ITERATIVE FILTERING OF PHONETIC TRANSCRIPTIONS OF PROPER NOUNS

Antoine Laurent[†§], *Teva Merlin*[†], *Sylvain Meignier*[†], *Yannick Estève*[†], *Paul Deléglise*[†]

[†]LIUM (Computer Science Research Center – Université du Maine) – Le Mans, France
[§]Spécinov – Trélazé, France
first.last@lium.univ-lemans.fr, a.laurent@specinov.fr

ABSTRACT

This paper focuses on an approach to enhancing automatic phonetic transcription of proper nouns by using an iterative filter to retain only the most relevant part of a large set of phonetic variants, obtained by combining rule-based generation with extraction from actual audio signals. Using this technique, we were able to reduce the error rate affecting proper nouns during automatic speech transcription of the ESTER corpus of French broadcast news. The role of the filtering was to ensure that the new phonetic variants of proper nouns would not induce new errors in the transcription of the rest of the words.

Index Terms**—** Speech recognition, Phonetic transcription, Proper nouns

1. INTRODUCTION

This work focuses on an approach to enhancing automatic phonetic transcription of proper nouns.

Proper nouns constitute a special case when it comes to phonetic transcription (at least in French, which was 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 guaranted to be pronounced the same way in two different proper nouns.

The lack of predictability also finds its roots in the wide array of origins proper nouns can be from: the more foreign the origin, the less predictable the pronunciation, with variations covering the whole range from the 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 they have to deal 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 the word error rate (WER): according to [1], 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 farther 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 from the frequency of their occurence—in other contexts such as in the case of automatic indexing of multimedia documents, or transcription of meetings.

Two common approaches to the problem of automatic phonetic transcription were proposed in the literature: the rule-based approach [2], and the statistic-based approach, including classification trees [3] and HMM-decoding-based methods [4, 5]. For the specific case of proper nouns, a study on dynamic generation of plausible distortions of canonical forms of proper nouns was proposed in [6].

We propose a method to build a dictionary of phonetic transcriptions of proper nouns by using an iterative filter to retain the most relevant part of a large set of phonetic variants, obtained by combining rule-based generation with extraction from actual audio signals. Rule-based generation of phonetic transcriptions is used to ensure that the most "common-sense" pronunciation variants are taken into account. It is combined with automatic extraction of phonetic variants from manually-annotated audio signals to enrich the set of transcriptions with those less predictable variants which actual people use. The iterative filter is then applied in order to reduce noise by invalidating the variants that are deemed irrelevant because too rarely used, and the ones that are found to be too prone to generate confusion with other words.

The intermediate (before filtering) and final sets of phonetic transcriptions were 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 [7].

First, we will present advantages and drawbacks of the generation and extraction methods. Next, we will explain how we combine them with the iterative filtering. Finally our results will be presented and commented on.

2. RULE-BASED GENERATION OF PHONETIC TRANSCRIPTIONS

A rule-based phonetic transcription system relies exclusively on the spelling of words to generate the possible corresponding chains of phones. It offers the advantage of providing phonetic variants even for words for which no speech signal is available. In the case of propers nouns, it serves to generate the most "common-sense" variants, *i.e.* the ones which people would use when they have no prior knowledge of the pronunciation of a particular proprer noun.

The rule-based generator we used was LIA_PHON [2]. During the ARC B3 evaluation campaign of French automatic phonetizers, 99.3 % of the phonetic transcriptions generated by LIA_PHON were correct. However, [2] reveals that transcription errors were not distributed evenly among the various classes of words: erroneous transcription of proper nouns represented 25.6 % of the errors even though proper nouns only represented 5.8 % of the test corpus, reflecting poorer performance by LIA_PHON on this class of words.

978-1-4244-2354-5/09/$25.00 ©2009 IEEE

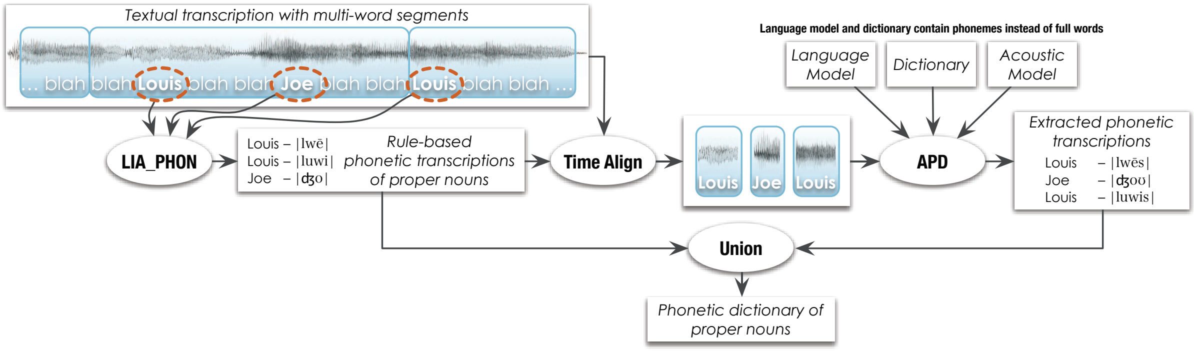
Textual transcription with multi-word segments
… blah blah Louis blah blah Joe blah blah Louis blah blah …
LIA_PHON
Louis – |lwē|
Louis – |luwi|
Joe – |ʤo|
Rule-based phonetic transcriptions of proper nouns
Time Align
Louis Joe Louis
Language model and dictionary contain phonemes instead of full words
Language Model
Dictionary
Acoustic Model
APD
Extracted phonetic transcriptions
Louis – |lwēs|
Joe – |ʤoʊ|
Louis – |luwis|
Union
Phonetic dictionary of proper nouns

Fig. 1. Illustration of the use of the acoustic-phonetic decoding system to extract phonetic transcriptions (transcriptions shown using the IPA)

Indeed, phonetic transcription of proper nouns has high and hardly predictable variability. It would be very difficult to establish the complete set of rules needed to automatically find all the possible phonetic transcriptions of every proper noun.

In order to do so, an ideal automatic system would have to be able to detect both the origin of the proper noun, and the various ways people, according to their own cultural and linguistic idiosyncrasies, might pronounce this noun. Unfortunately, both tasks are still open problems.

3. EXTRACTION OF PHONETIC TRANSCRIPTIONS USING ACOUSTIC-PHONETIC DECODING

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. Thus, proper nouns which are present several times in the corpus potentially get associated with several phonetic transcriptions each.

As is noted in [4], unconstrained phonetic decoding does not allow to obtain reliable phonetic transcriptions. Our own experiments lead us to the same conclusion.

The use of a language model allows some level of guidance for the speech recognition system: it does so by minimising the risk of having phoneme sequences with a very low probability appear in the transctiption results. We set constraints by using tied state triphones and a 3-gram language model as part of the decoding strategy, to generate the best path of phonemes. While this setup is close to a speech recognition system, here the dictionary and language model contain phonemes instead of full words. The trigram language model was trained using the phonetic dictionary used during the 2005 ESTER evaluation campaign. It contains about 65000 phonetic transcriptions of words, and was generated using BDLEX [8] and LIA_PHON. Only the words which were not part of the BDLEX corpus were phonetized automaticaly using LIA_PHON. Words which were identified as proper nouns were deleted from this dictionary before learning our 3-gram language model for phonemes.

As explained above, the first step consists in isolating the portions of signal corresponding to proper nouns using the textual transcription of the signal. However, in the manual transcription we used, words were not aligned with the signal: start and end times of individual words were not available; only longer segments (composed of several words) had their boundaries annotated. 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 figure 1).

The phonetic transcriptions used for proper nouns during this forced alignment were provided by LIA_PHON. Because of this, boundary detection was not very reliable. Portions of signal detected as proper nouns might overlap neighbor words. As a result, when applied to such portions of signal, the APD system might generate erroneous phonemes at the beginning and/or end of the proper nouns, which might in turn introduce errors when the flawed phonetic transcriptions are later used for decoding.

4. FILTERING OF PHONETIC TRANSCRIPTIONS

4.1. Motivation

The union of the generated transcriptions and the extracted transcriptions yields a high number of phonetic transcriptions per proper noun (specific figures for our experimental corpus can be found in section 5.4.1). This is expected to improve PNER.

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 occurences of the other categories of words is normally vastly superior to the number of occurences 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.

In order to minimize this risk, 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[1].

4.2. Iterative filtering

The goal pursued through this filtering is to detect and remove the phonetic variants of proper nouns that are the most likely to generate

[1] We already proposed a different approach to select phonetic transcriptions in a previous work [9]; however this early attempt was rendered impractical by its execution time which was directly proportional to the number of extracted phonetic transcriptions.

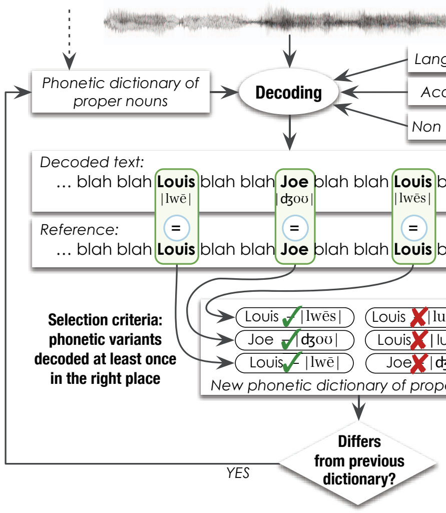
Phonetic dictionary of proper nouns
Decoding
Language Model
Acoustic Model
Non PN Dictionary
Decoded text:
… blah blah Louis blah blah Joe blah blah Louis blah Louis …
|lwē|
|ʤoʊ|
|lwēs|
|luwis|
=
=
=
≠
Reference:
… blah blah Louis blah blah Joe blah blah Louis blah blah …
Selection criteria: phonetic variants decoded at least once in the right place
Louis ✓ |lwēs|
Louis ✗ |luwis|
Joe ✓ |ʤoʊ|
Louis ✗ |luwi|
Louis ✓ |lwē|
Joe ✗ |ʤo|
New phonetic dictionary of proper nouns
Differs from previous dictionary?
YES

Fig. 2. 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.

confusion with other words. This is achieved by decoding the development corpus using the newly built phonetic dictionary (as well as a separate phonetic dictionary for all the other categories of words, of course).

Every 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.

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 phonetic transcriptions get removed from the dictionary.

This process is illustrated in figure 2, using the same example data as in figure 1.

5. EXPERIMENTS

5.1. Corpus

Experiments have been carried out on the ESTER corpus. ESTER was an evaluation campaign of French broadcast news transcription systems, which took place in January 2005 [7]. We divided the ESTER corpus into three parts: training, development and evaluation.

The training corpus used for the speech recognition system is composed of 81 hours of data recorded from four radio stations.

The development corpus, composed of 12.5 hours of data recorded from the same four radio stations, was used to generate and to filter the APD phonetic transcriptions.

The test corpus, used to evaluate the proposed methods, contains 10 hours from the same four radio stations plus two other stations, all of which were recorded 15 months after the development data.

Each corpus is annotated with named entities, allowing easy spotting of proper nouns.

5.2. Acoustic and language models

The decoding system is based on CMU Sphinx 3.6.

Our experiments were carried out using a one-pass decoding using 12 MFCC acoustic features plus the energy, completed with their primary and secondary derivatives. Acoustic models were trained on the ESTER training corpus. The trigram language model was trained using manual transcriptions of the corpus (1.35 M words). Articles from the French newspaper “Le Monde” were added, leading to a total of 319 M words.

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

5.3. Metric

The metrics used are the Word Error Rate (WER) and the Proper Noun Error Rate (PNER). The PNER is computed the same way as the WER but it is computed only for proper nouns and not for every word:

$$PNER = \frac{I + S + E}{N} \tag{1}$$

with I the number of wrong insertions of proper nouns, S the number of substitutions of proper nouns with other words, E the number of elisions of proper nouns, and N the total number of proper nouns.

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

5.4. Results

5.4.1. Number of phonetic transcriptions per proper noun

Figure 3 presents the number of phonetic transcriptions generated for the proper nouns present in the development corpus for each phonetic transcription system. The ESTER development corpus contains 1099 distinct proper nouns, appearing 4791 times.

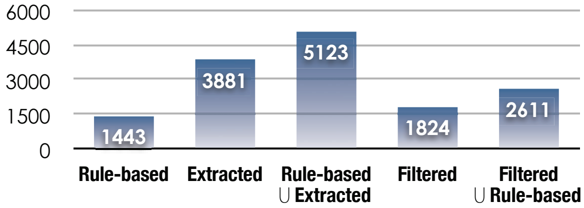

Fig. 3. Number of phonetic transcriptions generated by each method

The rule-based system generated 1443 differents transcriptions, *i.e.* an average of 1.31 phonetic transcriptions per proper noun.

From the same corpus, the APD system extracted 3881 phonetic transcriptions, which is an average of 3.53 variants for each proper noun. This number is more than 2.5 times the number of variants generated by LIA_PHON.

The union of the generated transcriptions and the extracted variants represents a total of 5123 transcriptions, *i.e.* an average of 4.66 variants per proper noun.

The iterative filtering removed 3539 phonetic transcriptions, leaving a total of 1824 phonetic transcription variants. Some proper nouns were completely removed. The number of proper nouns decreased from 1099 to 911. In order to be able to decode every proper noun, we merged the filtered dictionary with the rule-based generated dictionary. Doing so increased the number of phonetic transcriptions to 2611 (*i.e.* an average of 2.38 phonetic transcriptions per proper noun).

5.4.2. Error rates

Figure 4 unrolls the iterative filtering by presenting the PNER and the WER obtained when decoding the test corpus with the dictionary built at each step of the filtering process.

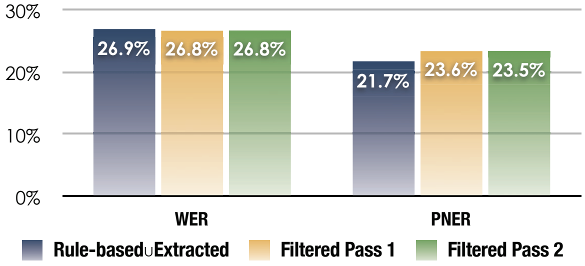

Fig. 4. PNER and WER of each filtering pass

The initial dictionary is the union of the phonetic transcriptions generated by LIA_PHON and the transcriptions extracted from the development corpus. It yields a WER of 26.9% and the best PNER (21.7 %). This can be explained by the high number of phonetic transcriptions (5123), which allows the correct decoding of many proper nouns, but generates noise in the rest of the decoding process.

After the first filtering pass, a decrease of the WER can be observed (26.8 %), while the decrease in the number of phonetic transcriptions (down to 1853) translates into an increase of the PNER (to 23.6 %).

After the second pass, the WER does not move, and the PNER slightly decreases (from 23.6 % to 23.5 %), probably as a result of some flawed phonetic transcriptions being eliminated.

The third pass dictionary is identical to the second pass dictionary, which signals the end of the filtering process.

Figure 5 compares the results (in terms of WER and PNER) of the reference phonetic dictionary (the rule-based generated variants) with those of the union of this dictionary with the filtered dictionary (after pass 2).

It confirms that the filtering method does not increase the global WER, while the unfiltered union of generated and extracted transcriptions did (as seen in figure 4: 26.9 %). The filtered dictionary also allows a gain of 3.8 % in terms of PNER.

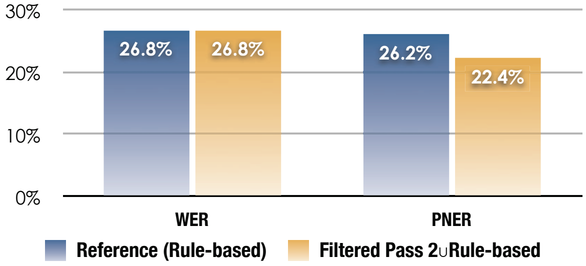

Fig. 5. PNER and WER on the reference system and on the filtered system

Other experiments, whose goal is to compute the WER on segments with and without proper nouns, show that the WER on segments that do not contain proper nouns are similar using both dictionaries.

On segments containing proper nouns, the union of the filtered and generated transcriptions yields a better WER than the generated transcriptions taken alone (the WER decreased by 0.5 %).

6. CONCLUSION

In this article, we proposed an iterative method to filter phonetic transcription variants by removing those which are the most likely to generate decoding errors. We applied this filtering method to a set of phonetic transcriptions of proper nouns obtained by combining rule-based generation with extraction from actual audio signals. The use of the resulting phonetic dictionary of proper nouns allows a gain in terms of PNER (Proper Noun Error Rate) and WER on the ESTER corpus. The WER on the segments that contain proper nouns decreased by 0.5 point and the PNER decreased by 3.8 points. As was expected, the rest of the corpus was unaffected, thanks to the filtering.

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.

7. REFERENCES

[1] R. Dufour, “From prepared speech to spontaneous speech recognition system: a comparative study applied to French language,” in *IEEE/ACM CSTST Student Workshop*, Cergy, France, October 2008.

[2] F. Béchet, “LIA_PHON : un système complet de phonétisation de textes,” in *TAL, Traitement Automatique des Langues*, 2001, pp. 47–67.

[3] R. I. Damper, Y. Marchand, M. J. Adamson, and K. Gustafson, “Automatic phonetic baseform determination,” in *Proc. of ESCA International Workshop on Speech Synthesis*, 1998, pp. 53–58.

[4] M. Bisani and H. Ney, “Breadth-first for finding the optimal phonetic transcription from multiple utterances,” in *Proc. of Eurospeech, European Conference on Speech Communication and Technology*, 2001.

[5] L. R. Bahl, S. Das, P. V. deSouza, M. Epstein, R. L. Mercer, B. Merialdo, D. Nahamoo, M. A. Picheny, and J. Powell, “Automatic phonetic baseform determination,” in *Proc. of ICASSP, International Conference on Acoustics, Speech, and Signal Processing*, December 1991, pp. 173–176.

[6] F. Béchet, R. de Mori, and G. Subsol, “Dynamic generation of proper name pronunciations for directory assistance,” in *Proc. of ICASSP, International Conference on Acoustics, Speech, and Signal Processing*, 2002, pp. 745–748.

[7] S. Galliano, E. Geoffrois, D. Mostefa, K. Choukri, J. F. Bonastre, and G. Gravier, “The ESTER phase II evaluation campaign for the rich transcription of French broadcast news,” in *Proc. of Eurospeech, European Conference on Speech Communication and Technology*, Lisbon, Portugal, September 2005.

[8] M. De Calmes and G. Perennou, “BDLEX: a lexicon for spoken and written French,” in *Proc. of LREC, International Conference on Language Resources and Evaluation*, 1998, pp. 1129–1136.

[9] A. Laurent, T. Merlin, S. Meignier, Y. Estève, and P. Deléglise, “Combined systems for automatic phonetic transcription of proper nouns,” in *LREC 2008*, Marrakech, Morocco, May 2008.