Towards Zero-Shot Learning for Automatic Phonemic Transcription

Xinjian Li, Siddharth Dalmia, David R. Mortensen, Juncheng Li, Alan W Black, Florian Metze
Language Technologies Institute, School of Computer Science
Carnegie Mellon University
{xinjianl, sdalmia, dmortens, junchenl, awb, fmetze}@cs.cmu.edu

Abstract

Automatic phonemic transcription tools are useful for low-resource language documentation. However, due to the lack of training sets, only a tiny fraction of languages have phonemic transcription tools. Fortunately, multilingual acoustic modeling provides a solution given limited audio training data. A more challenging problem is to build phonemic transcribers for languages with zero training data. The difficulty of this task is that phoneme inventories often differ between the training languages and the target language, making it infeasible to recognize unseen phonemes. In this work, we address this problem by adopting the idea of zero-shot learning. Our model is able to recognize unseen phonemes in the target language without any training data. In our model, we decompose phonemes into corresponding articulatory attributes such as *vowel* and *consonant*. Instead of predicting phonemes directly, we first predict distributions over articulatory attributes, and then compute phoneme distributions with a customized acoustic model. We evaluate our model by training it using 13 languages and testing it using 7 unseen languages. We find that it achieves 7.7% better phoneme error rate on average over a standard multilingual model.

Introduction

Over the last decade, automatic speech recognition (ASR) has achieved great successes in many rich-resourced languages such as English, French and Mandarin. On the other hand, speech resources are still sparse for the majority of other languages. They cannot thus benefit directly from recent technologies. As a result, there is an increasing interest in building speech processing systems for low-resource languages. In particular, phoneme transcription tools are useful for low-resource language documentation by improving workflow for linguists to analyze those languages (Adams et al. 2018; Michaud et al. 2018).

A more challenging task is to transcribe phonemes in the language with zero training data. This task has significant implications in documenting endangered languages and preserving the associated cultures (Gippert et al. 2006). This data setup has mainly been studied in the unsupervised speech processing field (Glass 2012; Versteegh et al. 2015; Hermann and Goldwater 2018), which typically uses an unsupervised technique to learn representations which can be used towards speech processing tasks.

However, those unsupervised approaches could not generate phonemes directly and there has been few works studying zero-shot learning for unseen phonemes transcription, which consist of learning an acoustic model without any audio data or text data for a given target language and unseen phonemes. In this work, we aim to solve this problem to transcribe unseen phonemes for unseen languages without considering any target data, audio or text.

The prediction of unseen objects has been studied for a long time in the computer vision field. For specific object classes such as *faces*, *vehicles* and *cats*, a significant number manually labeled data is usually available, but collecting sufficient data for every object human could recognize is impossible. Zero-shot learning attempts to solve this problem to classify unseen objects using mid-level side information. For example, *zebra* can be recognized by detecting attributes such as *stripped*, *black* and *white*. Inspired by approaches in computer vision research, we propose the Universal Phonemic Model (UPM) to apply zero-shot learning to acoustic modeling. In this model, we decompose the phoneme into its attributes and learn to predict a distribution over various articulatory attributes. For example, the phoneme /a/ can be decomposed into its attributes: *vowel*, *open*, *front* and *unrounded*. This can then be used to infer the unseen phonemes for the test language as the unseen phonemes can be decomposed into common attributes covered in the training phonemes.

Our approach is summarized in Figure 1. First, frames are extracted and a standard acoustic model is applied to map each frame into the acoustic space (or hidden space) $\mathcal{H}$. Next we transform it into the attribute space $\mathcal{P}$ which reflects the articulatory distribution of each frame (such as whether it indicates a *vowel* or a *consonant*). Then, we compute the distribution of phonemes for that frame using a predefined signature matrix S which describes relationships between articulatory attributes and phonemes in each language.

To evaluate our UPM approach, we trained the model on 13 languages and tested it on another 7 languages. We also trained a multilingual acoustic model as a baseline for com-

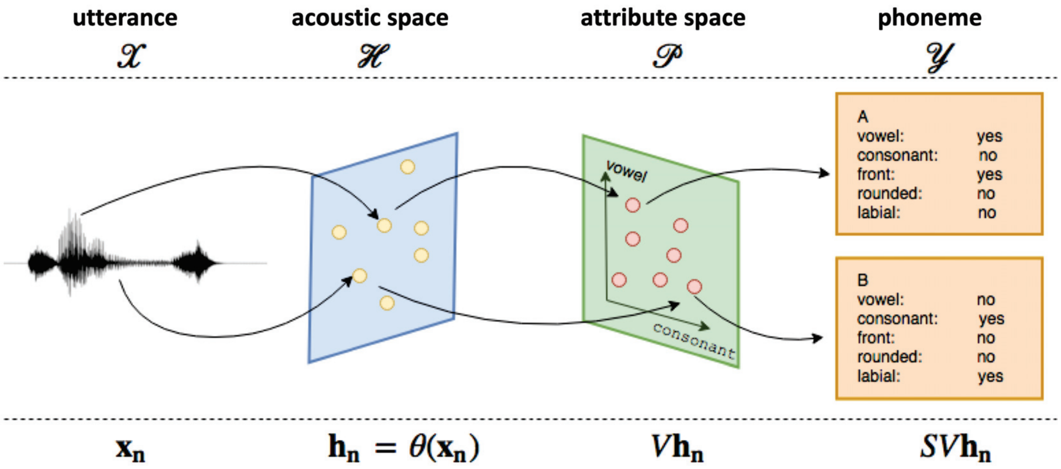

Figure 1: Illustration of the proposed zero-shot learning framework. Each utterance is first mapped into acoustic space (or hidden space) $\mathcal{H}$. Then we transform each point in the acoustic space into attribute space $\mathcal{P}$ with a linear transformation V. Finally phoneme distributions can be obtained by applying a signature matrix S

parison. The result indicates that we consistently outperform the baseline multilingual model, and we achieve 7.7% improvements in phoneme error rate on average.

The main contributions of this paper are as the followings:

1. We propose the Universal Phonemic Model (UPM) that can recognize unseen phonemes during training by incorporating knowledge from the phonetics/phonology domain.
2. We introduce a sequence prediction model to integrate a zero-shot learning framework for sequence prediction problem.
3. We show that our model is effective for 7 different languages, and our model gets 7.7% better phoneme error rate over the baseline on average.

Approach

This section explains the details of our Universal Phonemic Model (UPM). In the first section, we describe how we constructed a proper set of articulatory attributes for acoustic modeling. Next, we demonstrate how to assign attributes to each phoneme by giving an algorithm to parse X-SAMPA format. Finally we show how we integrate the phonetic information into the sequence model with a CTC loss (Graves et al. 2006).

Articulatory Attributes

Unlike attributes in the computer vision field, attributes of phonemes are independent of the corpus and dataset, they are well investigated and defined in the domain of articulatory phonetics (Ladefoged and Johnson 2014). Articulatory phonetics describes the mechanism of speech production such as the manner of articulation and placement of articulation, and it tends to describe phones using discrete features such as voiced, bilabial (made with the two lips) and fricative. These articulatory features have been shown to be useful in speech recognition (Kirchhoff 1998; Stüker et al. 2003b; Müller et al. 2017), and are a good choice for attributes for our purpose. We provide some categories of articulatory attributes below.

Consonants. Consonants are formed by obstructing the airstream through the vocal tract. They can be categorized in terms of the placement and the manner of this obstruction. The placements can be largely divided into three classes: *labial*, *coronal*, *dorsal*. Each of the class have more fine-grained classes. The manners of articulation can be grouped into: *stop*, *fricative*, *approximant* etc.

Vowel. In the production of vowels, the airstream is relatively unobstructed. Each vowel sound can be specified by the positions of lips and tongue (Ladefoged and Johnson 2014). For instance, the tongue is at its highest point in the front of the mouth for *front* vowels. Additionally, vowels can be characterized by properties such as whether the lips are rounding or not (*rounded*, *unrounded*).

Diacritics. Diacritics are small marks to modify vowels and consonants by attaching to them. For instance, *nasalization* marks a sound for which the velopharyngeal port is open and air can pass through the nose. To make the articulatory attribute set manageable, we assign attributes of diacritics to some existing consonants attributes if they share similar articulatory property. For example, *nasalization* is treated as the *nasal* attribute in consonants.

In addition to articulatory attributes mentioned above, we note that we also need to allocate an special attribute for blank in order to predict blank labels in CTC model, and backpropagate their gradients into the acoustic model. Thus, our articulatory attribute set A_{phone} is defined as the union

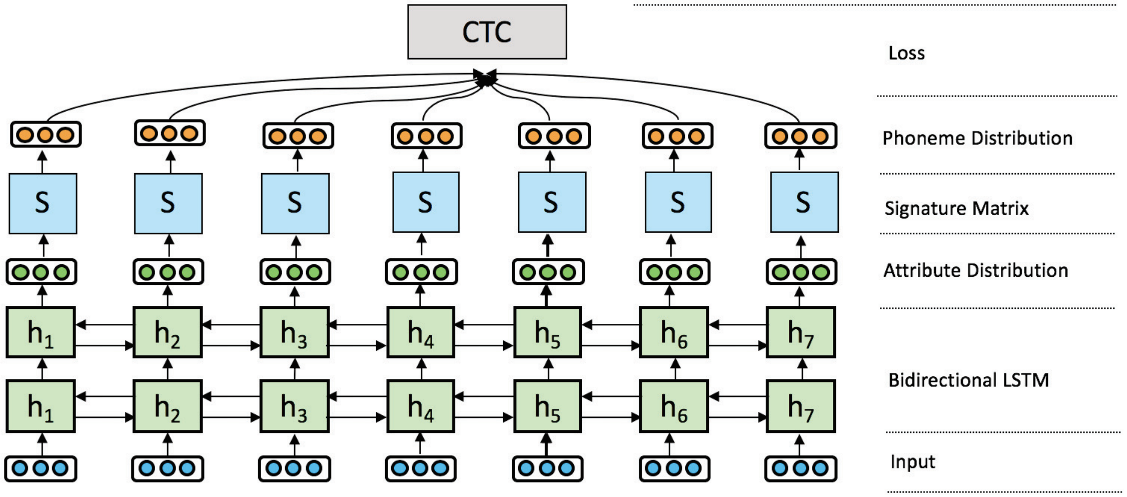

Figure 2: Illustration of the sequence model for zero-shot learning. The input layer is first processed with a Bidirectional LSTM acoustic model, and produces a distribution over articulatory attributes. Then it is transformed into a phoneme distribution by a language dependent signature matrix S

of these three domain attributes as well as the blank label,

$$A_{phone} = A_{consonants} \cup A_{vowels} \\ \cup A_{diacritics} \cup \{blank\}$$

Attribute Assignment Next, we need to assign each phoneme with appropriate attributes. There are multiple approaches to retrieve articulatory attributes. The simplest one is to use tools to collect articulatory features for each phoneme (Mortensen et al. 2016). However, those tools only provide coarse-grained phonological features but we expect more fine-grained and customized articulatory features. In this section, we propose a naive but useful approach for attribute assignment. We note that we use X-SAMPA format to denote each IPA in this work. X-SAMPA was devised to produce a computer-readable representation for IPA. Each IPA segment can be mapped to X-SAMPA with appropriate rule-based tools (Mortensen, Dalmia, and Littell 2018). For example, IPA /ə/ can be represented as /@/ in X-SAMPA.

```
input  : X-SAMPA representation of phoneme $p$
output: Articulatory attribute set $A \subseteq A_{phone}$ for $p$
$A \leftarrow$ empty set ;
while $p \notin P_{base}$ do
    find the longest suffix $p_s \in P_{base}$ ;
    Add $f|_{P_{base}}(p_s)$ to $A$ ;
    Remove suffix $p_s$ from $p$ ;
end
Add $f|_{P_{base}}(p)$ to $A$
```

Algorithm 1: A simple algorithm to assign attributes to phonemes

The assignment can be formulated as the problem to construct an assignment function $f : P_{xsampa} \rightarrow 2^{A_{phone}}$ where the domain P_{xsampa} is the set of all valid X-SAMPA phonemes, and the range $2^{A_{phone}}$ is a subset of articulatory attributes for each phoneme . The assignment function should map each phoneme into its corresponding subset of A_{phone}. To construct the function in the entire domain P_{xsampa}, we first manually map a small subset $P_{base} \subset P_{xsampa}$ and construct a restricted assignment function $f|_{P_{base}} : P_{base} \rightarrow 2^{A_{phone}}$. The mapping is customizable and has been verified with the IPA handbook (Decker and others 1999). Then for every phoneme $p \in P_{xsampa}$, we continue to remove diacritics suffix from it until it could be found in P_{base}. For example, to recognize /ts_>/, we can first match the suffix, /_>/ as an *ejective*, and then recognize /ts/ as a consonant defined in P_{base}. The Algorithm 1 summarizes our approach.

Sequence model for zero-shot learning

Zero-shot learning has rarely been applied to speech sequence prediction problems. Zero-shot translation is an example of applying zero-shot learning to a different type of sequence problems(Johnson et al. 2017). In the standard settings, the zero-shot translation means that the target language pair is not in the training dataset. However, both languages should be already seen in other training pairs. In contrast, we assume a harder problem here: there is no available training audio or text for the target language at all.

In this section we describe a novel sequence model architecture for zero-shot learning. We adapt a modified ESZSL architecture from (Romera-Paredes and Torr 2015). While the original architecture is devised to solve the classification problem with CNN(DECAF) features, our model aims to optimize a CTC loss over a sequence model as shown in Figure 2. We note our architecture is a general model, and it can also be used for other sequence prediction problems in zero-shot learning.

Language	Corpus Name	# Utterances	Language	Corpus Name	# Utterances
English	TED	268k	Mandarin	Hkust	197k
English	Switchboard	251k	Mandarin	OpenSLR 18	13k
English	Librispeech	281k	Mandarin	LDC98S73	36k
Amharic	OpenSLR 25	10k	Bengali	OpenSLR 37	196k
Cebuano	IARPA-babel301b-v2.0b	43k	Dutch	Voxforge	8k
Italian	Voxforge	10k	Javanese	OpenSLR35	185k
Kazakh	IARPA-babel302b-v1.0a	48k	Kurmanji	IARPA-babel205b-v1.0a	46k
Lao	IARPA-babel203b-v3.1a	66k	Turkish	IARPA-babel105b-v0.4	82k
Sinhala	openSLR52	185k			
German	Voxforge	41k	Mongolian	IARPA-babel401b-v2.0b	45k
Russian	Voxforge	8k	Spanish	Callhome Hub4	31k
Swahili	OpenSLR 25	10k	Tagalog	IARPA-babel106b-v0.2g	93k
Zulu	IARPA-babel206b-v0.1e	60k			

Table 1: Corpora of the training set and the test set used in the experiment. Both baseline model and proposed model are trained with 17 corpus across 13 languages, and tested on 7 corpus in 7 languages.

Given the training set $\{(\mathbf{x_n}, \mathbf{y_n}, \phi_n), n = 1...N\}$ where each input $\mathbf{x_n} \in \mathcal{X}$ is an utterance, ϕ_n is its language, and $\mathbf{y_n} \in \mathcal{Y}$ is the corresponding phoneme transcription. Suppose that $\mathbf{x_n} = (x_n^1, ..., x_n^T)$ is the input sequence where x_n^t is the frame of time step t, and T is the length of $\mathbf{x_n}$. Each frame x_n^t is first projected into a feature vector $h_n^t \in \mathbb{R}^d$ in the hidden space $\mathcal{H}$ with a Bidirectional LSTM model.

$$h_n^t = \theta(x_n^t; W_{\text{LSTM}}) \tag{1}$$

where W_{LSTM} is the parameter of the Bidirectional LSTM model. We assume that our phoneme inventory of ϕ_n consists of z phonemes in the training set, each of them having a signature of a attributes constructed as mentioned above. We can first represent our attributes in a constant signature matrix $S \in \{0,1\}^{z \times a}$ of ϕ_n. The (i, j) cell in the signature matrix is 1 if the i-th phoneme has been assigned the j-th attribute, otherwise it is assigned to 0. We note that while the signature matrix is constructed automatically in this work, it can be refined by linguists using phonology in each language. Then, we transform h_n^t into articulatory logits with the transformation matrix $V \in \mathbb{R}^{a \times d}$. Then it is further processed into the phoneme logits l_n^t with S.

$$l_n^t = SVh_n^t \tag{2}$$

The logits $\mathbf{l_n} = (l_n^1, ..., l_n^T)$ are then combined with $\mathbf{y_n}$ to compute the CTC loss (Graves et al. 2006). Additionally, regularizing V has been proved to be useful in the original ESZSL architecture (Romera-Paredes and Torr 2015). Eventually our target is to minimize the following loss function:

$$\underset{V, W_{\text{LSTM}}}{\text{minimize}}\, \text{CTC}(\mathbf{x_n}, \mathbf{y_n}; V, W_{\text{LSTM}}) + \Omega(V) \tag{3}$$

where $\Omega(V)$ is an simple ℓ^2 regularization. This objective can be easily optimized using standard gradient descent methods.

At the inference stage, we usually consider a new language ϕ_{test} with a new phoneme inventory. Suppose that the new inventory is composed of z' phonemes, then we can automatically create a new signature matrix $S' \in \{0,1\}^{z' \times a}$, and estimate probability distribution of each phoneme $P_{acoustic}(p|x_n^t)$ from logits using S' instead of S.

Experiments

Dataset

We prepare two datasets for this experiment. The training set consists of 17 corpora from 13 languages, and the test set is composed of corpora from 7 different languages. They are used by both our model and the baseline described later. Details regarding each corpus and each language are provided in Table 1.

We briefly describe our strategy of corpus selection in the experiment. To select the training corpus, the rich-resourced languages should be taken into account firstly to make sure the acoustic model can be fully trained. Therefore, we add three English corpora and three Mandarin corpora to the training set. Additionally, we expect both the baseline and our Universal Phonemic Model should be trained to recognize a variety of phonemes from different languages. Therefore we collect a number of corpora from different language families and diverse regions. Finally, we attempt to make the acoustic model robust to various channels and speech styles. For example, TED (Rousseau, Deléglise, and Esteve 2012) is the conference style, Switchboard (Godfrey, Holliman, and McDaniel 1992) is the spontaneous conversation style and Librispeech is the reading style (Panayotov et al. 2015). We note that 5 percent of the entire corpus was used as the validation set. The test corpora are selected in a similar style. They are selected from a variety of languages: not only from rich-resourced languages, but also low-resourced languages with stable audio alignments and reliable g2p models.

Language	# unseen phoneme	Baseline PER%	UPM PER%	Baseline Substitution%	UPM Substitution%
German	2	68.0	64.9	51.9	46.9
Mongolian	18	87.8	77.5	44.1	35.8
Russian	19	74.5	54.4	63.5	34.5
Swahili	2	55.7	48.9	27.4	26.6
Tagalog	0	60.7	57.0	27.2	20.1
Spanish	2	48.6	44.4	31.0	26.2
Zulu	8	73.1	67.9	36.2	33.5
Average	7.3	66.9	**59.2**	40.2	**31.9**

Table 2: Phoneme error rate and phoneme substitution rate of the baseline, and our approach. Our model (UPM) outperforms the baseline for all languages, by 7.7% (absolute) in phoneme error rate, and 8.3% in phoneme substitution error rate.

Experimental Settings

We use the EESEN framework for the acoustic modeling (Miao, Gowayyed, and Metze 2015). All the transcripts are transcribed into phonemes with Epitran (Mortensen, Dalmia, and Littell 2018). The input feature is 40 dimension high-resolution MFCCs, the encoder is a 5 layer Bidirectional LSTM model, each layer having 320 cells. The signature matrix is designed as we discussed above, different signature matrices are used for different languages. We train the acoustic model with stochastic gradient descent, using a learning rate of 0.005. In each iteration, we apply the uniform sampling (Li et al. 2019): first randomly select a corpus from the entire training set, and then randomly choose one batch from that corpus.

Our baseline model is the multilingual acoustic model with a shared phoneme inventory. This type of architecture is one of the standard approaches in the multilingual ASR community (Tong, Garner, and Bourlard 2017; Vu and Schultz 2013). In this architecture, all languages share a common acoustic model and a single output layer. The output layer is to predict phonemes in the universal phoneme inventory shared by all the training languages. In our experiment, the inventory consists of 131 distinct phonemes from 14 training languages. To compare the baseline with the proposed model, we also use the Bidirectional LSTM model as the encoder to compute phoneme distributions $P(p|x_n^t)$. Then we decode phonemes with greedy decoding as in our approach. We use the same configuration of LSTM architecture as well as the training criterion. As we focus on phonemic transcriptions in this work, we use phoneme error rate (PER) as the metric for evaluation.

Results

Our results are summarized in Table 2. As is shown, our approach consistently outperforms the baseline in terms of phoneme error rate. For example, the baseline achieves 55.7% phoneme error rate when evaluated with Swahili, and our approach obtains 48.9% in the same test set. For each language in our evaluation, we observe that we improve the phoneme error rate from 3.1% (German) to 20.1% (Russian) respectively. On average, the baseline has 66.9%, and our model gets 7.7 % better phoneme error rate.

The table also indicates the strong correlation between the number of unseen phonemes and the improvement in the phoneme error rate. For example, Russian achieves the largest improvement with our UPM: it improves significantly by 20.1% phoneme error rate. In our experiment, the Russian phoneme inventory has 48 phonemes in total out of which 19 of them are unseen during training. This suggests our model has a good generalization ability to adapt to languages whose acoustic contexts are rarely known. On the other hand, every phoneme in the Tagalog inventory has been covered by other languages in the training set. Therefore, the number of its unseen phoneme is 0 and the corresponding 3.7% phoneme error rate improvement is relatively limited. Similarly, the least improved language is German, which improved from 68.0% to 64.9% because there are only two unseen phonemes in German. This fact can also be explained by the relationship between German and English. German comes under the West Germanic branch in the Indo-European language family like English. As English is the largest training set in this experiment, phonemes of English are well-trained in the baseline and should be generalizing well to German. Therefore it is hard for UPM to outperform by a large margin. Additionally, we find that the correlation between the number of unseen phonemes and phoneme error rates is relatively weak. For example, Tagalog has 12% higher phoneme error rate compared with Spanish, even its unseen phonemes are less than Spanish. This might be explained by the discrepancy of the phoneme distribution between the target language and training languages. For example, even though in principle all the phonemes of Tagalog have been covered in the training languages, their relative frequencies are not similar, which would affect the quality of the results.

To further investigate the reason for improvements for our model, we computed the (phoneme) substitution error rate, shown in the two right columns of Table 2. It goes down from 40.2% in the baseline to 31.9% in our model. The numbers show that we have 8.3% improvement in substitution error rate. This result suggests that our model is good at improving confusions between phonemes. However, it also indicates that our model is not able to improve addition and deletion errors.

To understand how the number of training languages contributes to the performance in the experiment, we train different models by changing the numbers of training lan-

Language	Baseline unseen PER%	UPM unseen PER%	Baseline seen PER%	UPM seen PER%
German	100.0	100.0	63.9	61.9
Mongolian	100.0	91.9	86.8	78.6
Russian	100.0	96.1	69.5	51.7
Swahili	100.0	86.4	54.3	46.2
Tagalog	N.A.	N.A.	57.4	54.2
Spanish	100.0	58.0	45.2	41.7
Zulu	100.0	88.3	70.5	64.6
Average	100.0	**89.8**	64.2	**57.0**

Table 3: Phoneme error rate (%PER) of the seen phonemes and unseen phonemes in the baseline and our approach.

guages: we train those models with 2, 6, 10, 14 languages. The first two languages are English and Mandarin which are corresponding to the 6 well resourced corpus in Table.1. The other 4, 8, 12 languages are randomly selected from the remaining training languages.

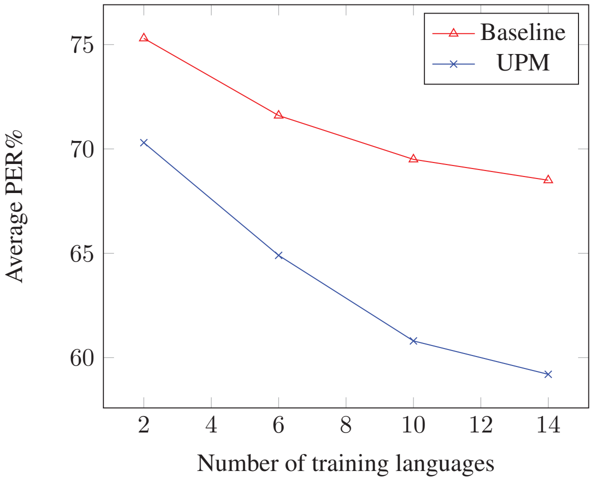

Figure 3: Illustration of the relationship between the number of training languages and the average phoneme error rate over 7 languages

Figure.3 demonstrates their performance: the red line (with triangular mark) and blue line (with cross mark) indicate the average PER of the baseline and UPM respectively. They suggest that increasing the number of training languages is helpful to reduce phoneme error rate for both models. For the baseline model, it indicates that the acoustic model get exposed to more diverse phonemes present in different languages. Therefore it learns to predict them with reduced error rates in the test set. Our UPM also improves by learning various acoustic contexts of broader articulatory attributes. The curves in Figure.3 show that UPM outperforms the baseline consistently with different training size. Additionally, the gap of phoneme error rate between the two models has increased when using more languages: the gap increased from 5.0 to 9.3. The results illustrate that our UPM is better at taking advantage of the diverse training languages. Our model can infer correlations between phonemes by using their shared articulatory attributes. This ability is helpful when a specific phoneme is rarely seen but its attributes have already been well-trained using other related phonemes. On the contrary, the baseline is not adapted well to those rare phonemes or unseen phonemes. It fails to predict those phonemes when their training data are limited.

Finally, to highlight the ability of our model, we compute the phoneme error rate for each phoneme, then classify them into the seen group and unseen group based on whether the phoneme is available in the training set. To compute phoneme error rate in this case, we align the expected phonemes with the predicted phonemes using their edit distance, the phoneme error rate here denotes the correction rate for each expected phone. Table.3 demonstrates the results of both the baseline and UPM, it suggests the UPM outperforms the baseline on both groups. On average, UPM would predict 10.2 % better for the unseen groups and 7.2 % better for the seen groups. The average numbers demonstrate that our approach has the ability to predict unseen phonemes and could even be adapted better to seen groups. The table also shows the difficulty of the task and the weakness of our approach: we could not predict any unseen phonemes for German. The two unseen phonemes of German are /pf/ and /C/, but the frequencies of both phonemes are less than 0.5 % in the test set, which makes the model extremely unstable when predicting those phonemes. On the other hand, the Spanish improvement of unseen PER is extremely significant, which can also be explained by the unstable prediction over low frequency unseen phonemes. Additionally, the 89.8 error rate of unseen groups is still not practical in the real-world production systems.

Related Work

We briefly outline several areas of related works, and describe their connections and differences with this paper. Zero-shot learning was first applied to recognize unseen objects during training in the computer vision field (Lampert, Nickisch, and Harmeling 2009; Palatucci et al. 2009; Socher et al. 2013). However those works rarely mention speech recognition.

Meanwhile there has been growing interests in zero-resource speech processing (Glass 2012; Jansen et al. 2013), most of the work focusing on tasks like acoustic unit discovery, unsupervised segmentation and spoken term discovery (Heck, Sakti, and Nakamura 2017). These models are useful for various extrinsic speech processing tasks like topic identification. However, the unsupervised concept cannot be directly grounded to actual phonemes, hence making it impracticable to do speech recognition or acoustic modeling. The usual intrinsic evaluations that these zero resource tasks are tested on is ABX discriminability task or the unsupervised word error rate which are good for quality estimates but not practical as they use an oracle or ground truth labels to assign cluster labels. In addition these approaches demands a modest size of audio corpus of targeting language (e.g: 2.5h to 40h). In contrast, our approach assumes no audio corpus and no text corpus for targeting languages. The idea of decomposing speech into concepts was also discussed by (Lake et al. 2014), where the authors propose a generative model to learn representations for spoken words which they then use to classify words with only one training sample available per word. Though this is in the same line as the zero-resource speech processing papers, we feel the motivation behind the decomposition is very similar to this work.

Another group of researchers explore adaptation techniques for multilingual speech recognition, especially for low resource languages. In these multilingual settings, the hidden layers are either HMM or DNN models which are shared by multiple languages, and the output layer is either language specific phone set or a universal IPA-based phone set (Tong, Garner, and Bourlard 2017; Vu and Schultz 2013; Thomas, Ganapathy, and Hermansky 2010; Chen and Mak 2015; Dalmia et al. 2018). However predictable phonemes are restricted to the phonemes in the training set, thus they fail to predict unseen phonemes in the test set. In contrast, our model can predict unseen phonemes by taking advantage of their articulatory attributes.

Articulatory features have been shown to be useful in speech recognition under several situation. Specifically, articulatory features has been used to improve robustness under noisy and reverberant environment (Kirchhoff 1998), compensate for crosslingual variability (Stüker et al. 2003b), improve word error rate in multilingual models (Stüker et al. 2003a), be beneficial for low resource languages (Müller, Stüker, and Waibel 2016), detecting spoken words (Prabhavalkar et al. 2013), clustering phoneme-like units for unwritten languages (Müller et al. 2017), recognizing unseen languages (Siniscalchi et al. 2011), developing phonological vocoder (Cernak and Garner 2016). There are also some attempts to predict articulatory features or distributions for clinical usages (Jiao, Berisha, and Liss 2017; Vásquez-Correa et al. 2019), but they do not provide a model to predict unseen phonemes.

We note that there are also several attempts to build acoustic models for unseen phonemes. For example, the authors in (Scharenborg et al. 2017) present an interesting method to predict unseen phonemes in Mboshi by mapping Dutch/Mboshi phonemes in the same space using an extrapolation approach. However starting phonemes used for extrapolation had to be manually assigned for every missing phoneme and every pair of languages. Compared with this work, our model proposes a much more generic algorithm to recognize unseen phonemes. Another previous work integrated articulatory attributes into the state-position based decision tree to predict unseen phones in their multilingual model (Knill et al. 2014), however the approach is limited to traditional HMM models and it is unclear how attributes are extracted and how it performs when predicting unseen phonemes.

Conclusion

In this work, we propose the Universal Phonemic Model to apply zero-shot learning to the automatic phonemic transcription task. Our experiment shows that it outperforms the baseline by 7.7 % phoneme error rate on average for 7 languages. While the performance of our approach is still not enough for the real-world production systems, it paves the way to tackle zero-shot learning of speech recognition with a new framework.

Acknowledgements

This project was sponsored 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.

References

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 *Proc. LREC*.

Cernak, M., and Garner, P. N. 2016. Phonvoc: A phonetic and phonological vocoding toolkit. In *Proc. Interspeech*.

Chen, D., and Mak, B. K.-W. 2015. Multitask learning of deep neural networks for low-resource speech recognition. *TASLP* 23(7):1172–1183.

Dalmia, S.; Sanabria, R.; Metze, F.; and Black, A. W. 2018. Sequence-based multi-lingual low resource speech recognition. In *Proc. ICASSP*.

Decker, D. M., et al. 1999. *Handbook of the International Phonetic Association: A guide to the use of the International Phonetic Alphabet*. Cambridge University Press.

Gippert, J.; Himmelmann, N.; Mosel, U.; et al. 2006. *Essentials of language documentation*, volume 178. Walter de gruyter.

Glass, J. 2012. Towards unsupervised speech processing. In *Proc. ISSPA*.

Godfrey, J. J.; Holliman, E. C.; and McDaniel, J. 1992. SWITCHBOARD: Telephone speech corpus for research and development. In *Proc. ICASSP*.

Graves, A.; Fernández, S.; Gomez, F.; and Schmidhuber, J. 2006. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proc. ICML*.

Heck, M.; Sakti, S.; and Nakamura, S. 2017. Feature optimized dpgmm clustering for unsupervised subword modeling: A contribution to zerospeech 2017. In *Proc. ASRU*.

Hermann, E., and Goldwater, S. 2018. Multilingual bottleneck features for subword modeling in zero-resource languages. In *Proc. Interspeech*.

Jansen, A.; Dupoux, E.; Goldwater, S.; Johnson, M.; Khudanpur, S.; Church, K.; Feldman, N.; Hermansky, H.; Metze, F.; Rose, R.; et al. 2013. A summary of the 2012 JHU CLSP workshop on zero resource speech technologies and models of early language acquisition. In *Proc. ICASSP*.

Jiao, Y.; Berisha, V.; and Liss, J. 2017. Interpretable phonological features for clinical applications. In *Proc. ICASSP*.

Johnson, M.; Schuster, M.; Le, Q. V.; Krikun, M.; Wu, Y.; Chen, Z.; Thorat, N.; Viégas, F.; Wattenberg, M.; Corrado, G.; et al. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *TACL* 5:339–351.

Kirchhoff, K. 1998. Combining articulatory and acoustic information for speech recognition in noisy and reverberant environments. In *Proc. ICSLP*.

Knill, K. M.; Gales, M. J.; Ragni, A.; and Rath, S. P. 2014. Language independent and unsupervised acoustic models for speech recognition and keyword spotting. In *Proc. Interspeech*.

Ladefoged, P., and Johnson, K. 2014. *A course in phonetics*. Nelson Education.

Lake, B.; Lee, C.-y.; Glass, J.; and Tenenbaum, J. 2014. One-shot learning of generative speech concepts. In *Proc. CogSci*.

Lampert, C. H.; Nickisch, H.; and Harmeling, S. 2009. Learning to detect unseen object classes by between-class attribute transfer. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, 951–958. IEEE.

Li, X.; Dalmia, S.; Black, A. W.; and Metze, F. 2019. Multilingual speech recognition with corpus relatedness sampling. In *Proc. Interspeech*.

Miao, Y.; Gowayyed, M.; and Metze, F. 2015. EESEN: End-to-end speech recognition using deep RNN models and WFST-based decoding. In *Automatic Speech Recognition and Understanding (ASRU), 2015 IEEE Workshop on*, 167–174. IEEE.

Michaud, A.; Adams, O.; Cohn, T. A.; Neubig, G.; and Guillaume, S. 2018. Integrating automatic transcription into the language documentation workflow: Experiments with na data and the persephone toolkit. *LD&C*.

Mortensen, D. R.; Littell, P.; Bharadwaj, A.; Goyal, K.; Dyer, C.; and Levin, L. 2016. Panphon: A resource for mapping ipa segments to articulatory feature vectors. In *Proc. COLING*.

Mortensen, D. R.; Dalmia, S.; and Littell, P. 2018. Epitran: Precision G2P for many languages. In *Proc. LREC*.

Müller, M.; Franke, J.; Stüker, S.; and Waibel, A. 2017. Improving phoneme set discovery for documenting unwritten languages. *Elektronische Sprachsignalverarbeitung (ESSV)* 2017.

Müller, M.; Stüker, S.; and Waibel, A. 2016. Towards improving low-resource speech recognition using articulatory and language features. In *Proc. IWSLT*.

Palatucci, M.; Pomerleau, D.; Hinton, G. E.; and Mitchell, T. M. 2009. Zero-shot learning with semantic output codes. In *Advances in neural information processing systems*, 1410–1418.

Panayotov, V.; Chen, G.; Povey, D.; and Khudanpur, S. 2015. Librispeech: an ASR corpus based on public domain audio books. In *Proc. ICASSP*.

Prabhavalkar, R.; Livescu, K.; Fosler-Lussier, E.; and Keshet, J. 2013. Discriminative articulatory models for spoken term detection in low-resource conversational settings. In *Proc. ICASSP*.

Romera-Paredes, B., and Torr, P. 2015. An embarrassingly simple approach to zero-shot learning. In *Proc. ICML*.

Rousseau, A.; Deléglise, P.; and Esteve, Y. 2012. TED-LIUM: an automatic speech recognition dedicated corpus. In *Proc. LREC*.

Scharenborg, O.; Ebel, P. W.; Ciannella, F.; Hasegawa-Johnson, M.; and Dehak, N. 2017. Building an asr system for mboshi using a cross-language definition of acoustic units approach. In *Proc. ICNLSSP*.

Siniscalchi, S. M.; Lyu, D.-C.; Svendsen, T.; and Lee, C.-H. 2011. Experiments on cross-language attribute detection and phone recognition with minimal target-specific training data. *TASLP* 20(3):875–887.

Socher, R.; Ganjoo, M.; Manning, C. D.; and Ng, A. 2013. Zero-shot learning through cross-modal transfer. In *Proc. NIPS*.

Stüker, S.; Metze, F.; Schultz, T.; and Waibel, A. 2003a. Integrating multilingual articulatory features into speech recognition. In *Proc. Eurospeech*.

Stüker, S.; Schultz, T.; Metze, F.; and Waibel, A. 2003b. Multilingual articulatory features. In *Proc. ICASSP*.

Thomas, S.; Ganapathy, S.; and Hermansky, H. 2010. Cross-lingual and multi-stream posterior features for low resource LVCSR systems. In *Proc. Interspeech*.

Tong, S.; Garner, P. N.; and Bourlard, H. 2017. An investigation of deep neural networks for multilingual speech recognition training and adaptation. In *Proc. Interspeech*.

Vásquez-Correa, J.; Klumpp, P.; Orozco-Arroyave, J. R.; and Nöth, E. 2019. Phonet: a tool based on gated recurrent neural networks to extract phonological posteriors from speech. In *Proc. Interspeech*.

Versteegh, M.; Thiolliere, R.; Schatz, T.; Cao, X. N.; Anguera, X.; Jansen, A.; and Dupoux, E. 2015. The zero resource speech challenge 2015. In *Proc. Interspeech*.

Vu, N. T., and Schultz, T. 2013. Multilingual multilayer perceptron for rapid language adaptation between and across language families. In *Proc. Interspeech*.