Automatic Error Recovery for Pronunciation Dictionaries

Tim Schlippe, Sebastian Ochs, Ngoc Thang Vu, Tanja Schultz

Cognitive Systems Lab, Karlsruhe Institute of Technology (KIT), Germany

tim.schlippe@kit.edu

Abstract

In this paper, we present our latest investigations on pronunciation modeling and its impact on ASR. We propose completely automatic methods to detect, remove, and substitute inconsistent or flawed entries in pronunciation dictionaries. The experiments were conducted on different tasks, namely (1) word-pronunciation pairs from the Czech, English, French, German, Polish, and Spanish *Wiktionary* [1], a multilingual wiki-based open content dictionary, (2) our *GlobalPhone* Hausa pronunciation dictionary [2], and (3) pronunciations to complement our Mandarin-English *SEAME* code-switch dictionary [3]. In the final results, we fairly observed on average an improvement of 2.0% relative in terms of word error rate and even 27.3% for the case of English *Wiktionary* word-pronunciation pairs.

Index Terms: pronunciation dictionaries, automatic error recovery, multilingual speech recognition

1. Introduction

With more than 6,900 languages in the world, the biggest challenge today is to rapidly port speech processing systems to new languages with low human effort and at reasonable cost. This includes the creation of qualified pronunciation dictionaries. The dictionaries provide the mapping from the orthographic form of a word to its pronunciation, which is useful in both text-to-speech and automatic speech recognition (ASR) systems. They are used to train speech processing systems by describing the pronunciation of words according to manageable units, typically phonemes [4]. Pronunciation dictionaries can also be used to build generalized grapheme-to-phoneme (g2p) models, for the purpose of providing pronunciations for words that do not appear in the dictionary [5].

The production of pronunciation dictionaries can be time-consuming and expensive if they are manually written by language experts. Therefore several approaches to automatic dictionary generation have been introduced in the past. [6] proposes heuristical and statistical methods. [7] apply letter-to-sound rules. Often, these methods still require post-editing by a human expert or leverage off another manually generated pronunciation dictionary [8][9][10]. Dictionary creation processes that combine machine learning with minimal human intervention were proposed by [11] and [12]. [13] and we [14][5] describe automatic methods to produce dictionaries using word-pronunciation pairs found in the World Wide Web.

As pronunciation dictionaries are so fundamental to speech processing systems, much care has to be taken to select a dictionary that is as free of errors as possible. For ASR systems, faulty pronunciations in the dictionary may lead to incorrect training of the system and consequently to a system that does not function to its full potential. Flawed or inadequate dictionary entries can originate from different subjective judgments, small typographical errors, and 'convention drift' by multiple annotators. As pronunciations from the World Wide Web often lack information about the corresponding word or language, it may happen that inappropriate word-pronunciation pairs are collected. Correct pronunciations that do not match the target domain or accent can also lead to worse ASR performance. For g2p extraction algorithms the correctness of the dictionary is equally important as each erroneous entry can cause an incorrect g2p model to be generated, thereby compromising the created dictionary.

Different approaches to detect flawed entries have been described in the past. [16] apply a stochastic g2p model to the task of dictionary verification and detect spurious entries, which can then be examined and corrected manually. [4] focus on mechanisms to identify incorrect entries that require limited human intervention. The techniques for verifying the correctness of a dictionary include word-pronunciation length relationships, g2p alignment, g2p rule extraction, variant modeling, duplicate pronunciations, and variant analysis. The automated correction of these entries is not investigated and erroneous entries are simply removed. [15] propose a semi-automated development and verification process. The extraction of g2p rules provides an immediate avenue for error detection: by cross-validating the dictionary errors made by the g2p predictor can be flagged for verification [18]. g2p rules themselves may also be able to identify highly irregular training instances [4] or provide an indication of the likelihood of a specific pronunciation [17][16] in order to flag possible errors. In [19] and [20], g2p accuracy is considered an indicator of dictionary consistency, especially where variants are concerned. Inconsistencies lead to unnecessarily complex pronunciation models, and consequently, suboptimal

generalization. [20] generate pronunciations with rules and flag pronunciations with alternative generated pronunciations. [18] describe a technique that does not only flag specific words for verification, but also presents verifiers with example words that produce pronunciation patterns conflicting with the flagged instances.

The previous approaches show separate single methods to detect inconsistent dictionary entries. To correct those entries, they substitute them manually in separate processes. The annotation of flagged entries may be still costly and time-consuming. Therefore we investigate the performance of different fully automatic data-driven methods to detect, remove and substitute such entries. We determine the thresholds for removing word-pronunciation pairs completely on the data in the dictionaries and generate pronunciations with validated g2p models where they have been removed. For better filtering, we experiment with single and 2-stage approaches.

Most approaches reported in related work have not been evaluated in ASR experiments. We investigate the performance of our methods on different tasks and check their impact on ASR: First we apply our methods to Czech, German, English, Spanish, French, and Polish *Wiktionary* word-pronunciation pairs that contain many different errors and inconsistencies and check the quality of resulting g2p models. Then we analyze their impact on the *GlobalPhone* Hausa pronunciation dictionary which had been manually cross-checked but still contains a few errors. Finally, we use our methods to select pronunciations from an additional dictionary to enhance the *SEAME* code-switch dictionary that contains entries of Mandarin and English with Singaporean and Malayan accent to transcribe Mandarin-English code-switching conversational speech.

2. Automatic rejection of inconsistent or flawed entries

Our investigated methods to filter erroneous word-pronunciation pairs fall into the following categories:

1. Length Filtering (Len)
   (a) Remove a pronunciation if the ratio of grapheme and phoneme tokens exceeds a certain threshold.
2. Epsilon Filtering (Eps)
   (a) Perform a 1-1 g2p alignment [4][7] which involves the insertion of graphemic and phonemic nulls (epsilons) into the lexical entries of words.
   (b) Remove a pron. if the proportion of graphemic and phonemic nulls exceeds a threshold.
3. m-n Alignment Filtering (*M2NAlign*)
   (a) Perform an M-N g2p alignment [4][7].
   (b) Remove a pronunciation if the alignment score exceeds a threshold.
4. g2p Filtering ($G2P$)
   (a) Train g2p models with "reliable" word-pronunciation pairs.
   (b) Apply the g2p models to convert a grapheme string into a most likely phoneme string.
   (c) Remove a pronunciation if the edit distance between the synthesized phoneme string and the pronunciation in question exceeds a threshold.

The threshold for each filtering method depends on the mean (μ) and the standard deviation (σ) of the measure in focus (computed on all word-pronunciation pairs), i.e. the ratio between the numbers of grapheme and phoneme tokens in Len, the ratio between the numbers of graphemic and phonemic nulls in Eps, the alignment scores in *M2NAlign*, and the edit distance between the synthesized phoneme string and the pronunciation in question in $G2P$. Those word-pronunciation pairs whose resulting number is shorter than $\mu-\sigma$ or longer than $\mu+\sigma$ are rejected.

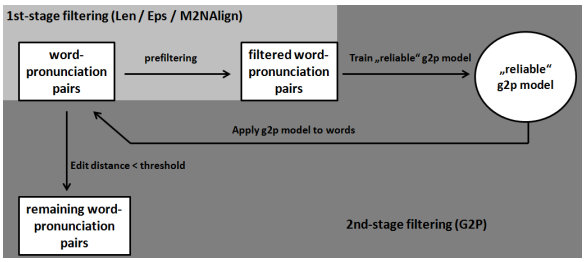

Figure 1: *2-Stage Filtering.*

To provide "reliable" pronunciations for $G2P$, we propose to prefilter the word-pronunciation pairs by applying Len, Eps or *M2NAlign*, as shown in Figure 1 (*1st stage*). On the remaining word-pronunciation pairs, $G2P$ is applied (*2nd stage*). Those *2-stage filtering* methods are represented as $G2P_{Len}$, $G2P_{Eps}$, and $G2P_{M2NAlign}$ in the following sections. Experiments with 2 stages in the prefiltering before $G2P$ (*3-stage filtering*) were too restrictive and did not leave enough training data for reliable g2p models. If a validated dictionary already exists, μ and σ can be computed on its entries. All our single and 2-stage data-driven methods expect more good than bad pronunciations in the data to obtain good estimates for μ and σ. With our pure statistical methods, no manual labor and linguistic knowledge is required.

3. Experiments and Results

We investigate the performance of our proposed methods by addressing the following 3 possible tasks occurring in the development of ASR systems: (1) We focus on *Wiktionary*, a source for online pronunciations to bootstrap dictionaries, as we expect most erroneous pronunciation entries there due to its crowdsourcing-based cre-

	baseline	$G2P$	Len	$G2P_{Len}$	Eps	$G2P_{Eps}$	*M2NAlign*	$G2P_{M2NAlign}$	∅ rel. improv.	best rel. improv.
cs	18.72	17.86	18.24	17.85	**17.74**	18.15	18.20	17.93	3.87	5.24
de	16.81	17.18	17.13	**16.79**	17.12	17.08	17.53	17.18	0.12	0.12
en	28.86	30.00	23.68	24.74	22.85	22.90	**20.97**	23.73	19.80	27.34
es	12.82	13.14	13.50	13.05	12.99	12.86	**12.25**	13.64	4.45	4.45
fr	25.79	25.62	25.48	25.59	**23.19**	25.44	25.70	25.03	2.39	10.08
pl	17.21	17.00	17.38	17.31	16.98	16.68	16.87	**16.57**	2.03	3.08

Table 1: WERs (%) with dictionaries built completely with g2p generated Wiktionary pronunciations.

	baseline	$G2P$	Len	$G2P_{Len}$	Eps	$G2P_{Eps}$	*M2NAlign*	$G2P_{M2NAlign}$	∅ rel. improv.	best rel. improv.
ha	23.49	23.68	23.20	**22.88**	23.30	23.15	23.17	23.11	1.51	2.60

Table 2: WERs (%) for Hausa.

ation. (2) An LVCSR dictionary which has been manually checked under supervision can still have a few errors and inconsistencies. We apply our methods on the *GlobalPhone* Hausa dictionary which represents such a dictionary. (3) The straightforward insertion of new pronunciation variants into an existing dictionary can lead to ASR performance degradations if the new pronunciations do no match the target domain or accent. We filter English pronunciation variants from a new dictionary that do not match the existing Singaporean/Malayan English pronunciations in our English-Mandarin code-switch dictionary. For all experiments, we report the error rates of the alignment that performed best on the particular dictionary.

3.1. *Wiktionary* Pronunciations

The World Wide Web has been increasingly used as a text data source for rapid adaptation of ASR systems and initial investigations to leverage off available pronunciations have been described [5][13][14]. g2p correspondences from the web-derived word-pronunciation pairs can be used to build statistical g2p models. With these models pronunciations for out-of-vocabulary words or pronunciation variants for existing words can be produced. To accumulate training data for g2p models, we downloaded dumps of 6 *Wiktionary* editions (cs (Czech), de (German), en (English), es (Spanish), fr (French), pl (Polish)) for which we have dictionaries from the *GlobalPhone* database [21] and parsed them for IPA notations.

After applying our filtering methods on the word-pronunciation pairs of each *Wiktionary* dump, we built g2p models with extracts from the remaining data (5 - 30k phoneme tokens with corresponding grapheme tokens to reflect a saturated g2p consistency [5]). Then we replaced the pronunciations in the dictionaries of the Czech, German, English, Spanish, French, and Polish *GlobalPhone* ASR systems with pronunciations generated with the *Wiktionary* g2p models. Finally, we trained and decoded the systems completely with those dictionaries. The *baseline* dictionaries were made with g2p models trained on randomly selected, unfiltered word-pronunciation pairs of an amount equal to the one used to train the filtered models. Table 1 shows that we are able to reduce the WER of all tested systems, while the success of each method differs among languages. We see improvements with $G2P_{Len}$, Eps, *M2NAlign*, and $G2P_{M2NAlign}$. With a WER reduction of 27.3% relative, most improvement is achieved on the English *Wiktionary* word-pronunciation pairs with *M2NAlign*. Samples in the original English data indicate a high number of pronunciations from other languages, pronunciations for stem or ending instead of the whole word or completely different pronunciations that result in a bad initial dictionary. Without this outlier, the average improvement is 2.5% relative. $G2P$ and Len do not improve the systems.

3.2. The *GlobalPhone* Hausa Dictionary

For the African language Hausa, we collected almost 9 hours of speech from 102 Hausa speakers reading newspaper articles as a part of our *GlobalPhone* corpus [2].

We evaluate our filtering methods on the initial *GlobalPhone* dictionary which has been created in a rule-based fashion and was then manually revised and cross-checked by native speakers and causes a WER of 23.49% (*baseline*). After filtering, we used the remaining word-pronunciation pairs to build new g2p models and applied them to the words with rejected pronunciations. Then we trained and decoded the Hausa system with each processed dictionary. Table 2 shows that we are able to reduce the WER with all filtered dictionaries but $G2P$ by 1.5% relative on average. $G2P_{Len}$ performs best with 2.6% relative improvement.

3.3. The English-Mandarin Code-Switch Dictionary

SEAME [3] contains 157 speakers and approximately 52k intra-sentential English-Mandarin code-switching utterances. The recorded speakers speak Singaporean/Malayan English which differs strongly from American English. In addition to our previous pronunciation dictionary (*prev*), our partners generated a new dictionary for Singaporean English (*new*) by applying 160 rules to the pronunciations in the American CMU dictionary which they had derived in a data-driven way. With

	prev	prev+new	*Len*	*Eps*	*G2P*	*M2NAlign1*	*M2NAlign2*	∅ rel. improv.	best rel. improv.
MERs	36.89	37.12	36.89	36.89	36.84	**36.79**	**36.79**	0.26	0.27
pronunciations/word	1.78	1.94	1.85	1.84	1.88	1.89	1.90		

Table 3: MERs (%) on the SEAME Mandarin-English Code-Switch Corpus development set.

this dictionary a WER of 16.89% was achieved on texts from the English Aurora 4 [22] [23] corpus which were read by Singaporean speakers. With the American CMU dictionary the WER was 75.19% on the same test set.

Our system with the existing dictionary (*prev*) has a mixed error rate (MER) [3] of 36.89% on SEAME. To improve *prev*, our goal was to enrich it with pronunciations from *new*. However, adding all almost 5k English pronunciations as pronunciation variants which are not equal to the pronunciations in *prev* for decoding led to a performance degradation of 0.23% absolute (*prev+new*). Therefore we applied our filtering methods to select only those pronunciations from *new* that fit the pronunciations which have been successfully used before. The mean (μ) and the standard deviation (σ) of the measure in focus were computed based on the word-pronunciation pairs of *prev* for Len, Eps, $G2P$, and *M2NAlign1*. The alignments of *M2NAlign2* were computed on those from *new*. Table 3 shows that we are able to slightly reduce the MER compared to *prev* by 0.2% relative on average with a decoding using the filtered new pronunciations. *M2NAlign1* and *M2NAlign2* marginally outperform the other methods and result in a MER reduction of 0.3% relative.

4. Conclusion and Future Work

We have presented completely automatic error recovery methods for pronunciation dictionaries. The methods are based on the means and deviations of certain characteristics computed on the word-pronunciation pairs of the dictionaries and on g2p model generation plus their application. We tested them on dictionaries from 7 languages (Czech, German, English, Spanish, French, Polish, Hausa), 1 accent (Singaporean/Malay English) and 3 tasks: (1) g2p model generation with web-derived pronunciations, (2) improve a manually cross-checked dictionary, and (3) enrich a dictionary with new pronunciation variants. Our methods improved the ASR performances in each language and task. Often the methods with a 2-stage filtering outperformed the separate single methods. Future work may include an analysis which method works how good on which kind of errors to find faster the best method for a dictionary in question. Furthermore our goal is to investigate approaches to combine the outputs of our methods.

5. Acknowledgements

This work was partly realized as part of the Quaero Programme, funded by OSEO.

6. References

[1] "Wiktionary - a wiki-based open content dictionary", Website, http://www.wiktionary.org.

[2] Schlippe, T., Komgang Djomgang, E. G., Vu, N. T., Ochs, S., and Schultz, T., "Hausa Large Vocabulary Continuous Speech Recognition", SLTU, 2012.

[3] Vu, T., Lyu, D.-C., Weiner, J., Telaar, D., Schlippe, T., Blaicher, F., Chng, E.-S., Schultz, T., and Li, H., "A First Speech Recognition System For Mandarin-English Code-Switch Conversational Speech", ICASSP, 2012.

[4] Martirosian, O. and Davel, M., "Error analysis of a public domain pronunciation dictionary", PRASA, 2007.

[5] Schlippe, T., Ochs, S., and Schultz, T., "Grapheme-to-Phoneme Model Generation for Indo-European Languages", ICASSP, 2012.

[6] Besling, S., "Heuristical and Statistical Methods for Grapheme-to-Phoneme Conversion", Konvens, 1994.

[7] Black, A. W., Lenzo, K., and Pagel, V., "Issues in Building General Letter to Sound Rules", ESCA Workshop on Speech Synthesis, 1998.

[8] Maskey, S., Tomokiyo, L., and Black, A. W., "Bootstrapping Phonetic Lexicons for new Languages", Interspeech, Jeju, Korea, October 2004.

[9] Davel M. and Barnard, E., "Bootstrapping for Language Resource Generation", PRASA, 2003.

[10] Mertens, P. and Vercammen, F., "Fonilex Manual", Tech. Rep., K. U. Leuven CCL, 1998.

[11] Kominek, J. and Black, A. W., "Learning Pronunciation Dictionaries: Language Complexity and Word Selection Strategies", HLT, 2006.

[12] Davel, M. and Barnard, E., "The Efficient Generation of Pronunciation Dictionaries: Human Factors during Bootstrapping", ICSLP, 2004.

[13] Ghoshal, A., Jansche, M., Khudanpur, S., Riley, M., and Ulinski, M., "Web-derived pronunciations", ICASSP, 2009.

[14] Schlippe, T., Ochs, S., and Schultz, T., "Wiktionary as a Source for Automatic Pronunciation Extraction", Interspeech, 2010.

[15] Davel, M. and Martirosian, O., "Pronunciation Dictionary Development in Resource-Scarce Environments", HLT, 2009.

[16] Vozila, P., Adams, J. and Thomas, R., "Grapheme to Phoneme Conversion and Dictionary Verification using Graphonemes", Eurospeech, 2003.

[17] Bisani, M. and Ney, H., "Joint Sequence Models for Grapheme-to-Phoneme Conversion", Speech Communication, 2008.

[18] Davel., M. and de Wet, F., "Verifying Pronunciation Dictionaries using Conflict Analysis", Interspeech, 2010.

[19] Wolff, M., Eichner, M., and Hoffmann, R., "Measuring the Quality of Pronunciation Dictionaries", PMLA, 2002.

[20] Davel, M. and Barnard, E., "Developing Consistent Pronunciation Variants", Interspeech, 2006.

[21] Schultz, T., "GlobalPhone: A Multilingual Speech and Text Database Developed at Karlsruhe University", ICSLP, 2002.

[22] Parihar, N., Picone, J., Pearce, D., Hirsch, H.G., "Performance analysis of the Aurora large vocabulary baseline system", European Signal Processing Conference, 2004.

[23] Au Yeung, S.-K. and Siu, M.-H."Improved performance of Aurora-4 using HTK and unsupervised MLLR adaptation", Int. Conference on Spoken Language Processing, 2004.