Differentiable Allophone Graphs for Language-Universal Speech Recognition

Brian Yan, Siddharth Dalmia, David R. Mortensen, Florian Metze, Shinji Watanabe
Differentiable Allophone Graphs for Language-Universal Speech Recognition
Brian Yan, Siddharth Dalmia, David R. Mortensen, Florian Metze, Shinji Watanabe

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:
- $a$: a phone is a unit of spoken sound within a universal set $\mathcal{N}$ which is invariant across all languages
- $w_i$: a phoneme 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

| MANY-TO-ONE | ONE-TO-MANY | MANY-TO-MANY |
|---|---|---|
| [a] | /a/ | [k] | /k/ | [s] | /s/ |

[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 confusion 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 $W = \{0, 1\}^{N \times M}$ for each language $l$
- each $(w_i, m_j)$ tuple in the mappings is represented by $a_{ij}^{(l)} = 1$
- and these AlloMatrices are fixed in value

AlloMatrix transforms a logit vector of phones, $p^{\mathcal{N}} = [p^{[a]}, ..., p^{[s]}]$, to a logit vector of phonemes, $p^{\mathcal{M}} = [p^{[k]}, ..., p^{[f]}]$, by the dot product:

$$p^{\mathcal{M}} = \sum_j w_j^{(l)} p_j^{\mathcal{N}}$$  \hspace{1cm} (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
- $\pi(n_i, m_j)$ giving each phone-to-phoneme transition and
- $w(n_i, m_j)$ giving likelihood that $n_i$ is the realization of $m_j$

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

$$E^{\mathcal{M}} = E^{\mathcal{N}} \circ G$$  \hspace{1cm} (2)

Additionally, a Universal Constraint enforces isometric transform:

$$\sum_{m_j \in \mathcal{M}_l} w(n_i, m_j) = 1$$  \hspace{1cm} (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:

**ONE-TO-MANY**

| [k] | 1.0 |
|-----|-----|
| /k/ |     |

**MANY-TO-MANY**

| [s] | 1.0 |
|-----|-----|
| /s/ |     |

(JAVANESE)

(TAGALOG)

Phoneme Error Rate

| Unseen Languages | Articulatory Feature Distance | Phone Error Rate |
|------------------|-----------------------------|------------------|
| N/A              | N/A                         | N/A              |

Qualitative Examples of Unseen Phone Recognition

Tuson phone recognition example with substitution errors in red:

| Model / Source | Phone Output | PER  | SER  | AFD  |
|----------------|--------------|------|------|------|
| AlloMatrix     | [بس] [ل]     | 90.0 | 50.0 | 15.4 |
| AlloGraph      | [توكوب] [ف]  | 70.0 | 50.0 | 5.6  |
| + UC           | [توكوب] [ف]  | 60.0 | 40.0 | 6.5  |
| Ground-Truth   | [توكوب] [ف]  | -    | -    | -    |

Example Application Towards Phone-based Lexicons

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

| Lang. | Word | Phonemic | Phonetic |
|-------|------|----------|----------|
| Eng   | hello | /həlo:/  | [həlo]   | [həlo]   |
| Tur   | alo   | /a.lo/   | [alo]    | -        |
| Tgl   | hello | /həlo/   | [həlo]   | [həlo]   |
| Vie   | a łu  | /lu lo/  | [lu lo]  | [lu lo]  |
| Kaz   | aəlo  | /əlo/    | [əlo]    | [əlo]    |
| Anh   | əəlo  | /həlo/   | [həlo]   | [həlo]   |
| Jav   | halo  | /həlo/   | [həlo]   | [həlo]   |

Pronunciations
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**  |  /həˈloʊ/  |  [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.

\[
\begin{align*}
\text{Many-to-One} & \quad \text{One-to-Many} & \quad \text{Many-to-Many} \\
[a] & \rightarrow /a/ & [k] & \rightarrow /k/ & [s] & \rightarrow /s/ \\
[a:] & \rightarrow /a/ & [k] & \rightarrow /k/ & [s] & \rightarrow /s/ \\
\end{align*}
\]
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, \( \mathbf{p}^{\mathcal{N}} = [p_1^{\mathcal{N}}, ..., p_{|\mathcal{N}|}^{\mathcal{N}}] \), to a logit vector of phonemes, \( \mathbf{p}^{\mathcal{M}^{(l)}} = [p_1^{\mathcal{M}^{(l)}}, ..., p_{|\mathcal{M}^{(l)}|}^{\mathcal{M}^{(l)}}] \) by the dot product:

\[
p_{j}^{\mathcal{M}^{(l)}} = \sum_{i} (a_{i,j}^{(l)}) (p_i^{\mathcal{N}})
\]
For each language \( \mathcal{L} \), 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 \( E^N \) and transduces them into **phoneme emission** probabilities \( E^M^{(l)} \) through **WFST composition**:

\[
E^M^{(l)} = E^N \circ G^{(l)}
\]
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:

![Diagram](image)
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.

| Model Type   | Model Name                        | Uses | Seen (Phoneme Error Rate %) | Unseen (Phone Error Rate %) |
|--------------|-----------------------------------|------|-----------------------------|-----------------------------|
|              |                                   |      | Eng | Tur | Tgl | Vie | Kaz | Amh | Jav | Total | Tusom | Inuktut | Total |
| Phoneme-Only | Multilingual-CTC [17]             | √    | 25.3 | 27.7 | 28.5 | 31.9 | 31.5 | 28.6 | 35.2 | **29.8** |        |         |       |
| 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**

| Model       | Tusom  |        | Inuktitut |
|-------------|--------|--------|-----------|
|             | PER    | SER    | AFD       | PER    | SER    | AFD       |
| AlloMatrix  | 91.2   | 65.6   | 12.3      | 96.7   | 75.3   | 12.4      |
| AlloGraph + UC | 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.

| Model          | Tusom Confusion | Inuktitut Confusion | AFD |
|----------------|-----------------|---------------------|-----|
| AlloMatrix     | [i] → [β]       | 15                  | [a] → [β] | 13 |
|                | [o] → [β]       | 13                  | [i] → [β] | 13 |
|                | [ɔ] → [s’]      | 17                  | [u] → [s’] | 23 |
| AlloGraph      | [i] → [iː]      | 2                   | [a] → [a] | 3  |
|                | [k] → [kp]      | 4                   | [u] → [o] | 4  |
|                | [a] → [aː]      | 2                   | [a] → [aː] | 2  |
| AlloGraph + UC | [a] → [e]       | 4                   | [q] → [k] | 2  |
|                | [o] → [e]       | 2                   | [a] → [e] | 4  |
|                | [a] → [o]       | 2                   | [i] → [i] | 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’s’p]          | 100.0        | 60.0 | 13.3 |     |
| AlloGraph                   | [akimu]           | 80.0         | 60.0 | 4.7  |     |
| + UC                        | [likru]           | 20.0         | 20.0 | 2.0  |     |
| Ground-Truth                | [likhr]           | -            | -    | -    | -   |
| AlloMatrix                  | [bs’bs’as]        | 83.3         | 83.3 | 12.2 |     |
| AlloGraph                   | [bengs’as]        | 66.6         | 66.6 | 8.3  |     |
| + UC                        | [bengyr]          | 50.0         | 50.0 | 4.0  |     |
| Ground-Truth                | [bengor]          | -            | -    | -    | -   |
| AlloMatrix                  | [bs’b’s’b]        | 90.0         | 50.0 | 15.4 |     |
| AlloGraph                   | [lokubu:fe]       | 70.0         | 50.0 | 5.6  |     |
| + UC                        | [lokubu:fe]       | 60.0         | 40.0 | 6.5  |     |
| Ground-Truth                | [lokxuka:fe]      | -            | -    | -    | -   |

| **UNSEEN LANGUAGE: Inuktitut** | **Model / Source** | Phone Output | PER  | SER  | AFD |
|---------------------------------|--------------------|--------------|------|------|-----|
| AlloMatrix                      | [ks’ks’ks’k]       | 60.0         | 60.0 | 18.3 |     |
| AlloGraph                       | [kimu:k kimu]     | 50.0         | 30.0 | 6.0  |     |
| + UC                            | [knu:k knujuk]    | 30.0         | 30.0 | 2.7  |     |
| Ground-Truth                    | [knjuk knjuk]     | -            | -    | -    | -   |
| AlloMatrix                      | [ks’ks’ks’k]       | 80.0         | 70.0 | 9.7  |     |
| AlloGraph                       | [sikak su:kak]    | 60.0         | 60.0 | 2.3  |     |
| + UC                            | [sukak sukaq]     | 50.0         | 50.0 | 2.8  |     |
| Ground-Truth                    | [sukaq sukaq]     | -            | -    | -    | -   |
| AlloMatrix                      | [s’ks’t s’ks’t]   | 87.5         | 75.0 | 13.8 |     |
| AlloGraph                       | [ikikik ikikik]   | 75.0         | 75.0 | 2.7  |     |
| + UC                            | [ikip ikipiq]     | 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.

| Lang. | Word | Phonemic | Pronunciations | Phonetic |
|-------|------|----------|----------------|----------|
| Eng   | hello| /hələw/  | [halo] 54%     | [hələw] 8% |
| Tur   | alo  | /aːlo/   | [aːlo] 100%    | -        |
| Tgl   | hello| /hələʊ/  | [hello] 99%    | [hello] 1% |
| Vie   | a lō | /ʔa lō/  | [ʔa lō] 100%   | -        |
| Kaz   | aillo| /ʔillo/  | [ʔillo] 75%    | [ʔillo] 20% |
| Amh   | እሎ | /hело/   | [ɦəlo] 99%     | [ɦəlo] 1% |
| Jav   | halo | /həlo/   | [halo] 88%     | [halo] 11% |
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                 | ✓                  | [#bə] #bə #bɪ         |
| [β] → /b/        | 29.7                 | ✓                  | [eβe] [eβi] #βi        |
| [ə] → /ə/        | 32.7                 | ✓                  | [nəw] [dəh] dət        |
| [ɑː] → /ɔː/      | 29.2                 | ✗                  | [ʔəl] [səl] səm        |
| [ɛ] → /ɵ/        | 16.4                 | ✓                  | [ɡɛɾ] [bɛɾ] lɛt        |
| [ɔ] → /ɔ/        | 13.8                 | ✓                  | [ʔɔw] [ʔɔj] ɔn         |
Thank You!

Brian Yan
byan@cs.cmu.edu