COMBINING GRAPHEME-TO-PHONEME CONVERTER OUTPUTS FOR ENHANCED PRONUNCIATION GENERATION IN LOW-RESOURCE SCENARIOS

Tim Schlippe, Wolf Quaschningk, Tanja Schultz

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

ABSTRACT

For pronunciation dictionary creation, we propose the combination of grapheme-to-phoneme (G2P) converter outputs where low resources are available to train the single converters. Our experiments with German, English, French, and Spanish show that in most cases the phoneme-level combination approaches validated reference pronunciations more than the single converters. In case of only little training data, the impact of the fusion is high which shows their great importance for under-resourced languages. We detected that the output of G2P converters built with web-derived word-pronunciation pairs can further improve pronunciation quality. With 23.1% relative in terms of phoneme error rate to the reference dictionary, we report the largest improvement for the scenario where only 200 French word-pronunciation pairs and web data are given as training data. In additional automatic speech recognition experiments we show that the resulting dictionaries can lead to performance improvements.

Index Terms— pronunciation dictionary, pronunciation modeling, low-resource scenarios, multilingual speech recognition, rapid language adaptation

1. INTRODUCTION

With more than 7,000 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 speech synthesis and automatic speech recognition (ASR) systems. 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 [1]. The manual production of dictionaries can be time-consuming and expensive. Therefore knowledge-based and data-driven grapheme-to-phoneme (G2P) conversion approaches for the automatic dictionary generation have been introduced (see Section 2). As pronunciation dictionaries are so fundamental to speech processing systems, much care has to be taken to create 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 dictionary entries can originate from G2P converters with shortcomings.

Our goal is to investigate if a combination of G2P converter outputs outperforms the single converters. This is particularly important for the rapid bootstrapping of speech processing systems if not many manual created example word-pronunciation (*W-P*) pairs are available and therefore a single G2P converter has a poor performance. In the case of semi-automatic pronunciation generation, enhanced pronunciations derived from the combination would reduce the editing effort and speed up the annotation process. We combine the G2P converter outputs based on a voting scheme at the phoneme-level. Our motivation is that the converters are reasonably close in performance but at the same time produce an output that differs in their errors. This provides complementary information which leads in combination to performance improvements. With the phoneme error rate (PER), we evaluate how close the resulting pronunciations come to pronunciations which have been successfully used in speech processing projects.

For training the G2P converters, we select different amounts of English, French, German, and Spanish *W-P* pairs to simulate scenarios with small amounts of *W-P* pairs, since our intention is that our approach can be applied to languages with very limited lexical resources and differing grade in G2P regularity. In additional ASR experiments we investigate the impact of the phoneme-level combination on ASR performance, especially in the context of confusion network combinations.

This paper is structured as follows: Section 2 gives an overview of knowledge-based, data-driven and semi-automatic G2P conversion approaches. In Section 3 we present the G2P converters we conduct our experiments with. We describe our experiments in Section 4. In Section 5 we conclude our work and propose further steps.

2. RELATED WORK

Knowledge-based approaches with rule-based G2P conversion systems were developed which can typically be expressed as finite-state automata [2] [3]. Often, these methods require specific linguistic skills and exception rules formulated by human experts. In contrast, data-driven G2P conversion approaches predict the pronunciation of unseen words purely by analogy. The benefit of the data-driven approach is that it trades the time- and cost-consuming task of designing rules, which requires linguistic knowledge, for the much simpler one of providing example pronunciations. [3] propose Classification and Regression Trees (CART) to the G2P task. In [4], the alignment between graphemes and phonemes is generated using a variant of the Baum-Welch expectation maximization algorithm. [5], [6] and [7] use a joint-sequence model. [8] and [9] utilize weighted finite-state transducers (WFSTs) for decoding as a representation of the joint-sequence model. [10], [11], and [12], apply statistical machine translation (SMT)-based methods for the G2P conversion. A good overview of state-of-the-art G2P methods is given in [13]. Methods to leverage off pronunciations from the World Wide Web have been introduced [14] [15] [16] [1] [17]. Furthermore several methods to generate pronunciations in a semi-automatic way have been presented [18][19][20][21][22].

3. EXPERIMENTAL SETUP

3.1. Grapheme-to-Phoneme Converters

We analyze five common G2P conversion approaches and their combination:

- SMT-based with Moses Package [23] [24] (*Moses*)
- Graphone-based with Sequitur G2P [25] (*Sequitur*)
- WFST-driven with Phonetisaurus [8] (*Phonetisaurus*)
- CART-based with t2p [3] (*Carttree*)
- Simple G2P conversion based only on the mostly uttered phoneme for each grapheme[1] (*Rules*).

3.2. Data

As our methods should work for languages with different grade of regularity in G2P relationship, our experiments are conducted with German (*de*), English (*en*), Spanish (*es*), and French (*fr*). G2P accuracy is a measure of the regularity of the G2P relationship of a language and [1] showed that the G2P accuracy for *en* is very low, for *es* it is very high, whereas *de* and *fr* are located in between.

For evaluating our G2P conversion methods, we use *GlobalPhone* dictionaries for *de* and *es* as reference data since they have been successfully used in LVCSR [26]. For *fr*, we employ a dictionary developed within the Quaero Programme. The *en* dictionary is based on the CMU dictionary[2]. All dictionaries contain words from the broadcast news domain. For each language, we randomly selected 10k *W-P* pairs from the dictionary for testing. From the remainder, we extracted 200, 500, 1k, and 5k *W-P* pairs for training to investigate how well our methods perform on small amounts of data. To evaluate the quality of the G2P converters' outputs, we apply them to the words in the test set and compute their phoneme error rate (PER) to the original pronunciations.

For the ASR experiments, we replace all pronunciations in the dictionaries of our *de* and *en GlobalPhone*-based speech recognizers with pronunciations generated with the G2P converters. Thereby we replace 39k pronunciations for German for *de* and 64k for *en*. Then we use them to build and decode LVCSR systems. The transcribed audio data and language models for *de* come from the *GlobalPhone* project [26], those for *en* from the *WSJ0* corpus [27]. Finally, we decode their test sets with the resulting systems.

As we are able to find *en*, *fr*, *de*, and *es W-P* pairs in *Wiktionary*[3], we additionally built a G2P converter with these data for each language. The quality of web-derived pronunciations is usually worse than handcrafted pronunciations. However, the *W-P* pairs from the Web can include complementary information than our given training data and we can find *W-P* pairs even for languages with no or very limited lexical resources as we have shown in [17].

4. EXPERIMENTS AND RESULTS

4.1. Analysis of the G2P Converters' Output

For all G2P converters, we use context and tuning parameters that result in lowest PERs on the test set with 1k training data. Figure 1 demonstrates the PERs of the single G2P converter outputs. The converters with lowest PER are marked with arrows. They serve as a baseline for us and we compute the relative PER change compared to their PER in Section 4.2 and 4.3. We observe lower PERs with increasing amount of training data. Lowest PERs are achieved with *Sequitur* and *Phonetisaurus* for all languages and data sizes. *Carttree* results in worse performance. *Moses* is always worse than *Sequitur* and *Phonetisaurus*, even it is very close for *de*. For 200 *en* and *fr W-P* pairs, *Rules* outperforms *Moses*. To show that the G2P converters produce different outputs, we present the edit distances at the phoneme-level between the G2P converter outputs trained with 1k *W-P* pairs in Table 1. How much they differ depends on the similarity of the corresponding technique. For example, the smallest distances are between *Sequitur* and *Phonetisaurus*, while *Rules* has the highest distances to the other approaches. It is also dependent on the G2P relationship: While the *en* outputs differ most for all

[1] In contrast to the other data-driven approaches, it represents a knowledge-based approach.

[2] http://www.speech.cs.cmu.edu/cgi-bin/cmudict

[3] http://www.wiktionary.org

	Sequitur	Phonetisaurus	Carttree	Moses
Phonetisaurus	10.8 / 4.4 / 4.8 / 1.0	X	X	X
Carttree	23.4 / 11.9 / 9.4 / 2.2	23.6 / 12.3 / 10.9 / 2.2	X	X
Moses	35.8 / 7.7 / 25.9 / 7.6	35.7 / 7.6 / 26.3 / 7.7	39.8 / 12.5 / 27.9 / 7.1	X
Rules	45.8 / 34.4 / 32.6 / 11.3	45.6 / 34.3 / 32.7 / 11.5	46.1 / 34.3 / 33.5 / 11.3	40.3 / 34.7 / 35.7 / 10.7

Table 1. Edit Distances at the Phoneme-Level between G2P Converter Outputs (en / de / fr / es).

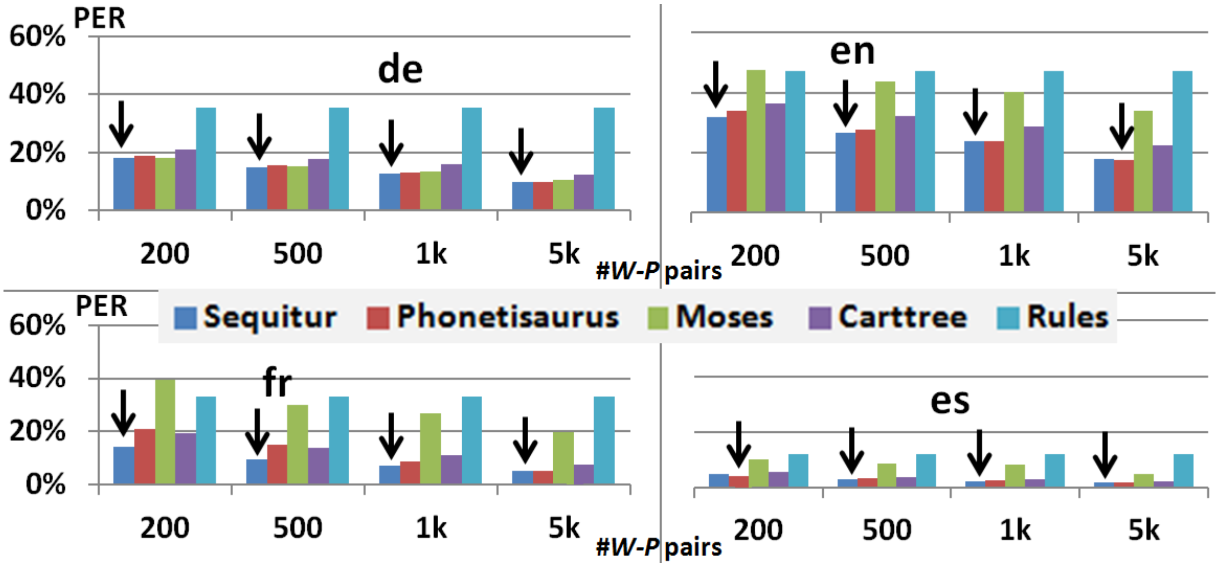

Fig. 1. PER of Single G2P Converter Outputs to Reference Pronunciations over Amount of Training Data.

amounts of training data, the *es* ones are closest. The distances of *fr* and *de* are located in between.

4.2. Phoneme-Level Combination: Combining the G2P Converters' Output

For the phoneme-level combination (*PLC*), we apply *nbest-lattice* at the phoneme-level which is part of the SRI Language Modeling Toolkit [28]. From each G2P converter we select the most likely output phoneme sequence (1st-best hypothesis). Then we use *nbest-lattice* to construct a phoneme lattice from all converters' 1st-best hypotheses and extract the path with the lowest expected PER. We detected that in some cases the combination of subsets of G2P converter outputs improved PER slightly. In other cases single much worse 1st-best G2P converter outputs even helped to improve quality. As in a real scenario the impact is not clear, we continued our experiments with the combination of all 1st-best converter outputs.

The left blue bars in Figure 2 (*PLC-w/oWDP*) show the change in PER compared to the G2P converter output with the highest quality. In 10 of 16 cases the combination performs equal or better than the best single converter. For *de*, we observe improvements for all training data sizes, for *en* slight improvements in four of five cases. Therefore we selected these languages for our ASR experiments (see Section 4.4). For *es*, the language with the most regular G2P relationship, the combination never results in improvements. While for *de* the improvement is higher with less training data, the best *fr* improvement can be found with 5k training data. Further approaches of weighting the 1st-best G2P converter outputs could only reach the quality of the best single converter and not outperform it.

4.3. Adding Web-driven G2P Converters' Output

We used *Sequitur* to build additional G2P converters based on pronunciations which we found in *Wiktionary* together with corresponding words (*WDP*) and analyzed their impact to the combination quality. The single *de* web-driven converter trained with unfiltered *W-P* pairs has a PER of 16.74%, the *en* one 33.18%, the *fr* one 14.96%, and the *es* one 10.25%. The *de* one trained with filtered *W-P* pairs has a PER of 14.17%, the *en* one 26.13%, and the *fr* one 13.97%. However, the PER of the *es* one slightly increased to 10.90%.

Figure 2 shows the changes without (*PLC-w/oWDP*) and with additional converter outputs compared to the best single converter. First we built G2P converters after we extracted *W-P* pairs from *Wiktionary* without any filtering (*PLC-unfiltWDP*). Second we filtered them before we built the G2P converters as described in [29] (*PLC-filtWDP*).

For *de*, the web-driven G2P converter's optimal training

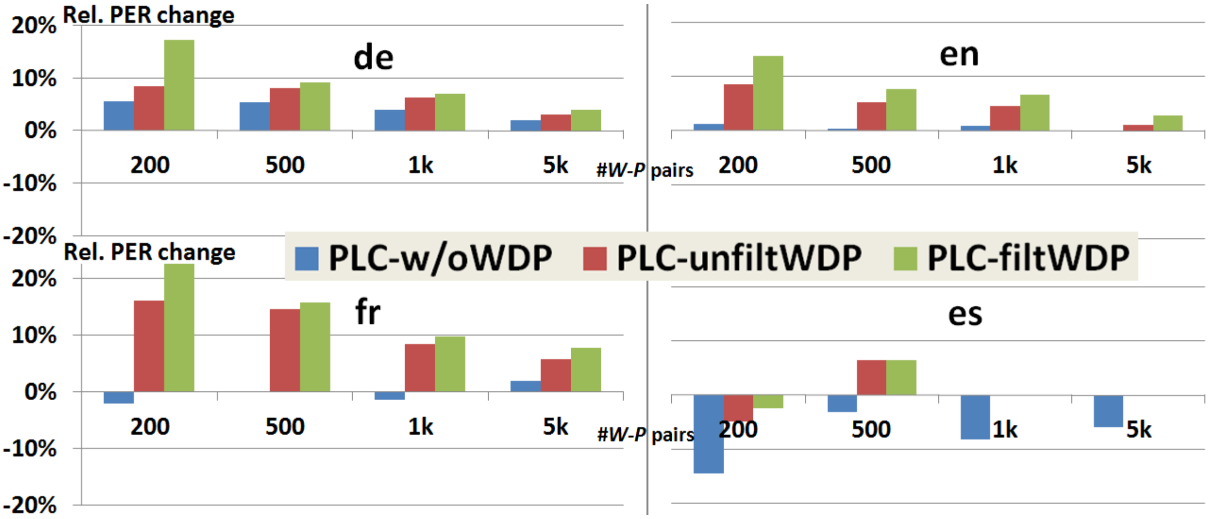

Fig. 2. Rel. PER Change to Reference Pronunciations with *PLC* using Converters Trained with Web-derived Pronunciations

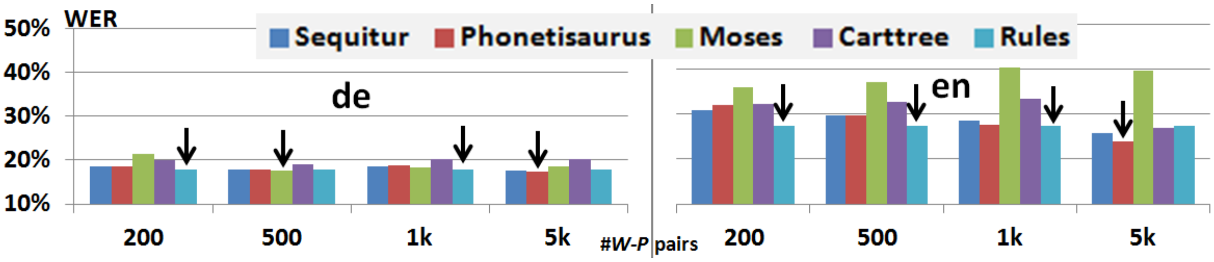

Fig. 3. WER with Dictionaries from Single G2P Converter Outputs over Amount of G2P Training Data.

data were obtained with a *2-stage filtering* ($G2P_{Len}$): First we compute the mean (μ_{Len}) and the standard deviation (σ_{Len}) of the ratio of grapheme and phoneme tokens over all *W-P* pairs. Then we remove *W-P* pairs whose ratio of grapheme and phoneme tokens was shorter than $\mu_{Len} - \sigma_{Len}$ or longer than $\mu_{Len} + \sigma_{Len}$. With the remaining *W-P* pairs we train a G2P model and apply it to convert the grapheme strings of the remaining words into a most likely phoneme strings. Then we compute the mean (μ_{G2P}) and the standard deviation (σ_{G2P}) of the edit distances between the synthesized phoneme strings and the pronunciations from the Web. Finally, we remove a *W-P* pair if the edit distance between a synthesized phoneme string and the pronunciation from the Web is shorter than $\mu_{G2P} - \sigma_{G2P}$ or longer than $\mu_{G2P} + \sigma_{G2P}$.

For *en* and *es*, the web-driven G2P converter's optimal training data were obtained with the *m-n Alignment Filtering* (*M2NAlign*). For that we perform an M-N G2P alignment [30] [3] to the web-derived *W-P* pairs. Then we compute the mean ($\mu_{M2NAlign}$) and the standard deviation ($\sigma_{M2NAlign}$) of the alignment scores. Finally, we remove *W-P* pairs whose alignment score is shorter than $\mu_{M2NAlign} - \sigma_{M2NAlign}$ or longer than $\mu_{M2NAlign} + \sigma_{M2NAlign}$.

Best quality for *fr* was achieved with *Epsilon Filtering* (Eps). For that we perform a 1-1 g2p alignment [30][3] which involves the insertion of graphemic and phonemic nulls (epsilons) into the lexical entries of words. Then we remove a *W-P* if the proportion of graphemic and phonemic nulls is shorter than $\mu_{Eps} - \sigma_{Eps}$ or longer than $\mu_{Eps} + \sigma_{Eps}$.

With each filtering method, 15% of the inconsistent web-derived word-pronunciation pairs were removed. More information about our filtering methods is described in in [29].

We observe that *PLC-unfiltWDP* outperforms the best single converter output in 15 of 16 cases. In all cases it is better than *w/oWDP*. Like *PLC-unfiltWDP*, *PLC-filtWDP* outperforms the best single method in 15 cases. However, it is in all cases better than *PLC-unfiltWDP* and better than *PLC-w/oWDP*. With 23.1% relative PER improvement, we report the largest improvement for *fr* where only 200 French *W-P* pairs and web data are given as training data.

Where our *PLC* methods improves PER, a linguist or native speaker has to change less phonemes to meet a validated pronunciation quality. Therefore *PLC* has potentials to enhance the processes of semi-automatic pronunciation dictionary creation described in [18], [19], [20], [21], and [22].

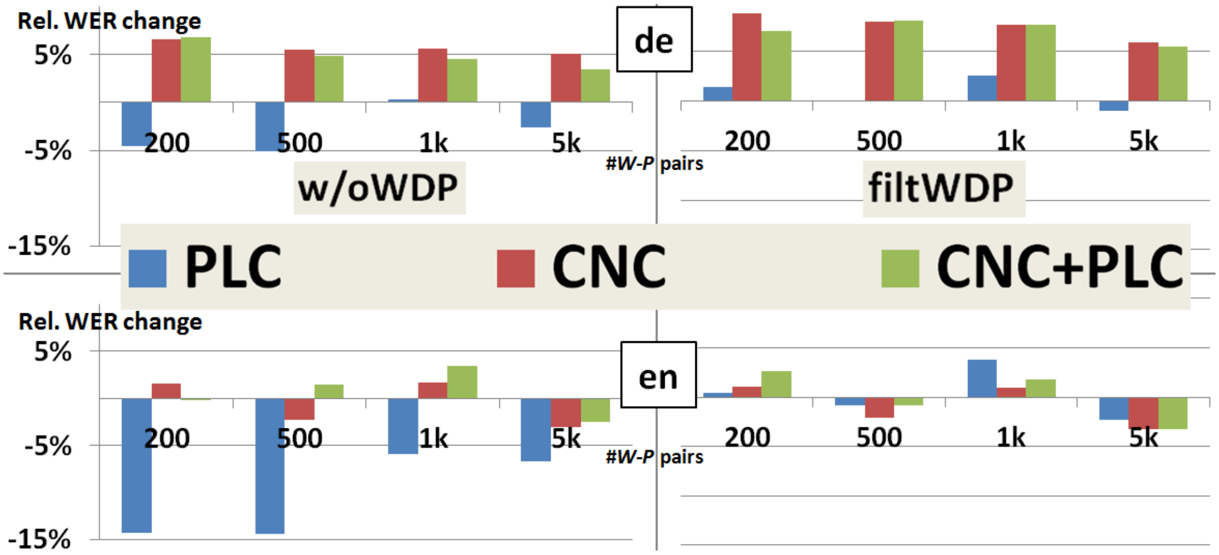

Fig. 4. Rel. WER (%) Change over Training Data Size With and Without Web-derived Data.

4.4. ASR Experiments

For *de* and *en*, we have illustrated that we can approximate validated pronunciations using *PLC*, which can be helpful for speech synthesis and to lower the editing effort in the semi-automatic dictionary generation. In the following sections we investigate if the impact of the phoneme-level combination additionally has immediately a positive impact on ASR performance. Furthermore we compare *PLC* (*early fusion*) to a combination at lattice-level (*late fusion*) from the output of individual ASR systems.

For the evaluation we built separate ASR systems for each single G2P converter as follows: We replaces the pronunciations for all words in our *de* and *en* reference dictionary (39k for *de* and 64k for *en*) with pronunciations generated with the G2P converters. Then we used them to build and decode the systems. For each *W-P* pairs size, the best performing single system serves as baseline. Then we evaluated the combination approaches with the relative change in word error rate (WER) compared to the best performing system that is trained with a dictionary that has been built with a single converter. We marked those baseline systems with arrows in Figure 3.

Figure 3 depicts variations in WER with increasing amounts of training data, even if there is a general decrease with more training data using our data-driven G2P converters except for *en* with *Moses*. As in our PER evaluation, *Sequitur* and *Phonetisaurus* outperform the other approaches in most cases. However, *Rules* results in lowest WERs for most scenarios with less than 1k training data.

ASR Systems with Dictionaries from Phoneme-Level Combination

Figure 4 shows the WER changes compared to the best single converter, using dictionaries generated from *PLC* without (*PLC-w/oWDP*) and with (*PLC-filtWDP*) the additional web-driven G2P converter outputs. *PLC-w/oWDP* is only in one case better than the best single method, whereas *PLC-filtWDP* outperforms the best single system in four cases. This shows that the web data can also have a positive impact on improve ASR performance.

Furthermore we learn that getting closer to the qualified pronunciations with *PLC* does not mean that the WER of the ASR systems improves. Figure 5 indicates that the correlation between the percentage of insertion and deletion errors (*I*+*D*) to the reference pronunciations at the phoneme-level correlates stronger with the WER than the PER to the reference pronunciations. We believe that ASR systems usually deal better with substitution errors in the pronunciations than insertion and deletion errors due to the acoustic model. Additionally, the fact that errors in the pronunciations of words that occur frequent in training and test set have a bigger impact on the word error rate than less frequent ones blurs the correlation between WER and PER.

Confusion Network Combinations

ASR system combination methods are known to lower the WER of ASR systems [31]. They require the training of systems that are reasonably close in performance but at the same time produce an output that differs in their errors. This provides complementary information which leads to performance improvements. As the individual ASR systems with the dictionaries from different G2P converter outputs are

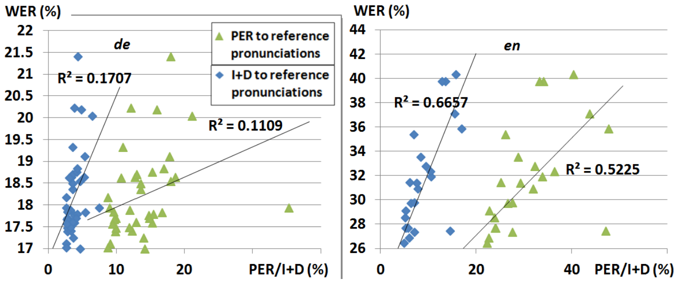

Fig. 5. Correlation betw. PER/I+D to qualified dictionary and WER.

close in performance, we combine them with a Confusion Network Combination (CNC) (*late fusion*) and compare it to the *PLC* performance.

Figure 4 illustrates that a *late fusion* with *CNC* outperforms the *early fusion* approach with *PLC*. Including the systems with pronunciation dictionaries that have been built with *PLC* to *CNC* (*CNC*+*PLC*), outperformed *CNC* in six systems. While for *de CNC* gave improvement for all amounts of G2P training material, it outperformed the best single system in only half of the cases for *en*. With 8.8% relative WER improvement, we report the largest improvement for *de* where only 200 German *W-P* pairs and web data are given as training data. We believe that the advantage of *CNC* is that language model information is available which lacks in the *PLC* approach.

5. CONCLUSION AND FUTURE WORK

We have analyzed the G2P converter output combination of four languages with differing grade in G2P regularity and simulated scenarios with small amounts of *W-P* pairs. First we showed that the different converters produce different pronunciations which are close in performance. We have evaluated the phoneme-level combination approach with the phoneme error rate to qualified pronunciations and conducted additionally ASR experiments for German and English.

The output of G2P converters built on web-derived word-pronunciation pairs could further improve pronunciation quality. Filtering the web data enhances the resulting pronunciation quality and the ASR performance. Our phoneme-level combination has potentials to enhance the processes of semi-automatic pronunciation dictionary creation by reducing the human editing effort.

The positive impact of the combination in terms of lower PERs compared to qualified pronunciations had only little influence on the WERs of our ASR systems - more for *de* than for *en*. Including the systems with pronunciation dictionaries that have been built with the phoneme-level combination to confusion network combinations led to improvement in six systems.

We plan to investigate our approaches for further under-resourced languages and enhance the combination at the phoneme-level.

6. REFERENCES

[1] T. Schlippe, S. Ochs, and T. Schultz, "Grapheme-to-Phoneme Model Generation for Indo-European Languages," in *The 37th International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2012)*, Kyoto, Japan, March 2012.

[2] R. M. Kaplan and M. Kay, "Regular Models of Phonological Rule Systems," in *Computational Linguistics*, 1994, vol. 20, pp. 331–378.

[3] A. W. Black, K. Lenzo, and V. Pagel, "Issues in Building General Letter to Sound Rules," in *3rd ESCA Workshop on Speech Synthesis*, Jenolan Caves, Australia, November 1998, International Speech Communication Association.

[4] R. Kneser, "Grapheme-to-Phoneme Study," Tech. Rep. WYT-P4091/00002, Philips Speech Processing, Germany, 2000.

[5] S. F. Chen, "Conditional and Joint Models for Grapheme-to-Phoneme Conversion," in *8th European Conference on Speech Communication and Technology (Eurospeech)*, Geneva, Switzerland, September 2003.

[6] P. Vozila, J. Adams, Y. Lobacheva, and T. Ryan, "Grapheme to Phoneme Conversion and Dictionary Verification using Graphonemes," in *8th European Conference on Speech Communication and Technology (Eurospeech)*, Geneva, Switzerland, September 2003.

[7] S. Jiampojamarn, G. Kondrak, and T. Sherif, "Applying Many-to-Many Alignments and Hidden Markov Models to Letter-to-Phoneme Conversion," in *HLT: Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*, Rochester, NY, April 2007.

[8] J. Novak, "Phonetisaurus: A WFST-driven Phoneticizer," 2011.

[9] J. Novak, N. Minematsu, and K. Hirose, "WFST-based Grapheme-to-Phoneme Conversion: Open Source Tools for Alignment, Model-Building and Decoding," in *International Workshop on Finite State Methods and Natural Language Processing*, Donostia-San Sebastián, Spain, July 2012.

[10] M. Gerosa and M. Federico, "Coping with Out-of-Vocabulary Words: Open versus Huge Vocabulary ASR, booktitle = International Conference on Acoustics, Speech, and Signal Processing (ICASSP), year = 2009, address = Taipei, Taiwan, month = April," .

[11] A. Laurent, P. Deléglise, and S. Meignier, "Grapheme to Phoneme Conversion Using an SMT System," in *10th Annual Conference of the International Speech Communication Association (Interspeech)*, Brighton, UK, September 2009.

[12] P. Karanasou and L. Lamel, "Comparing SMT Methods for Automatic Generation of Pronunciation Variants," in *7th International Conference on Advances in Natural Language Processing (IceTAL'10)*, Reykjavik, Iceland, 2010.

[13] S. Hahn, P. Vozila, and M. Bisani, "Comparison of Grapheme-to-Phoneme Methods on Large Pronunciation Dictionaries and LVCSR Tasks," in *The 13th Annual Conference of the International Speech Communication Association (Interspeech 2012)*, Portland, Oregon, September 2012.

[14] A. Ghoshal, M. Jansche, S. Khudanpurv, M. Riley, and M. Ulinski, "Web-derived Pronunciations," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Taipei, Taiwan, April 2009.

[15] D. Can, E. Cooper, A. Ghoshal, M. Jansche, S. Khudanpur, B. Ramabhadran, M. Riley, M. Saraclar, A. Sethy, M. Ulinski, and C. White, "Web Derived Pronunciations for Spoken Term Detection," in *32nd Annual International ACM SIGIR Conference*, 2009.

[16] T. Schlippe, S. Ochs, and T. Schultz, "Wiktionary as a Source for Automatic Pronunciation Extraction," in *The 11th Annual Conference of the International Speech Communication Association (Interspeech)*, Makuhari, Japan, 2010.

[17] T. Schlippe, S. Ochs, and T. Schultz, "Web-based tools and methods for rapid pronunciation dictionary creation," *Speech Communication*, vol. 56, no. 0, pp. 101 – 118, 2014.

[18] S. R. Maskey, A. W. Black, and L. M. Tomokiyo, "Bootstrapping Phonetic Lexicons for New Languages," in *International Conference of Spoken Language Processing (ICSLP)*, Jeju, Korea, 2004.

[19] M. Davel and O. Martirosian, "Pronunciation Dictionary Development in Resource-scarce Environments," in *10th Annual Conference of the International Speech Communication Association (Interspeech)*, Brighton, UK, September 2009, pp. 2851–2854.

[20] J. Kominek, *TTS From Zero: Building Synthetic Voices for New Languages*, Ph.D. thesis, Language Technologies Institute, School of Computer Science, Carnegie Mellon University, 2006.

[21] T. Schultz, A. W. Black, S. Badaskar, M. Hornyak, and J. Kominek, "SPICE: Web-based Tools for Rapid Language Adaptation in Speech Processing Systems," in

Annual Conference of the International Speech Communication Association (Interspeech), Antwerp, Belgium, August 2007.

[22] S. L. Davis, S. Fetters, B. Gustafson, L. Loney, and D. E. Schulz, "System and Method for Preparing a Pronunciation Dictionary for a Text-to-speech Voice," Tech. Rep. US Patent 7630898 B1, AT&T, September 2005.

[23] P. Koehn, H. Hoang, A. Birch an C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst, "Moses: Open Source Toolkit for Statistical Machine Translation.," in *Annual Meeting of ACL, demonstration session*, Prag, Czech Republic, June 2007.

[24] F. J. Och and H. Ney, "A Systematic Comparison of Various Statistical Alignment Models," *Computational Linguistics*, vol. 29, no. 1, pp. 19–51, 2003.

[25] M. Bisani and H. Ney, "Joint-Sequence Models for Grapheme-to-Phoneme Conversion," *Speech Communication*, vol. 50, no. 5, pp. 434 – 451, 2008.

[26] T. Schultz, N. T. Vu, and T. Schlippe, "GlobalPhone: A Multilingual Text Speech Database in 20 Languages," in *The 38th International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Vancouver, Canada, May 2013.

[27] John Garofalo, David Graff, Doug Paul, and David Pallett, "Continous Speech Recognition (CSR-I) Wall Street Journal (WSJ0) News, Complete," Tech. Rep., Linguistic Data Consortium, Philadelphia, 1993.

[28] A. Stolcke, "SRILM - An Extensible Language Modeling Toolkit," in *International Conference on Spoken Language Processing (ICSLP)*, Denver, Colorado, September 2002.

[29] T. Schlippe, S. Ochs, and T. Schultz, "Automatic Error Recovery for Pronunciation Dictionaries," in *The 13th Annual Conference of the International Speech Communication Association (Interspeech)*, Portland, Oregon, September 2012.

[30] O. Martirosian and M. Davel, "Error analysis of a public domain pronunciation dictionary," in *18th Annual Symposium of the Pattern Recognition Association of South Africa (PRASA)*, Pietermaritzburg, South Africa, November 2007, pp. 13–16.

[31] S. Stüker, C. Fügen, S. Burger, and M. Wölfel, "Cross-System Adaptation and Combination for Continuous Speech Recognition: The Influence of Phoneme Set and Acoustic Front-End," in *Ninth International Conference on Spoken Language Processing (Interspeech - ICSLP)*, Pittsburgh, PA, September 2006.