ZERO-SHOT PRONUNCIATION LEXICONS FOR CROSS-LANGUAGE ACOUSTIC MODEL TRANSFER

Matthew Wiesner[†], *Oliver Adams*[†], *David Yarowsky*[†], *Jan Trmal*[†‡], *Sanjeev Khudanpur*[†‡]

[†]Center for Language and Speech Processing, The Johns Hopkins University, USA
[‡]Human Language Technology Center of Excellence, The Johns Hopkins University, USA

ABSTRACT

Existing acoustic models can be transferred to any language with a pronunciation lexicon (lexicon) that uses the same set of sub-word units as in training. Unfortunately such lexicons are not readily available in many low-resource languages. We bypass this requirement and create lexicons by training a grapheme-to-phoneme (G2P) transducer on a subset of words from other languages for which pronunciations are available. The subset of words is selected based on how representative it is of target language text. We find that cross-language acoustic model transfer using our selection strategy outperforms selection based on language similarity, and results in ASR performance approaching that of hand-crafted rule based lexicons in the majority of cases.

Index Terms— Pronunciation Lexicon, Cross-language transfer, Submodularity

1. INTRODUCTION

Pronunciation lexicons (lexicons) are crucial components in many automatic speech recognition (ASR) and text-to-speech (TTS) systems. Yet for most of the 3 to 4 thousand languages in the world with writing systems [1], this resource does not exist. In many of these languages, there is also not enough transcribed speech to train deep learning-based ASR systems. In such cases existing acoustic models trained on higher resource languages can be transferred by means of an appropriately constructed pronunciation lexicon.

For languages with segmental orthographies, one approach to lexicon creation is to derive pronunciations for new words via grapheme-to-phoneme conversion (G2P) trained using a seed lexicon [2, 3]. When a seed lexicon does not exist, a trained professional and/or native speaker can construct a lexicon using grapheme-to-phoneme rules. Tools and algorithms designed to facilitate this process include `Epitran` [4], `SPICE` [5], the active learning approach of [6], or the tool described in [2].

However, these approaches require human effort. One way to reduce this effort is to find a small subset of words to annotate, on which G2P conversion can be trained with minimal degradation in performance [7]. [8, 9] eliminate this effort entirely by training the G2P on lexicons from nearby languages selected using a language similarity metric. Underpinning this approach is the fact that most low-resource languages have shallow orthographies often borrowed or adapted from nearby higher-resource languages.

In this paper, we unify these two approaches by extending the subset selection task to a cross-lingual scenario. We start with *target*-language text on which we estimate a distribution of character n-grams. Then, from a pool of words from *other* languages, we greedily select words so that the sample distribution over n-gram features in the selected subset is close to that of the target language text. Thus, the algorithm will tend to select similar words across disparate languages, including loan words, place names, or shared lexemes. We select words using constrained submodular maximization. A problem framed in these terms can be solved using efficient greedy algorithms and has theoretical performance guarantees [10]. These words are then used for training G2P in a target language.

We study the effect of using lexicons generated entirely by G2P in cross-lingual ASR. The word error rate (WER) of these systems on actual speech indicates the quality of the pronunciation lexicons. For languages without lexicons this is one of the few ways of evaluating the quality of generated pronunciations.

Our contributions are: 1. We re-frame the state of the art G2P subset selection problem as a submodular optimization problem, allowing us to scale to much larger cross-lingual candidate pools. 2. We show that the same procedure can be applied in a cross-lingual setting, without the need for language-based distance criteria. 3. We show the resulting lexicons are viable for cross-lingual ASR and evaluate competing lexicon generation methods on languages for which no ground-truth lexicons are available.

2. RELATED WORK

The works most similar to ours are [7], and [8]. [7] presents a greedy selection scheme based on feature coverage for G2P,

This work was supported by DARPA LORELEI Grant № HR0011-15-2-0024

while [8] proposes a cross-lingual transfer of trained G2P models using language similarity and phoneme similarity metrics to adapt their trained G2P models. [11, 12, 13] examine the relationship between G2P performance and ASR quality on high-resource languages with large seed lexicons.

A wealth of work on submodular optimization for problems in ASR exists, but this particular problem has not yet been addressed to our knowledge. [14], for instance, uses a trained ASR model to obtain frame-level features used in submodular selection of a smaller subset with performance similar to using the full set in order to speed up computation; [15] also use submodular optimization in order to select phonetically balanced sentences for ASR training.

There has been extensive work in cross-language transfer for ASR, but to our knowledge none of it addresses the case where neither transcribed speech nor pronunciation lexicon is available. [16, 17] among others studied cross-language transfer without transcribed speech. However, they assume a lexicon exists, and explore how to map from the (sub)phonemic units of one language to another using small amounts of transcribed speech. [18] examines unsupervised adaptation approaches that use untranscribed speech and a lexicon in a new language to build a Czech-language ASR from scratch. [19], however, assumes that no pronunciation lexicon is available, but does have access to a small amount of transcribed speech. This can be used to train a probabilistic lexicon mapping prexisting acoustic units to graphemes in a new language.

3. SUBMODULAR SELECTION FOR G2P

The key idea behind our approach is to select G2P training words that are most orthographically representative of target language words. We frame this subset selection as a constrained *submodular optimization* problem.

A submodular function is a set function with the property of diminishing returns. A function f is called submodular if $\forall j \leq i$,

$$f\left(S_i \cup \{s_n\}\right) - f\left(S_i\right) \leq f\left(S_j \cup \{s_n\}\right) - f\left(S_j\right)$$

where $S_j \subseteq S_i \subseteq V$ is an arbitrary subset of a candidate pool V containing i elements, and $s_n \in V \setminus S_i$. When furthermore

$$f\left(S_i \cup \{s_n\}\right) - f\left(S_i\right) \geq 0,\ \forall\, s_n \in \mathcal{V} \setminus S_i,\ S_i \subset \mathcal{S}$$

the function is known as monotone. The cardinality-constrained maximization of a monotone submodular function can be approximately solved using a greedy algorithm with a lower-bound approximation error factor of $\left(1 - e^{-1}\right)$ [14], and for which submodularity can be further exploited to yield a lazy version with almost linear time complexity. [14] introduced the class of feature-based functions (FBF) of a subset $\mathcal{S} \subseteq V$

$$f_{\text{fea}}\left(\mathcal{S}\right) = \sum_{u \in \mathcal{U}} \omega_u g\left(m_u\left(\mathcal{S}\right)\right). \tag{1}$$

Here, $\mathcal{U}$ is the set of all features, $m_u\left(\mathcal{S}\right)$ is the count (possibly weighted) of feature u in the set $\mathcal{S} \subseteq V$, $g\left(\right)$ is a concave, non-negative, non-decreasing monotone, function, and ω_u is a non-negative weight.

In this work we use character n-gram counts as our feature set. [15, 20] show that when $g\left(x\right) = \log\left(1 + x\right)$ and $\omega = \{\omega_u\}$ is a probability distribution, the constrained maximization of such functions is equivalent to the constrained minimization of the KL-divergence between ω and the empirical distribution of features in the selected subset. Thus, if we set ω as the sample distribution over n-grams in our target language we can interpret our algorithm as selecting a subset of words whose n-gram frequencies approach those seen in our target language.

3.1. G2P Selection Algorithm

To our knowledge the state-of-the art selection algorithm for the G2P word selection task was the Feature Coverage Maximization (FC) method presented in [7]. They propose greedily selecting words to maximize the character n-gram feature coverage of their currently selected subset. We use their objective function, but modified it to be the submodular function shown below.[1]

$$f(\mathcal{S}) = \sum_{u \in \mathcal{U}} C_u \left(1 - \eta^{-m_u(\mathcal{S})}\right). \tag{2}$$

Note that $\omega_u = C_u$ is a non-negative weight and $g\left(x\right) = 1 - \eta^{-x}$ is a non-negative, non-decreasing, monotone concave function for $\eta \geq 1$ (we use $\eta = 8.0$).

In this way we can exploit the lazy greedy algorithm [21] which significantly speeds up the selection procedure and allows us to use much larger candidate pools. We also can use a simple knapsack constraint on the length of words to discourage selecting longer words as in [22]. To do so we compute the conditional gain of adding an element s_n with cost $c(s_n)$ to the set $\mathcal{S}$ as

$$\frac{f\left(\mathcal{S} \cup \{s_n\}\right) - f\left(\mathcal{S}\right)}{c\left(s_n\right)^r}.$$

r is a weighting factor tuned to balance the two scores (we used $r = 1.0$). In our experiments we use all character 4-grams and optionally all the lower order character n-grams as well.

4. MONOLINGUAL SELECTION

We first demonstrate our submodular selection in a monolingual setting across 24 languages found in the IARPA BABEL Language Packs (LP) as well as English. We use the lexicons provided in the LPs to perform our selection and as reference

[1] We replace the term $\frac{I(m_u(\mathcal{S})<C_u)}{\eta^{m_u(\mathcal{S})}}$, which is non-smooth and non-concave, with $\eta^{-m_u(\mathcal{S})}$.

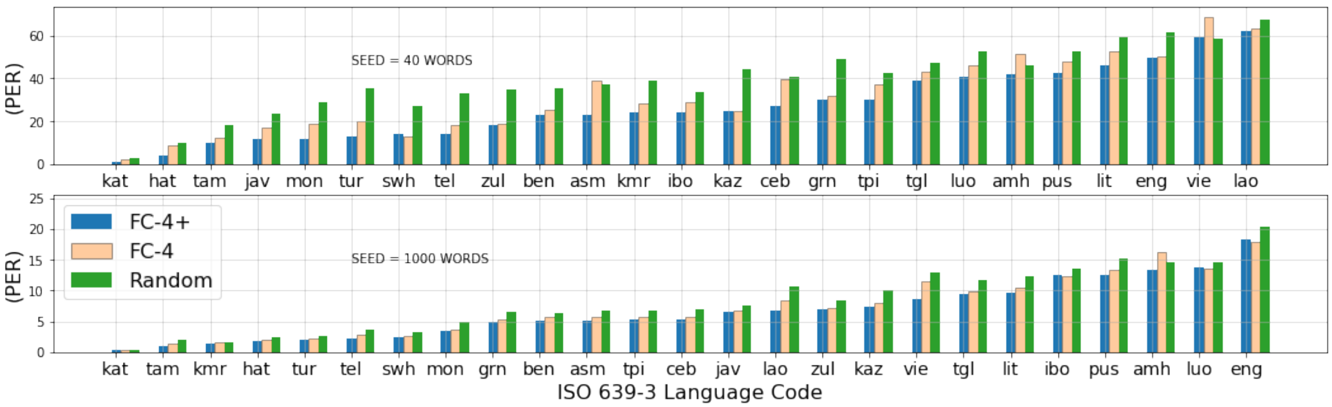

Fig. 1. G2P selection experiments in 25 languages. For each language we report the phoneme error rate (PER) of the lexicons obtained by using G2P pronunciations for all words absent from the training set. The PER is the total phoneme string edit distance across all words in the reference lexicon. FC-4 is the greedy selection in [7]. FC-4+ uses our submodular speedup, and lower order character n-gram features.

lexicons. For English, we use CMU-DICT [23]. We compare our selection strategy (FC-4+), which includes all lower order n-grams, to our implementation of the greedy selection method presented in [7], and to a random baseline averaged over 20 trials.

These experiments are not "zero-shot". They are included to demonstrate that the new selection approach performs similarly to prior approaches, on a much wider variety of languages than previously reported on, and significantly faster.

For each language we select words and their corresponding pronunciations from the ground-truth lexicon to include in G2P training. To construct a lexicon we use the ground-truth pronunciations for all selected words and for all the remaining words in the lexicon we use the G2P derived pronunciations. We then compute the phoneme error rate (PER) of the induced lexicon. In the event that a word has multiple pronunciations in the reference lexicon, the pronunciation resulting in the lowest PER is chosen as reference. In a few rare cases the G2P is not able to produce a pronunciation for a word and we use the average number of phonemes seen in word's reference pronunciations to compute the number of reference/deleted phonemes. We train the G2P in all our experiments using `Phonetisaurus` [24].

Figure 1 shows the performance of three selection strategies across all languages using different seed lexicon sizes in training. We note that both selection strategies significantly outperform random selection with the exception of the 40-word Vietnamese seed lexicon. For consistency we showed results on all languages using the same selection strategy, however, most Vietnamese words are shorter than 4 characters and hence selection based on coverage of character 4-grams was sub-optimal. Our modified selection strategy (FC-4+) outperforms the greedy selection strategy (FC) of [7], in almost every case, while training significantly faster (60x when selecting 500 words).

5. ZERO-SHOT CROSS-LINGUAL LEXICAL MODELING

Re-framing the Feature Coverage (FC) objective as a submodular function enables its use in a cross-lingual scenario since it allows us to scale our algorithm to much larger candidate pools. Cross-lingual transfer of G2P systems assumes a strong relationship between graphemes and phonemes across languages. Prior work assumes that *all* words have have useful information [9] or that all words in a given set of similar languages are useful [8]. Our selection strategy instead selects training examples on a word-by-word basis from any language.

As a preliminary test we examine two selection strategies on French, Spanish, and Vietnamese (see Figure 2). For the French, and Spanish lexicons, we scraped Wiktionary using `Wikt2pron`.[2] For Vietnamese, we use the lexicon provided in the BABEL Vietnamese LP [25]. We used the remaining 23 lexicons from the BABEL languages with segmental writing systems as our candidate pool in addition to scraped lexicons in Russian, English, Italian, Portuguese, Ukrainian, and Catalan.

We use the KL-divergence minimum between sample distributions of grapheme features in the selected subset and the test set to estimate the optimal subset size. We first select a large number of words and then pick the subset size with the smallest KL-divergence. Figure 3 shows the a typical behavior of KL-divergence as a function of subset size.[3] Plots for French and Vietnamese were similar.

In these preliminary experiments we included lower order n-gram features. However, we noticed that in a cross-lingual setting this resulted in prioritizing the selection of shorter words from other languages and worse performance in PER.

[2] `https://github.com/abuccts/wikt2pron`

[3] We scaled the KL-divergence appropriately for illustration only.

For this reason in all subsequent cross-lingual selection experiments, we only use character 4-grams as features.

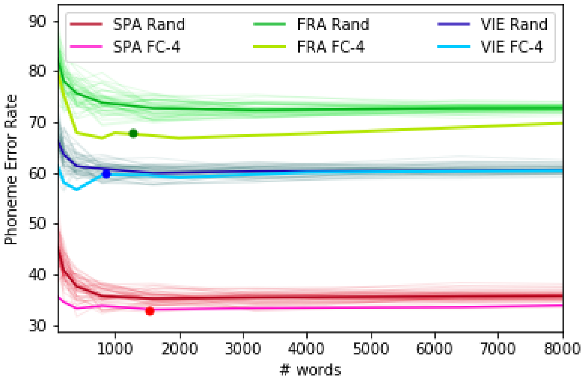

Fig. 2. Random cross-lingual selection compared with FC selection. Note that the optimal subset does not include all the data. The circles represent the chosen lexicons subset sizes.

We use an approach closely resembling that of [8] as a zero-shot comparison to our algorithm. We create G2P training data from languages similar to the target language. We used the pronunciation lexicons of the top 3 most similar languages from our pool of training data, using the `lang2lang` similarity metric of [8]. Table 1 compares our selection method with selection via `lang2lang`. We mapped phoneme outputs to the closest phoneme in a target language's inventory using the phonetic distance metric, `phon2phon`, as in [8]. While the output mapping improved phoneme error rate, we saw no significant difference in performance between our methods. This suggests that our selection strategy is working comparably with the method proposed in [8], but without using an explicit language-similarity metric.

6. CROSS-LANGUAGE ACOUSTIC MODEL TRANSFER

Our goal is to study the performance of cross-language acoustic model transfer in the absence of a lexicon in a new language. To this end, we train acoustic models on 300h of data from 25 languages all sharing a common phonemic representation. This includes about 10h in 21 different languages from the IARPA BABEL corpora, a 20h subset of the Wall Street Journal, Hub4 Spanish Broadcast news, and the Russian and French portions of the Voxforge corpus. We then transfer these models to a new language by rebuilding the decoding graph using a pronunciation lexicon with the same phonetic representation as used in training, and a language model estimated from target language text. This also provides a natural way of measuring the quality of a pronunciation lexicon. We note that since the same pronunciation lexicons that form the pool of candidate words are used in acoustic model training, we completely avoid the problem of unseen phonemes in a new language. As in [26] we split diphthongs and triphthongs to increase phoneme coverage.

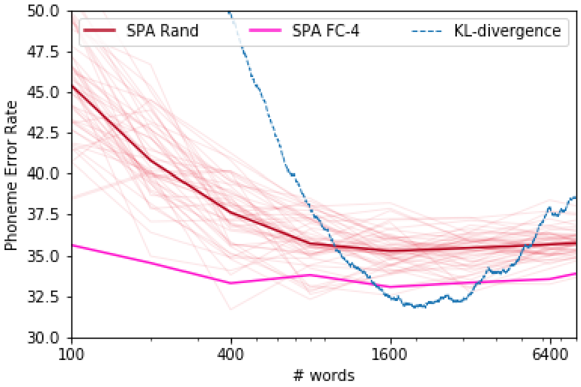

Fig. 3. Our proposed feature coverage (FC-4) approach compared to 20 trials of random selection (Rand). KL-divergence stopping criteria indicates when phoneme error rate is likely to be at a minimum.

Our model is a hybrid ASR system trained in KALDI [27]. The neural network is an 11-layer TDNN [28] with 512 hidden units per layer, interleaved with batchnorm and dropout layers trained under the LF-MMI objective [29]. We combine all the transcribed speech and pronunciation lexicons together for training. Since we trained on mixed bandwidth data (8 and 16 kHz) we chose to down-sample the wide-band data to 8 kHz when learning alignments. However, when training the TDNN, we use MFCC+Pitch features up-sampled to 16 kHZ to match the evaluation data. Save these modifications, we trained our system using the BABEL s5d recipe in KALDI.

We evaluate our lexicons by decoding speech from the CMU Wilderness Multilingual Speech Dataset [30]. The data consists of dramatized readings of the bible in around 700 languages, many without lexicons. For each language, we create training, development, and evaluation sets using 80-10-10 splits respectively. All WER results are reported on the evaluation sets.

To contextualize the results from cross-language acoustic model transfer, we also report the WER of a speech recognizer trained on the training set of each language for which we had ground-truth lexicons (See table 2). For these monolingual systems, we trained a 6-layer TDNN with 512 hidden units each. We use a smaller model for the monolingual systems since we are training on a relatively small amount of data compared to the multilingual models.

	FC		L2L	
	-	map	-	map
Spanish	38.51	**29.47**	42.69	40.16
Cebuano	60.46	58.67	56.01	**55.08**
Tagalog	64.53	64.24	58.91	**58.88**
Javanese	57.51	**30.64**	68.22	35.41
Russian	72.92	**57.22**	63.73	59.75
Kazakh	82.33	80.29	82.35	**80.02**
Average	62.71	**53.42**	62.0	54.38

Table 1. Phoneme error rates against a reference lexicon of the proposed method (Feature coverage; FC), and a language-similarity based approach (L2L) inspired by [8]. For each method, an alternative variation maps OOV phonemes to the closest target language phoneme. The discrepancy between the Spanish PER in figure 3 and this table is because the lower order character n-gram features were no longer used.

Language	Spa	Ceb	Tag	Jav	Rus	Kaz
WER	5.97	7.23	7.99	23.16	16.56	31.83

Table 2. Monolingual ASR results across 6 languages for which ground truth lexicons exist.

7. EXPERIMENTS

We compare three methods for generating pronunciations for a given phoneme-like unit set: Epitran (Epi), the approach of [8] that selects G2P training subsets based on language similarity (L2L), and our cross-lingual submodular selection based off of the Feature Coverage (FC) approach in [7]. The Epitran approach relies on ground-truth knowledge of the new language's phoneme inventory as well as the most important grapheme-to-phoneme relationships in the new language. Since these lexicons are extremely close to ground-truth lexicons we can use them to determine how much performance degradation we incur when using the cross-lingual lexicons. L2L serves as a baseline alternative method for cross-lingual G2P transfer. In every experiment we evaluate ASR performance by rebuilding the decoding graph using the appropriate lexicons. For both cross-lingual methods we examine to what extent knowledge about the target language's phoneme inventory can help in cross-lingual acoustic model transfer by again mapping all phonemes in the G2P pronunciations to the closest phoneme in the new phoneme inventory using `phon2phon` from [8].

We note that most of the languages in the CMU Wilderness Multilingual Speech Dataset do not have supporting Epitran modules. We chose our evaluation languages in order to be able to make this comparison, but for most languages this approach requires first creating an Epitran module.

		map		no map		
	Epi	L2L	FC	L2L	FC	All
Malay	**7.3**	13.1	12.6	11.6	11.1	17.8
Indonesian	**23.2**	30.1	26.2	28.3	27.4	34.7
Hausa	32.9	40.2	35.0	37.9	**30.8**	31.5
Swedish	78.5	84.0	75.2	72.2	**71.5**	72.0
Spanish†	**8.5**	57.2	32.6	49.7	35.2	36.9
Cebuano†	**21.1**	35.2	29.7	30.0	27.2	29.5
Tagalog†	**24.6**	32.0	30.0	30.3	29.8	32.4
Russian†	**45.8**	68.4	71.7	72.1	74.4	75.2
Javanese†	**47.9**	63.4	51.3	55.0	52.8	57.6
Kazakh†	**57.1**	79.1	79.2	81.1	80.9	78.1
Avg.	**33.6**	50.3	44.3	46.8	44.1	46.6
– {Spa,Rus,Kaz}	**32.1**	42.6	37.1	37.9	35.8	39.4
{Spa,Rus,Kaz}	**37.1**	68.2	61.2	67.6	63.5	63.4

Table 3. Decoding results (WER) using the universal phoneset ASR with different pronunciation lexicons for a variety of languages. `Map` indicates application of `phon2phon`. *Epi* is an Epitran model and is supervised. The rest are zero-shot: *L2L* is trained on lexicons from similar languages, *FC* uses submodular selection to match n-gram distributions, and *all* uses all training data. Languages marked with † were seen in training, though from a different corpus. Underlined is the best cross-lingual zero-shot approach. – {Spa,Rus,Kaz} is the averaged result when removing Spanish, Russian, and Kazakh. {Spa,Rus,Kaz} are the average results of only those languages.

# Words	50	100	500	1000	2000	All	FC
Spanish	26.53	20.24	19.91	18.63	16.67	13.44	35.2
Cebuano	40.45	26.45	25.69	25.47	27.20	28.89	27.2
Tagalog	49.33	36.31	27.98	31.97	31.70	34.61	29.8
Russian	87.49	65.54	49.00	48.87	48.22	47.39	74.4
Javanese	44.41	44.09	45.47	44.22	44.12	44.46	52.8
Kazakh	64.05	54.14	54.83	54.03	53.51	54.85	80.9

Table 4. WER of Acoustic model transfer using G2P lexicons where the G2P is trained on subsets of words from the target language itself. We compare subset sizes ranging from 50 words to the whole ground-truth lexicon to our FC selection. We see that for all languages except Cebuano and Tagalog, only a few G2P training examples are necessary to outperform cross-lingual subset selection. However, for both Cebuano and Tagalog, about 100-500 training examples are needed to outperform FC selection.

Table 3 shows cross-lingual ASR results. We note our selection method outperforms L2L (contrary to evaluation against ground-truth lexicons). However, the `phon2phon` mapping *degrades*, rather than helps performance. This could indicate that knowledge of the ground-truth phoneme inventory may not be useful. In practice, we saw that the post-processing applied by phon2phon either corrected phoneme substitutions that would not have resulted in ASR errors or on occasion introduced detrimental errors.

8. ANALYSIS

In 7 of the 10 evaluation languages the cross-lingual approaches perform on average only slightly worse than when using the Epitran lexicons. However, for Spanish, Russian, and Kazakh there is a large performance gap. A significant portion of this gap can be attributed to bad pronunciations for the most frequent words. In the mapped FC Spanish lexicon, for instance, the pronunciation for the word "que" was "q { w e". Simply replacing this single pronunciation with the one found in the Epitran lexicon reduced the WER from 32.6 to 28.1.

Finally in order to compare the value of cross-lingual selection to monolingual selection, we select different training example subset sizes from the ground-truth lexicons themselves. Since we are evaluating the selected subsets for G2P training we do *not* use the ground-truth pronunciations for selected words. We instead use the G2P derived pronunciations for all words. Table 4, shows that the same 3 languages that exhibited poor performance in the cross-lingual experiments benefit greatly from from monolingual selection. We suspect that knowledge of the ground-truth phoneme inventory and a better way of preventing the G2P from producing pronunciations outside this inventory might greatly improve results on these languages.

9. CONCLUSION

We presented a method for creating pronunciation lexicons by selecting words from other languages that are similar to the target language. Evaluation without a reference lexicon demonstrates their efficacy in a downstream low-resource ASR task using a universal phoneset acoustic model. Our submodular selection creates usable pronunciation lexicons with no human effort, outperforming alternative zero-shot selection strategies. Future work should focus on automatic discovery of phonemic inventory from untranscribed speech and a tighter integration of this inventory with pronunciation generation than the current phon2phon post-processing approach.

10. REFERENCES

[1] M Paul Lewis, Gary F Simons, and Charles D Fennig (eds.), *Ethnologue: Languages of the World, Eighteenth edition*, SIL International, Dallas, Texas, 2015.

[2] Sameer Maskey, Alan Black, and Laura Tomokiya, "Boostrapping phonetic lexicons for new languages," in *Eighth International Conference on Spoken Language Processing*, 2004.

[3] Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur, "Acoustic data-driven pronunciation lexicon generation for logographic languages," in *Proc. ICASSP*, 2016.

[4] David R Mortensen, Siddharth Dalmia, and Patrick Littell, "Epitran: Precision G2P for many languages," in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*, 2018.

[5] Tanja Schultz, Alan W Black, Sameer Badaskar, Matthew Hornyak, and John Kominek, "Spice: Web-based tools for rapid language adaptation in speech processing systems," in *Eighth Annual Conference of the International Speech Communication Association*, 2007.

[6] John Kominek and Alan W Black, "Learning pronunciation dictionaries: language complexity and word selection strategies," in *Proc. NAACL*, 2006.

[7] Young-Bum Kim and Benjamin Snyder, "Optimal data set selection: An application to grapheme-to-phoneme conversion," in *Proc. NAACL*, 2013.

[8] Aliya Deri and Kevin Knight, "Grapheme-to-phoneme models for (almost) any language," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2016, vol. 1, pp. 399–408.

[9] Ben Peters, Jon Dehdari, and Josef van Genabith, "Massively multilingual neural grapheme-to-phoneme conversion," in *Proceedings of the First Workshop on Building Linguistically Generalizable NLP Systems*, 2017, pp. 19–26.

[10] A. Krause and D. Golovin, "Submodular function maximization," *Tractability: Practial Approaches to Hard Problems*, February 2014.

[11] Maximilian Bisani and Hermann Ney, "Multigram-based grapheme-to-phoneme conversion for LVCSR," in *Eighth European Conference on Speech Communication and Technology*, 2003.

[12] Denis Jouvet, Dominique Fohr, and Irina Illina, "Evaluating grapheme-to-phoneme converters in automatic speech recognition context," in *Proc. ICASSP*, 2012.

[13] Tim Schlippe, Sebastian Ochs, and Tanja Schultz, "Grapheme-to-phoneme model generation for indo-european languages," in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2012, pp. 4801–4804.

[14] Kai Wei, Yuzong Liu, Katrin Kirchhoff, and Jeff Bilmes, "Unsupervised submodular subset selection for speech data," in *Proc. ICASSP*, 2014.

[15] Yusuke Shinohara, "A submodular optimization approach to sentence set selection," in *Proc. ICASSP*, 2014.

[16] Tanja Schultz and Alex Waibel, "Experiments on cross-language acoustic modeling," in *Seventh European Conference on Speech Communication and Technology*, 2001.

[17] Tanja Schultz and Alex Waibel, "Language-independent and language-adaptive acoustic modeling for speech recognition," *Speech Communication*, vol. 35, no. 1-2, pp. 31–51, 2001.

[18] Ngoc Thang Vu, Franziska Kraus, and Tanja Schultz, "Cross-language bootstrapping based on completely unsupervised training using multilingual a-stabil," in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2011, pp. 5000–5003.

[19] Ramya Rasipuram and Mathew Magimai-Doss, "Acoustic and lexical resource constrained asr using language-independent acoustic model and language-dependent probabilistic lexical model," *Speech Communication*, vol. 68, pp. 23–40, 2015.

[20] Jeffrey Bilmes and Wenruo Bai, "Deep submodular functions," *arXiv preprint arXiv:1701.08939*, 2017.

[21] Michel Minoux, "Accelerated greedy algorithms for maximizing submodular set functions," *Optimization Techniques*, pp. 234–243, 1978.

[22] Hui Lin and Jeff Bilmes, "Multi-document summarization via budgeted maximization of submodular functions," in *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 2010, pp. 912–920.

[23] "The CMU Pronouncing Dictionary," `http://www.speech.cs.cmu.edu/cgi-bin/cmudict`, Accessed: 2019-01-04.

[24] Josef Robert Novak, Nobuaki Minematsu, and Keikichi Hirose, "Phonetisaurus: Exploring grapheme-to-phoneme conversion with joint n-gram models in the wfst framework," *Natural Language Engineering*, vol. 22, no. 6, pp. 907–938, 2016.

[25] Tony Andrus et al., "IARPA Babel Vietnamese Language Pack IARPA-babel107b-v0.7 LDC2017S01. IARPA (Intelligence Advanced Research Projects Activity) Babel program, distributed via LDC," Tech. Rep., ISLRN 401-277-958-467-7, 2017.

[26] Kate M Knill, Mark JF Gales, Anton Ragni, and Shakti P Rath, "Language independent and unsupervised acoustic models for speech recognition and keyword spotting," in *Proc. Interspeech*, 2014.

[27] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukáš Burget, Ondřej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlíček, Yanmin Qian, Petr Schwarz, et al., "The Kaldi speech recognition toolkit," in *Proc. ASRU*, 2011.

[28] Vijayaditya Peddinti, Daniel Povey, and Sanjeev Khudanpur, "A time delay neural network architecture for efficient modeling of long temporal contexts," in *Proc. Interspeech*, 2015.

[29] Daniel Povey, Vijayaditya Peddinti, Daniel Galvez, Pegah Ghahremani, Vimal Manohar, Xingyu Na, Yiming Wang, and Sanjeev Khudanpur, "Purely sequence-trained neural networks for ASR based on lattice-free MMI," in *Proc. Interspeech*, 2016.

[30] Alan W Black, "CMU Wilderness Multilingual Speech Dataset," in *IEEE ICASSP*, 2019.