Available online at www.sciencedirect.com
ScienceDirect
Computer Speech and Language xxx (2014) xxx–xxx
COMPUTER SPEECH AND LANGUAGE
www.elsevier.com/locate/csl

Improving recognition of proper nouns in ASR through generating and filtering phonetic transcriptions☆

Antoine Laurent[a,b,*], Sylvain Meignier[a], Paul Deléglise[a]

[a] LIUM, Computer Science Research Department – Université du Maine, Le Mans, France
[b] Spécinov, Trélazé, France

Received 4 December 2012; received in revised form 13 December 2013; accepted 27 February 2014

Abstract

Accurate phonetic transcription of proper nouns can be an important resource for commercial applications that embed speech technologies, such as audio indexing and vocal phone directory lookup. However, an accurate phonetic transcription is more difficult to obtain for proper nouns than for regular words. Indeed, phonetic transcription of a proper noun depends on both the origin of the speaker pronouncing it and the origin of the proper noun itself.

This work proposes a method that allows the extraction of phonetic transcriptions of proper nouns using actual utterances of those proper nouns, thus yielding transcriptions based on practical use instead of mere pronunciation rules.

The proposed method consists in a process that first extracts phonetic transcriptions, and then iteratively filters them. In order to initialize the process, an alignment dictionary is used to detect word boundaries. A rule-based grapheme-to-phoneme generator (LIA_PHON), a knowledge-based approach (JSM), and a Statistical Machine Translation based system were evaluated for this alignment. As a result, compared to our reference dictionary (BDLEX supplemented by LIA_PHON for missing words) on the ESTER 1 French broadcast news corpus, we were able to significantly decrease the Word Error Rate (WER) on segments of speech with proper nouns, without negatively affecting the WER on the rest of the corpus.

Keywords: Speech recognition; Phonetic transcription; Proper nouns; SMT; Moses; G2P

1. Introduction

Proper nouns constitute a special case when it comes to phonetic transcription, at least in French, which is the language used for this study. Indeed, there is much less predictability in how proper nouns may be pronounced than for regular words. This is partly due to the fact that, in French, pronunciation rules are much less normalized for proper nouns than for other categories of words: a given sequence of letters is not guaranteed to be pronounced the same way in two different proper nouns.

☆ This paper has been recommended for acceptance by Saraclar Murat.

* Corresponding author at: LIUM, Computer Science Research Department – Université du Maine, Le Mans, France. Tel.: +33 243833859.
E-mail addresses: antoine.laurent@lium.univ-lemans.fr, antoinelaurent@me.com (A. Laurent), sylvain.meignier@lium.univ-lemans.fr (S. Meignier), paul.deleglise@lium.univ-lemans.fr (P. Deléglise).
URL: http://www-lium.univ-lemans.fr (A. Laurent).

http://dx.doi.org/10.1016/j.csl.2014.02.006
0885-2308/

The lack of predictability also finds its roots in the wide array of origins proper nouns can come from: the more foreign the origin, the less predictable the pronunciation, with variations covering the whole range from correct pronunciation in the original language to a Frenchified interpretation of the spelling.

The high variability induced by this low predictability is a source of difficulty for Automatic Speech Recognition (ASR) systems when dealing with proper nouns. For an ASR system, being confronted with a proper noun pronounced using a phonetic variant very remote from any variant present in its dictionary is a situation similar to encountering an unknown word, if the language model cannot compensate for the acoustic gap. Such errors can have a strong impact on word error rate (WER): according to a comparative study of out-of-vocabulary impact of words in spontaneous and prepared speech (Dufour, 2008) the recognition error on an out-of-vocabulary word propagates through the language model to the surrounding words, causing a WER of about 50% within a window of 5 words to the left and to the right (again, in French). This highlights that the influence of the quality of the phonetic dictionary of proper nouns extends further than just the recognition of proper nouns themselves. It is particularly true in the case of applications where proper nouns are frequently encountered, such as transcription of broadcast news. However, aside from its potential impact on WER, accurate recognition of proper nouns can also be very important—independently of the frequency of their occurrence—in other contexts such as in the case of automatic indexing of multimedia documents, or transcription of meetings.

Setting up a phonetic dictionary of proper nouns (or any other class of words) requires grapheme to phoneme (G2P) conversion, be it manual or automatic. Automatic G2P conversion techniques are widely studied in the literature. Strik and Cucchiarni (1999) present an overview of techniques in 1999 and propose to classify the G2P systems into two categories: the knowledge-based approaches, which use existing linguistic knowledge to derive pronunciations, and the data-driven approaches, which derive pronunciation models from acoustic data. Knowledge-based approaches are further divided between formalized (*e.g.* rule based) and non-formalized (*e.g.* dictionary lookup). de Calmès and Pérennou (1998) propose a dictionary look-up strategy (non-formalized knowledge-based). Béchet (2001), Tihoni and Pérennou (1991), and Réveil et al. (2012) present rule-based knowledge-based techniques. Réveil et al. (2012) propose a rule-based strategy that integrates different type of features (orthographic, syllabic, morphological, . . .) to describe the rule context. A large variety of knowledge-based techniques are proposed in the literature: Torkkola (1993), Ma and Randolph (2001), Jensen and Riis (2000), and Seng et al. (2011) propose local classification strategies and Galescu and Allen (2001), Bellegarda (2005), and Bisani and Ney (2008) propose some pronunciation-by-analogy approaches. Many data-driven (acoustic-based) strategies can also be found in the literature (Holter and Svendsen, 1999; Byrne et al., 1998; Deligne and Mangu, 2003; Svendsen et al., 1995).

We propose an acoustic-based method to build a dictionary of phonetic transcriptions of proper nouns by using an iterative filter to retain the most relevant parts of a large set of phonetic variants, the latter being obtained by combining three G2P methods with extraction from actual audio signals (Laurent et al., 2009).

In this work for French, we compare three different G2P systems to initialize the process, and we use a two-level iteration to converge on the best filtered dictionary. In order to reduce noise, an iterative filter is applied to invalidate the variants that are deemed irrelevant because never used, and the ones that are found to be too prone to generating confusion with other words.

First, related works will be presented. After proposing an overview of the method, we will focus on the grapheme-to-phoneme systems used to initialize the process. In the next part, the proposed method will be described, and before concluding, experiments and results will be introduced and commented on. The intermediate (before filtering) and final sets of phonetic transcriptions are evaluated in terms of Word Error Rate (WER) and Proper Noun Error Rate (PNER), computed over the corpus of French broadcast news from the ESTER evaluation campaign (Galliano et al., 2005).

2. Related works

Many G2P systems are presented in the literature. Several names are attributed to this task: grapheme-to-phoneme conversion (Andersen et al., 1996; Bellegarda, 2005), phonetic pronunciation modeling (Riley et al., 1999), letter-to-sound translation (Pagel et al., 1998), letter-to-phoneme conversion (Rama et al., 2009; Seng et al., 2011), phonetic baseform generation (Bahl et al., 1991; Ramabhadran et al., 1998), phonetic transcription (Bisani and Ney, 2001), text-to-phoneme mapping (Suontausta and Häkkinen, 2000), among others.

The simplest strategy to get phonetic transcriptions of a word is the dictionary look-up, which consists in searching in a human-made phonetic dictionary. Making such a dictionary is costly and time-intensive. We have at our disposal the

Please cite this article in press as: Laurent, A., et al., Improving recognition of proper nouns in ASR through generating and filtering phonetic transcriptions. Comput. Speech Lang. (2014), http://dx.doi.org/10.1016/j.csl.2014.02.006

BDLEX dictionary (de Calmès and Pérennou, 1998). This dictionary has the advantage of providing a very complete and accurate set of transcriptions for each word it contains. However, it only contains a limited number of entries, and more importantly for our case, it does not contain any proper noun.

Rule-based conversion techniques have been developed in order to obtain phonetic transcriptions of new words automatically. A rule-based phonetic transcription system generates the possible chains of phones by relying exclusively on the spelling of words. It offers the advantage of providing phonetic variants even for words for which no speech signal is available. In the case of proper nouns, it generates the most "common-sense" variants, *i.e.* the ones which people would use when they have no *a priori* knowledge of the pronunciation of a particular proper noun. It would be prohibitively difficult to establish the complete set of rules needed to automatically find all the commonly used phonetic transcriptions of every proper noun.

In order to do so, an ideal automatic system would have to be able to detect not only the origin of the proper noun, but also the various ways people might pronounce this noun according to their own cultural and linguistic background. Unfortunately, both tasks are still open problems.

In the rest of this section, we will focus on a third approach: *data-driven* G2P conversion systems based on the use of acoustic data. A thorough description of the other methods can be found in Bellegarda (2005) and in Bisani and Ney (2008).

Haeb-Umbach et al. (1995), based on an approach developed by Bahl et al. (1993), propose a method that first extracts a phonetic transcription for each utterance of a word, and then selects, with a heuristic function, the one that maximizes the likelihood from the set of acoustic representations. This method is based on the assumption that one phonetic transcription only is enough to represent a word.

Svendsen et al. (1995) improved that approach by computing that heuristic from the best path of every acoustic representation. Wu and Gupta (1999) propose to simultaneously decode multiple utterances to derive one or more phonetic transcriptions for each word by using a word-network-based algorithm. Mokbel and Jouvet (1999) develop a method that consists in searching the k best phonetic transcriptions from a set of extracted pronunciations. Two decision criteria are tested. The first criterion is based on transcription occurrence frequency, and the second on the maximization of likelihood. The method that gives the best results is the one based on likelihood maximization. For each acoustic realization, the n-best list (with n set experimentally to 50) is constructed and constrained by the likelihood maximization of the union of those lists. Sloboda (1995) uses the first criterion: the selection of the k most frequently extracted phonetic transcriptions.

Bisani and Ney (2001) propose a beam search approach with a two-level (*intra*-arc, arc) pruning criteria. At least 10 samples are needed to get a reasonable (between 5 and 10%) Phoneme Error Rate (PER). The PER is the average edit distance between the found phonetic transcription and the reference phonetic transcription.

Deligne and Mangu (2003) and Deligne et al. (2001) develop a method based on an acoustic phonetic decoding for the addition of words to the personalized vocabulary of their users. To do this, users have to repeat one or two times every word they want to add to their lexicon. Rose and Lleida (1997) and Ramabhadran et al. (1998) describe an almost similar acoustic–phonetic decoding system, which requires the user to repeat the various words to phonetize. Every user has to pronounce twelve different proper nouns and has to call 10 times from different phones (cellular and landline) and in several different acoustic environments (hall, cafeteria, ...). The decoding strategy is based on the combination of speaker-independent acoustic models and a language model that represents the transition probabilities between various phonemes.

The work presented in Galescu and Allen (2001) is based on the use of a bi-directional n-gram joint sequence model. This model can be used to get a phonetic transcription of a word thanks to its spelling or by using an acoustic representation of it.

3. Overview of the proposed method

We propose a strategy that allows the extraction of phonetic transcriptions of proper nouns from utterances. It is a multi-step, iterative process. The first step consists in isolating portions of the signal corresponding to proper nouns, using the textual transcription of the audio and a forced alignment to get word boundaries. For this step, we need a dictionary that contains every word present in the textual transcription. During our study, we noticed that the dictionary used for this initialization step had a great influence on our results. The use of poor phonetic transcriptions results in boundary detection errors. In this study, we compare a rule-based phonetic transcription generator (LIA_PHON, Béchet,

Please cite this article in press as: Laurent, A., et al., Improving recognition of proper nouns in ASR through generating and filtering phonetic transcriptions. Comput. Speech Lang. (2014), http://dx.doi.org/10.1016/j.csl.2014.02.006

2001), a joint-sequence model based method (JSM, Bisani and Ney, 2008), and a Statistical Machine Translation based grapheme-to-phoneme converter (SMT, Laurent et al., 2009).

We consider a proper noun to be each word (token) composing the name of a person. For example, the name of a person composed of a first and a last name is considered as two proper nouns in the rest of this paper. Portions of the speech signal assumed to be corresponding to proper nouns are then extracted and fed to an APD (Acoustic Phonetic Decoding) system to obtain their phonetic transcription. Thus, proper nouns which are present several times in the corpus potentially get associated with several distinct phonetic transcriptions. APD yields a high number of phonetic transcriptions per proper noun (specific figures for our experimental corpus can be found in Section 8.1). However some of the extracted transcriptions may be flawed: often, some phonemes of neighboring words are added or deleted at the end or at the beginning of the phonetic transcription, and some wrong phonemes are inserted in noisy conditions. Also, the high number of transcriptions increases the risk of generating confusion with other words. Proper nouns could erroneously appear in the ASR output instead of words from other categories. Therefore, it can negatively impact the quality of the decoding for the rest of the corpus. In order to avoid these problems, the result of the extraction is filtered to discard unfit phonetic transcriptions and the other phonetic transcriptions are kept in the "filtered" dictionary.

The proposed method for filtering is iterative: the filtered dictionary of each iteration is used again to carry out the alignment step, and the process starts again. This process is repeated until two consecutive filtered dictionaries are exactly the same. At least one phonetic transcription of each proper noun is always kept in the proper noun dictionary (*i.e.* there is no out-of-vocabulary word in the ASR lexicon). The method was trained and evaluated using broadcast news in French composed of French, European and world news reports. These data contain a high number of foreign journalist names.

4. Initial dictionary generation

4.1. Rule-based generation of phonetic transcriptions

The rule-based generator we used is LIA_PHON (Béchet, 2001). LIA_PHON is available under the GPL license. It participated in the ARC B3 evaluation campaign of French automatic phonetizers, in which phonetic transcriptions generated by the systems were compared with phonetizations made by human experts. This campaign was held in 1997, and results were published by Yvon et al. (1998) in 1998. Error rate was calculated according to the same principle as for the classical word error rate used in speech recognition. Compared to human-made phonetic transcriptions, 99.3% of the transcriptions generated by LIA_PHON were correct (for a total of 86,938 phonemes) (This measure is computed at the phone level). However, results reveal that transcription errors were not distributed evenly among the various classes of words: erroneous transcription of proper nouns represented 25.6% of the errors generated by LIA_PHON even though proper nouns only represented 5.8% of the test corpus. This reflects poorer performance by LIA_PHON on this class of words.

4.2. Data-driven conversion techniques

In this section, we describe a G2P system based on the use of joint-sequence models (JSM) and a conversion technique based on the use of a Statistical Machine Translation (SMT) system. Both these systems need a bitext corpus for the training step.

4.2.1. Bitext corpus format for data-driven methods

Q3 To convert graphemes to phonemes, a bitext associates sequences of letters with sequences of phonemes. Fig. 1 shows examples of two representations of the bitext corpus, denoted by A and B. In representation A, the sequence of letters corresponds to a word. In representation B, the sequence of letters corresponds to a group of words. A symbol is added to mark the boundary of each word and each phonemic representation of the words. This representation allows to differentiate inter- and intra-word influence. In order to build a bitext corpus for representation B, every sequence of words of the training corpus between two fillers (silence, music, laughter, hesitation, ...) is aligned using the baseline acoustic models and the baseline dictionary. Our baseline dictionary contains variants that take into account the interword coarticulation influence (liaisons in French). Indeed, because it makes a gap in the speech flow, we

Please cite this article in press as: Laurent, A., et al., Improving recognition of proper nouns in ASR through generating and filtering phonetic transcriptions. Comput. Speech Lang. (2014), http://dx.doi.org/10.1016/j.csl.2014.02.006

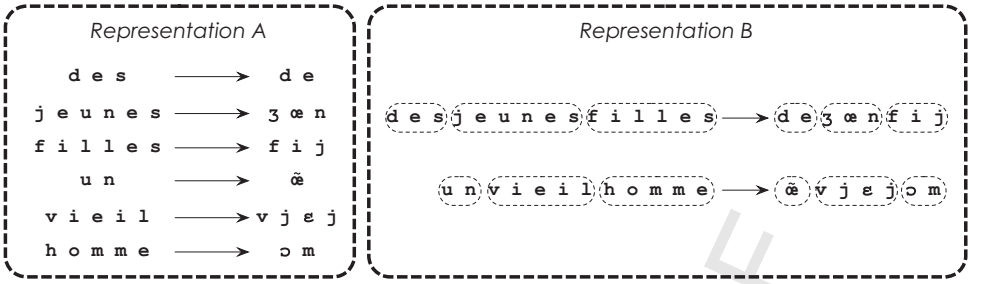

Fig. 1. Examples of representations A and B of the bitext corpus.

formed the hypothesis that the influence of a word on the pronunciation of its neighbors is negligible when they are separated by a filler.

4.2.2. Joint-sequence models (JSM)

This system is a *data-driven* conversion system, available under the GPL license. The system is based on the idea that, given enough examples, it should be possible to predict the pronunciation of unseen words, purely by analogy. The use of joint-sequence models to convert graphemes to phonemes (Bisani and Ney, 2008) will be denoted by JSM in the rest of this article. JSM being a data-driven conversion system means that we have to give it pronunciation examples in order to train it. Training takes a pronunciation dictionary and creates new model files successively, starting with unigram models and up to 6-gram models. Model files can then be used to transcribe words that were not in the dictionary. The fundamental idea of joint multigram model is that for each word, its orthographic form and its pronunciation are generated by a common sequence of *graphones*. A *graphone*, or grapheme-phoneme joint multigram is a pair $q=(g, \varphi)$ of a letter sequence g and a phoneme sequence φ of possibly different length. For example, the pronunciation of "jeunes" may be regarded as a sequence of three graphones:

"jeunes"		j	eu	nes
ʒ œ n	=	ʒ	œ	n

The procedure for having the alignment between graphemes and phonemes is described in Bisani and Ney (2002). The joint probability distribution $p(\varphi, g)$ is modeled using a standard M-gram:

$$p(q_1^L) = \prod_{i=1}^{L+1} p(q_i | q_{i-1}, \ldots, q_{i-M+1}) \quad (1)$$

Phonetic transcriptions are then obtained from words by searching the most likely graph sequence matching the given spelling and projecting it onto the phonemes.

Because computing time on representation B is very expensive using JSM, it is trained only on representation A. Using this representation, JSM is not able to learn intra-word pronunciation influence.

4.2.3. Grapheme to phoneme conversion using Statistical Machine Translation (SMT)

We proposed a method in Laurent et al. (2009), based on the open source Moses toolkit (Koehn et al., 2007) to convert graphemes to phoneme sequences.

A Statistical Machine Translation system (SMT) is used to transform text from a source language into a target language. The training step needs a data corpus which is composed of bitext data: source language sentences associated with their translation in the target language.

SMT is commonly used to translate data in which the elementary unit is the word in both the source and target parts

The training of a grapheme-to-phoneme translation model is similar to the training of a translation model as described in the Moses documentation.

Please cite this article in press as: Laurent, A., et al., Improving recognition of proper nouns in ASR through generating and filtering phonetic transcriptions. Comput. Speech Lang. (2014), http://dx.doi.org/10.1016/j.csl.2014.02.006

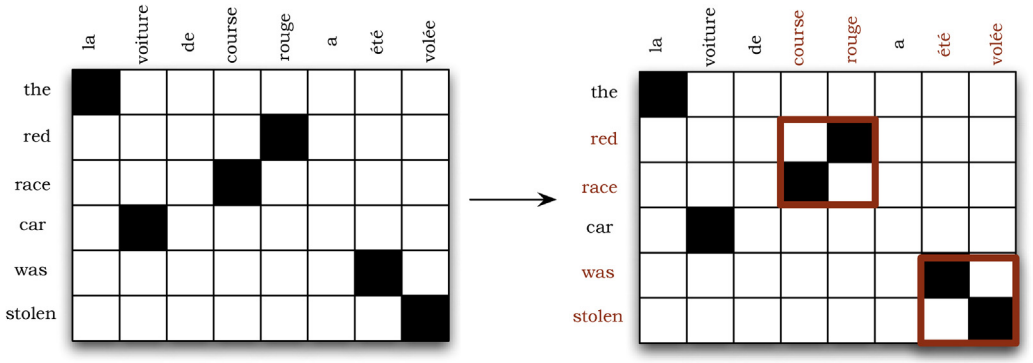

Fig. 2. Bitext alignment matrix: the extracted phrase pairs ("red race", "course rouge") and ("was stolen", "été volée") are in discontinuous order, the extracted phrase pairs ("red", "rouge") and ("race", "course") are in swap order, extracted phrase pairs ("was", "été") and ("stolen", "volée") are in monotone order.

4.2.3.1. SMT models. First, the bitext corpus has to be aligned at word level in both directions (source to target and target to source). The phrase pairs are extracted using some heuristics known as *diag-grow-final* which start from the intersection of the two alignments and then adds additional alignment points. After extraction, the phrase pairs are scored. Assuming that we want to convert a sequence of graphemes $\tilde{g}$ to a sequence of phonemes $\tilde{p}$, a standard translation model contains 5 different scores, namely direct and inverse phrase translation probabilities $\varphi(\tilde{p}|\tilde{g})$ and $\varphi(\tilde{g}|\tilde{p})$, direct and inverse lexical weightings $lex(\tilde{p}|\tilde{g})$ and $lex(\tilde{g}|\tilde{p})$ and a phrase insertion penalty (always set to e). Standard SMT systems allow reordering between phrases (sequence of words), which is controlled by the distortion limit. Setting this limit to zero implies monotone decoding, which seems to be convenient for G2P task. This reordering is handled by the introduction of a distortion model and a lexical reordering model. A distortion model penalizes phrases proportionally to the amount of reordering. As presented in Fig. 2, the lexical reordering model takes into account three different features corresponding to three kinds of reordering, namely monotone (phrase pairs are adjacent and in the same order), swap (phrase pairs are adjacent and in the reverse order), and discontinuous (the phrase pairs are not adjacent). For each phrase pair, the relative frequency of each kind of reordering is calculated (a smoothing technique is applied to avoid zero probabilities for unseen orientations). The last main component of a SMT system is the language model which is trained on the target side of the bitexts and all available monolingual data in target language.

Fig. 3 shows how the SMT is trained and used for the translation of graphemes to phonemes. We trained a 4-gram language model composed of phonemes learned from a phonemic forced alignment of the ESTER 1 training corpus. The bitext corpus is used to produce a translation model. However two training strategies are proposed: the first one corresponds to the standard Moses training framework based on the maximization of BLEU (Papineni et al., 2002). The second, based on the Levenshtein metric, minimizes insertion, deletion, and substitution errors of phonemes.

4.2.3.2. BLEU score. The BLEU score is commonly used for the optimization in order to have the best translation system according to this measure. Training reserves 3% of the corpus for optimization of the parameters according to the BLEU score. Experiments show that the best score (in terms of WER computed over our development corpus) is obtained by allowing reordering for representation A, while for representation B, the best score is obtained by setting the distortion limit to zero. Weights of the different models where trained using the Minimum Error Rate Training (Och, 2003).

4.2.3.3. Levenshtein score. For our task, we decided to try using a normalized Levenshtein edit distance for parameter optimization.

At the end of a training iteration, 3-best phonetic transcriptions for each training example (sequence of letters) are generated using the current translation model. When using the Levenshtein score optimization, we only optimized the five weights of the translation model (the distortion limit was set to zero, disabling distortion and lexical reordering models).

Please cite this article in press as: Laurent, A., et al., Improving recognition of proper nouns in ASR through generating and filtering phonetic transcriptions. Comput. Speech Lang. (2014), http://dx.doi.org/10.1016/j.csl.2014.02.006

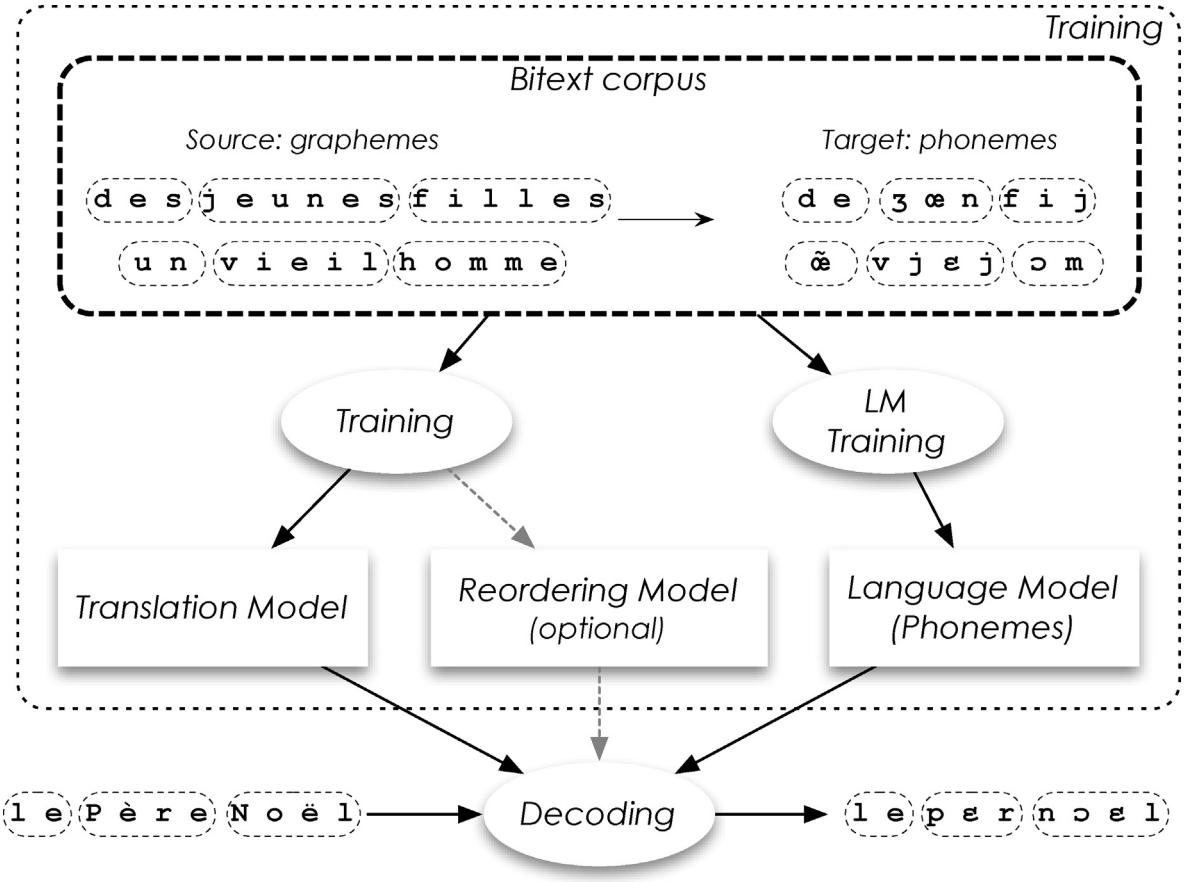

Fig. 3. SMT training and use for grapheme-to-phoneme translation.

The sum of the normalized Levenshtein measures, S, is computed between the phonetic transcriptions and the references (Eq. (2)).

$$S = \sum_{e \in E} \log\left(1 - \min\left(\forall n \in [1,3] \frac{d(p_e^n, r_e)}{max(l_{p_e^n}, l_{r_e})}\right)\right) \quad (2)$$

where p_e^n is the phonetic transcription n of the example e. As stated before, we consider the 3-best phonetic transcriptions, thus n vary from 1 to 3. $d(p_e^n, r_e)$ is the edit distance of Levenshtein of the phonetic transcription p_e^n, with r_e the reference phonetic transcription for example e. E is the set of the generated phonetic transcriptions. $l_{p_e^n}$ is the length of the phonetic transcription p_e^n of the example e and l_{r_e} is the length of the reference phonetic transcription (r_e). Every log argument is floored at 10^{-7} to avoid that just one bad phonetic transcription could impact the measure of the entire database.

Until getting the highest S over all the training examples, a simplex framework[1] is used to tune the model parameters.

When using the Levenshtein optimization, the language model weight is set to 0.1 and the word penalty weight is set to 0. The word penalty weight was determined experimentally, and, in order to optimize each of the 6 remaining weights, we fixed the language model weight to 0.1 and optimized the 5 others according to it.

For the task of grapheme-to-phoneme conversion, the best results were obtained by using the Levenshtein optimization and representation B. Learning time on our training corpus (ESTER 1 Training corpus, see Section 7.1 for details) is about 13 times more for JSM (175.5 h) than for SMT (13.5 h).

[1] Thanks to the Condor toolkit (Vanden Berghen et al., 2005).

Please cite this article in press as: Laurent, A., et al., Improving recognition of proper nouns in ASR through generating and filtering phonetic transcriptions. Comput. Speech Lang. (2014), http://dx.doi.org/10.1016/j.csl.2014.02.006

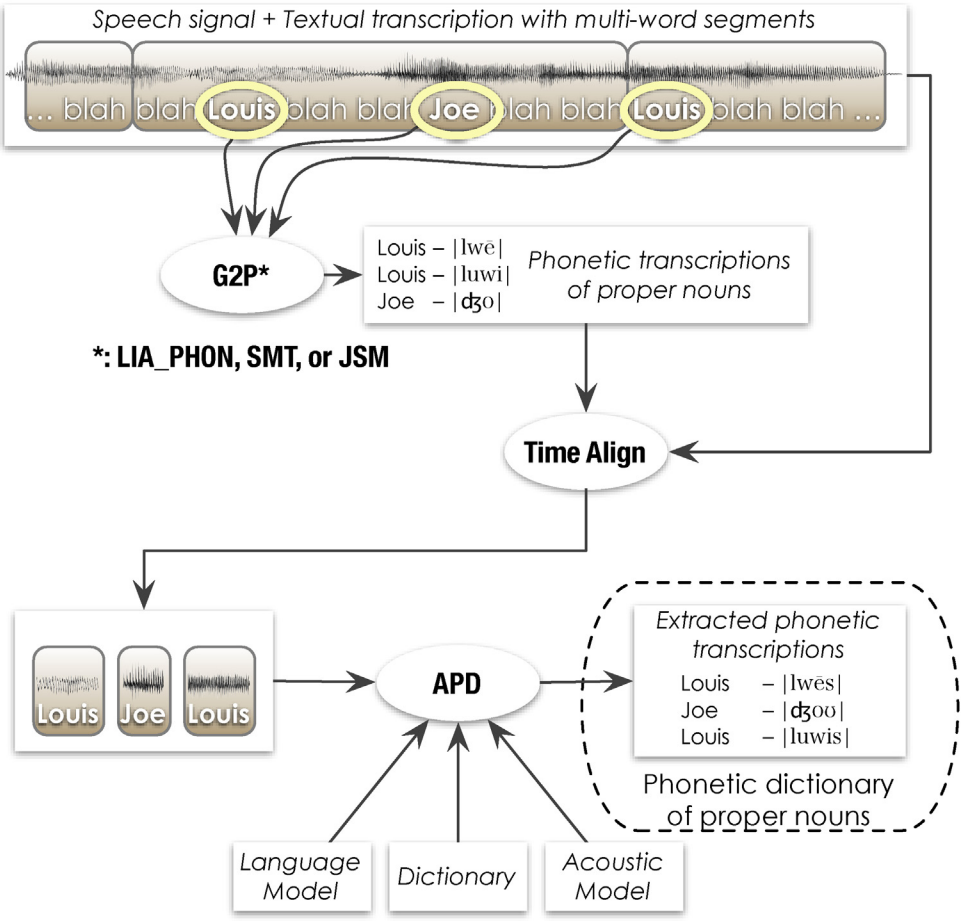

Fig. 4. Use of the acoustic–phonetic decoding system.

5. Extraction of phonetic transcriptions using acoustic–phonetic decoding

5.1. Method

In order to enrich the set of phonetic transcriptions of proper nouns with some less predictable variants, we gather actual utterances of proper nouns by actual people. This process relies on an acoustic–phonetic decoding system (APD), which generates a phonetic transcription of the speech signal.

In a corpus consisting of speech with a manual word transcription, portions of the speech signal corresponding to proper nouns are extracted. They are then fed to the APD system to obtain their phonetic transcription. Since the phonetic decoding results of various utterances can be different, proper nouns which are present several times in the corpus potentially get associated with several different phonetic transcriptions each.

5.2. Proper noun boundaries

As explained above, the first step consists in isolating the portions of signal corresponding to proper nouns using word transcription. However, in the manual transcription we used, words were not aligned with the signal. Therefore, the start and end times of each word of the transcription had to be determined by aligning the words with the signal, using a speech recognition system (see Fig. 4—"Time Align" step) with the various G2P systems presented previously (see Fig. 4—"G2P*" step).

Please cite this article in press as: Laurent, A., et al., Improving recognition of proper nouns in ASR through generating and filtering phonetic transcriptions. Comput. Speech Lang. (2014), http://dx.doi.org/10.1016/j.csl.2014.02.006

In Fig. 4, we have two different boxes that contain phonetic transcriptions. The first one represents phonemes that we get directly by using one of our three different G2P systems (LIA_PHON, SMT, or JSM). The second box represents the phonetic transcriptions that we get from the signal, at the output of our Acoustic Phonetic Decoding system.

5.3. APD based phonetic transcription

When boundaries of the proper nouns have been determined, APD is applied to the corresponding portions of the signal. The decoding path gives a series of phonemes considered as the phonetic transcription of the proper noun.

As noted in Bisani and Ney (2001), unconstrained phonetic decoding does not allow the system to obtain reliable phonetic transcriptions. Our own experiments lead us to the same conclusion. The use of a language model allows for some level of guidance for the speech recognition system: it does so by minimizing the risk of having phoneme sequences with a very low probability appear in the transcription results. We set constraints by using tied state triphones and a 3-gram phone language model as part of the decoding strategy, to generate the best sequence of phonemes. While this setup is close to a speech recognition system, the dictionary and phone language model contain phonemes instead of full words. The trigram phone language model was trained using the phonetic dictionary used during the 2005 ESTER evaluation campaign (Deléglise et al., 2005). It contains about 65,000 lexical entries of words, and was generated using BDLEX and LIA_PHON. Only the words which were not part of the BDLEX corpus were phonetised automatically using LIA_PHON. Words which were identified as proper nouns were deleted from this dictionary before learning our 3-gram phone language model.

6. Filtering of phonetic transcriptions

6.1. Motivation

As previously mentioned, we call each token composing the name of a person a “proper noun”. The extraction of phonetic transcriptions for utterances yields an average of 6 phonetic transcriptions per proper noun (token) in our experiments (complete results for our experimental corpus can be found in Section 8.1).

However, as stated in the previous section, some of the extracted transcriptions may be flawed. Also, the high number of transcriptions increases the risk of some phonetic transcriptions of proper nouns being erroneously used to decode words of another type. Therefore, it can negatively impact the quality of the decoding for the rest of the corpus. Given that the number of occurrences of the other categories of words is expected to be much higher than the number of occurrences of proper nouns, there is a risk of seeing any gain in performance for proper nouns being outbalanced by a negative impact on the rest of the corpus and on the global WER. The goal of this filtering is to detect and remove the phonetic variants of proper nouns that are the most likely to generate confusion with other words.

6.2. Iterative filtering

In order to minimize the risk of negatively affecting the global WER, it is desirable to filter the set of phonetic transcriptions and keep only the most appropriate. We propose an iterative filtering method to select only those transcriptions deemed to be reliable enough. We have already proposed a different approach to select phonetic transcriptions in previous work (Laurent et al., 2008); however this early attempt was rendered impractical because of its execution time which was directly proportional to the number of extracted phonetic transcriptions. For a proper noun present in s segments, with v phonetic transcriptions, it was necessary to decode $v \times s$ segments to validate or invalidate the overall set of phonetic variants for this proper noun.

In the present work, we have managed to detect and remove phonetic variants of proper nouns generating confusion with other words by decoding our training corpus using the newly built phonetic dictionary (as well as a separate phonetic dictionary for all the other categories of words, of course). Any phonetic transcription that was never used to decode the corresponding proper noun in the right place gets removed from the dictionary, since it either caused an error or was not used at all. However, a heuristic is set in order to keep at least one phonetic transcription for each proper noun. The process then gets repeated: the corpus is decoded again using the modified dictionary, which then gets filtered according to the results of this decoding. The whole decoding/filtering process is repeated until no more

Please cite this article in press as: Laurent, A., et al., Improving recognition of proper nouns in ASR through generating and filtering phonetic transcriptions. Comput. Speech Lang. (2014), http://dx.doi.org/10.1016/j.csl.2014.02.006

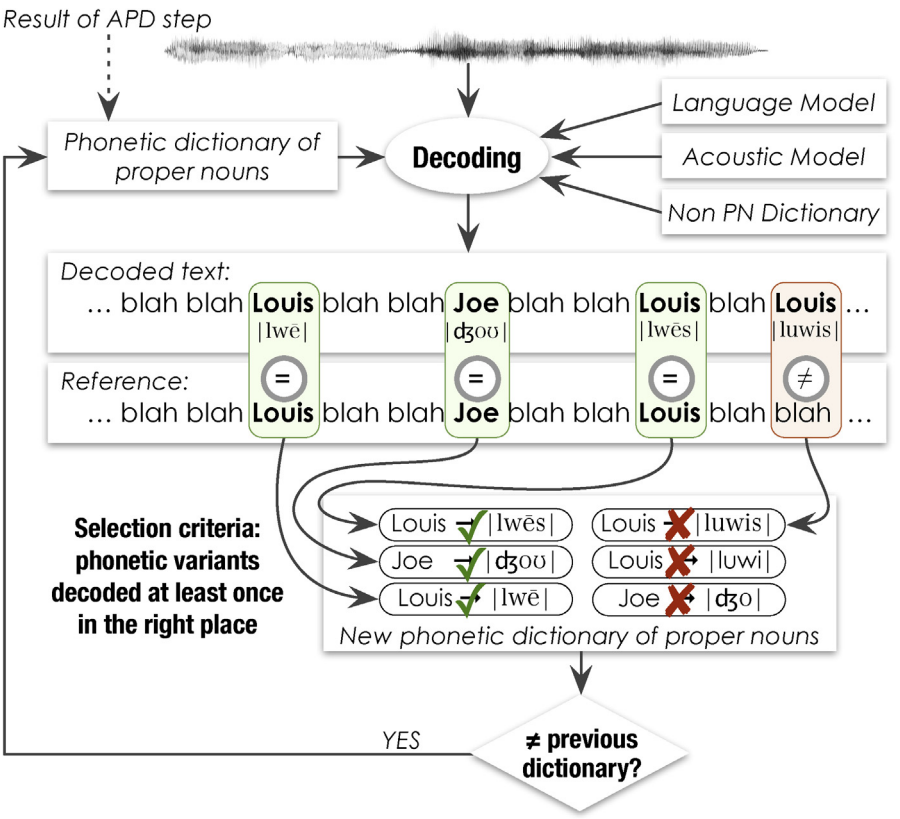

Fig. 5. Illustration of iterative filtering of phonetic transcriptions. The initial value of the phonetic dictionary of proper nouns is the union of rule-based and extracted transcriptions.

phonetic transcriptions are removed from the dictionary. This process is illustrated in Fig. 5, using the same example data as in Fig. 4.

6.3. Two-level iterative filtering

As stated earlier, the alignment dictionary used to initialize the process has a strong impact on the accuracy of the phonetic transcriptions generated. For this reason, we have decided to rerun the whole process, this time using the iteratively filtered dictionary (the output of the iterative filtering described above) instead of G2P systems to get boundaries of proper nouns inside the audio data during the forced alignment step. This allows the system to call proper noun boundaries into question with the newly built dictionary. Convergence to a dictionary, independently of the one used for the initialization, is always done after three complete iterations. Fig. 8 in Section 8 shows the coverage between the final dictionaries.

This extraction+filtering cycle, illustrated in Fig. 6, is repeated until two consecutive iteratively filtered dictionaries are exactly identical.

7. Experiments

7.1. Corpus

Our experiments were carried out on the ESTER 1 corpus. ESTER is an evaluation campaign of French broadcast news transcription systems which took place in January 2005 (Galliano et al., 2005). The ESTER corpus was divided into three parts: training, development, and evaluation. The training (81 h) and the development (12.5 h) corpora are composed of data recorded from four radio stations in French (France Inter, France Info, Radio France Internationale,

Please cite this article in press as: Laurent, A., et al., Improving recognition of proper nouns in ASR through generating and filtering phonetic transcriptions. Comput. Speech Lang. (2014), http://dx.doi.org/10.1016/j.csl.2014.02.006

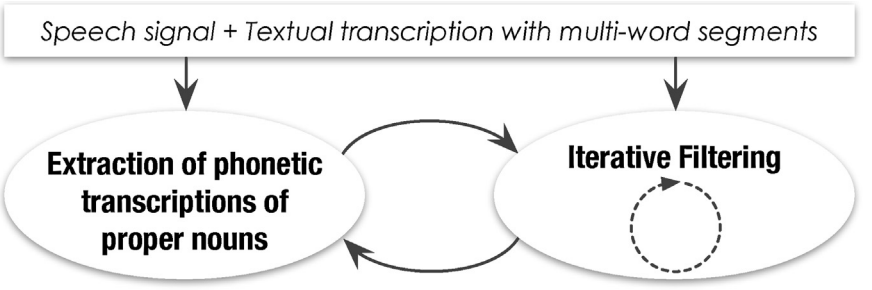

Fig. 6. Overview of the double iterative process: the filtered dictionary is used for the initialization of the next extraction+filtering cycle, until the result is stable.

and Radio TV Maroc). The test corpus is composed of 10 h coming from the same four radio stations plus two other stations (France Culture and Radio Classique), all of which were recorded 15 month after the development data. Each corpus is annotated with named entities, allowing easy spotting of proper nouns.

The training corpus was used to learn our automatic speech recognition system. The training corpus and the development corpus are jointly employed to extract phonetic transcriptions and to filter them. The JSM and SMT grapheme-to-phoneme converters were also trained over the ESTER 1 training corpus.

7.2. Metrics

The intermediate and final sets of phonetic transcriptions were evaluated in terms of Word Error Rate (WER) and Proper Noun Error Rate (PNER). PNER is computed the same way as the WER, but it is computed only for proper nouns and not for every word:

$$\mathrm{PNER} = \frac{I + S + E}{N} \tag{3}$$

with I being the number of wrong insertions of proper nouns, S the number of substitutions of proper nouns with other words (where the reference word is a proper noun), E the number of elisions of proper nouns, and N the total number of proper nouns.

The use of PNER as a metric reflects the goal of this work, which is to enhance the recognition of proper nouns, and not merely have an accurate chain of phonemes.

While PNER allows to evaluate the quality of the detection of proper nouns, WER is used to evaluate the impact of the new phonetic transcriptions on the whole test corpus.

7.3. Acoustic and language models

The decoding system is based on CMU Sphinx 3.6 (Ravishankar et al., 2000).

Our experiments were carried out using a one-pass decoding coming from the LIUM ESTER 1 system (Deléglise et al., 2005), using 12 MFCC acoustic features plus the energy, completed with their primary and secondary derivatives. Acoustic models were trained on the ESTER training corpus. These models are composed of 5500 tied states, each state being modeled by a mixture of 22 diagonal Gaussians. Decoding employs tied-state word-position dependent triphone acoustic models which are made gender- and bandwidth-dependent through MAP adaptation of means, covariances and weights. The trigram language model was learned using three different data sources:

- The manual transcriptions of ESTER 1, available for the training corpus (81 h of data) and the development corpus (12.5 h) recorded from the four radio stations presented in Section 7.1. These transcriptions contain about 1.35 M occurrences of 34k distinct words.
- The articles coming from the French newspaper “Le Monde” from the year 2003 (19M occurrences of 220k distinct words).
- The articles coming from the French newspaper “Le Monde” from 1987 to 2002 (300M word occurrences).

Table 1
Number of phonetic transcriptions generated by each method.

Method	Generated variants (G2P)	Extracted variants (APD)	After 1 iteration (all process)	After 2 iterations (all process)	After 3 iterations (all process)
LIA_PHON	4364	20,218	6776	6524	6502
SMT	7031	20,184	7065	6813	6802
JSM	3626	20,008	6876	6711	6708
Average	5007	20,137	6906	6683	6671

Three 3-gram language models were learned: one using the 81 h of the ESTER 1 training corpus, and the others on the two other data sources ("Le Monde" 2003 and "Le Monde" from 1987 to 2002). A linear interpolation was performed to minimize perplexity on the remaining 12.5 h of data coming from the development corpus. The vocabulary contains all of the 34k distinct word of the manual transcriptions, and words appearing more than ten times in the 2003 articles (about 19k words). The most frequent words in the rest of the articles from "Le Monde" (from 1987 to 2002) are used to complete the vocabulary, up to 65k words.

To estimate each of the three language models, the SRILM toolkit (Stolcke, 2002) is employed using the modified Kneser-Ney discounting method. Unigrams and bigrams are all kept, but trigrams occurring only once are discarded.

The language model includes all the proper nouns present in the training corpus. All the dictionaries contain the same proper nouns, with only their phonetic transcriptions varying.

8. Results

8.1. Number of phonetic transcriptions per proper noun

Table 1 presents the number of phonetic transcriptions generated with the three G2P methods. The ESTER 1 training corpus contains 3348 distinct proper nouns, appearing 28,866 times.

On average, the number of phonetic transcriptions between G2P generation and APD extraction grows from 5k to 20k (column "generated variants" compared to column "extracted variants"). We only consider the best hypothesis generated by the APD.

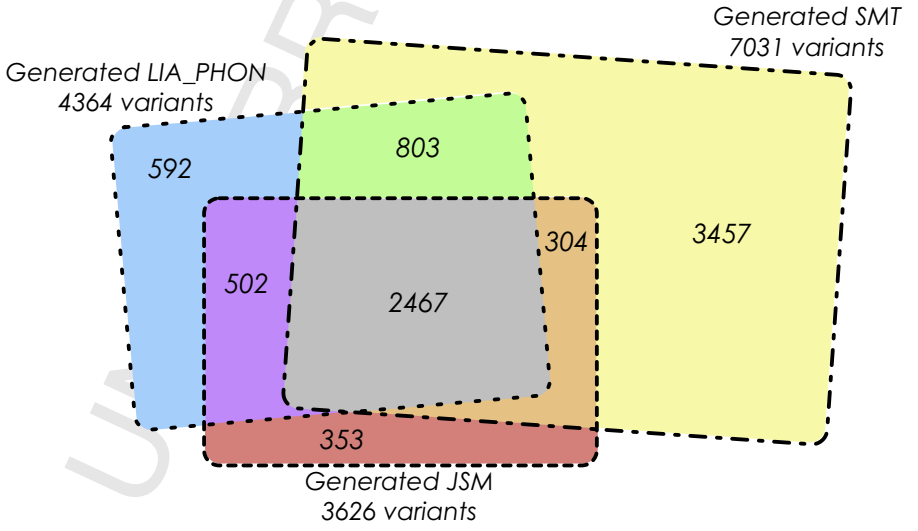

Fig. 7. Overlap among the 3 initialization dictionaries.

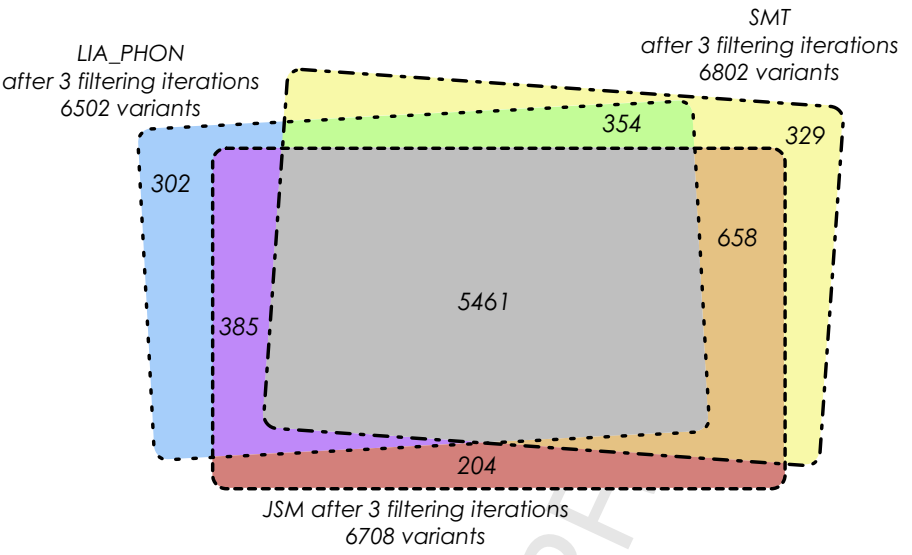

Fig. 8. Overlap among the final dictionaries.

One pass of iterative filtering keeps about 7k phonetic transcription variants from the 20k variants generated by the APD. For each of our three grapheme-to-phoneme strategies, filtering is done in 3 iterations. The number of variants contained in the final filtered dictionary slightly decreases compared to the first iteration. The number of variants in the final dictionaries is stable across methods used for the initialization. This is due to the fact that almost the same number of variants are extracted with the APD regardless of the initial G2P used. Fig. 7 shows the overlap among the generated dictionaries. As we can see, there are 2467 variants common to the 3 generated dictionaries. Fig. 8 shows the overlap among the final dictionaries. As we can see, there is more overlap: 5461 phonetic transcriptions are common to the 3 final dictionaries.

An analysis of the best filtered dictionary (generated with SMT and after 3 filtering iterations) shows the following composition: 65.6% of its phonetic variants were initially present in the generated SMT dictionary (4460 variants, 2043 from the 3457 pure SMT phonetic transcriptions, 2307 from the 2467 common variants, 103 from the 803 SMT ∪ LIA_PHON variants, and 7 from the 304 SMT ∪ JSM variants); 30.9% (2100) are new variants; and 3.6% (242) are phonetic transcriptions that were present in either the LIA_PHON or the JSM initialization dictionaries.

8.2. *Results of the first iteration*

This section compares the results obtained by directly using the three G2P methods with the use of the extraction and filtering of proper nouns.

Fig. 9 shows the PNER obtained using the filtering method after the first iteration for each G2P system on the ESTER test corpus.

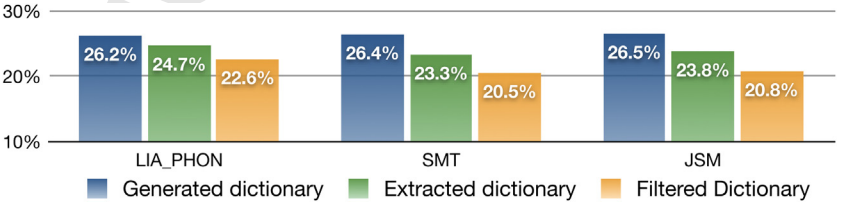

Fig. 9. PNER using each G2P method (ESTER test corpus).

Please cite this article in press as: Laurent, A., et al., Improving recognition of proper nouns in ASR through generating and filtering phonetic transcriptions. Comput. Speech Lang. (2014), http://dx.doi.org/10.1016/j.csl.2014.02.006

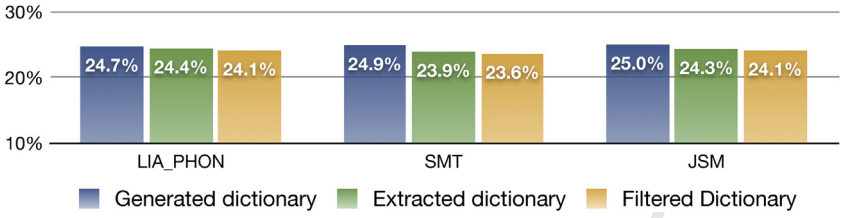

Fig. 10. WER on segments with proper nouns in the test corpus.

These results show that the filtering method produces significant gains in terms of PNER for every G2P system. As we can see, the APD method supplemented by the SMT-based grapheme-to-phoneme conversion system is the one that yields the lowest PNER.

As explained previously, phonetic transcriptions for non-proper nouns are taken from the BDLEX database, or generated by the rule-based grapheme-to-phoneme tool LIA_PHON for words which are not in the database. The generated dictionaries (SMT, JSM, and LIA_PHON) include the non-proper noun dictionary, supplemented by the phonetic transcriptions of all proper nouns generated using SMT, JSM, or LIA_PHON. Fig. 10 compares the results obtained using the three generated dictionaries (SMT, JSM, and LIA_PHON) to initialize the method, in term of WER computed only over segments that contain proper nouns.

Figs. 9 and 10 show the interest of filtering: it reduces both the PNER and the WER on segments with proper nouns.

8.3. Using iterative acoustic-based phonetic transcription

Table 2 shows the results obtained with the full iterative process initialized with LIA_PHON, SMT, and JSM G2P systems. The results in bold are those with the best gain in terms of WER and PNER. The results inside the parentheses are the difference, in terms of WER and PNER, with the baseline, LIA_PHON generated, dictionary. WER and PNER are computed on segments that contain proper nouns. We can see a small gap between the first filtering iteration and the last one. Using LIA_PHON to initialize our method, the WER decreased from 24.1% (after the first filtering iteration) to 24.0% (at the end) and the PNER decreased from 22.6% to 22.5%. With SMT, there is a gain of 0.2 point in terms of WER and a gain of 0.3 point in terms of PNER between the first and the last filtering iterations. Finally, when using JSM, the gains are of 0.2 point in terms of WER and 0.3 point in terms of PNER.

Fig. 11 shows the WER obtained on the whole ESTER 1 test corpus. The test corpus contains 11,087 segments. 1412 of them contain proper nouns. With no filtering, extracted dictionaries, while improving the WER on segments

Table 2
WER and PNER using the full iterative process.

Dictionary	WER (segments with PN)	PNER
LIA_PHON	24.7%	26.2%
SMT generated	24.9% (+0.2)	26.4% (+0.2)
JSM generated	25.0% (+0.3)	26.5% (+0.3)
Two-level filtering iteration 1		
LIA_PHON	24.1% (−0.6)	22.6% (−3.6)
SMT	23.6% (−1.1)	20.5% (−5.7)
JSM	24.1% (−0.6)	20.8% (−5.4)
Two-level filtering iteration 2		
LIA_PHON	24.1% (−0.6)	22.6% (−3.6)
SMT	23.5% (−1.2)	20.3% (−5.9)
JSM	24.0% (−0.7)	20.5% (−5.7)
Two-level filtering iteration 3		
LIA_PHON	24.0% (−0.7)	22.5% (−3.7)
SMT	23.4% (**−1.3**)	20.2% (**−6**)
JSM	23.9% (−0.8)	20.5% (−5.7)

Please cite this article in press as: Laurent, A., et al., Improving recognition of proper nouns in ASR through generating and filtering phonetic transcriptions. Comput. Speech Lang. (2014), http://dx.doi.org/10.1016/j.csl.2014.02.006

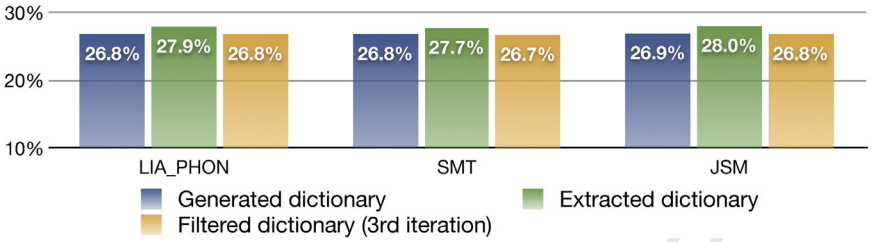

Fig. 11. WER on every segment in the test corpus.

Table 3
Some examples of phonetic transcriptions.

Proper nouns	SMT	Proposed method (initialized with SMT)
Jintao	ʒintao	jintao
Roger	ʁ oʒe	ʁ odʒ œʁ
Decaens	dœkɛn	dœkɑ̃
Fatima	fatima	fatma
Rumsfeld	ʁ ymsfɛld	ʁ œmsfɛld
Yahia	jaja	jaʁ ja
Ahmed	amɛd	aʁ kmɛd

that contain proper nouns, also increase the global WER. Errors are introduced: other words are substituted by proper nouns, and some proper nouns are wrongly inserted. The results show that with the filtering step, our method does not generate new errors with other word classes. The WER on segments with no proper nouns remains the same using filtered dictionaries as it is with the generated dictionaries. With the SMT generated dictionary as well as with our baseline dictionary, we get a WER of 27.1% on segments without proper nouns. With the SMT extracted dictionary, the WER is 28.2% on the same segments, and with the final dictionary (SMT after 3 filtering iterations), we get the same WER as with the generated dictionary (27.1%). This highlights the role of filtering, which removes confusable variants from the lexicon.

8.4. Analysis of the results

In our evaluation corpus, 640 different proper nouns are present, with a total of 2080 occurrences. The proposed method decreases the PNER for 152 proper nouns, and increases the PNER for 26 of them. Most of those 152 proper nouns are foreign, therefore they do not follow the usual rules of pronunciation used in French. Examples of those nouns are: Jiantao, Fatima, Rumsfeld, Yahia, Ahmed.

When pronounced in Arabic (Radio TV Maroc station), certain proper nouns contain phonemes that are not present in our French phoneme set. When decoding portions of signal corresponding to those nouns, those phonemes are replaced with close French phonemes as shown in Table 3.

As explained earlier, during filtering, a rule was set in order to avoid eliminating the last phonetic transcription variant of each noun. The average number of phonetic transcriptions per proper noun is about 2. It is only 1.3 for the 26 proper nouns for which the PNER is increased. This actually corresponds to 20 proper nouns with only one variant, which would have been eliminated without this heuristic. If every utterance in the training data is noisy, the extracted phonetic transcriptions may be wrong and eliminated by the iterative filtering. Maybe we should consider replacing, at the end of the process, the remaining variant that should have been eliminated by the generated phonetic transcriptions of this proper noun (with one of the presented G2P system).

9. Conclusion

In this article, we proposed an iterative, two-step acoustic-based process for phonetic transcription generation, and applied it to the specific case of proper nouns.

Please cite this article in press as: Laurent, A., et al., Improving recognition of proper nouns in ASR through generating and filtering phonetic transcriptions. Comput. Speech Lang. (2014), http://dx.doi.org/10.1016/j.csl.2014.02.006

The first step adopts a data-driven approach of building a dictionary of phonetic transcriptions, aiming for a closer match to actual usage of proper nouns than knowledge-based approaches can provide. This is accomplished through extraction of phonetic variants from actual audio signals, which is used to filter and enrich an initial set of phonetic transcriptions generated by a knowledge-base grapheme-to-phoneme system—filtering out unused variants and adding variants that the G2P system could not generate.

The second step of our method consists in filtering the resulting dictionary in order to avoid a negative impact on the other classes of words. Indeed, the extraction of phonetic transcriptions for proper nouns in the first step yields a high number of phonetic variants, which generates noise during the decoding. Many of these phonetic variants are too close to the pronunciation of other words of the dictionary. As a result, when used directly, this dictionary has a negative impact on the WER on segments that do not contain any proper noun. The goal of the iterative filtering process is the detection and removal of the phonetic variants that are the most likely to generate confusion with words from other classes.

The method loops, rerunning steps one and two over the resulting dictionary, iterating until stability is reached.

The use of the resulting phonetic dictionaries of proper nouns yields a significant gain in terms of PNER (Proper Noun Error Rate) and WER on the ESTER corpus. The best results are obtained by using an SMT (Statistical Machine Translation, Laurent et al., 2009) system to generate the initial proper noun dictionary for the process. The WER on segments that contain proper nouns decreased by 1.3 points and the PNER decreased by 6.0 points compared to the simpler, rule-based system. As was expected, with the filtering step, the WER on segments without proper nouns is unaffected, thus allowing the global WER to improve slightly thanks to better detection of proper nouns.

Even though the impact on the global WER is only minor on a corpus such as ESTER, improved detection of proper nouns is crucial for some tasks. An interesting field where the proposed method is useful is named speaker identification, which consists in the automatic extraction of speaker identities (first name and last name) from the transcription (El-Khoury et al., 2012; Canseco-Rodriguez et al., 2005). The new phonetic transcriptions generated by the proposed method should contribute to render detection easier by improving the decoding of proper nouns. A good transcription of proper nouns is important for this task: as mentioned by Poignant et al. (2013), if we consider that we are able to find who is speaking when their name is pronounced inside the current, previous or next speech turn, we can name more than 50% of the speakers inside TV shows of the REPERE (Giraudel et al., 2012) corpus. In the case of information retrieval, it could be interesting to be able to find where a proper noun is pronounced inside automatic speech transcription outputs, and low PNER is important for this task.

Finally, one of the advantages of the filtering method described here is that its execution time is not linked to the size of the set of transcriptions to be filtered. This opens up the possibility of applying it to other, larger classes of words. For example, it could be applied to frequently poorly decoded words in order to try to automatically find a better set of phonetic transcriptions based on existing utterances.

Uncited reference

Jousse et al. (2009).

Acknowledgement

Special thanks to Dr. Teva Merlin for his help with this work.

References

Andersen, O., Kuhn, R., Lazaridès, A., Dalsgaard, P., Haas, J., Nöth, E., 1996. Comparison of two tree-structured approaches for grapheme-to-phoneme conversion. In: Proceedings of the International Conference on Spoken Language Processing (ICSLP 96), vol. 3, Philadelphia, PA, USA, pp. 1700–1703.

Béchet, F., 2001. LIA_PHON: un système complet de phonétisation de textes. Traitement Automatique des Langues 42, 47–67.

Bahl, L., Das, S., deSouza, P., Epstein, M., Mercer, R., Merialdo, B., Nahamoo, D., Picheny, M., Powell, J., 1991. Automatic phonetic baseform determination. In: Proceedings of the International Conference on Acoustics, Speech and Signal Processing (IEEE, ICASSP 91), vol. 1, Toronto, Canada, pp. 173–176.

Please cite this article in press as: Laurent, A., et al., Improving recognition of proper nouns in ASR through generating and filtering phonetic transcriptions. Comput. Speech Lang. (2014), http://dx.doi.org/10.1016/j.csl.2014.02.006

Bahl, L., Brown, P., de Souza, P., Mercer, R., Picheny, M., 1993. A method for the construction of acoustic Markov models for words. IEEE Trans. Speech Audio Process. 1, 443–452.
Bellegarda, J.R., 2005. Unsupervised, language-independent grapheme-to-phoneme conversion by latent analogy. Speech Commun. 46, 140–152.
Bisani, M., Ney, H., 2001. Breadth-first for finding the optimal phonetic transcription from multiple utterances. In: Proceedings of the International Conference on Speech Communication and Technology (ISCA, Interspeech 2001), vol. 2, Aalborg, Denmark, pp. 1429–1432.
Bisani, M., Ney, H., 2002. Investigations on joint-multigram models for grapheme-to-phoneme conversion. In: Proceedings of International Conference on Spoken Language Processing (ISCA, ICSLP 2002), vol. 1, Denver, CO, USA, pp. 105–108.
Bisani, M., Ney, H., 2008. Joint-sequence models for grapheme-to-phoneme conversion. Speech Commun. 50, 434–451.
Byrne, W., Finke, M., Khudanpur, S., McDonough, J., Nock, H., Riley, M., Saraclar, M., Wooters, C., Zavaliagkos, G., 1998. Pronunciation modeling using a hand-labelled corpus for conversational speech recognition. In: Proceedings of International Conference on Acoustics Speech and Signal Processing (IEEE, ICASSP 98), vol. 1, Seattle, WA, USA, pp. 313–316.
Canseco-Rodriguez, L., Lamel, L., Gauvain, J.-L., 2005. A comparative study using manual and automatic transcriptions for diarization. In: Proceedings of the Workshop on Automatic Speech Recognition and Understanding (IEEE, ASRU 2005), vol. 1, Puerto Rico, USA, pp. 415–419.
de Calmès, M., Pérennou, G., 1998. BDLEX: a lexicon for spoken and written French. In: Language Evaluation and Resources Conference (LREC 1998), Grenada, Spain, pp. 1129–1136.
Deléglise, P., Estève, Y., Meignier, S., Merlin, T., 2005. The LIUM speech transcription system: a CMU Sphinx III-based system for French broadcast news. In: Proceedings of International Conference on Speech Communication and Technology (ISCA, Interspeech 2005), Lisbon, Portugal, pp. 1653–1656.
Deligne, S., Mangu, L., 2003. On the use of lattices for the automatic generation of pronunciations. In: Proceedings of the International Conference on Acoustics, Speech and Signal Processing (IEEE, ICASSP 2003), vol. 1, Hong Kong, China, pp. 204–207.
Deligne, S., Maison, B., Gopinath, R., 2001. Automatic generation and selection of multiple pronunciations for dynamic vocabularies. In: Proceedings of the International Conference on Acoustics, Speech and Signal Processing (IEEE, ICASSP 2001), vol. 1, Salt Lake City, UT, USA, pp. 565–568.
Dufour, R., 2008. From prepared speech to spontaneous speech recognition system: a comparative study applied to French language. In: IEEE/ACM CSTST Student Workshop, vol. 1, Cergy, France, pp. 595–599.
El-Khoury, E., Laurent, A., Meignier, S., Petitrenaud, S., 2012. Combining transcription-based and acoustic-based speaker identifications for broadcast news. In: Proceedings of the International Conference on Acoustics, Speech and Signal Processing (IEEE, ICASSP 2012), Kyoto, Japan, pp. 4377–4380.
Galescu, L., Allen, J.F., 2001. Bi-directional conversion between graphemes and phonemes using a joint n-gram model. In: Proceedings of 4th ISCA Tutorial and Research Workshop on Speech Synthesis, Perthshire, Scotland.
Galliano, S., Geoffrois, É., Mostefa, D., Choukri, K., Bonastre, J.-F., Gravier, G., 2005. The ESTER phase II evaluation campaign for the rich transcription of French broadcast news. In: Proceedings of International Conference on Speech Communication and Technology (ISCA, Interspeech 2005), vol. 1, Lisbon, Portugal, pp. 1149–1152.
Giraudel, A., Carré, M., Mapelli, V., Kahn, J., Galibert, O., Quitard, L., 2012. The REPERE corpus: a multimodal corpus for person recognition. In: The 8th International Conference on Language Resources and Evaluation (LREC 2012), Istanbul, Turkey, pp. 1102–1107.
Haeb-Umbach, R., Beyerlein, P., Thelen, E., 1995. Automatic transcription of unknown words in a speech recognition system. In: Proceedings of the International Conference on Acoustics, Speech and Signal Processing (IEEE, ICASSP 95), vol. 1, Detroit, MI, USA, pp. 840–843.
Holter, T., Svendsen, T., 1999. Maximum likelihood modelling of pronunciation variation. Speech Commun. 29, 171–191.
Jensen, K., Riis, S., 2000. Self-organizing letter code-book for text-to-phoneme neural network model. In: Proceedings of the International Conference on Spoken Language Processing (ISCA, ICSLP 2000), vol. 3, Beijing, China, pp. 318–321.
Jousse, V., Petitrenaud, S., Meignier, S., Estève, Y., Jacquin, C., 2009. Automatic named identification of speakers using diarization and ASR systems. In: Proceedings of International Conference on Acoustics Speech and Signal Processing (IEEE, ICASSP 2009), Taipei, Taiwan, pp. 4557–4560.
Koehn, P., Hoang, H., Birch, A., Calisson-Burch, C., Federico, M., Bertholdi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., Herbst, E., 2007. Moses: open-source toolkit for statistical machine translation. In: Proceedings of the Association for Computational Linguistics.
Laurent, A., Merlin, T., Meignier, S., Estève, Y., Deléglise, P., 2008. Combined systems for automatic phonetic transcription of proper nouns. In: Language Evaluation and Resources Conference (LREC 2008), Marrakech, Morocco.
Laurent, A., Merlin, T., Meignier, S., Estève, Y., Deléglise, P., 2009. Iterative filtrering of phonetic transcriptions of proper nouns. In: Proceedings of the International Conference on Acoustics, Speech and Signal Processing (IEEE, ICASSP 2009), vol. 1, Taipei, Taiwan, pp. 4265–4268.
Laurent, A., Deléglise, P., Meignier, S., 2009. Grapheme-to-phoneme conversion using an SMT system. In: Proceedings of the 10th Annual Conference of the International Speech Communication Association (ISCA, Interspeech 2009), Brighton, England, pp. 708–711.
Ma, C., Randolph, M.A., 2001. An approach to automatic phonetic baseform generation based on Bayesian networks. In: Proceedings of the International Conference on Speech Communication and Technology (ISCA, Interspeech 2001), vol. 1, Aalborg, Denmark, pp. 1453–1456.
Mokbel, H., Jouvet, D., 1999. Derivation of the optimal set of phonetic transcriptions for a word from its acoustic realizations. Speech Commun. 29, 49–64.
Och, F.J., 2003. Minimum error rate in statistical machine translation. In: Proceedings of the Association for Computational Linguistics, pp. 160–167.
Pagel, V., Lenzo, K., Black, A.W., 1998. Letter to sound rules for accented lexicon compression. In: Proceedings of the International Conference on Spoken Language Processing (ICSLP 98), Sydney, Australia, pp. 2015–2018.
Papineni, K., Roukos, S., Ward, T., Zhu, W.-J., 2002. BLEU: a method for automatic evaluation of machina translation. In: Proceedings of the Association for Computational Linguistics.

Please cite this article in press as: Laurent, A., et al., Improving recognition of proper nouns in ASR through generating and filtering phonetic transcriptions. Comput. Speech Lang. (2014), http://dx.doi.org/10.1016/j.csl.2014.02.006

Poignant, J., Besacier, L., Le, V.-B., Rosset, S., Quénot, G., 2013. Unsupervised naming of speakers in broadcast TV: using written names, pronounced names or both? In: Proceedings of International Conference on Speech Communication and Technology (ISCA, Interspeech 2013), Lyon, France, pp. 1462–1466.

Réveil, B., Martens, J.-P., van den Heuvel, H., 2012. Improving proper name recognition by means of automatically learned pronunciation variants. Speech Commun. 54, 321–340.

Rama, T., Singh, A.K., Kolachina, S., 2009. Modeling letter-to-phoneme conversion as a phrase based statistical machina translation problem with minimum error rate training. In: Proceedings of North American Chapter of the Association for Computational Linguistics – Human Language Technologies (NAACL HLT) 2009 conference, vol. 1, Boulder, CO, USA, pp. 90–95.

Ramabhadran, B., Bahl, L., deSouza, P., Padmanabhan, M., 1998. Acoustics-only based automatic phonetic baseform generation. In: Proceedings of the International Conference on Acoustics, Speech and Signal Processing (IEEE, ICASSP 98), vol. 1, Seattle, WA, USA, pp. 309–312.

Ravishankar, M., Singh, R., Raj, B., Stern, R.M., 2000. The 1999 CMU 10x real time broadcast news transcription system. In: Proceedings of the DARPA Workshop on Automatic Transcription of Broadcast News, Washington, DC, USA.

Riley, M., Byrne, W., Finke, M., Khudanpur, S., Ljolje, A., McDonough, J., Nock, H., Saraclar, M., Wooters, C., Zavaliagkos, G., 1999. Stochastic pronunciation modelling from hand-labelled phonetic corpora. Speech Commun. 29, 209–224.

Rose, R.C., Lleida, E., 1997. Speech recognition using automatically derived baseforms. In: Proceedings of the International Conference on Acoustics, Speech and Signal Processing (IEEE, ICASSP 97), vol. 2, Munich, Germany, pp. 1271–1274.

Seng, K., Iribe, Y., Nitta, T., 2011. Letter-to-phoneme conversion based on two-stage neural network focusing on letter and phoneme contexts. In: Proceedings of the 12th Annual Conference of the International Speech Communication Association (ISCA, Interspeech 2011), Florence, Italy, pp. 1885–1888.

Sloboda, T., 1995. Dictionary learning: performance through consistency. In: Proceedings of the International Conference on Acoustics, Speech and Signal Processing (IEEE, ICASSP 95), vol. 1, Detroit, MI, USA, pp. 453–456.

Stolcke, A., 2002. SRILM-an extensible language modeling toolkit. In: Proceedings of International Conference on Spoken Language Processing (ISCA, ICSLP 2002), vol. 2, Denver, CO, USA, pp. 901–904.

Strik, H., Cucchiarni, C., 1999. Modeling pronunciation variation for ASR: a survey of the literature. Speech Commun. 29, 224–246.

Suontausta, J., Häkkinen, J., 2000. Decision tree based text-to-phoneme mapping for speech recognition. In: Proceedings of the International Conference on Spoken Language Processing (ISCA, ICSLP 2000), vol. 2, Beijing, China, pp. 831–834.

Svendsen, T., Soong, F., Purnhagen, H., 1995. Optimizing baseforms for HMM-based speech recognition. In: Proceedings of the European Conference on Speech Communication and Technology (ESCA, Eurospeech 95), Madrid, Spain, pp. 783–786.

Tihoni, J., Pérennou, G., 1991. Phonotypical transcription through the GEPH expert system. In: Proceedings of the European Conference on Speech Communication and Technology (ESCA, Eurospeech 1991), vol. 1, Genoa, Italy, pp. 767–770.

Torkkola, K., 1993. An efficient way to learn English grapheme-to-phoneme rules automatically. In: Proceedings of the International Conference on Acoustics, Speech and Signal Processing (IEEE, ICASSP 93), vol. 2, Minneapolis, MN, USA, pp. 199–202.

Vanden Berghen, F., Bersini, H., 2005. CONDOR, a new parallel, constrained extension of Powell's UOBYQA algorithm: experimental results and comparison with the DFO algorithm. J. Comput. Appl. Math. 181, 157–175.

Wu, J., Gupta, V., 1999. Application of simultaneous decoding algorithms to automatic transcription of known and unknown words. In: Proceedings of the International Conference on Acoustics, Speech and Signal Processing (IEEE, ICASSP 99), vol. 2, Phoenix, AZ, USA, pp. 589–592.

Yvon, F., Boula De Mareuil, P., D'Alessandro, C., Aubergé, V., Bagin, M., Bailly, G., Béchet, F., Foukia, S., Goldman, J.-P., Keller, E., O'Shaughnessy, D., Pagel, V., Sannier, F., Véronis, J., Zellner, B., 1998. Objective evaluation of grapheme to phoneme conversion for text-to-speech synthesis in French. Comput. Speech Lang. 12 (4), 393–410.

Antoine Laurent obtained his Ph.D. degree in 2010 from Université du Maine, Le Mans, France, in the field of automatic speech recognition. He is currently R&D project manager at Spécinov (Angers, France), as well as part-time associate professor in the Language and Speech technology team at LIUM (the computer science research department of Université du Maine, Le Mans, France). His research focuses on automatic adaptation of the ASR system.

Sylvain Meignier received his Ph.D. degree in computer science from Université d'Avignon et des Pays de Vaucluse, Avignon, France, in 2002. His work was about speaker recognition. In 2003, he was with LIMSI-CNRS, Orsay, France, in the Spoken Language Processing Group as a Researcher. Since 2004, he has been an associate professor at Université du Maine, where he works on speech processing in the Language and Speech Technology team of the LIUM laboratory.

Please cite this article in press as: Laurent, A., et al., Improving recognition of proper nouns in ASR through generating and filtering phonetic transcriptions. Comput. Speech Lang. (2014), http://dx.doi.org/10.1016/j.csl.2014.02.006

Paul Deléglise received his Ph.D. in computer science from Pierre & Marie Curie University (Paris, France) in 1983 and his Doctorat d'État in 1991. He worked in the Signal Laboratory of École Nationale Supérieure des Télécommunications (ENST) on automatic speech recognition from 1985 to 1992. Since October 1992, he is full professor at Université du Maine where he works in the LIUM laboratory on data fusion applied to audio-visual speech recognition, and leads the Language and Speech Technology team. Since 2004, he has been working on large vocabulary speech recognition.

Please cite this article in press as: Laurent, A., et al., Improving recognition of proper nouns in ASR through generating and filtering phonetic transcriptions. Comput. Speech Lang. (2014), http://dx.doi.org/10.1016/j.csl.2014.02.006