Evaluating grapheme-to-phoneme converters in automatic speech recognition context

Denis Jouvet, Dominique Fohr, Irina Illina

EVALUATING GRAPHEME-TO-PHONEME CONVERTERS IN AUTOMATIC SPEECH RECOGNITION CONTEXT

Denis Jouvet, Dominique Fohr, Irina Illina

Speech Group, INRIA - LORIA, 615 rue du Jardin Botanique, 54602 Villers les Nancy, France

ABSTRACT

This paper deals with the evaluation of grapheme-to-phoneme (G2P) converters in a speech recognition context. The precision and recall rates are investigated as potential measures of the quality of the multiple generated pronunciation variants. Very different results are obtained whether or not we take into account the frequency of occurrence of the words. Since G2P systems are rarely evaluated on a speech recognition performance basis, the originality of this paper consists in using a speech recognition system to evaluate the G2P pronunciation variants. The results show that the training process is quite robust to some errors in the pronunciation lexicon, whereas pronunciation lexicon errors are harmful in the decoding process. Noticeable speech recognition performance improvements are achieved by combining two different G2P converters, one based on conditional random fields and the other on joint multigram models, as well as by checking the pronunciation variants of the most frequent words.

Index Terms— Grapheme-to-phoneme, pronunciation lexicon, speech recognition

1. INTRODUCTION

Speech recognition systems rely on three main components: the set of acoustic models, the pronunciation lexicons and the language models. Acoustic models and language models are built automatically from large speech and text databases using data driven processes. However, the pronunciation lexicons typically result from human expertise in developing either the pronunciation lexicon itself or a more or less complex set of rules for grapheme-to-phoneme conversion. This required human expertise for creating good pronunciation lexicons make it difficult to develop or adapt a speech recognition system to a new language. Moreover, even when manually developed pronunciation lexicons exist, they are always of finite size, and, consequently, when moving to a new speech recognition task, the new words that appear are not always present in the available pronunciation lexicons. These are two examples where grapheme-to-phoneme converters are useful. However the quality of a grapheme-to-phoneme converter strongly impacts speech recognition performance.

Grapheme-to-phoneme (G2P) converters have been studied for a long time. The first ones were essentially rule-based. They were developed by taking into account linguistic expertise. This led to efficient systems that were able to handle some linguistic information for pronunciation disambiguation whenever necessary [1],[2]. Other G2P systems for deriving word pronunciations rely on data driven techniques and can be trained automatically from an initial list of pronunciation examples. In the Default&Refine approach [3], the training is supervised by a human verifier. On the other hand, some other data driven approaches are fully automatic. This includes the Joint-Multigram Model (JMM) [4],[5], the statistical machine translation approach [6], and also Conditional Random Field (CRF) models [7],[8]. In this paper we shall focus on the CRF-based and JMM-based G2P converters.

Many evaluations of G2P converters rely on evaluating only a single pronunciation of each word. However, speech recognition systems need multiple pronunciations whenever relevant. An evaluation criterion based on the recall and precision measures was initially proposed in [9] for estimating the quality of multiple pronunciation generation. Extensions of this measure are discussed in this paper. In the literature, G2P systems are rarely evaluated from a speech recognition performance point of view. Hence, the originality of this paper consists in using a speech recognition system to evaluate the G2P pronunciation variants.

Data-driven G2P methods need an initial pronunciation lexicon for training. Since getting directly good and large pronunciation lexicons is not always possible, some studies have investigated collecting pronunciation data from the web, as for example [10]. Some other studies have investigated the usage of Wiktionary data [11] for adding extra variants in a currently available lexicon [12]. In the paper, the impact of the G2P converters on speech recognition performance will be analyzed using G2P converters trained on the French BDLex [13] pronunciation lexicon.

The paper is organized as follows. Section 2 provides a brief presentation of the G2P converters that are used: JMM-based and CRF-based. Section 3 discusses their evaluation on a reference pronunciation database, namely BDLex. A particular focus is placed on evaluating the quality of multiple pronunciation variants. Section 4 investigates the impact of using G2P generated pronunciation lexicons in automatic speech recognition systems, for training the acoustic models, decoding speech signals, or both. Finally a conclusion ends the paper.

2. GRAPHEME-TO-PHONEME CONVERSION

Two G2P conversion systems are considered. One is based on Conditional Random Field (CRF) modeling, and the other one relies on Joint-Multigram Model (JMM).

2.1. Joint-multigram model approach

The Joint-Multigram Model approach is a state of the art approach for grapheme-to-phoneme conversion [5]. The JMM approach relies on using joint sequences, where each joint sequence is actually composed of a sequence of graphemes and its associated sequence of phonemes. A language model is applied on the joint sequences. The training algorithm aims at determining the optimal

set of joint sequences as well as the associated language model. The training proceeds in an incremental way. An initial pass creates a very simple model. Then each new training pass refines the model by enlarging the joint sequences whenever it is relevant to do so (i.e. it optimizes some training criteria).

In the following experiments, the Sequitur G2P software was used [14], and 8 training (refinement) passes were carried out on the training data. Then, the model that provides the smallest error rate on a given development set was chosen. In the reported experiments, it typically corresponds to the model obtained from the 6th training pass.

2.2. Conditional random field approach

The CRF-based approach for grapheme-to-phoneme conversion [7],[8] is more recent than the JMM-based approach. It relies on the probabilistic framework and discriminative training offered by CRFs for labeling structured data such as sequences.

However, training the CRFs requires a labeled database. That means that for grapheme-to-phoneme conversion, a preliminary alignment of the phonemes with the letters has to be carried out on the training data. In our approach [8], discrete HMMs are used to determine this letter-to-phoneme alignment.

The advantage of the CRF approach is its ability to handle various features, that is an arbitrary window length of letters, and possibly additional information such as word category. The CRF++ software [15] was used. It is a customizable and open source implementation of CRFs for segmenting and labeling sequential data. Following previous experiments, bigram features were used, and the letter context was set to 9, that is the current letter, plus 4 letters before and 4 letters after.

When used for predicting the phonemes of a given word, the CRF, as well as the JMM, can generate the n-best sequences of phonemes for that word, with associated probabilities. This characteristic will be used later to control the number of multiple pronunciations that are generated for each word.

3. EVALUATION ON PRONUNCIATION DATA

In this section, the grapheme-to-phoneme conversion systems are evaluated in a standard way, that is on a phonetic (pronunciation) reference database: the French BDLex pronunciation dictionary.

3.1. BDLex French pronunciation lexicon

BDLex is a French pronunciation dictionary that was developed at IRIT, Paul Sabatier University [13]. It contains lexical, phonological and morphological information. BDLex contains inflected forms, corresponding to about 49 000 canonical forms (lemma). Besides the spelling and pronunciation fields, each lexicon entry also contains the lemma. The phone set consists of 38 phonemes.

The BDLex lexicon was split into 3 sets according to lemmas: training set (75%), development set (5%) and test set (20%). All inflected forms of a lemma are kept together in the same set. This yields 263 473 entries in the training set, 17 814 entries in the development set, and 70 625 entries in the test set.

3.2. Single pronunciation variant

The G2P conversion systems trained on the BDLex training set were first evaluated on the BDLex test set when generating a single pronunciation variant per word. For single pronunciation variant generation, the word pronunciation (phoneme sequence) error rate is 4.3% for the JMM-based approach. The CRF-based approach leads to somewhat better results, 3.22% pronunciation error rate.

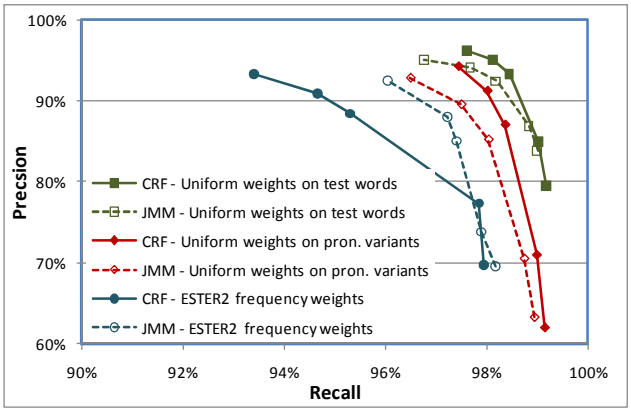

Fig. 1. Precision and recall rates of the two G2P converters evaluated on the BDLex test set using different weights (see text for details).

3.3. Multiple pronunciation variants

For automatic speech recognition systems, multiple pronunciation variants are necessary, at least for some words. Both G2P conversion systems previously described can produce an n-best list of pronunciation variants. Thus, for each word an initial list of up to 10 pronunciation variants was computed, such that the sum of the probabilities of the pronunciation variants exceeded 0.995. Then several lists of pronunciation variants were extracted by varying the minimum threshold on the probabilities of the pronunciation variants (i.e. only the pronunciation variants having a probability above the given threshold are kept). The average number of pronunciation variants generated on the test set, according to the minimum probability threshold are reported in the following table. It is interesting to notice that both G2P conversion systems provide a similar average number of pronunciation variants.

Table 1. Average number of pronunciation variants per word.

Probability threshold	0.200	0.100	0.050	0.010	0.005
With JMM-based G2P	1.07	1.12	1.19	1.45	1.62
With CRF-based G2P	1.07	1.11	1.17	1.44	1.66

The precision and recall measures should be good indicators of the quality of the pronunciation variants as they give the percentage of expected pronunciation variants that are actually predicted (recall) and the percentage of generated pronunciation variants that are correct (precision). This measure was proposed in [9] for comparing JMM-based and CRF-based G2P converters, as well as studying the impact of the training data size. This measure considers the set of all pronunciation variants, and gives the same weight to each pronunciation variant. This measure corresponds to the red curves (diamonds) in Fig. 1. The various points are obtained by varying the minimum probability threshold when generating the pronunciation variants, as explained before.

However, when using G2P generated pronunciation variants in a speech recognition system, it quickly becomes evident that all

G2P errors do not have the same impact. For example, an error on a frequently used word is more harmful to speech recognition performance than an error occurring on a rarely used word. This leads to a definition of alternate measures. The first modification consists in computing the precision and recall measures for each word individually (using the generated and expected pronunciation variants of each word), and then averaging these measures over the test set, giving the same weight to each word (green curves - squares - in Fig. 1).

The second modification consists in giving higher weight to frequent words when averaging the individual word precision and recall measures. This was achieved by using, for each word of the test set, a weight equal to the frequency of the corresponding word observed in the ESTER2 training corpus (described in the next section). The drawback of these frequency weights is that many of them are equal to zero, as they correspond to words that were not observed in the ESTER2 training corpus. Only 8790 words of the BDLex test set were observed in the ESTER2 training corpus. Nevertheless the corresponding blue curves (circles) in Fig. 1 are worse than the curves corresponding to the other measures, which means that more G2P errors are present in frequent words, than in rarely used words.

These evaluations show that different weightings of the pronunciation variants lead to very different results in terms of global precision and recall measures. Hence the importance of using weights that are consistent with the anticipated application that will be made of the G2P derived pronunciations.

4. EVALUATION IN AN ASR FRAMEWORK

Automatic Speech Recognition (ASR) experiments were conducted to evaluate the G2P converters. French broadcast news data from the ESTER2 evaluation campaign [16] were used. In this section the training was carried out on the ESTER2 training data (about 190 hours) and the recognition results are reported for a large subset of the ESTER2 development data, about 4h30 of audio signal corresponding to 36 800 running words. Finally the impact of checking the pronunciation variants of the most frequent words is discussed.

4.1. Experimental setting

All experiments were conducted using the Sphinx speech recognition toolkit [17]. For each experiment, the training of the acoustic models was performed from scratch. Hence, because of the full training required for each experiment, and in order to keep within reasonable processing time, the evaluations were conducted using only trigram language models and environment (studio quality vs. telephone quality) and speaker gender specific acoustic models. This corresponds typically to the first recognition pass of speech transcription systems, before applying further passes that use discriminative (LDA, MPE, ...) and adapted (MLLR, SAT, ...) acoustic models, and as well as larger language models, e.g. [18]. In our experiments, the pronunciation lexicon used for speech recognition contains about 64 000 entries.

For each pronunciation lexicon built for training, acoustic models were trained, from scratch, for the studio quality data (16 kHz) and telephone quality data (8 kHz) using the corresponding pronunciation lexicon. All acoustic models have 4500 senones (shared densities), and 64 Gaussian components per mixture. They are then adapted to the speaker gender.

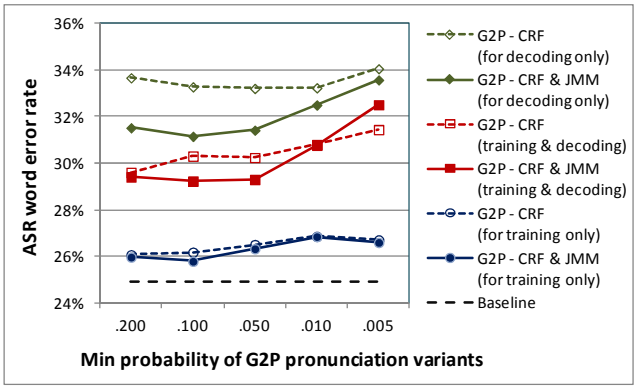

Fig. 2. ASR word error rates on the development set using G2P pronunciation variants (in training lexicons, or in decoding lexicons, or both); with more or fewer pronunciation variants per word (a smaller minimum probability threshold means more pronunciation variants).

For the baseline, the same training procedure was applied using an in-house lexicon derived from the BDLex pronunciation lexicon.

4.2. Using only G2P pronunciations in ASR lexicons

Evaluations were first carried out using pronunciation lexicons that were obtained entirely with the G2P converters. Results are reported for the CRF-based G2P pronunciation variants, and also when combining the pronunciation variants generated by both the JMM-based and the CRF-based G2P converters. Those G2P generated pronunciation lexicons are used in the decoding process only, in the training process only, or both. Speech recognition error rates are reported in Fig. 2, along with the baseline results.

The top curves exhibit the ASR word error rates achieved when G2P-based lexicons are used for decoding with the acoustic models trained using baseline pronunciation variants. Results show that setting the average number of pronunciation variants too high is harmful for the decoder. It is interesting to note that a significant reduction of the word error rate is achieved by merging the pronunciation variants generated by the two G2P converters (JMM-based and CRF-based). Nevertheless the word error rate is much higher than the one achieved by the baseline system.

The bottom curves are obtained using the G2P-based lexicons in the full training process, and the decoding is performed with the resulting acoustic HMMs and the baseline pronunciation lexicon. Results shows that using G2P pronunciations for training the acoustic HMMs does not have too great an impact on the quality of acoustic models.

Finally, the middle curves report the results using G2P pronunciation lexicons for both training and testing. The consistency between the training and decoding pronunciation lexicons helps recover some errors of the pronunciation lexicons.

These experiments show that the quality of the pronunciation lexicons is more crucial for decoding than for training.

4.3. Checking pronunciation of frequent words

Since the results reported in section 3.3. show a large impact due to pronunciation errors on frequent words, it seems interesting to investigate the benefit one can expect from checking and correcting the pronunciation variants of the most frequent words.

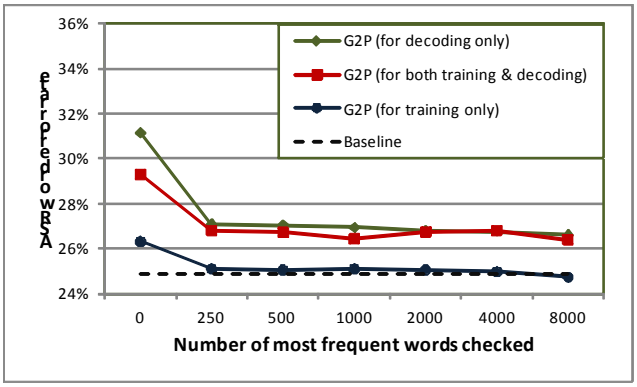

Fig. 3. Word error rates on the development set using G2P pronunciation variants (in training lexicons, decoding lexicons, or both). The horizontal axis indicates the number of most frequent words for which the pronunciations were checked.

This was achieved here simply by using directly, for the N most frequent words, the pronunciation variants extracted from the baseline lexicon. These partially checked pronunciation lexicons were then used for training, testing, or both. The results are reported in Fig. 3.

The results show that a large performance improvement is achieved by checking and correcting only the 250 most frequent words. Correcting other frequent words still provides small performance improvements. Further analysis is planned to determine the part of the errors due to proper names with respect to the part of the errors due to common names. It is important to note that the G2P converters were trained on a large subset of the BDLex pronunciation dictionary which does not contain proper names. Moreover the pronunciation of proper names does not always follow standard pronunciation rules, as many proper name variants are related to the foreign origin of the name.

6. CONCLUSION

This paper has investigated the evaluation of pronunciation variants generated by grapheme-to-phoneme (G2P) converters. Two systems were used, one based on conditional random fields, the other one on joint multigram models. These systems can produce more or fewer pronunciation variants per word, the number of which can be controlled through several parameters such as the total number of variants generated per word, and/or by setting a minimum probability threshold for the pronunciation variants to be kept.

In order to evaluate the quality of the multiple pronunciation variants generated by such G2P converters, the precision and recall rates evaluated on a test set seem to be a good indicator. Indeed, such a measure indicates the number of pronunciation variants that are correctly produced, and the number of incorrect variants generated. However, it appears that this measure is not directly related to the quality of the generated pronunciation variants when used in automatic speech recognition, as it does not take into account the frequency of the words for which incorrect pronunciation variants occur.

As the two G2P systems do not rely on the same principles and do not systematically make the same errors, their combination leads to improved speech recognition performance with respect to the usage of a single approach. It also appears that the quality of the G2P generated pronunciation variants is more critical when they are used in the decoding process, and that the training process is quite tolerant with respect to pronunciation variant errors.

Finally, experiments show that checking and correcting the pronunciation variants of the most frequently used words leads to noticeable performance improvements. The largest gain is achieved by checking and correcting the 250 most frequent words. Hence, the best usage of G2P generated pronunciations is when used for processing new words outside a generic lexicon.

11. REFERENCES

[1] K. Bartkova, D. Larreur & I. Metayer, "About choosing and adapting a grapheme-to-phoneme converter for automatic speech recognition", in French "Choix et adaptation d'un phonétiseur pour la reconnaissance automatique de la parole", in *Proc. JEP'1994*, Trégastel, France, pp. 181-186, 1994.

[2] F. Bechet, "LIA PHON: a complete system for grapheme-to-phoneme conversion of texts", in French "LIA PHON: un système complet de phonétisation de textes", *Traitement Automatique des Langues*, vol. 42, n° 1, pp. 47-67, 2001.

[3] M. Davel & E. Barnard, "Pronunciation prediction with Default&Refine", *Computer Speech and Language*, vol. 22, 2008.

[4] S. Deligne, F. Yvon & F. Bimbot, "Variable-length sequence matching for phonetic transcription using joint multigrams", in *Proc. EUROSPEECH'1995*, Madrid, Spain, pp. 2243-2246, 1995.

[5] M. Bisani & H. Ney, "Joint-sequence models for grapheme-to-phoneme conversion", *Speech Communications*, vol. 50, n° 5, 2008.

[6] A. Laurent, P. Deléglise & S. Meignier, "Grapheme to phoneme conversion using an SMT system", in *Proc. INTERSPEECH'2009*, Brighton, UK, Sept. 2009.

[7] D. Wang & S. King, "Letter-to-sound Pronunciation Prediction Using Conditional Random Fields", *IEEE Signal Processing Letters*, vol. 18, n° 2, pp. 122-125, 2011.

[8] I. Illina, D. Fohr & D. Jouvet, "Grapheme-to-Phoneme Conversion using Conditional Random Fields", *in Proc. INTERSPEECH'2011*, Florence, Italy, Aug. 2011.

[9] I. Illina, D. Fohr & D. Jouvet, "Multiple pronunciation generation using grapheme-to-phoneme conversion based on conditional random fields", *in Proc. SPECOM'2011*, Kazan, Russia, 2011.

[10] A.Ghoshal, M. Jansche, S. Khudanpur, M. Riley & M. Ulinski, "WEB-derived pronunciations", in *Proc. ICASSP'2009*, Taipei, Taiwan, April 2009.

[11] (2011) Wiktionary - a wiki based open content dictionary. [Online], Available http://www.wiktionary.org.

[12] T. Schlippe, S. Ochs & T. Schultz, "Wiktionary as a source for automatic pronunciation extraction", in *Proc. INTERSPEECH'2010*, Makuhari, Chiba, Japan, Sept. 2010.

[13] M. De Calmes & G. Pérennou, "BDLEX: a lexicon for spoken and written French", in *Proc LREC'1998*, Grenade, Spain, pp. 1129-1136, 1998.

[14] (2011) Sequitur G2P. [Online]. Available: http://www-i6.informatik.rwth-aachen.de/web/Software/g2p.html.

[15] (2011) CRF++: Yet Another CRF toolkit. [Online]. Available: http://crfpp.sourceforge.net/.

[16] S. Galliano, G. Gravier & L. Chaubard, "The ESTER 2 evaluation campaign for rich transcription of French broadcasts", in *Proc. INTERSPEECH'2009*, Brighton, UK, 2009.

[17] (2011) Sphinx. [Online] Available: http://cmusphinx.sourceforge.net.

[18] P. Deléglise, Y. Estève, S. Meignier & T. Merlin, "Improvements to the LIUM French ASR system based on CMU Sphinx: what helps to significantly reduce the word error rate?", in *Proc. INTERSPEECH'2009*, Brighton, UK, Sept. 2009.