HOW PHONOTACTICS AFFECT MULTILINGUAL AND ZERO-SHOT ASR PERFORMANCE

Siyuan Feng1†, Piotr Želasko2,3†, Laureano Moro-Velázquez2, Ali Abavisani4, Mark Hasegawa-Johnson1, Odette Scharenborg1, Najim Dehak2,3

1Multimedia Computing Group, Delft University of Technology, Delft, The Netherlands 2Center for Language and Speech Processing, 3Human Language Technology Center of Excellence, Johns Hopkins University, Baltimore, MD, USA 4Department of Electrical and Computer Engineering, University of Illinois at Urbana-Champaign, IL, USA

s.feng@tudelft.nl, piotr.andrzej.zelasko@gmail.com

ABSTRACT
The idea of combining multiple languages’ recordings to train a single automatic speech recognition (ASR) model brings the promise of the emergence of universal speech representation. Recently, a Transformer encoder-decoder model has been shown to leverage multilingual data well in IPA transcriptions of languages presented during training. However, the representations it learned were not successful in zero-shot transfer to unseen languages. Because that model lacks an explicit factorization of the acoustic model (AM) and language model (LM), it is unclear to what degree the performance suffered from differences in pronunciation or the mismatch in phonotactics. To gain more insight into the factors limiting zero-shot ASR transfer, we replace the encoder-decoder with a hybrid ASR system consisting of a separate AM and LM. Then, we perform an extensive evaluation of monolingual, multilingual, and crosslingual (zero-shot) acoustic and language models on a set of 13 phonetically diverse languages. We show that the gain from modeling crosslingual phonotactics is limited, and imposing a too strong model can hurt the zero-shot transfer. Furthermore, we find that a multilingual LM hurts a multilingual ASR system’s performance, and retaining only the target language’s phonotactic data in LM training is preferable.

Index Terms— Automatic Speech Recognition, Phonotactics, Multilingual, Zero-shot learning

1. INTRODUCTION
Automatic speech recognition (ASR) technologies have achieved enormous advancement during the past decade [1]. Typically, a well-trained deep neural network (DNN) based ASR system [2] requires hundreds to thousands of hours of transcribed speech data. This requirement can be challenging to most of the world’s spoken languages that are considered low-resourced [3]. Most languages, including ethnic minority languages in most countries outside Europe, lack transcribed speech data, digitized texts, and digitized pronunciation lexicons. Consequently, high-performance ASR systems are only available for a small number of major languages, with English and Mandarin being prominent examples.

To facilitate the development of ASR technologies in low-resource languages, knowledge transfer across languages has been proposed in acoustic modeling [4–9]. In [4], a DNN acoustic model (AM) is pretrained with the speech from non-target languages and further retrained with a small amount of speech from the target language. In [5], a multilingual DNN is trained by simultaneously adopting speech of target and non-target languages. A very recent study [6] proposed an approach to building a nearly-universal phone recognition model. By employing 11 resource-rich languages in multilingual training, the model could achieve 17% phone error rate (PER) improvements in the low-resource languages studied, compared to a system without using such a nearly-universal phone model. These results indicate that speech resources from other languages benefit acoustic modeling for a target low-resource language. Presumably, the advantage of multilingual acoustic modeling can be explained by the possibility of sharing phonetic representations between languages. Most past research reports performance improvements from monolingual to multilingual training in terms of overall PER or word error rate (WER). It is not clear, however, what aspects of phonetic representations a multilingual AM is learning and whether the benefit of adopting multilingual acoustic modeling is consistent for every phone. To gain a deeper understanding of how phonetic representations are shared and transferred across languages, our recent work [7] adopted the end-to-end (E2E) ASR architecture [8] to perform International Phonetic Alphabet (IPA) token sequence recognition tasks. Extensive analyses of the results revealed that all phone tokens, including language-unique ones, can benefit from multilingual training. Analyses also showed that the representations learned in [7] were less effective in transferring to unseen languages.

The question is to what degree the ASR performance degradation to unseen languages [7] suffered from differences in pronunciation or the mismatch in phonotactics [1]. Here, we use the hybrid ASR system to investigate the role of language-specific phonotactic information on phone recognition performance, especially under a zero-shot scenario (which we refer to as crosslingual ASR in this paper) where speech and linguistic resources are unknown for a target language. In an encoder-decoder ASR system, phonotactic and acoustic information are learned within a single model, making it difficult to separate the effects of phonotactic and acoustic information; a hybrid ASR, by contrast, makes such experiments natural and effective.

In the present study, we strive to answer the above question by adopting the hybrid DNN-hidden Markov model (DNN-HMM) ASR architecture to conduct a series of IPA phone recognition experiments. In our hybrid ASR systems, we model basic acoustic units or

1Phonotactics defines restrictions in a language on the permissible combinations of phonemes.
phones. We will use a phone categorization assuming that a phone can be represented by one or several IPA symbols (letters and modifiers), corresponding to a single phone. We define these phones as **IPA phones**. For instance, [a] and [aː] will be considered as two different IPA phones, independently of the language. This is different from **Georgian**, in which each IPA symbol (including modifiers, e.g., the tone symbols) was modeled as a base unit in the E2E ASR systems. To measure the effect of phonotactic information, phonotactic language models (LMs) of different capacities are trained to decode alongside the hybrid DNN AM. Besides, a word LM which could impose a much stronger phonotactic constraint, is trained and compared with the phonotactic LMs. The experiments are carried out with 13 phonetically diverse languages, based on the selection criteria in [1]. Three ASR training scenarios are considered: monolingual as a baseline; multilingual, to investigate the effect of merging data from all the 13 languages; and crosslingual, namely a zero-shot learning task, to investigate whether universal phone representations exist that can be used for recognition in an unknown language.

## 2. EXPERIMENTAL SETUP

### 2.1. Databases

There are in total 13 languages chosen in this study to mimic the setup of [1], as shown in Table 1. Table 1 reports the number of distinct IPA phones used to describe the phone inventories of each language. Note that for tone languages, phonemes that differ only in tone levels are described by distinct IPA phones. For instance, the high-, rising-, dipping- and falling-tone /i/ in Mandarin are described by four different IPA phones, namely [i], [iː], [iːː], and [iːːː], respectively.

Speech recordings of the 13 languages are taken from two corpora: GlobalPhone [10] and IARPA Babel. GlobalPhone has a smaller number of recordings and is characterized by limited noise and reverberation. Babel contains more hours of training data and is significantly more challenging to recognize due to more naturalistic recording conditions. We follow train, development, and evaluation data splits adopted in our previous work [1]. Note that the high number of vowels in tone languages is due to different combinations of tone modifiers (e.g., Vietnamese and Cantonese each have six tones per vowel).

### 2.2. Implementation of hybrid ASR systems

We adopt the hybrid DNN-HMM architecture [11] for training the IPA phone recognition systems in this study. To obtain the IPA phonetic transcriptions for ASR system training and evaluation, LanguageNet G2P models [12] are leveraged to convert orthographic transcriptions in the GlobalPhone and Babel corpora into IPA phonetic sequences.

The hybrid ASR AM is implemented using the Kaldi toolkit [13]. It is a factorized time-delay neural network (TDNNF) AM [14], consisting of 12 layers, with a hidden dimension of 1024 and bottleneck dimension of 128 and Resnet-style skip connections. The TDNNF AM is trained with the lattice-free maximum mutual information (LF-MMI) criterion [15] for 4 epochs. Frame-level phone alignments used as supervision for the TDNNF model training are obtained by forced-alignment with a GMM-HMM AM trained beforehand. The input features consist of 40-dimensional high-resolution MFCCs, 3-dimension pitch features [16] and 100-dimension i-vectors. The 43-dimension MFCC-pitch feature and its left and right neighboring feature are appended and further complemented with i-vectors to form the TDNNF AM input. In TDNNF training, we use the hyperparameters from Kaldi’s Wall Street Journal recipe [17] without further tuning in our experiments. An identical TDNNF structure is used for each of the experiments in this study.

One possible concern about the validity of our approach is that a sequence-level criterion (e.g. LF-MMI) might “leak” the language model into the acoustic model. In principle, this is possible – however, we believe that LF-MMI is less likely to cause that (as compared to e.g. CTC) since it imposes a (weak) language model supervision during training, allowing the model to spend less effort trying to learn it. In either case, the acoustic model networks frame-wise predictions are conditionally independent of each other, limiting the opportunities for learning a language model.

The LM in the hybrid ASR system is an n-gram phonotactic LM estimated using the SRILM toolkit [17]. IPA phonetic transcriptions converted from the GlobalPhone and Babel training graphemic transcripts are used to train the phonotactic LM. We implement uni-gram, bi-gram, and tri-gram LMs and compare their performances in scoring the TDNNF AM to explore the effect of imposing different amounts of phonotactic information on a hybrid ASR system in recognizing IPA phone sequences. We also experiment with decoding using a word-level training LM and converting the word lattices to phone lattices before scoring. This way we can impose a much stronger phonotactic constraint on the ASR model than would be possible with phone-level n-grams.

There are three experimental scenarios considered in this work. In the monolingual scenario, a TDNNF AM and an LM are trained using data from one language, and are tested on the same language’s evaluation set. In the crosslingual (zero-shot) scenario, for each experiment, one language that is going to be tested is taken out from training, and data for the remaining 12 languages are merged to train a crosslingual AM and a crosslingual LM. We also test scoring the crosslingual AM using a monolingual LM trained with the left-out language’s data. In this case, it is not zero-shot transfer anymore, as the monolingual LM encodes the phonotactic information of the target language. It allows us to remove the phonotactics from the equation and check the transferability of the AM alone. In the multilingual scenario, all the 13 languages’ training data are merged to train a multilingual TDNNF AM, which is tested with the evaluation sets of each of the 13 languages. There are two types of LMs used to score the multilingual TDNNF AM: one multilingual LM, which

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Language</th>
<th>Train</th>
<th>Eval</th>
<th>Vow</th>
<th>Con</th>
<th>Uniq</th>
</tr>
</thead>
<tbody>
<tr>
<td>GlobalPhone</td>
<td>Czech</td>
<td>24.0</td>
<td>3.8</td>
<td>10</td>
<td>25</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>French</td>
<td>22.8</td>
<td>2.0</td>
<td>19</td>
<td>26</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td>Spanish</td>
<td>11.5</td>
<td>1.2</td>
<td>6</td>
<td>24</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>Mandarin</td>
<td>14.9</td>
<td>1.6</td>
<td>60</td>
<td>33</td>
<td>41</td>
</tr>
<tr>
<td></td>
<td>Thai</td>
<td>22.9</td>
<td>0.4</td>
<td>108</td>
<td>35</td>
<td>17</td>
</tr>
<tr>
<td>Babel</td>
<td>Cantonese</td>
<td>126.6</td>
<td>17.7</td>
<td>110</td>
<td>25</td>
<td>88</td>
</tr>
<tr>
<td></td>
<td>Bengali</td>
<td>54.5</td>
<td>9.8</td>
<td>19</td>
<td>33</td>
<td>13</td>
</tr>
<tr>
<td></td>
<td>Vietnamese</td>
<td>78.2</td>
<td>10.9</td>
<td>204</td>
<td>25</td>
<td>130</td>
</tr>
<tr>
<td></td>
<td>Lao</td>
<td>58.7</td>
<td>10.5</td>
<td>106</td>
<td>61</td>
<td>83</td>
</tr>
<tr>
<td></td>
<td>Zulu</td>
<td>54.4</td>
<td>10.4</td>
<td>15</td>
<td>54</td>
<td>19</td>
</tr>
<tr>
<td></td>
<td>Amharic</td>
<td>38.8</td>
<td>11.6</td>
<td>8</td>
<td>52</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td>Javanese</td>
<td>40.6</td>
<td>11.3</td>
<td>11</td>
<td>23</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Georgian</td>
<td>45.3</td>
<td>12.3</td>
<td>5</td>
<td>30</td>
<td>2</td>
</tr>
</tbody>
</table>

3wsj/a5/local/chain/tuning/run_cdisn_1g.sh
3. EXPERIMENTAL RESULTS

Experimental results of the IPA phone recognition experiments under monolingual, crosslingual, and multilingual scenarios are listed in Table 2. These results are reported in phone error rate (PER). The effect of imposing different amounts of phonotactic information in hybrid phone-level ASR is shown in Table 3. We first compare different LMs of tri-gram. It can be observed that when the multilingual TDNNF is adopted as the AM, scoring with the tri-gram multilingual phonotactic LM (multi-tg) achieves worse — or at least not better — performance than the monolingual LM counterpart (mono-tg) in 10 of the 13 languages. There is a similar pattern in the word-level LM experiments. This shows that the phonotactic information from other languages is harmful for phone recognition in a target language with a hybrid ASR system.

3.2. The effect of phonotactics

The impact of imposing different amounts of phonotactic information in hybrid phone-level ASR is shown in Table 2. We first compare different LMs of tri-gram. It can be observed that when the multilingual TDNNF is adopted as the AM, scoring with the tri-gram multilingual phonotactic LM (multi-tg) achieves worse — or at least not better — performance than the monolingual LM counterpart (mono-tg) in 10 of the 13 languages. There is a similar pattern in the word-level LM experiments. This shows that the phonotactic information from other languages is harmful for phone recognition in a target language with a hybrid ASR system.

Results in Table 2 suggest that scoring the crosslingual TDNNF AM with mono-tg resulted in PER reductions in 12 of the 13 languages, compared to scoring the same AM but with cross-tg. This improvement is gained by imposing phonotactics knowledge of the target language. This observation demonstrates the importance of phonotactics in the phone-level ASR task and reflects a significant limitation of zero-shot phone recognition: we would never have known the phonotactics of a previously unstudied language.

We further compare phonotactic LMs of different n-gram orders and the word LM in scoring the hybrid TDNNF AM. For the monolingual and multilingual TDNNF AMs, the tri-gram phonotactic model outperforms the bi-gram and uni-gram phonotactic models. The word LMs (mono-wtg and multi-wtg) outperform the phonotactic LMs for 8 out of 11 languages for the monolingual and multilingual TDNNF AMs. As expected, the above observations show that the hybrid phone-level ASR system benefits from imposing more substantial phonotactic constraints of the target language (word LMs comparing to phonotactic LMs) during decoding. In contrast, in the crosslingual scenario (i.e., when the phonotactic LM does not contain phonotactics of the target language), scoring with cross-wtg is worse than that with cross-tg in all the languages, except for Spanish and Cantonese. This shows that too strong phonotactic constraints (cross-wtg comparing to cross-tg) from other languages applied to the target language decoding are harmful to phone recognition. On the other hand, there are benefits to be gained from leveraging some degree of phonotactic knowledge from the non-target languages, as shown by the cross-tg performance compared to cross-wtg and cross-bg. Overall, to infer phone inventories for an unwritten language, leveraging non-target languages to build an LM is desirable at the phone level rather than at the word level.
phonotactics causes the stronger LM (is not valid for the zero-shot system - the lack of target languages’ shifts the error types from deletions towards insertions. The same languages’ base phone inventories with a zero-shot ASR model. suggests that disposing of the modifier symbols could help discover languages, achieves 75.2% PER in such a lenient scoring scheme. It cross-tg of this phenomenon, we scored the system separately with all the tem outputs only the base vowel symbol. To investigate the scope fact, we noticed that many modifiers are simply deleted, and the sys-
vowels correctly but gets their suprasegmentals and tones wrong. In
nario. In other words, the zero-shot ASR system tends to recognize errors account for a noticeable portion of PER in the crosslingual sce-

<table>
<thead>
<tr>
<th>AM</th>
<th>LM</th>
<th>PER</th>
<th>Insertions</th>
<th>Deletions</th>
<th>Substitutions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mono</td>
<td>mono-wtg</td>
<td>30.0</td>
<td>17.6</td>
<td>34.0</td>
<td>48.4</td>
</tr>
<tr>
<td>Mono</td>
<td>mono-tg</td>
<td>37.7</td>
<td>11.5</td>
<td>39.6</td>
<td>48.9</td>
</tr>
<tr>
<td>Multi</td>
<td>multi-wtg</td>
<td>34.2</td>
<td>18.3</td>
<td>29.2</td>
<td>52.5</td>
</tr>
<tr>
<td>Multi</td>
<td>multi-tg</td>
<td>43.0</td>
<td>9.6</td>
<td>40.2</td>
<td>50.3</td>
</tr>
<tr>
<td>Cross</td>
<td>cross-wtg</td>
<td>85.1</td>
<td>3.2</td>
<td>40.6</td>
<td>56.1</td>
</tr>
<tr>
<td>Cross</td>
<td>cross-tg</td>
<td>82.6</td>
<td>2.3</td>
<td>42.1</td>
<td>55.6</td>
</tr>
</tbody>
</table>

Table 3. Phone error rates (%) computed across all languages for mono(lingual), multi(lingual), and cross(lingual) systems and their breakdown into three error type.

It is worth noting that when scoring the hybrid ASR systems, we searched for the LM weights that minimize the PER in the range of 2 ~ 17, which is a broader range than the typical Kaldi setup (7 ~ 17) to give the system more freedom in down-weighting the LM if it is not useful. The most frequently chosen weights were 6 ~ 7 for the word LMs and 4 ~ 5 for the phonotactic LMs, suggesting that it is beneficial to give more emphasis to the acoustic model when the phonotactics are uncertain.

4. DISCUSSION

We present the errors’ breakdown in terms of the balance between the insertions, deletions, and substitutions in Table 2. We observe that substitutions dominate across all experiments with a contribution of around 50%, closely followed by deletions. Imposing a strong phonotactic prior with a word tri-gram LM (wtg) enhances the system performance when target language text data is available for training (21% relative improvement in mono and multi) and shifts the error types from deletions towards insertions. The same is not valid for the zero-shot system - the lack of target languages’ phonotactics causes the stronger LM (cross-wtg) to fail.

The large number of deletions in ASR is generally concerning, and there may be multiple factors behind it. Firstly, we observe from the phonotactic LM vs. word LM experiment that a weak phonotactic prior is a significant contributor (see Table 3). Secondly, we suspect another factor - both GlobalPhone and Babel transcripts contain a (silence) pseudo-word, which sometimes occurs in the middle of an utterance. It is possible that as the alignments are re-estimated in each GMM training step, the silence model is learned using some of the speech regions. In our preliminary experiments, silence frames were removed before training the AM, using a re-segmentation strategy based on alignments, however we found out it did not consistently improve the phone recognition performance (not reported in this paper). Finally, using data augmentation in training would likely help as well, increasing overall AM robustness. However, our preliminary experiments with 3-way speed perturbation [19] have not shown a significant change in the crosslingual systems’ PER.

A preliminary analysis reveals that the IPA modifier symbol errors account for a noticeable portion of PER in the crosslingual scenario. In other words, the zero-shot ASR system tends to recognize vowels correctly but gets their suprasegmentals and tones wrong. In fact, we noticed that many modifiers are simply deleted, and the system outputs only the base vowel symbol. To investigate the scope of this phenomenon, we scored the system separately with all the IPA modifiers stripped from the reference and hypothesis transcripts. The cross-tg system, which scores 82.6% PER aggregated over all languages, achieves 75.2% PER in such a lenient scoring scheme. It suggests that disposing of the modifier symbols could help discover languages’ base phone inventories with a zero-shot ASR model.

The boxplots in Figure 1 show individual phones’ error rates in the crosslingual experiments, depending on how many languages share that phone, in the crosslingual experiments.

While the comparison of the hybrid ASR performance with that of a Transformer encoder-decoder ASR used in [7] is not the primary goal of this work, the setup is similar enough to warrant a brief discussion. Notably, the error rate numbers are not directly comparable, as the Transformer setup used phone tokens as the base units for both prediction and scoring. However, we see that the hybrid monolingual system performance (especially with word LM) is much stronger than that of the Transformer; on the other hand, the Transformer was able to leverage the multilingual data to gain massive performance improvements, which is not true for the hybrid system. It is difficult to say whether the Transformer multi-system outperformed the hybrid mono systems because of the scoring differences and close error rate values. However, it is clear that the Transformer managed to somehow separate the influence of the non-target languages’ phonotactics. Neither system performed well in the zero-shot scenario.

5. CONCLUSIONS

We hypothesized that the inclusion of non-target languages’ phonotactics in training hurts multilingual and crosslingual ASR systems’ performance. We supported that hypothesis with an extensive evaluation of mono-, multi-, and cross-lingual hybrid ASR systems that use either target-only (mono) or joint (multi and cross) phonotactic LMs of different strength (phonotactic uni-, bi-, tri- grams or word tri-grams). We show that while the mono and multi AM systems benefit from a stronger phonotactic constraint, the cross system performs best with a weak phonotactic prior. We also assert that a system’s performance with multi AM degrades when combined with multi, instead of mono, LM. We concluded that zero-shot ASR benefits from weakening the phonotactic constraints, which is easily achieved in a hybrid ASR system; it is not yet clear how to similarly weaken phonotactic constraints in an E2E model that learns acoustics and phonotactics jointly. Furthermore, our findings clearly show the deficiencies of zero-shot AM representation transfer, and demonstrate that shared phones suffer less crosslingually.
6. REFERENCES


