universal Speech Recognition

Brian Yan, Siddharth Dalmia, David R. Mortensen, Florian Metze, Shinji Watanabe

Language Technologies Institute, Carnegie Mellon University, USA

{byan,sdalmia}@cs.cmu.edu

Abstract

Building language-universal speech recognition systems entails producing phonological units of spoken sound that can be shared across languages. While speech annotations at the language-specific phoneme or surface levels are readily available, annotations at a universal phone level are relatively rare and difficult to produce. In this work, we present a general framework to derive phone-level supervision from only phonemic transcriptions and phone-to-phoneme mappings with *learnable* weights represented using weighted finite-state transducers, which we call *differentiable allophone graphs*. By training multilingually, we build a universal phone-based speech recognition model with interpretable probabilistic phone-to-phoneme mappings for each language. These phone-based systems with learned allophone graphs can be used by linguists to document new languages, build phone-based lexicons that capture rich pronunciation variations, and re-evaluate the allophone mappings of seen language. We demonstrate the aforementioned benefits of our proposed framework with a system trained on 7 diverse languages.

Index Terms: universal phone recognition, differentiable WFST, multilingual ASR, phonetic pronunciation, allophones

1. Introduction

The objective of language-universal speech recognition is to indiscriminately process utterances from anywhere in the world and produce intelligible transcriptions of what was said [1, 2]. In order to be truly universal, recognition systems need to encompass not only speech from many languages, but also intra-sentential code-switched speech [3, 4], speech with accents or otherwise non-standard pronunciations [5, 6], and speech from languages without known written forms [7, 8].

Language-universal speech recognition requires phonological units that are agnostic to any particular language such as articulatory features [9–11] or global phones [12, 13], which can be annotated through examination of audio data. While recent advancements in the related field of multilingual speech recognition have significantly improved the language coverage of a single system [14, 15], these works differ in that they operate on language-specific levels of surface vocabulary units [16] or phonemic units that are defined with reference to the unique phonological rules of each language [17]. Prior works have avoided universal phone level annotation by implicitly incorporating this knowledge in shared latent representations that map to language-specific phonemes with neural nets [17–19].

Another approach is to learn explicit universal phone representations by relating language-specific units to their universal phonetic distinctions. Instead of relying on phone annotations, these prior works approximate universal phonological units through statistical acoustic-phonetic methods [1] or phone-to-phoneme realization rules [13, 20]. Unlike the implicit latent approach, this method allows for language-universal prediction. However, performance is dependent on the clarity of the phone-phoneme dynamics in the selected training languages [13, 21].
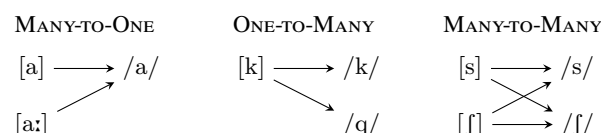

Figure 1: *Examples showing three types of manifold mappings of phones (in brackets) to phonemes (in slashes). Many-to-one describes allophones of a phoneme. One-to-many describes a duplicitous phone that maps to multiple phonemes. Many-to-many consists of both allophones and duplicitous phones.*

We are interested in systems that can incorporate the strengths of both the implicit and explicit approaches to representing universal phones. In particular, we are interested in language-universal automatic speech recognition (ASR) systems that can 1) explicitly represent universal phones and language-specific phonemes, 2) be built using only automatically generated grapheme-to-phoneme annotations and phone-to-phoneme rules, 3) resolve naturally ambiguous phone-to-phoneme mappings using information from other languages, and 4) learn interpretable probabilistic weights of each mapping.

In this work, we seek to incorporate these desiderata in a phone-based speech recognition system. We first propose a general framework to represent phone-to-phoneme rules as *differentiable allophone graphs* using weighted finite-state transducers [22–27] to probabilistically map phone realizations to their underlying language-specific phonemes (§3.1). We then incorporate these differentiable allophone graphs in a multilingual model with a universal phone recognizing layer trained in an end-to-end manner, which we call the AlloGraph model (§3.2). We show the efficacy of the AlloGraph model in predicting phonemes for 7 seen languages and predicting phones for 2 unseen languages with comparison to prior works (§5). More importantly, we show that our model resolves the ambiguity of manifold phone-to-phoneme mappings with an analysis of substitution errors and an examination of the interpretable allophone graph weights (§5.1). Finally we demonstrate our phone-based approach in two linguistic applications: pronunciation variation and allophone discovery (§5.2).

2. Background and Motivation

In this section, we first introduce phone-to-phoneme mappings for manufacturing phone supervision from phoneme annotations (§2.1). Then we discuss short-comings of a baseline method representing mappings as a pass-through matrix (§2.2) to motivate our graph-based framework in the subsequent section (§3).

2.1. Phonological Units

2.1.1. Language-Specific Phonemes vs. Universal Phones

A phone n is a unit of spoken sound within a universal set $\mathcal{N}$ which is invariant across all languages, where $\mathcal{N} = \{n_1, ..., n_{|N|}\}$ consists of $|\mathcal{N}|$ total phones [12]. In contrast,

a phoneme $m^{(l)}$ is a unit of linguistically contrastive sound for a given language l within a language specific set, where $\mathcal{M}^{(l)} = \{m_1^{(l)}, ..., m_{|\mathcal{M}^{(l)}|}^{(l)}\}$ consists of $|\mathcal{M}^{(l)}|$ total phonemes [28]. Phonemes defined for different languages describe different underlying sounds. Multilingual systems that conflate phonemes across languages have been shown to perform worse than those that treat phonemes as language-specific [13, 21].

2.1.2. Phone-to-Phoneme Mappings

For each language, the phone-to-phoneme mappings are defined as a series of tuples, $(n_i, m_j^{(l)})$, where $m_j^{(l)} \in \mathcal{M}^{(l)}$ and $n_i \in \mathcal{N}' \subseteq \mathcal{N}$ for some subset $\mathcal{N}'$ of phones that occur as realizations in the language. Each phoneme has one or more phone realization and not all universal phones are necessarily mapped to a phoneme grounding in a particular language. Note that mappings may be imperfect in our resources [20].

Phone-to-phonemes can be one-to-one mappings, but often the relationships are manifold. As shown in Figure 1, many-to-one mappings are found in scenarios where multiple phones are allophones, or different realizations, of the same phoneme. This is the prototypical mapping type. One-to-many mappings also occur for duplicitous phones that are mapped to multiple phonemes.[1] Furthermore, many-to-one and one-to-many mappings can occur together in various many-to-many forms.

2.1.3. Manufacturing Phone-Level Supervision

Since phones are fine-grained distinctions of spoken sounds in the universal space, phonemes are only fuzzy approximations. Multilingual sharing between diverse languages is required to properly learn phonetic distinctions. Consider the following:

One-to-One: If a phone is mapped one-to-one with a phoneme, then the learned phone representation will directly correspond to one supervising phoneme. In the multilingual setting, these direct mappings help other languages disambiguate this phone.

One-to-Many: If a phone is mapped to many phonemes, then each phoneme provides supervision in proportion to their prior distributions. If the learned phonemes representations are mapped from the learned phone, phoneme confusions occur if the one-to-many mappings are not disambiguated. This ambiguity persists despite information sharing from other languages.

Many-to-One: If many phones are mapped to a phoneme, each phone receives the same supervision. A second language with complementary mappings is required to learn distinct phones.

Many-to-Many: When one-to-many and many-to-one mappings occur together, they can take various forms. Generally, the many-to-one portions can be resolved through multilingual sharing but the one-to-many portions would still be problematic.

2.2. Encoding Phone-to-Phoneme as Pass-through Matrix

Prior works have shown that phone-to-phoneme mappings can be encoded as pass-through layers that convert a phone distribution into a phoneme distribution [13]. This phone-to-phoneme encoding, which we call AlloMatrix, is a sparse matrix $A^{(l)} = \{0, 1\}^{|\mathcal{N}| \times |\mathcal{M}^{(l)}|}$ where each $(n_i, m_j^{(l)})$ tuple in the mappings desribed in §2.1.2 is represented by $a_{i,j}^{(l)} = 1$. The AlloMatrix

transforms a logit vector of phones, $\mathbf{p}^{\mathcal{N}} = [p_i^{\mathcal{N}}, ..., p_{|\mathcal{N}|}^{\mathcal{N}}]$, to a logit vector of phonemes, $\mathbf{p}^{\mathcal{M}^{(l)}} = [p_j^{\mathcal{M}^{(l)}}, ..., p_{|\mathcal{M}^{(l)}|}^{\mathcal{M}^{(l)}}]$ by the dot product of the jth column of $A^{(l)}$ with each phone logit $p_i^{\mathcal{N}}$:

$$p_j^{\mathcal{M}^{(l)}} = \sum_i^{|\mathcal{N}|} (a_{i,j}^{(l)})(p_i^{\mathcal{N}}) \tag{1}$$

In the many-to-one approach, this amounts to summing the phone contributions which is in accordance with our desired mapping of allophones in §2.1.2. However, in one-to-many mappings a phone logit broadcast equally to each of the phonemes. This disagrees with the definition of phone realization. Rather we state that a realized phone in an utterance is grounded to each of the mapped phonemes with probability.

3. Proposed Framework

3.1. Encoding Phone-to-Phoneme as WFST

We define the allophone graph for language l, denoted by $G^{(l)}$, to be a single state weighted finite-state transducer (WFST) with a transition function $\pi(n_i, m_j^{(l)})$ giving each phone-to-phoneme mapping and a corresponding weight function $w(n_i, m_j^{(l)})$ giving the likelihood that n_i is the phonetic realization of $m_j^{(l)}$ for each transition. The allophone graph $G^{(l)}$ accepts phone emission probabilities $E^{\mathcal{N}}$ and transduces them into phonemes $E^{\mathcal{M}^{(l)}}$ through WFST composition [22], which is denoted as $\circ$.

$$E^{\mathcal{M}^{(l)}} = E^{\mathcal{N}} \circ G^{(l)} \tag{2}$$

This WFST is an analogous data structure to the aforementioned matrix in §2.2, but this graphical representation of phone-to-phoneme mappings as arcs in a probabilistic transduction allows us to make two key intuitive determinations. First, many-to-one mappings are transductions of several phones into the same phoneme and therefore the phoneme posterior is given by summing over the input phone posteriors, as is also done in §2.2. Second, one-to-many mappings are transductions splitting the posterior of a single phone to several phoneme posteriors, depending on how likely those phonemes are to be groundings of the phone. In §2.2, the broadcasting method fails to do this probabilistic splitting in one-to-many scenarios, creating ambiguity.

3.2. Phone Recognition with Allophone Graphs

In this section, we apply the allophone graphs as differentiable WFST [22–27] layers in phone-based ASR systems optimized with only multilingual phoneme supervision.

In this work, we use the connectionist temporal classification network (CTC) [29, 30] where a language-universal ENCODER maps input sequence $\mathbf{x} = [\mathbf{x}_t, ..., \mathbf{x}_T]$ to a sequence of hidden representations $\mathbf{h} = [\mathbf{h}_t, ..., \mathbf{h}_T]$, where $\mathbf{h}_t \in \mathbb{R}^d$. The phone emission probabilities $E^{\mathcal{N} \cup \varnothing}$ are given by the affine projection of $\mathbf{h}$ followed by the softmax function, denoted as SOFTMAXOUT.[2] To handle the blank token $\varnothing$ used in CTC to represent the null emission [29], we add the $\varnothing \rightarrow \varnothing$ transition as an additional arc in the language-specific allophone graphs $G^{(l)}$. Phone and phoneme emissions are thus given by:

$$\mathbf{h} = \text{ENCODER}(\mathbf{x}) \tag{3}$$

$$E^{\mathcal{N} \cup \varnothing} = \text{SOFTMAXOUT}(\mathbf{h}) \tag{4}$$

$$E^{\mathcal{M}^{(l)} \cup \varnothing} = E^{\mathcal{N} \cup \varnothing} \circ G^{(l)} \tag{5}$$

---

[1] These occur in resources like [20] when the source conflates allophonic and morphophonemic alternations, in instances of archiphonemic underspecification and neutralization (e.g. treating Japanese [m] as a realization of both /m/ and /N/ or English [ɾ] as a realization of both /t/ and /d/ as in *writer* [ɹaɪɾɹ̩] and *rider* [ɹaːɪɾɹ̩]), or—spuriously—when the grapheme-phoneme mapping is complex.

[2] In training, logits corresponding to unmapped phones in a particular language are masked prior to being softmax normalized similar to [31].

Table 1: *Results presenting the performances of our proposed AlloGraph models with our implementations of Phoneme-Only and Al-loMatrix baselines, as measured by language-specific phoneme error-rate (%) for seen languages and universal phone error-rate (%) for unseen languages. Performances on unseen languages were evaluated using phone-level annotations for the Tusom and Inuktitut corpora. Note that while our proposed AlloGraph and our baseline AlloMatrix models produce both phone and phoneme-level predictions, the Phoneme-Only approach only recognizes language-specific phonemes. The averaged totals across unseen/seen are shown in bold and the best performing models in each category are shown in bold underlined.*

Model Type	Model Name	Uses Phones	Seen (Phoneme Error Rate %)								Unseen (Phone Error Rate %)		
			Eng	Tur	Tgl	Vie	Kaz	Amh	Jav	Total	Tusom	Inuktitut	Total
Phoneme-Only	Multilingual-CTC [17]	✗	25.3	27.7	28.5	31.9	31.5	28.6	35.2	**29.8**	*No Phone Predictions*		
AlloMatrix	Allosaurus [13]	✓	26.5	27.6	33.1	32.0	31.9	28.2	39.0	**31.2**	91.2	96.7	**94.0**
AlloGraph	Our Proposed Model	✓	26.0	28.6	28.2	31.9	32.5	29.1	36.2	**30.5**	81.2	85.8	**84.1**
AlloGraph	+ Universal Constraint (UC)	✓	27.3	28.7	29.9	32.5	35.1	30.9	36.6	**31.6**	80.5	79.9	**80.2**

ONE-TO-MANY   MANY-TO-MANY
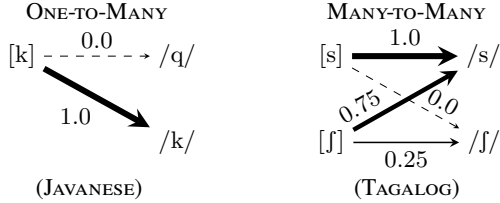
(JAVANESE)   (TAGALOG)

Figure 2: *Examples of disambiguated phone-to-phoneme mappings using the interpretable weights of our AlloGraph + UC model, where each [phone] is probabilistically mapped to a /phoneme/. In the one-to-many example from Javanese, [k] is predominantly a realization of /k/. In the many-to-many example from Tagalog, [s] is predominantly a realization of /s/ while [ʃ] is a realization of /s/ 75% of the time and /ʃ/ otherwise.*

Equation 5 shows the CTC specific form of the general phone-to-phoneme emission transduction shown in Equation 2. During training, we maximize the likelihood of the ground-truth phonemes $y = [y_1, ..., y_S]$, where $y_s \in \mathcal{M}^{(l)}$ and S is the length of the ground-truth which is at most the length of the input T, by marginalizing over all possible CTC alignments using the forward-backward computation [29, 30].

We refer to this multilingual CTC architecture with allo-phone graphs as *our proposed AlloGraph* model. In the vanilla AlloGraph, we allow the weights of $G^{(l)}$ to freely take on any values. This is a loose-coupling of phone and phoneme emissions where each $G^{(l)}$ may amplify or reduce the phone posteriors; for instance, this allows $G^{(l)}$ to learn cases where a phone is universally rare but is a prominent realization in language l.

While loose-coupling of phone and phoneme emissions is beneficial to language-specific phoneme recognition, it dilutes supervision to the universal phone layer. We address this by enforcing a tight-coupling of phone and phoneme emissions such that the phone posterior is only isometrically transformed: $\sum_{m^{(l)} \in \mathcal{M}'^{(l)}} w(n_i, m) = 1$, where $\mathcal{M}'^{(l)}$ is the subset of phonemes $\mathcal{M}^{(l)}$ that n_i is mapped to in language l. Now, Equation (5) exactly sums phone posteriors for many-to-one and splits phone posteriors for one-to-many in the manner that we desire, as stated in §3.1. We call this tightly-coupled variant the *AlloGraph + Universal Constraint (UC)* model.

4. Data and Experimental Setup

Data: We use the English LDC Switchboard Dataset [32–34] and 6 languages from the IARPA BABEL Program: Turkish, Tagalog, Vietnamese, Kazakh, Amharic and Javanese [35]. These datasets contain 8kHz recordings of conversational speech each containing around 50 to 80 hours of training data,

with an exception of around 300 hours for English. We also consider two indigenous languages with phone level annotations, Tusom and Inukitut, during evaluation only. We obtain phonemic annotations using Epitran for auto grapheme-to-phoneme [28] and phone-to-phoneme rules from Allovera [20].

Experimental Setup: All our models were trained using the ESPnet toolkit [36] with differentiable WFSTs implemented using the GTN toolkit [26]. To prepare our speech input features we first upsample the audio to 16kHz, augment it by applying a speed perturbation of 0.9 and 1.1, and then extract global mean-variance normalized 83 log-mel filterbank and pitch features. Input frames are processed by an audio encoder with convolutional blocks to subsample by 4 [36] before feeding to 12 transformer-encoder blocks with a feed-forward dim of 2048, attention dim of 256, and 4 attention heads. We augment our data with the Switchboard Strong (SS) augmentation policy of SpecAugment [37] and apply a dropout of 0.1 for the entire network. We use the Adam optimizer to train 100 epochs with an inverse square root decay schedule, a transformer-lr scale [36] of 5, 25k warmup steps, and an effective batchsize of 768.

5. Results

In Table 1, we show the results of our AlloGraph and AlloGraph + UC models. As mentioned in §4, we use Tusom and Inuktitut as two unseen languages with phone level annotations to evaluate our language-universal predictions; since these languages are unseen our model does not know their phoneme sets or which phones appear as realizations, allowing us to assess how universal our phone-based predictions are. On these two unseen languages our AlloGraph model outperforms our AlloMatrix baseline based on [13] by an average of 9.9 phone error-rate (%). When using the Universal Constraint described in §3.2, our approach gains an additional 3.9 phone error-rate improvement. The AlloGraph models make fewer substitution errors than the AlloMatrix baseline, and the substitutions are also less severe. While the AlloMatrix frequently confuses vowels for plosives, the AlloGraph models tend to only confuse related phones.

We also show the language-specific phoneme level performance of the AlloGraph model on 7 seen languages. Note that these languages are annotated with phonemes as described in §4 but not at the phone level. Here our AlloGraph model slightly outperforms the AlloMatrix baseline, but both show degradation compared to our Phoneme-Only[3] baseline based on [17]. We observe that models placing emphasis on learning universal phones do so with some cost to the language-specific level.

---

[3]Phoneme-Only [17] directly maps the shared ENCODER hidden states to language-specific phoneme level SOFTMAXOUT, replacing the shared phone level in Equation (4). Thus there are no phone predictions.
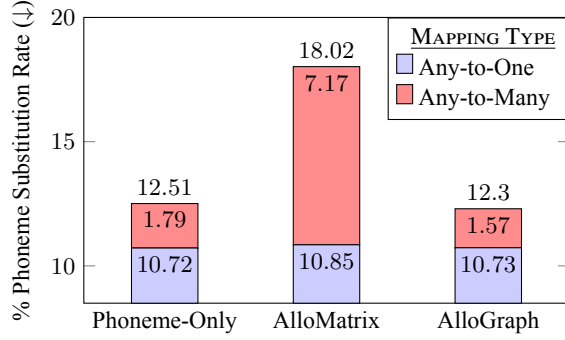
Figure 3: *Results comparing the performances of our baseline Phoneme-Only, baseline AlloMatrix, and proposed AlloGraph models on a high phone-to-phoneme complexity language, Tagalog, as measured by phoneme substitution error-rate (%). The any-to-one category includes phonemes in one-to-one and many-to-one mappings, and any-to-many includes phonemes in one-to-many and many-to-many mappings.*

The AlloGraph is advantageous in jointly modeling phones and phonemes compared to the AlloMatrix baseline due to learned disambiguations of phone-to-phoneme mappings; we examine this benefit further in the next section.

5.1. Probabilistic Phone-to-Phoneme Disambiguation

An added benefit of our model is the ability to interpret the weights of learned AlloGraphs, which show disambiguations of ambiguous phone-to-phoneme mappings. As shown in Figure 2, our AlloGraph + UC model distributes phone emissions to multiple phonemes in the one-to-many and many-to-many scenarios. These probabilities can be interpreted as prior distributions of each mapping captured by the allophone graph and can be used to determine the relative dominance of each arc in manifold mappings that can be otherwise difficult to explain.

The performance of AlloGraph + UC on languages with complex phone-phoneme mappings, such as Tagalog and Javanese, is greatly improved over the AlloMatrix baseline. In these languages, phones are frequently defined as realizations of multiple ostensive phonemes and there are many allophones of each phoneme. As shown in Figure 3 these ambiguous mappings are especially detrimental to the AlloMatrix model, which produces a high number of phoneme substitution errors compared to our AlloGraph model and Phoneme-Only baseline.

5.2. Linguistic Applications

In this section, we demonstrate the efficacy of phone-based predictions from our AlloGraph + UC model in two applications.

As shown in Table 2, our AlloGraph + UC model produces different phonetic realizations of a single phonemic pronunciation. By collecting all of the phonetic realizations for correct phonemic transcriptions of the word 'hello' uttered by numerous speakers across test sets in our conversational corpora, we automatically identified the most frequent phonetic pronunciations. These qualitative examples suggest that dynamic methods for building lexicons using universal phone recognition systems can capture diverse pronunciations that can bolster knowledge sets [5]. This may benefit pronunciation-sensitive tasks like code-switched [4] or accented speech recognition [38].

Since the AlloGraph + UC model produces joint alignments of phones and phonemes for seen languages, it can also dis-

Table 2: *Results showing the pronunciations of the word 'hello' across the 7 languages discovered by our AlloGraph + UC model, as shown in phonemic and phonetic forms. Pronunciation variations between different speakers in our conversational test set are captured at the phonetic level. We present the 3 most frequent phone-based pronunciations and their percentages.*

Lang.	Word	Phonemic	Phonetic					
Eng	hello	/həlow/	[halo]	54%	[həlow]	8%	[hɛlow]	8%
Tur	alo	/alo/	[aːɫo]	100%	-		-	
Tgl	hello	/hello/	[hello]	99%	[hellu]	1%	-	
Vie	a lô	/ʔa lo/	[ʔa lo]	100%	-		-	
Kaz	алло	/ɑllo/	[ɑllo]	75%	[ɑɒll̩ o]	20%	[bɒll̩ o]	5%
Amh	ሄሎ	/helo/	[fielo]	99%	[helo]	1%	-	
Jav	halo	/halo/	[halo]	88%	[hɔlo]	11%	[helo]	1%

Table 3: *Results showing the most frequent triphone contexts and realization rates of various phones mapped to the phonemes /b/ and /ə/ in Amharic, as discovered by our AlloGraph + UC model on our test corpus. Phones that are not mapped to any phoneme, such as [ɐ] in Amharic, can still appear as hypothesized realizations suggesting new phone-to-phoneme mappings.*

Phone-to-Phoneme	Realization Rate (%)	Predefined Mapping	Frequent Triphone Contexts		
[b] → /b/	64.5	✓	[#bɐ]	[#bə]	[#bɪ]
[β̞] → /b/	29.7	✓	[əβe]	[əβɦ]	[#βɪ]
[ə] → /ə/	32.7	✓	[nəw]	[dəfi]	[dət]
[ɐ] → /ə/	29.2	✗	[ʔɐl]	[sɐl]	[sɐm]
[ɛ] → /ə/	16.4	✓	[gɛr]	[bɛr]	[lɛt]
[ɔ] → /ə/	13.8	✓	[ʔɔw]	[ʔɔj]	[ʔɔn]

cover the allophone realization rates and triphone contexts in test corpora (Table 3). Our method can also hypothesize new allophones such as the the phone [ɐ] which is not mapped to any of the phonemes in Amharic [20]. One important step in language documentation is discovering and defining the relationship between phones and phonemes [7], ensuring that mappings are exhaustive but devoid of spurious pairs. Automatic, data-driven methods to generate phone-phoneme mappings allow linguists to discover these relationships more effectively.

6. Conclusion and Future Work

We present differentiable allophone graphs for building universal phone-based ASR using only language-specific phonemic annotations and phone-to-phoneme rules. We show improvements in phone and phoneme prediction over prior works. More importantly, our framework enables model interpretability and unique linguistic applications, such as phone-based lexicons and allophone discovery. In future work, we will seek to incorporate contextually dynamic phone-to-phoneme mappings using convolutional or attention-based WFST weights. We hope that our insights stimulate research on learnable representations of other linguistic rules, such as articulatory features [11], phonotactics [39], and cross-lingual mappings [40] in multilingual ASR.

7. Acknowledgements

8. References

[1] J. Köhler, "Multilingual phone models for vocabulary-independent speech recognition tasks," *Speech Communication*, 2001.

[2] T. Schultz and A. Waibel, "Language-independent and language-adaptive acoustic modeling for speech recognition," *Speech Communication*, 2001.

[3] B. E. Bullock and A. J. E. Toribio, *The Cambridge handbook of linguistic code-switching.* Cambridge University Press, 2009.

[4] K. Li, J. Li, G. Ye, R. Zhao, and Y. Gong, "Towards code-switching asr for end-to-end ctc models," in *Proc. ICASSP*, 2019.

[5] N. Coupland, *Style: Language variation and identity.* Cambridge University Press, 2007.

[6] S. Sun, C.-F. Yeh, M.-Y. Hwang, M. Ostendorf, and L. Xie, "Domain adversarial training for accented speech recognition," in *Proc. ICASSP*, 2018.

[7] N. P. Himmelmann *et al.*, "Language documentation: What is it and what is it good for," *Essentials of language documentation*, 2006.

[8] S. Hillis, A. P. Kumar, and A. W. Black, "Unsupervised phonetic and word level discovery for speech to speech translation for unwritten languages." in *Proc. Interspeech*, 2019.

[9] S. Stuker, F. Metze, T. Schultz, and A. Waibel, "Integrating multilingual articulatory features into speech recognition," in *Eighth European Conference on Speech Communication and Technology*, 2003.

[10] K. Livescu, P. Jyothi, and E. Fosler-Lussier, "Articulatory feature-based pronunciation modeling," *Computer Speech & Language*, 2016.

[11] X. Li, S. Dalmia, D. Mortensen, J. Li, A. Black, and F. Metze, "Towards zero-shot learning for automatic phonemic transcription," in *Proc. AAAI*, 2020.

[12] T. Schultz, "Globalphone: a multilingual speech and text database developed at karlsruhe university," in *Seventh International Conference on Spoken Language Processing*, 2002.

[13] 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 *Proc. ICASSP*, 2020.

[14] O. Adams, M. Wiesner, S. Watanabe, and D. Yarowsky, "Massively multilingual adversarial speech recognition," in *Proceedings of NAACL-HLT*, 2019.

[15] 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," *Proc. Interspeech*, 2020.

[16] B. Li, Y. Zhang, T. Sainath, Y. Wu, and W. Chan, "Bytes are all you need: End-to-end multilingual speech recognition and synthesis with bytes," in *Proc. ICASSP*, 2019.

[17] S. Dalmia, R. Sanabria, F. Metze, and A. W. Black, "Sequence-based multi-lingual low resource speech recognition," in *Proc. ICASSP*, 2018.

[18] A. Stolcke, F. Grezl, M.-Y. Hwang, X. Lei, N. Morgan, and D. Vergyri, "Cross-domain and cross-language portability of acoustic features estimated by multilayer perceptrons," in *Proc. ICASSP*, 2006.

[19] K. Veselỳ, M. Karafiát, F. Grézl, M. Janda, and E. Egorova, "The language-independent bottleneck features," in *Proc. SLT*, 2012.

[20] D. R. Mortensen, X. Li, P. Littell, A. Michaud, S. Rijhwani, A. Anastasopoulos, A. W. Black, F. Metze, and G. Neubig, "Allovera: a multilingual allophone database," in *Proc. LREC*, 2020.

[21] J. Kohler, "Multi-lingual phoneme recognition exploiting acoustic-phonetic similarities of sounds," in *Proc. ICSLP*, 1996.

[22] M. Mohri, F. Pereira, and M. Riley, "Weighted finite-state transducers in speech recognition," *Computer Speech & Language*, 2002.

[23] N. Moritz, T. Hori, and J. L. Roux, "Semi-supervised speech recognition via graph-based temporal classification," *Proc. ICASSP*, 2021.

[24] P. Doetsch, A. Zeyer, P. Voigtlaender, I. Kulikov, R. Schlüter, and H. Ney, "Returnn: The rwth extensible training framework for universal recurrent neural networks," in *Proc. ICASSP*, 2017.

[25] Y. Shao, Y. Wang, D. Povey, and S. Khudanpur, " PyChain: A Fully Parallelized PyTorch Implementation of LF-MMI for End-to-End ASR," in *Proc. Interspeech 2020*, 2020.

[26] A. Hannun, V. Pratap, J. Kahn, and W.-N. Hsu, "Differentiable weighted finite-state transducers," *arXiv preprint arXiv:2010.01003*, 2020.

[27] D. Povey, F. Kuang, H. Qiu *et al.*, "k2 fsa and fst autograd integration," https://github.com/k2-fsa/k2, 2021.

[28] D. R. Mortensen, S. Dalmia, and P. Littell, "Epitran: Precision G2P for many languages." in *LREC*, 2018.

[29] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *Proc. ICML*, 2006.

[30] Y. Miao, M. Gowayyed, and F. Metze, "Eesen: End-to-end speech recognition using deep rnn models and wfst-based decoding," in *Proc. ASRU*, 2015.

[31] S. Dalmia, X. Li, A. W. Black, and F. Metze, "Phoneme level language models for sequence based low resource asr," in *Proc. ICASSP*, 2019.

[32] J. Godfrey and E. Holliman, "Switchboard-1 Release 2 LDC97S62," *Web Download. Philadelphia: Linguistic Data Consortium*, 1993.

[33] L. D. Consortium, "2000 HUB5 English Evaluation Transcripts LDC2002T43," *Philadelphia: Linguistic Data Consortium*, 2002.

[34] ——, "2000 HUB5 English Evaluation Speech LDC2002S09," *Philadelphia: Linguistic Data Consortium*, 2002.

[35] "Full Language Packs (FLP) released by the IARPA Babel Research Program (IARPA-BAA-11-02): IARPA-babel105b-v0.4, IARPA-babel106-v0.2g, IARPA-babel107b-v0.7, IARPAbabel302b-v1.0a, IARPAbabel307b-v1.0b, IARPAbabel402b-v1.0b."

[36] S. Watanabe, T. Hori, S. Karita, T. Hayashi, J. Nishitoba, Y. Unno, N. Enrique Yalta Soplin, J. Heymann, M. Wiesner, N. Chen, A. Renduchintala, and T. Ochiai, "ESPnet: End-to-end speech processing toolkit," in *Proc. Interspeech*, 2018.

[37] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition," in *Proc. Interspeech*, 2019.

[38] T. Viglino, P. Motlicek, and M. Cernak, "End-to-end accented speech recognition." in *Proc. Interspeech*, 2019.

[39] S. Feng, P. Żelasko, L. Moro-Velázquez, A. Abavisani, M. Hasegawa-Johnson, O. Scharenborg, and N. Dehak, "How Phonotactics Affect Multilingual and Zero-shot ASR Performance," in *Proc. ICASSP*, 2021.

[40] K. Hu, A. Bruguier, T. N. Sainath, R. Prabhavalkar, and G. Pundak, "Phoneme-Based Contextualization for Cross-Lingual Speech Recognition in End-to-End Models," in *Proc. Interspeech*, 2019.