AlloVera: a multilingual allophone database

David R Mortensen, Xinjian Li, Patrick Littell, Alexis Michaud, Shruti Rijhwani, Antonios Anastasopoulos, Alan Black, Florian Metze, Graham Neubig

Proceedings of LREC 2020: 12th Language Resources and Evaluation Conference, Marseille.

AlloVera: A Multilingual Allophone Database

David R. Mortensen[*], Xinjian Li[*], Patrick Littell[†], Alexis Michaud[‡], Shruti Rijhwani[*], Antonios Anastasopoulos[*], Alan W. Black[*], Florian Metze[*], Graham Neubig[*]

[*]Carnegie Mellon University; [†]National Research Council of Canada; [‡]CNRS-LACITO

[*]5000 Forbes Ave, Pittsburgh PA 15213, USA;

[†]1200 Montreal Rd, Ottawa ON K1A0R6, Canada; [‡]7 rue Guy Môquet, 94800 Villejuif, France

[*]{dmortens, xinjianl, srijhwan, aanastas, awb, fmetze, neubig}@cs.cmu.edu;

[†]patrick.littell@nrc-cnrc.gc.ca; [‡]alexis.michaud@cnrs.fr

Abstract

We introduce a new resource, AlloVera, which provides mappings from 218 allophones to phonemes for 14 languages. Phonemes are contrastive phonological units, and allophones are their various concrete realizations, which are predictable from phonological context. While phonemic representations are language specific, phonetic representations (stated in terms of (allo)phones) are much closer to a universal (language-independent) transcription. AlloVera allows the training of speech recognition models that output phonetic transcriptions in the International Phonetic Alphabet (IPA), regardless of the input language. We show that a "universal" allophone model, Allosaurus, built with AlloVera, outperforms "universal" phonemic models and language-specific models on a speech-transcription task. We explore the implications of this technology (and related technologies) for the documentation of endangered and minority languages. We further explore other applications for which AlloVera will be suitable as it grows, including phonological typology.

1. Introduction

Speech can be represented at various levels of abstraction (Clark et al., 2007; Ladefoged and Johnson, 2014). It can be recorded as an acoustic signal or an articulatory score. It can be transcribed with a panoply of detail (a NARROW transcription), or with less detail (BROAD transcription). In fact, it can be transcribed retaining only those features that are *contrastive* within the language under description, or with abstract symbols that stand for contrastive units. This latter mode of representation is what is called a PHONEMIC representation while the finer-grained range of representations are PHONETIC representations. Most NLP technologies that represent speech through transcription do so at a phonemic level (that is, words are represented as strings of PHONEMES). For language-specific models and questions, such representations are often adequate and may even be preferable to the alternatives. However, in multilingual models, the language-specific nature of phonemic abstractions can be a liability. The added phonetic realism of even a broad phonetic representation moves transcriptions closer to a universal space where categories transcend the bounds of a particular language.

This paper describes AlloVera[1], a resource that maps between the phonemic representations produced by many NLP tools—including grapheme-to-phoneme (G2P) transducers like our own (Mortensen et al., 2018)—and broad phonetic representations. Specifically, it is a database of phoneme-allophone pairs (where an allophone is a phonetic realization of a phoneme—see § 1.1. below) for 14 languages. It is designed for notational compatibility with existing G2P systems. The phonetic representations are relatively broad, a consequence of our sources, but they are phonetically realistic enough to improve performance on a speech-to-phone recognition task, as shown in § 3.

This resource has applications beyond universal speech-to-phone recognition, including approximate search and speech synthesis (in human language technologies) and phonetic/phonological typology (in linguistics). The usefulness of AlloVera for all purposes will increase as it grows to cover a broad range of the languages for which phonetic and phonological descriptions have been completed. However, to illustrate the usefulness of AlloVera, we will rely primarily on the zero-shot, universal ASR use-case in the evaluation in this paper.

1.1. Phonemes and Allophones

There have been various attempts at universal ASR: "designing a universal phone recognizer which can decode a new target language with neither adaptation nor retraining" (Siniscalchi et al., 2008). This goal is up against major challenges. To begin with, defining the relevant units is no trivial task. Some research teams use grapheme-to-phoneme transducers to map orthography into a universal representational space. But in fact, as the name implies, these models typically yield *phonemes* as their output and phonemes are, by their nature, language specific. Consider Figure 1.

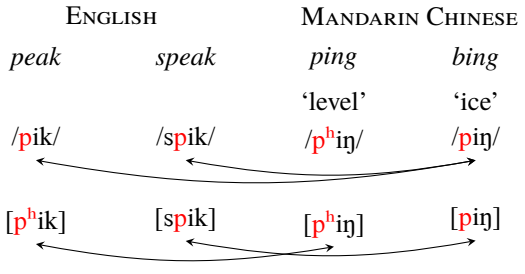

Figure 1: Words, phonemes (slashes), and phones (square brackets) in English and Mandarin Chinese

In English there is a single /p/ phoneme which is realized two different ways depending on the context in which it occurs (Ladefoged, 1999). These contextual realizations are

[1]https://github.com/dmort27/allovera

allophones. The aspirated allophone [pʰ] occurs word initially and at the beginning of stressed syllables. The unaspirated allophone [p] occurs in most other contexts. This is to be contrasted with Mandarin Chinese where there are distinct /pʰ/ and /p/ phonemes that are "in contrast" (that is, exchanging one for the other can result in a new morpheme) (Norman, 1988). Mandarin /pʰ/ has one allophone, [pʰ], and Mandarin /p/ has one allophone, [p]. Thus, English and Mandarin have the same two "p" (allo)phones, but organize these into phonemes in different ways[2].

1.2. AlloVera and Multilingual Models

For reasons stated above, multilingual training of speech models on English and Mandarin Chinese phonemes is problematic. A /pʰ/ phoneme in Chinese is always going to be roughly the same, phonetically, but a /p/ phoneme in English could be either [p] or [pʰ]. Once data from these two languages is combined, the contextual information separating the two sets of phones in English is erased and the result is a very noisy model. This problem is frequent when blending data from different languages.

A thoroughly different way to go about doing multilingual training is defining a (universal) set of features to describe sounds in articulatory, acoustic or perceptual terms (Siniscalchi et al., 2008; Johny et al., 2019). But defining these features raises considerable epistemological difficulties. There are cogent proposals from phoneticians and phonologists for transcribing sounds by means of their defining properties, rather than through International Phonetic Alphabet symbols (Vaissière, 2011). While these proposals appear promising in the mid/long run, they are not currently tractable to computational implementation in any straightforward way. Furthermore, the information that would be needed to implement such a system simply is not currently available to us in any form that we can consume.

The method explored here takes the middle ground: we create a database of allophones—that is to say, phonetic representations (referred to in phonetics/phonology as *broad* phonetic transcriptions) rather than phonemic representations[3]. This simplifies the annotation task: curators simply translate a set of relations among IPA symbols given in text to a simple allophone-phoneme mapping table. A curator can learn to do this with only a few hours of training. A multilingual model can then use these mappings in conjunction with speech data transcribed at the phonemic level to build a representation of each phone in the set. This resource is described in the following sections.

[2] In fact, the situation is even more complicated: [b], [p], and [pʰ] exist on a continuum called "voice onset time" or VOT. A sound transcribed as [p] in one language may have a voice onset time relatively close to [pʰ]. In another language, it may be similarly close to [b]. These categories, too, are to some degree language-specific (Abramson and Whalen, 2017). They are simply a step closer to phonetic reality than phonemic representations.

[3] An *ideal* solution to such a problem might be to construct rule-based phoneme-to-allophone transducers (perhaps as FSTs) for each language in the training set. Then phonetic representations could be derived by first applying G2P to the orthographic text, then applying the appropriate transducer to the resulting phonemic representation. However, constructing such a resource is expensive, requires several specialized skills on the part of curators—who must encode the phonological environments in which allophones occur—and requires information that is often omitted from phonological descriptions of languages.

2. Description

The AlloVera database is publicly available (via GitHub) at `https://github.com/dmort27/allovera` under an MIT license. In the following section, we explain the contents of the database, how they were curated, and give details about the data format and metadata provided for each language.

2.1. Sources and Languages

This resource consists of mappings between phonemic and broad phonetic representations for 14 languages with diverse phonologies. Languages were chosen based on three conditions:

- There is significant annotated speech data available for the language variety
- There is an existing G2P system for the language variety or resources for adding support for that language variety to Epitran (Mortensen et al., 2018).
- There is a description of the phonology of the language variety including allophonic rules

The languages released in final form are listed in Table 1. Additionally, there are currently several languages in alpha and beta states. Current alpha languages are Bengali (Indian; ben), Sundanese (sun), Swahili (swa), Portuguese (por), Cantonese (yue), Haitian (hat), and Zulu (zul). Current beta languages are Nepali (nep), Bengali (Bangladesh; ben), Korean (kor), Mongolian (mon), Greek (tsd), and Catalan (cat). We view AlloVera as an open-ended project which will continue expanding in the future.

2.2. Curation Practices

Most mappings were initially encoded by non-experts with a few hours of training, but all were subsequently checked by the first author, a professional linguist with graduate training in phonetics and phonology.

Our policy, in creating the mappings, was to use—where available—the "Illustrations of the IPA" series published in the *Handbook of the International Phonetic Association* (International Phonetic Association, 1999) and the *Journal of the International Phonetic Association* as our primary references. When that was not possible, we used other references, including Wikipedia summaries of research on the relevant languages. Each mapping was designed to be used with a particular G2P model. Curators mapped each phone in the description to the relevant phoneme using a spreadsheet. The phonemes from the standard (in IPA) were then mapped to the phonemes output by the G2P system (typically in X-SAMPA). When there was imperfect alignment between these sets, changes were typically made to the G2P model, expanding or restricting its range of outputs. However, in some cases, phonemes output by the G2P system could be shown to occur extra-systemically (for example, in loanwords) and the phoneme set was expanded to accommodate it. In these cases, we used equivalent IPA/X-SAMPA

Language	Phonemes	Phones	Sources
Amharic	49	57	Hayward and Hayward (1999)
English (American)	38	44	Ladefoged (1999)
French	36	38	Fougeron and Smith (1993)
German	40	42	Kunkel-Razum and Dudenredaktion (Bibliographisches Institut) (2005)
Italian (Standard)	41	45	Rogers and d'Arcangeli (2004)
Japanese	30	47	Wikipedia contributors (2019a)
Javanese	31	34	Suharno (1982, 4–6)
Kazakh	41	45	McCollum and Chen (2018)
Mandarin	31	41	Norman (1988)
Russian	45	62	Yanushevskaya and Bunčić (2015)
Spanish	30	39	Martínez-Celdrán et al. (2003; Wikipedia contributors (2019b)
Tagalog	29	42	Wikipedia contributors (2019c)
Turkish	30	43	Zimmer and Orgun (1992)
Vietnamese (Hanoi)	34	42	Kirby (2011)

Table 1: Languages included in AlloVera

symbols for the phonemic and broad phonetic representations.

Languages show considerable internal variation. For example, the system of fricatives differs significantly between various varieties of Spanish. In some cases, our speech data is from a specific variety (e.g. Castilian Spanish). In other cases, it may be polydialectal. Where possible (as with Spanish), we have made the mappings general, so that they admit phoneme-allophone mappings from a variety of dialects. In other cases (as with German), however, our resources describe a single "standard" variety and the range of phonetic variation present in colloquial speech is not necessarily reflected in the mappings. Dialectal variation also posed a challenge when the datasets did not have sufficient information about the speakers and the dialect used in the recorded speech. In these cases, we resorted to asking native speakers to identify the appropriate variant and create the mappings based on their analysis. However, we observed that this task is sometimes difficult, even for life-long speakers of the language. For example, to differentiate between Indian and Bangladeshi variants of Bengali, multiple examples had to be presented to the native speakers in order to get a reasonably confident analysis of the dataset. In future work, we plan to increase the generality of as many of the mappings as possible by incorporating information from scholarly resources on phonetic variation in the relevant languages.

There were a few recurring challenges facing curators. These include descriptions that do not distinguish between allophonic and morphophonemic (=morphophonological) rules, or between allophonic rules and orthographic rules. In these cases, curators were told to ignore any abstraction above the level that would be produced by the G2P system. On a related front, some G2P systems—like the one we use for Japanese—generate archiphonemes, as with the Japanese moraic nasal *N*. In these cases, we allowed an archiphonemic analysis even though it deviated from the phonemic ideal assumed by most of the mappings.

2.3. Data format

The data is distributed as a set of JSON files (one per language variety) and a BibTeX file containing source information. Each file contains the following metadata:

- The ISO 639-3 code for the language
- The Glottocode(s) for the supported lect(s) (Hammarström et al., 2019)
- The primary source for the mapping (as a BibTeX cite key)
- The secondary sources for the mapping (as BibTeX cite keys)
- If the mapping is constructed to be used with Epitran (Mortensen et al., 2018), the associated Epitran language-script code.
- If the mapping is made for use with another G2P engine, an identifier for this engine.

The data itself is represented as an array (rather than an object, in order to allow many-to-many mappings) Each element in this array has the following required fields:

- Phonemic representation (IPA)
- Phonetic representation (IPA)

It may have the following optional fields:

- Environment (verbal description of the phonological environment in which an allophone occurs)
- Source (when source for mapping differs from primary source)
- Glottocodes, if the mapping only applies to a subset of the lects listed in the global metadata
- Notes

An excerpt from one of these files is given in Figure 2.

2.4. Summary of the Contents

The database currently defines 218 phones which are associated with one of 148 phoneme symbols. This falls far short of the total number of phones in the languages of the world—the PHOIBLE database has 3,183 phones (Moran and McCloy, 2019)—but AlloVera has good representation of the most common phones, as shown in Table 2.

2.5. Limitations

Currently, AlloVera does not support tone, stress or other suprasegmentals, despite including mappings for two tone languages (Mandarin Chinese and Vietnamese) and one language with a pitch accent or restricted tone system (Japanese), as well as several languages with contrastive

```
{
    "iso": "spa",
    "glottocode": [
        "amer1254",
        "cast1244"
    ],
    "primary src": "Martinez-Celdran-et-al:2003-illustration",
    "secondary srcs": ["Wiki:2019-spanish-language"],
    "epitran": "spa-Latn",
    "mappings": [
    ...
        {
         "phone": "χ",
         "phoneme": "x",
         "environment": "optionally, before a back vowel",
         "glottocodes": [
             "cast1244"
         ]
        },
    ...
    ]
}
```


Figure 2: Fragment of the JSON object for Spanish.

PHOIBLE Set	Intersection with AlloVera Phone Set
Top 50	44
Top 100	72
Top 200	107

Table 2: Representation of top PHOIBLE phones (by number of languages) in AlloVera

stress (e.g. English). This is due, in large part, to the complexity of representing these phonemes—which are "spread out" over multiple segments—in terms of IPA strings. There are two separate standards for representing tone within IPA (International Phonetic Association, 1999), one of which is used primarily by linguists working on East and Southeast Asian languages (Chao tone letters written at the beginning or end of a syllable) and one of which is used by linguists working on languages elsewhere in the world (diacritics written over the nuclear vowel of a syllable). To achieve the multilingual aims of AlloVera, it would be necessary to have a single scheme for representing tone across languages.

3. Experiments

To highlight one application of AlloVera, we implement an allophone speech recognition system, Allosaurus. We compare its performance with standard universal phoneme model and language-specific model. The results suggest that Allovera helps to improve the phoneme error rate on both the training languages and two unseen languages.

3.1. Multilingual Allophone Model

The standard multilingual speech recognition models can be largely divided into two types as shown in Figure 3. The *shared phoneme model* is a universal multilingual model which represents phonemes from all languages in a shared space. The underlying assumption of this architecture is that the same phoneme from different languages should be treated similarly in the acoustic model. This assumption is, however, not an very accurate approximation as the same phonemes from different languages are often realized by different allophones as described in § 1.1..

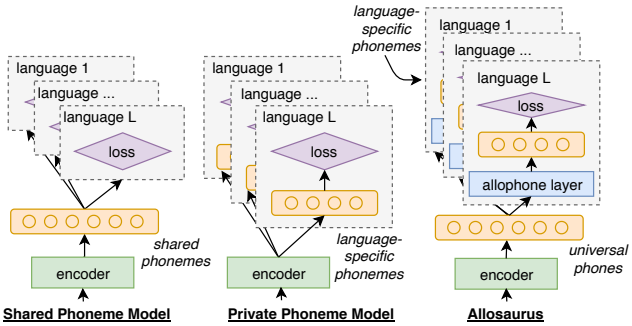

Figure 3: Traditional approaches predict phonemes directly, either for all languages (left) or separately for each language (middle). On the contrary, our approach (right) predicts over a shared phone inventory, then maps into language-specific phonemes with an allophone layer.

The second standard approach is the *private phoneme model*, which is shown in the middle of the Figure 3. The model applies a language-specific classifier, which distinguishes the phonemes from different languages. This approach consists of a shared multilingual encoder and language-specific projection layer. This approach tends to perform better than the shared phoneme model, however, it fails to consider associations between phonemes across the languages. For example, /p/ in English and /p/ in Mandarin are treated as two completely distinct phonemes despite the fact that their surface realizations overlap with each other. Additionally, it is difficult to derive language-independent phones or phonemes from this approach.

In contrast, the Allosaurus model described on the right side of Figure 3 can overcome both issues of those standard models by taking advantage of AlloVera. Instead of constructing a shared phoneme set, the Allosaurus model constructs a shared phone set by taking the union of all 218 phones covered in the AlloVera dataset. The shared encoder first predicts the distribution over the phone set, then transforms the phone distribution into the phoneme distribution in each language using the allophone layer. The allophone layer is implemented by looking up the language-specific phone-phoneme correspondences as annotated in Allovera. By adopting this approach, the Allosaurus model overcomes the disadvantages of the two standard models: the phone representation is a much more appropriate choice for the language-independent representation than the shared phoneme representation, and this phone representation can be implemented without sacrificing language-specificity. For example, the language-independent phone [p] is first learned and then projected into English phoneme /p/ and Mandarin phoneme /p/.

3.2. Results

To investigate how AlloVera improves multilingual speech recognition, we implemented three multilingual models mentioned above and compared their performance. In particular, we selected 11 languages from AlloVera taking into consideration the availability of those languages in our training speech corpus. For each language, we selected a

		Amh	Eng	Ger	Ita	Jap	Man	Rus	Spa	Tag	Tur	Vie	Average
Full	**Shared Phoneme PER**	78.4	71.7	71.6	62.9	65.9	76.5	76.9	62.6	74.1	76.6	82.7	73.8
	Private Phoneme PER	37.1	22.4	17.6	26.2	17.6	17.9	21.3	18.5	47.6	35.8	56.5	25.6
	Allosaurus PER	36.0	20.5	18.8	23.7	23.8	17.0	26.3	19.4	57.4	35.3	57.3	25.0
Low	**Shared Phoneme PER**	80.4	73.3	74.3	72.2	77.1	83.0	83.2	72.8	84.8	84.4	84.5	78.4
	Private Phoneme PER	55.4	50.6	41.9	31.6	36.8	37.0	47.9	36.7	62.3	54.5	73.6	43.8
	Allosaurus PER	54.8	47.0	41.5	37.4	40.5	33.4	45.0	35.9	70.1	53.6	72.5	41.8

Table 3: Three models' phoneme error rates on 11 languages. The top half shows the results when training with full datasets. The bottom half shows the low-resource results in which only 10k utterances are used for training from each dataset.

Language	**Corpora**	**Utt.**
English	voxforge, Tedlium (Rousseau et al., 2012), Switchboard (Godfrey et al., 1992)	1148k
Japanese	Japanese CSJ (Maekawa, 2003)	440k
Mandarin	Hkust (Liu et al., 2006), openSLR (Bu et al., 2017; Dong Wang, 2015)	377k
Tagalog	IARPA-babel106b-v0.2g	93k
Turkish	IARPA-babel105b-v0.4	82k
Vietnamese	IARPA-babel107b-v0.7	79k
Kazakh	IARPA-babel302b-v1.0a	48k
German	voxforge	40k
Spanish	LDC2002S25	32k
Amharic	openSLR25 (Abate et al., 2005)	10k
Italian	voxforge	10k
Russian	voxforge	8k
Inukitut	private	1k
Tusom	private	1k

Table 4: Training corpora and size in utterances for each language. Models are trained and tested with 12 rich resource languages (top) and 2 low resource unseen languages (bottom).

training corpus from voxforge[4], openSLR[5] and other resources. The source and size of the data sets used in these experiments are given in Table 4. To evaluate the model, we used 90% of each corpus as the training set, and the remaining 10% as the testing set. The evaluation metric is the phoneme error rate (PER) between the reference phonemes and hypothesis phonemes. For all three models, we applied the same bidirectional LSTM architecture as the encoder. The encoder has 6 layers and each layer has a hidden size of 1024. Additionally, the private phoneme model has a linear layer to map the hidden layer into language specific phoneme distributions and the Allosaurus model applies AlloVera to project the universal phone distribution into the language specific phoneme distributions. The loss function is CTC loss in all three models. The input features are 40-dimensional MFCCs.

Table 3 show the performance of the three models under two different training conditions. The row tagged with Full means that the whole training set was used to train the multilingual model. In contrast, the row with tag Low is trained under a low resource condition in which we only select 10k utterances from each training corpus. This low resource condition is useful when building speech recognizers for new languages since training sets of most new languages are very limited. As Table 3 suggests, the private phoneme model significantly outperforms the shared phoneme on all languages—the average PER of the shared phoneme model is 73.8% and the private phoneme model has 25.6% PER in the full training condition. During the evaluation process, we find that the performance of the shared phoneme model decreases significantly when increasing the number of training languages. This can be explained by the fact that phoneme assignment schemes are different across languages. Therefore, adding more languages can confuse the model, leading it to assign incorrect phonemes. In contrast, AlloVera provides a consistent assignment across languages by using allophone inventories. Comparing Allosaurus and the private phoneme model, we find that Allosaurus further improves from the private phoneme model by 0.6% under the full condition and 2.0% under the low resource condition. While the improvement is relatively limited in the full training case, it suggests AlloVera would be valuable for creating speech recognition models for low resource languages.

AlloVera gives Allosaurus another important capability—the ability to generate phones from the universal phone inventory. As Figure 3 shows, the layer before the allophone layer represents the distribution over universal phone inventory. The universal phone inventory consists of all allophones in AlloVera. In contrast, the shared phoneme model could only generate inconsistent universal phonemes and the private phoneme model could only generate language-specific phonemes. Table 5 highlights the generalization ability of Allosaurus and AlloVera over two unseen languages: Inuktitut and Tusom. The table suggests that Allosaurus and AlloVera improve the performance over both the shared phoneme model and the private phoneme model substantially.

4. Applications

Currently, we intend to integrate AlloVera and Allosaurus (or other future systems trained using AlloVera) into three practical downstream systems for very-low-resource languages, addressing tasks identified as development priorities in recent surveys of indigenous and other low-resource

[4] http://www.voxforge.org/

[5] https://openslr.org/

	Inuktitut	Tusom
Shared Phoneme PER	94.1	93.5
Private Phoneme PER	86.2	85.8
Allosaurus PER	73.1	64.2

Table 5: Comparisons of phone error rates in two unseen languages

language technology (Thieberger, 2016; Levow et al., 2017; Littell et al., 2018).

In our experience, the most requested speech technology for very-low-resource languages is **transcription acceleration**, an application of speech recognition for decreasing the workload of transcribers. Many low-resource and endangered languages do already have extensive *untranscribed* speech collections, in the form of recorded radio broadcasts, linguists' field recordings, or other personal recordings. Transcribing these collections is a high priority for many speech communities, as an untranscribed corpus is difficult to use in either research or education (Adams et al., 2018; Foley et al., 2019). AlloVera and Allosaurus were originally and primarily intended for use in transcription acceleration, although we will also be exploring other practical applications.

Another priority technology is **approximate search** of speech databases. While the aforementioned untranscribed speech collections can straightforwardly be made *available* online, they are not especially *accessible* as such. A researcher, teacher, or student cannot in practice listen to years' worth of radio recordings in search of a particular word or topic. AlloVera and Allosaurus, by making an approximate text representation of the corpus, open up the possibility for efficient approximate phonetic search through otherwise-untranscribed speech databases. Previous work has demonstrated the feasibility of such an approach (Anastasopoulos et al., 2017; Boito et al., 2017), but the quality of the search results can be significantly boosted by improvements in a first-pass phonetic transcription (Ondel et al., 2018).

We are also planning on integrating AlloVera and Allosaurus into a language-neutral **forced-alignment** pipeline. While forced-alignment is a task that is already commonly done in a zero-shot scenario (by manually mapping target-language phones to the vocabulary of a pretrained acoustic model, often an English one), the extensive phonetic vocabulary of AlloVera means that many phones are already covered. This greatly expands the number of languages that can be aligned without the need for an extensive transcribed corpus or manual system configuration.

5. Related Work

AlloVera builds on work in three major areas: phonetics and theoretical phonology, phonological ontologies, and human language technologies.

The term ALLOPHONE was coined by Benjamin Lee Whorf in the 1920s and was popularized by Trager and Block (1941). However, the idea goes back much further, to Baudoin de Courtenay (1894). The idea of allophony most relevant to our work here comes from American Structuralist linguists like Harris (1951), but we also invoke the concept of the archiphoneme, associated with the Prague Circle (Trubetskoy, 1939). In the 1950s and 1960s, the structuralist notions of the "taxonomic" phoneme and of allophones came under attack by generative linguists (Chomsky and Halle, 1968; Halle, 1962; Halle, 1959), but they have retained their importance both in linguistic practice and linguistic theory.

Various resources containing phonological information, especially phonological inventories, have been compiled. An early resource was UCLA's UPSID. A more recent resource that combines UPSID and other segment inventories in a unified ontology is PHOIBLE (Moran and McCloy, 2019). However, due to the nature of PHOIBLE's sources, it is not always clear what level of representation is intended within a segment inventory and PHOIBLE does not consistently establish relationships between abstract segments—phonemes—and concrete segments—(allo)phones. In these respects, it is complementary to AlloVera.

6. Conclusion

AlloVera embraces the fact that it is useful to analyze the sounds of language at different levels. It allows scientists and engineers to build models that are based on phones using tools that generate phonemic representations. It also captures allophonic relations in a way that is more generally useful but which does not require a highly specialized notation (for example, for stating phonological environments). We have demonstrated its usefulness, in its current form, for a valuable task (zero-shot speech-to-phone recognition).

The resource will become even more useful as more languages are added. What we have produced so far should be seen as a proof of concept. As we develop the resource further, the accuracy of our recognizers will go up, our approximate search and forced alignment models will improve, and new avenues of research will be opened.

Achieving this goal will require participation from more than just our research team: we invite linguists and language scientists who have special knowledge or a particular interest in a language to contribute their knowledge to AlloVera in the form of a simple allophone-to-phoneme mapping (preferably with natural language descriptions of the environments in which each allophone occurs). With international participation, AlloVera can go from a database that is merely useful to a resource that is indispensable for speech and language research.

7. Acknowledgements

This project was sponsored in part by the Defense Advanced Research Projects Agency (DARPA) Information Innovation Office (I2O), program: Low Resource Languages for Emergent Incidents (LORELEI), issued by DARPA/I2O under Contract No. HR0011-15-C-0114. This material is also based in part upon work supported by the National Science Foundation under grant 1761548. Shruti Rijhwani is supported by a Bloomberg Data Science Ph.D. Fellowship. Alexis Michaud gratefully acknowledges grants ANR-10-LABX-0083 and ANR-19-CE38-0015.

8. Bibliographical References

Abate, S. T., Menzel, W., and Tafila, B. (2005). An amharic speech corpus for large vocabulary continuous speech recognition. In *INTERSPEECH-2005*.

Abramson, A. S. and Whalen, D. H. (2017). Voice Onset Time (VOT) at 50: Theoretical and practical issues in measuring voicing distinctions. *Journal of Phonetics*, 63:75–86.

Adams, O., Cohn, T., Neubig, G., Cruz, H., Bird, S., and Michaud, A. (2018). Evaluating phonemic transcription of low-resource tonal languages for language documentation. In *Proceedings of LREC 2018 (Language Resources and Evaluation Conference)*, pages 3356–3365, Miyazaki.

Anastasopoulos, A., Bansal, S., Chiang, D., Goldwater, S., and Lopez, A. (2017). Spoken term discovery for language documentation using translations. In *Proc. Workshop on Speech-Centric NLP*, pages 53–58.

Baudoin de Courtenay, J. (1894). Próba teorii alternacji fonetycznych. *Rozprawy Wydziału Filologicznego*, 20:219–364.

Boito, M. Z., Bérard, A., Villavicencio, A., and Besacier, L. (2017). Unwritten languages demand attention too! word discovery with encoder-decoder models. In *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 458–465. IEEE.

Bu, H., Du, J., Na, X., Wu, B., and Zheng, H. (2017). Aishell-1: An open-source mandarin speech corpus and a speech recognition baseline. In *Oriental COCOSDA 2017*, page Submitted.

Chomsky, N. and Halle, M. (1968). *The Sound Pattern of English*. Harper & Row, New York, NY.

Clark, J., Yallop, C., and Fletcher, J. (2007). *An Introduction to Phonetics and Phonology*. Blackwell Publishing, Oxford, 3rd edition.

Dong Wang, Xuewei Zhang, Z. Z. (2015). Thchs-30 : A free chinese speech corpus.

Foley, B., Rakhi, A., Lambourne, N., Buckeridge, N., and Wiles, J. (2019). Elpis, an accessible speech-to-text tool. In *Proceedings of Interspeech 2019*, pages 306–310, Graz.

Fougeron, C. and Smith, C. (1993). Illustrations of the IPA: French. *Journal of the International Phonetic Association*, 23(2):73–76.

Godfrey, J. J., Holliman, E. C., and McDaniel, J. (1992). SWITCHBOARD: Telephone speech corpus for research and development. In *Acoustics, Speech, and Signal Processing, 1992. ICASSP-92., 1992 IEEE International Conference on*, volume 1, pages 517–520. IEEE.

Halle, M. (1959). *The Sound Pattern of Russian*. Mouton, The Hague.

Halle, M. (1962). Phonology in generative grammar. *Word*, 18:54–72.

Hammarström, H., Forkel, R., and Haspelmath, M. (2019). Glottolog 4.0. Max Planck Institute for the Science of Human History.

Harris, Z. (1951). *Structural Linguistics*. University of Chicago Press, Chicago.

Hayward, K. and Hayward, R. J. (1999). Amharic. In *Handbook of the International Phonetic Association*, pages 45–50. Cambridge University Press, Cambridge.

International Phonetic Association. (1999). *Handbook of the International Phonetic Association*. Cambridge University Press, Cambridge.

Johny, C. C., Gutkin, A., and Jansche, M. (2019). Cross-lingual consistency of phonological features: An empirical study. In *Interspeech 2019*, Graz.

Kirby, J. P. (2011). Illustrations of the IPA: Vietnamese. *Journal of the International Phonetic Association*, 41(3):381–392.

Kunkel-Razum, K. and Dudenredaktion (Bibliographisches Institut). (2005). *Duden: die Grammatik : unentbehrlich für richtiges Deutsch*. Der Duden in zwölf Bänden : das Standardwerk zur deutschen Sprache. Dudenverlag, Mannheim.

Ladefoged, P. and Johnson, K. (2014). *A Course in Phonetics*. Cengage Learning, Boston, 7th edition.

Ladefoged, P. (1999). American English. In *Handbook of the International Phonetic Association*, pages 41–44. Cambridge University Press, Cambridge.

Levow, G.-A., Bender, E. M., Littell, P., Howell, K., Chelliah, S., Crowgey, J., Garrette, D., Good, J., Hargus, S., Inman, D., Maxwell, M., Tjalve, M., and Xia, F. (2017). STREAMLInED challenges: Aligning research interests with shared tasks. In *Proceedings of the 2nd Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 39–47, Honolulu, March. Association for Computational Linguistics.

Littell, P., Kazantseva, A., Kuhn, R., Pine, A., Arppe, A., Cox, C., and Junker, M.-O. (2018). Indigenous language technologies in Canada: Assessment, challenges, and successes. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2620–2632, Santa Fe, New Mexico, USA, August. Association for Computational Linguistics.

Liu, Y., Fung, P., Yang, Y., Cieri, C., Huang, S., and Graff, D. (2006). Hkust/mts: A very large scale mandarin telephone speech corpus. In *Chinese Spoken Language Processing*, pages 724–735. Springer.

Maekawa, K. (2003). Corpus of spontaneous japanese: Its design and evaluation. In *ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition*.

Martínez-Celdrán, E., Fernández-Planas, A. M., and Carrera-Sabatè, J. (2003). Illustrations of the IPA: Castillian Spanish. *Journal of the International Phonetic Association*, 33(2):255–259.

McCollum, A. G. and Chen, S. (2018). Illustrations of the IPA: Kazakh. Ms. Univerity of California, San Diego.

Steven Moran et al., editors. (2019). *PHOIBLE 2.0*. Max Planck Institute for the Science of Human History, Jena.

Mortensen, D. R., Dalmia, S., and Littell, P. (2018). Epitran: Precision G2P for many languages. In Nicoletta Calzolari (Conference chair), et al., editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Paris, France, May. European Language Resources Association (ELRA).

Norman, J. (1988). *Chinese*. Cambridge University Press, Cambridge.

Ondel, L., Godard, P., Besacier, L., Larsen, E., Hasegawa-Johnson, M., Scharenborg, O., Dupoux, E., Burget, L., Yvon, F., and Khudanpur, S. (2018). Bayesian models for unit discovery on a very low resource language. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5939–5943. IEEE.

Rogers, D. and d'Arcangeli, L. (2004). Illustrations of the IPA: Italian. *Journal of the International Phonetic Association*, 34(1):117–121.

Rousseau, A., Deléglise, P., and Esteve, Y. (2012). TED-LIUM: an automatic speech recognition dedicated corpus. In *LREC*, pages 125–129.

Siniscalchi, S. M., Svendsen, T., and Lee, C.-H. (2008). Toward a detector-based universal phone recognizer. In *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4261–4264. IEEE.

Suharno, I. (1982). *A Descriptive Study of Javanese*. Pacific Linguistics Press, Canberra.

Thieberger, N. (2016). Documentary linguistics: methodological challenges and innovatory responses. *Applied Linguistics*, 37(1):88–99.

Trager, G. L. and Block, B. (1941). The syllabic phonemes of english. *Language*, 17(3):223–246.

Trubetskoy, N. (1939). *Grundzüge der Phonologie*, volume VII of *Travaux du Cercle Linguistique de Prague*. Cercle Linguistique de Prague, Prague.

Vaissière, J. (2011). Proposals for a representation of sounds based on their main acoustico-perceptual properties. In Elizabeth Hume, et al., editors, *Tones and Features*, pages 306–330. De Gruyter Mouton, Berlin.

Wikipedia contributors. (2019a). Japanese phonology — Wikipedia, the free encyclopedia. [Online; accessed 12-Nov-2019].

Wikipedia contributors. (2019b). Spanish language in the Americas — Wikipedia, the free encyclopedia. [Online; accessed 26-Mar-2019].

Wikipedia contributors. (2019c). Tagalog phonology — Wikipedia, the free encyclopedia. [Online; accessed 26-Mar-2019].

Yanushevskaya, I. and Bunčić, D. (2015). Illustrations of the IPA: Russian. *Journal of the International Phonetic Association*, 45(2):221–228.

Zimmer, K. and Orgun, O. (1992). Illustrations of the IPA: Turkish. *Journal of the International Phonetic Association*, 22(1–2):43–45.