Differentiable Allophone Graphs for Language-Universal Speech Recognition

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

Carnegie Mellon University
Language Technologies Institute

Differentiable Allophone Graphs for Language-Universal Speech Recognition

Carnegie Mellon University

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

Language Technologies Institute

At a Glance

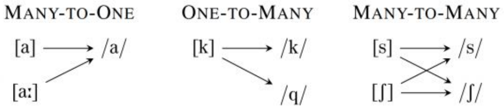
We present a general framework to derive phone-level supervision from only phonemic transcriptions and phone-to-phoneme mappings with *learnable* weights represented using weighted finite-state transducers, which we call *differentiable allophone graphs*

Language-Specific Phonemes vs. Universal Phones

Definitions of phonological units discussed in this work:

- a phone n is a unit of spoken sound within a universal set $\mathcal{N}$ which is invariant across all languages
- a phoneme $m^{(l)}$ is a unit of linguistically contrastive sound for a given language l within a language specific set $\mathcal{M}^{(l)}$
- allophones are distinct phones which appear as realizations of the same phoneme in a language

Phone-to-Phoneme Mappings



[Phone]-to-/phoneme/ relationships are often manifold:

- *One-to-One* - direct mapping; unambiguous
- *One-to-Many* - can cause confusions in phoneme prediction
- *Many-to-One* - can cause confusions in phone prediction
- *Many-to-Many* - phone and phoneme confusions likely

Phone-to-Phoneme as Pass-Through Matrices

As a baseline, consider a pass-through layer as follows:

- a sparse matrix $A^{(l)} = \{0,1\}^{|\mathcal{N}|\times|\mathcal{M}^{(l)}|}$ for each language l
- where each $(n_i, m_j^{(l)})$ tuple in the mappings is represented by $a_{i,j}^{(l)} = 1$
- and these AlloMatrices are fixed in value

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:

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

Phone-to-Phoneme as Differentiable WFSTs

Allophone graph for language l, denoted by $G^{(l)}$, is:

- a single state weighted finite-state transducer (WFST), with
- $\pi(n_i, m_j^{(l)})$ giving each phone-to-phoneme transition and
- $w(n_i, m_j^{(l)})$ giving likelihood that n_i is the realization of $m_j^{(l)}$

Allophone graph $G^{(l)}$ accepts phone emission probabilities $E^{\mathcal{N}}$ and transduces them into phonemes $E^{\mathcal{M}^{(l)}}$ through WFST composition:

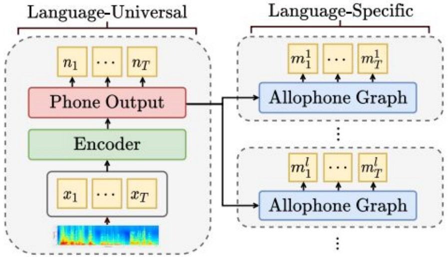
$$E^{\mathcal{M}^{(l)}} = E^{\mathcal{N}} \circ G^{(l)} \tag{2}$$

Additionally, a *Universal Constraint* enforces isometric transform:

$$\sum_{m^{(l)} \in \mathcal{M}^{(l)}} w(n_i, m) = 1 \tag{3}$$

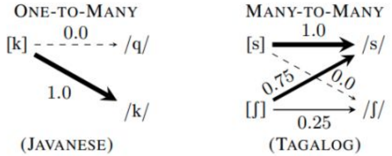
Phone Recog. via Multilingual Phoneme Supervision

1) Shared encoder maps speech to phone emissions 2) allophone graphs transduce phone emissions to phoneme emissions 3) CTC loss maximizes the likelihood of phoneme ground-truths



Learned Allophone Graph Weights

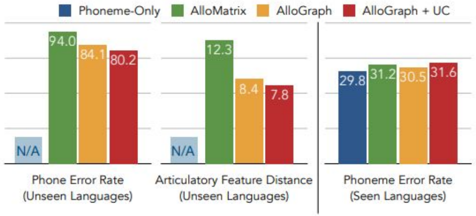
Graphs capture relative dominance of arcs in manifold mappings:



Phone and Phoneme Recognition Results

AlloGraph models vs. AlloMatrix and Phoneme-only baselines:

- makes fewer errors in phone recognition on unseen languages
- and the errors remaining are less severe as measured by AFD
- plus both AlloGraph and AlloMatrix maintain phoneme recognition



Qualitative Examples of Unseen Phone Recognition

Tusom phone recognition example, with substitution errors in red:

Model / Source	Phone Output	PER	SER	AFD
AlloMatrix	[βks'bs'β]	90.0	50.0	15.4
AlloGraph	[ʔoku:bu:ʃe:]	70.0	50.0	5.6
+ UC	[ʔokubu:ʃe:]	60.0	40.0	6.5
Ground-Truth	[ʔukxukəʃue]	-	-	-

Example Application Towards Phone-based Lexicons

Discovered phone-based pronunciations of the word "hello":

		Pronunciations			
Lang.	Word	Phonemic	Phonetic		
Eng	hello	/həlow/	[halo]	[həlow]	[hɛlow]
Tur	alo	/alo/	[a:ɫo]	-	-
Tgl	hello	/hello/	[hello]	[hellu]	-
Vie	a lô	/ʔa lo/	[ʔa lo]	-	-
Kaz	алло	/ɑllo/	[ɑ̱l̪l̪o]	[ɑ̱ɒl̪l̪ o]	[ɒl̪l̪ o]
Amh	ሄሎ	/helo/	[ɦelo]	[helo]	-
Jav	halo	/halo/	[halo]	[hɔlo]	[helo]

Outline

- Language-Universal ASR

- Allophone Graphs for Language-Universal ASR
  - Phone-to-Phoneme Mappings
  - Encoding Phone-to-Phoneme as WFST
  - Phone Recognition with Allophone Graphs

- Linguistic Applications
  - Phone-based Pronunciations
  - Allophone Discovery

What is Language-Universal Speech Recognition?

Objective: indiscriminately process utterances from anywhere in the world and produce intelligible transcriptions of what was said

To be truly universal, recognition systems should encompass:

- speech from **any language**
- speech with intrasentential **code-switching**
- speech with **accents** or otherwise **non-standard pronunciations**
- speech from languages **without known written forms**
- … and many more variations

Multilingual ≠ Universal. We care about all of the above variations in speech!

Language-Specific vs Universal Units

Most ASR systems are built to predict **language-specific units**

- **Surface-level** units like characters or words are language-specific
- **Phonemes** only distinguish sounds that are linguistically contrastive in a particular language

Alternatively, systems can predict units that are **agnostic to any particular language**

- **Phones** are units of spoken sound that are invariant across all languages **(our focus)**
- **Articulatory features** can also be defined to be invariant across all languages

Surface-Level	Phoneme	Phone
hello	/həlow/	[halo]

Challenges in Universal Phone-Based ASR

Problem: How can we obtain supervision at the phone level?

One approach is to **manually annotate** at the phone level (Schultz 2002)

- But this is **labor intensive** and thus scaling can become **cost prohibitive**

Another approach is to **approximate phone-level supervision** from phoneme annotations + phone-to-phoneme mappings (Kohler 2001, Li et al. 2020)

- But performance is **dependent on the clarity** of the phone-to-phoneme mappings
- And phone-to-phoneme mappings are **naturally ambiguous** for many languages

Outline

- ❖ Language-Universal ASR
- ❖ Allophone Graphs for Language-Universal ASR
  - ➢ Phone-to-Phoneme Mappings
  - ➢ Encoding Phone-to-Phoneme as WFST
  - ➢ Phone Recognition with Allophone Graphs
- ❖ Linguistic Applications
  - ➢ Phone-based Pronunciations
  - ➢ Allophone Discovery

Allophone Graphs for Language-Universal ASR

In this work, we seek to build Language-Universal ASR systems are:

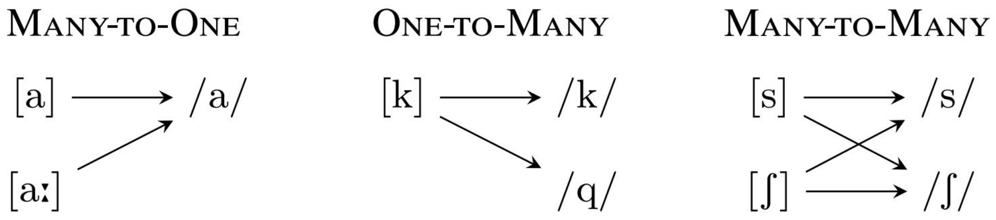
1. **Phone-based:** jointly representing **phones** and **phonemes**
2. **Scalable:** using automatic **grapheme-to-phoneme** annotations & **phone-to-phoneme rules**
3. **Adaptable:** using multilingual sharing to **resolve ambiguous phone-to-phoneme mappings**
4. **Interpretable:** by learning **interpretable probabilistic weights** of each mapping

Phone-to-Phoneme Mappings

Linguists can define phone **realizations** of phonemes for each language

But **manifold mappings** of [phones] to /phonemes/ occur naturally in many languages

- **One-to-Many** mappings can cause **phoneme confusions**
- **Many-to-One** mappings can cause **phone confusions**
- **Many-to-Many** mappings combine the complexities of both One-to-Many and Many-to-One



Encoding Phone-to-Phoneme as Pass-Through Layer

As a baseline, consider a **pass-through layer** as follows:

- **a sparse matrix** $A^{(l)} = \{0, 1\}^{|\mathcal{N}| \times |\mathcal{M}^{(l)}|}$ for each language l
- where each $(n_i, m_j^{(l)})$ tuple in the mappings is represented by $a_{i,j}^{(l)} = 1$
- And all of these AlloMatrices are **fixed** in value

AlloMatrix transforms a logit vector of phones, $\mathbf{p}^{\mathcal{N}} = [p_i^{\mathcal{N}}, \ldots, p_{|\mathcal{N}|}^{\mathcal{N}}]$, to a logit vector of phonemes, $\mathbf{p}^{\mathcal{M}^{(l)}} = [p_j^{\mathcal{M}^{(l)}}, \ldots, p_{|\mathcal{M}^{(l)}|}^{\mathcal{M}^{(l)}}]$ by the dot product:

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

Encoding Phone-to-Phoneme as WFST

For each language l, we define an **allophone graph** $G^{(l)}$ as a single-state WFST with

- **Transition function** giving each phone-to-phoneme mapping as a transduction
- **Weight function** giving the likelihood that a phone is the realization of a phoneme

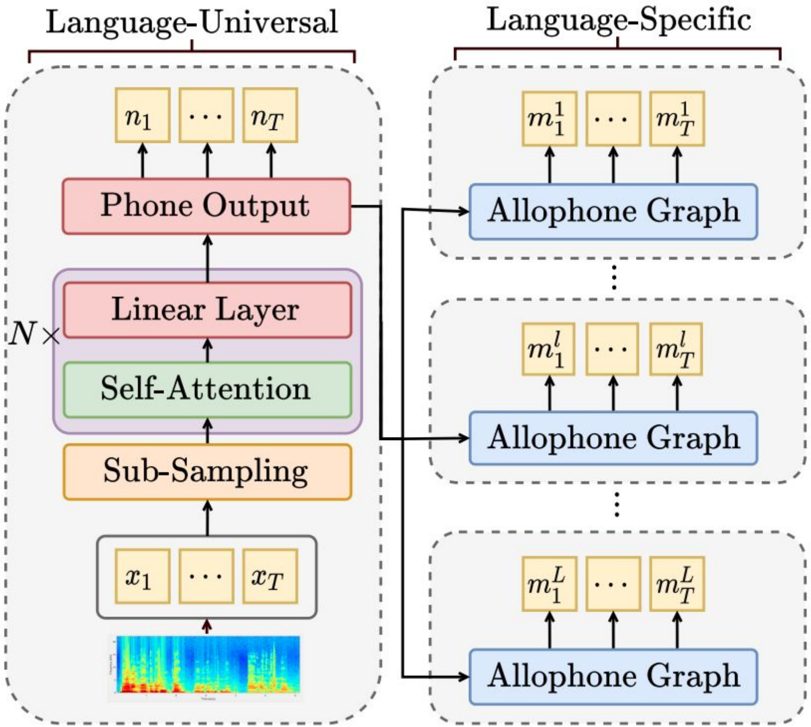
The allophone graph $G^{(l)}$ accepts **phone emission** probabilities $E^{\mathcal{N}}$ and transduces them into **phoneme emission** probabilities $E^{\mathcal{M}^{(l)}}$ through **WFST composition:**

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

Phone Recognition with Allophone Graphs

We learn a phone-based model using multilingual phoneme supervision in which:

- A **CTC encoder** maps input sequence of speech to **universal phone emission probabilities**
- An **allophone graph** for each language transduces phone emissions to **phoneme emissions**
- CTC loss is applied to maximize the likelihood of the **phoneme ground-truth**

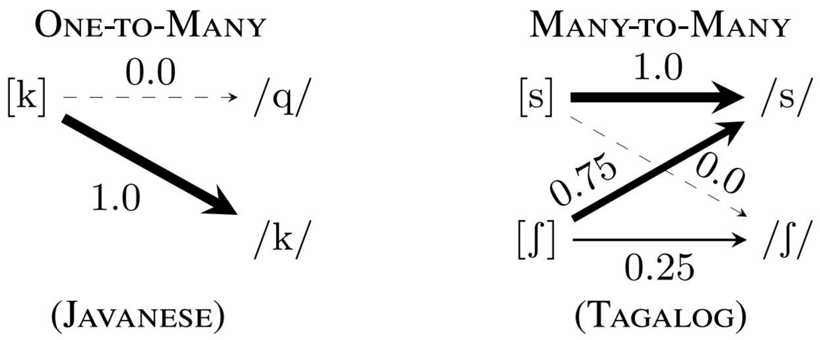


Phone Recognition with Allophone Graphs

The learned probabilistic weights of the allophone graphs are **interpretable**

Allophone graphs capture the **prior distributions** of phone-to-phoneme mappings

This prior shows the **relative dominance** of each arc in manifold mappings, which can be otherwise difficult to explain:



Phone Recognition with Allophone Graphs

We compare our **AlloGraph** model to **Phoneme-Only** and **AlloMatrix** (fixed pass-through matrix method of representing phone-to-phoneme mappings) baselines

The AlloGraph + Universal Constraint variant places greater emphasis on phone level

Our approach **improves phone-based ASR**, evaluated on difficult **unseen** languages, while maintaining performance at the **phoneme-level** on the **seen** languages

		Uses	Seen (Phoneme Error Rate %)								Unseen (Phone Error Rate %)		
Model Type	Model Name	Phones	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**

Phone Recognition with Allophone Graphs

Improvements in phone recognition for unseen langs. via **reduced substitution errors**

The **articulatory feature distance** between substitutions that remain is also reduced

The errors made by AlloGraph are **fewer and also less severe**

	Tusom			Inuktitut		
Model	PER	SER	AFD	PER	SER	AFD
AlloMatrix	91.2	65.6	12.3	96.7	75.3	12.4
AlloGraph	81.2	56.8	8.7	85.8	65.8	8.4
+ UC	**80.5**	**54.9**	**7.8**	**79.9**	**59.9**	**7.8**

Phone Recognition with Allophone Graphs

The **3 most frequent confusion pairs** of the AlloMatrix show degenerate behavior

Vowels and plosives are very distant in articulatory feature space

AlloGraph’s most frequent confusions are between **related phones**; much less severe

	Tusom		Inuktitut	
Model	Confusion	AFD	Confusion	AFD
AlloMatrix	[ɨ] → [β̞]	15	[a] → [β̞]	13
	[ə] → [β̞]	13	[i] → [β̞]	13
	[ə] → [s’]	17	[u] → [s’]	23
AlloGraph	[i] → [iː]	2	[a] → [ɑ̠]	3
	[k] → [k͡p]	4	[u] → [o]	4
	[a] → [aː]	2	[a] → [aː]	2
AlloGraph + UC	[a] → [ɐ]	4	[q] → [k]	2
	[ə] → [ɐ]	2	[a] → [ɐ]	4
	[a] → [ɑ]	2	[i] → [ɪ]	2

Phone Recognition with Allophone Graphs

Qualitative examples show that the AlloGraph produces intelligible transcriptions

UNSEEN LANGUAGE: Tusom				
Model / Source	Phone Output	PER	SER	AFD
AlloMatrix	[s’s’β̞]	100.0	60.0	13.3
AlloGraph	[əkɨɹu]	80.0	60.0	4.7
+ UC	[ʔɨkru]	20.0	20.0	2.0
Ground-Truth	[ʔɨkʰru]	-	-	-
AlloMatrix	[bs’β̞g̤s’ɹ]	83.3	83.3	12.2
AlloGraph	[bɐŋg̤s’ɹ]	66.6	66.6	8.3
+ UC	[bɐŋg̤ɣr]	50.0	50.0	4.0
Ground-Truth	[baŋgor]	-	-	-
AlloMatrix	[β̞ks’bs’β̞]	90.0	50.0	15.4
AlloGraph	[ʔoku:bu:ʃe:]	70.0	50.0	5.6
+ UC	[ʔokubu:ʃe:]	60.0	40.0	6.5
Ground-Truth	[ʔukxukəʃue]	-	-	-

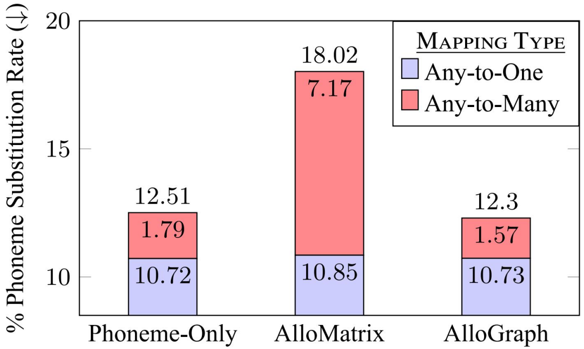
UNSEEN LANGUAGE: Inuktitut				
Model / Source	Phone Output	PER	SER	AFD
AlloMatrix	[ks’βs’k ks’βs’k]	60.0	60.0	18.3
AlloGraph	[kimuckʰ kimu]	50.0	30.0	6.0
+ UC	[kɪŋok kɪŋuk]	30.0	30.0	2.7
Ground-Truth	[kiŋuk kiŋuk]	-	-	-
AlloMatrix	[ʃβs’k ʃβks’]	80.0	70.0	9.7
AlloGraph	[sɨka:k su:ka:k]	60.0	60.0	2.3
+ UC	[s̺ukʌk sukʌk]	50.0	50.0	2.8
Ground-Truth	[sukaq sukaq]	-	-	-
AlloMatrix	[s’ks’tʔ s’ks’t]	87.5	75.0	13.8
AlloGraph	[i:ki:kʰ i:ki:kʰ]	75.0	75.0	2.7
+ UC	[ikɪp ikɪpq]	62.5	50.0	6.5
Ground-Truth	[ikiq ikiq]	-	-	-

Phone Recognition with Allophone Graphs

Due to the naturally ambiguous nature of phone-to-phoneme mappings, the fixed **AlloMatrix** method results in a high rate of **phoneme substitution errors**

These errors are greatly pronounced in the **ambiguous Any-to-Many** mappings

The learnable phone-to-phoneme mappings in AlloGraph resolve this ambiguity:



Outline

- ❖ Language-Universal ASR
- ❖ Allophone Graphs for Language-Universal ASR
  - ➢ Phone-to-Phoneme Mappings
  - ➢ Encoding Phone-to-Phoneme as WFST
  - ➢ Phone Recognition with Allophone Graphs
- ❖ Linguistic Applications
  - ➢ Phone-based Pronunciations
  - ➢ Allophone Discovery

Phone-Based Pronunciations

Our AlloGraph model can **discover phonetic pronunciations** and their relative frequencies, useful towards building a universal phone-based lexicon

Phone-based pronunciations capture **richer variation** than the traditional phoneme-based method which may benefit pronunciation-sensitive tasks such as code-switched or accented speech recognition

		Pronunciations						
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/	[ɑ̠l̠l̪o]	75%	[ɑ̠ɒl̠l̪ o]	20%	[ɒl̠l̪ o]	5%
Amh	ሄሎ	/helo/	[ɦelo]	99%	[helo]	1%	-	
Jav	halo	/halo/	[halo]	88%	[hɔlo]	11%	[helo]	1%

Allophone Discovery

Our AlloGraph model can **discover new phone realizations**, or allophones of the same phoneme, useful towards defining / updating the phone-to-phoneme mappings of languages

The AlloGraph model can also **contextualize phone realizations**

These types of **automatic, data-driven insights** may benefit tasks such as language documentation

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

Thank You!

Brian Yan

byan@cs.cmu.edu