Acoustics-Based Phonetic Transcription Method for Proper Nouns

Antoine Laurent[†§], *Sylvain Meignier*[†], *Teva Merlin*[†], *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 improve automatic phonetic transcription of proper nouns. The method is based on a two-level iterative process that extract the phonetic variants from the audio signals before filtering the irrelevant variants. The evaluation of the method shows a decreasing of the Word Error Rate (WER) on segments of speech with proper nouns, without affecting negatively the WER on the rest of the corpus (ESTER corpus of French broadcast news).

Index Terms: Speech recognition, Phonetic transcription, Proper nouns

1. Introduction

This work focuses on an approach to enhancing automatic phonetic transcription of proper nouns. Accurate phonetic transcription of proper nouns is a difficult task. A proper noun with a given spelling is not guaranteed to be pronounced the same way depending on both the speaker and the geographic origin of that noun.

The domain of grapheme to phoneme (G2P) conversion is well covered in the literature. The most popular techniques are: dictionary look-up strategies [1], rule-based approaches [2, 3], and knowledge-based approaches. The latter set can be subdivided into three categories: top-down strategies (or local classification) [4, 5, 6], bottom-up approaches (or pronunciation by analogy) [7, 8, 9] and acoustics-based strategies [10, 11, 12].

We propose an acoustics-based approach based on a previous work to extract phonetic transcriptions of proper nouns from audio signals.

In manual transcriptions, words are not aligned with the signal: start and end times of individual words are not available. The time stamps are detected using a forced alignment of each word in the signal. When proper nouns are isolated in the signal, an acoustic-phonetic decoding (APD) system generates a set of phonetic variants. However this set is large and noisy, so it is filtered to invalidate 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.

In a previous work [13], we only employed the rule-based system LIA_PHON [2] as forced alignment dictionary and we emphasized the fact that using not very reliable phonetic transcriptions to get the forced alignment produces boundary detection errors. In the present article, we address this problem by comparing three different G2P systems to initialize the process, and we use a two-level iteration that employs the best filtered dictionary to re-initialize the process. This process gets repeated until two consecutive filtered dictionaries are exactly the same.

The generated 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 [14].

2. G2P methods to build initial dictionary

In this paper, three different G2P systems are evaluated to detect the boundaries of proper nouns. These initial dictionaries impact the quality of the boundaries detections by generating erroneous phonemes at the beginning and/or end of the proper nouns during the ADP decoding.

2.1. Dictionary look-up

The simplest strategy is a dictionary look-up. It consists in searching in a human-made phonetic dictionary. We used the BDLEX dictionary [1], which is very efficient but contains a finite number of entries (limited coverage). This dictionary does not contain any proper noun.

2.2. Rule-based

Rule-based conversion techniques do not exhibit coverage limits. Using a set of human-written rules, they rely on the spelling of words to generate the possible corresponding chains of phones.

In the case of propers nouns, it serves to generate the most "common-sense" variants, *i.e.* the ones people would use when they have no *a priori* knowledge of the pronunciation of a particular proper noun.

The rule-based generator used is LIA_PHON, an accurate French grapheme-to-phoneme converter. As noted in [2] it performs less well with proper nouns than with words of other classes.

Indeed, natural languages frequently exhibit irregularities, and phonetic transcription of proper nouns has high and hardly predictable variability. It would be impossible to establish the complete set of rules needed to automatically find all the possible phonetic transcriptions of every proper noun, and irregularities would have to be captured by exception rules.

2.3. Joint-Sequence models (JSM)

This system is a *data-driven* conversion system that 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 [9] will be denoted as JSM in the rest of this article. Being a data-driven conversion system, JSM must be fed pronunciation examples in order to be trained. Training takes a pronunciation dictionary and create new model files successively, starting with a unigram model and up to a 6-gram model. The model files can then be used to transcribe words that were not in the dictionary.

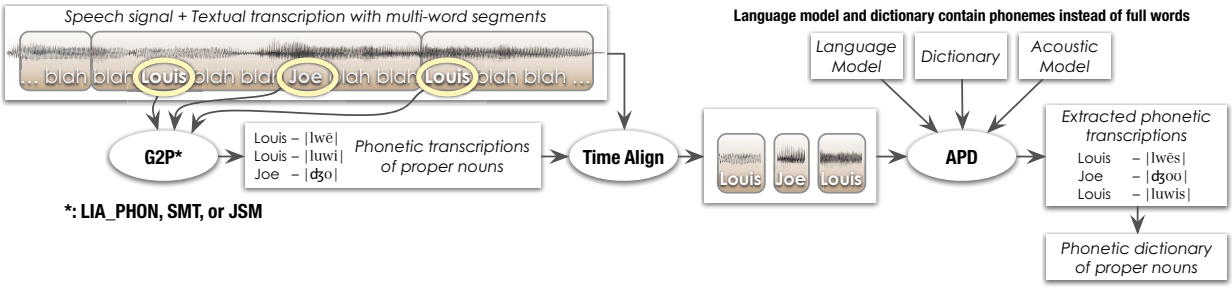

Figure 1: Illustration of the use of the acoustic-phonetic decoding system to extract phonetic transcriptions (transcriptions shown using the IPA)

2.4. Statistical machine translation (SMT)

We proposed a method in [15], based on the use of a statistical machine translation (SMT) system to convert graphemes to phonemes. A SMT system is commonly used to translate word sequences of a source language into a target language. We used it in order to rewrite sequence of letters into sequence of phonemes. The training step needs a data corpus which is composed of bitext data. In our case, a bitext would associate sequences of letters with sequences of phonemes. We chose a representation that allows to take into account inter- and intra-word influences.

Optimization of the translation model parameters is commonly based on the maximization of the BLEU score [16]. Our proposed method is based on the minimization of the Levenshtein edit distance, which gives better results when SMT is used to convert graphemes to phonemes.

3. Extraction of phonetic variants using APD

3.1. Extraction of segments that contain proper nouns

In order to enrich the set of phonetic transcriptions of proper nouns with some less predictable variants, we used an APD system on speech segments that correspond to utterances of proper nouns.

The boundaries 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 first forced alignment employed a word dictionary generated by one of the three grapheme to phoneme conversion methods described in section 2. When start and end times of segments that contains proper nouns are determined, 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.

3.2. Acoustic-phonetic decoding

As noted in [17, 13], unconstrained phonetic decoding does not allow to obtain reliable phonetic transcriptions. The decoding strategy uses tied state triphones and a 3-gram language model to allow some level of guidance during the decoding.

While the APD system is very close to an ASR system, the dictionary and language model used for this task contain phonemes instead of full words. Our trigram language model is trained by using our baseline phonetic dictionary in which every word identified as a proper nouns is removed.

Our baseline phonetic dictionary contains about 65,000 phonetic transcriptions of words drawn from the BDLEX dictionary. Phonetic transcriptions of words that are not present in BDLEX are generated using LIA_PHON.

4. Filtering irrelevant variants

4.1. Motivation

The APD system extracts a high number of phonetic variants per proper noun (the numbers are reported in Table 1). Because occurrences of the other categories of words are normally much more frequent than occurrences of proper nouns, there is a risk of seeing any improvement in PNER (Proper Noun Error Rate) being outbalanced by a negative impact on other words of the corpus and on the global WER.

4.2. Method

In order to minimize this risk, the generated phonetic transcriptions are filtered. The filtering removes phonetic variants of proper nouns that are the most likely to generate confusion with other words. We decode the training corpus using the proper noun phonetic dictionary that we want to filter, augmented with the phonetic transcriptions of non proper noun words.

Every phonetic transcription that is never used to decode the corresponding proper noun is removed from the dictionary. Indeed, 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 and 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. Iterative acoustic-based phonetic transcription method

The parameter that we want to tune in this paper is the dictionary used to make the forced alignment between signal and textual reference.

Once the filtering process is over, the filtered dictionary is used instead of the G2P dictionary used at the beginning. The full process (alignment/APD/filtering) is repeated until we obtain the same filtered dictionary twice (see Figure 3).

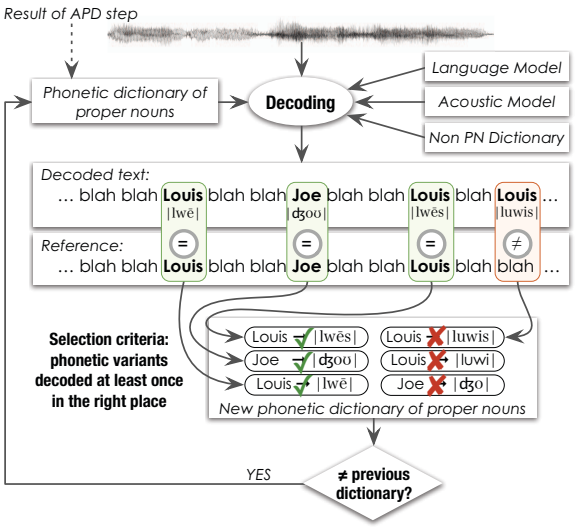

Figure 2: Illustration of filtering of phonetic transcriptions.

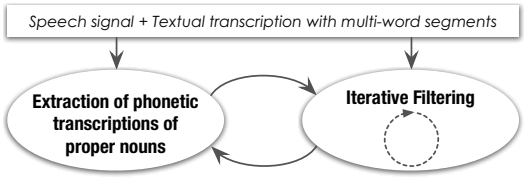

Figure 3: Iterative process using APD and filtering

6. Experiments

6.1. Corpus

Experiments have been carried out on the ESTER corpus. ESTER is an evaluation campaign of French broadcast news transcription systems which took place in January 2005 [14]. The ESTER corpus was divided into three parts: training, development and evaluation. The training (81 hours) and the development (12.5 hours) corpora are composed of data recorded from four radio stations in French. The test corpus is composed of 10 hours coming from the same four radio stations plus two other stations, all of which recorded 15 month after the development data. The training corpus was used to learn our automatic speech recognition system. The training corpus and the development corpora are jointly employed to extract and to filter them. JSM and SMT grapheme to phoneme converters were also trained over the ESTER 1 training corpus.

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

6.2. 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 train corpus and articles from the French newspaper "Le Monde".

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.

6.3. Metric

The metrics used are based on the Word Error Rate (WER) and on the Proper Noun Error Rate (PNER). The PNER is computed the same way as the WER but it is computed only for proper nouns: $PNER = (I + S + E)/(N)$ where I is 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.

6.4. Results

6.4.1. Number of phonetic transcriptions per proper noun

Table 1 presents the number of phonetic transcriptions generated by each system. The ESTER development plus training corpus contains 3,348 distinct proper nouns, appearing 28,866 times.

Table 1: *Number of phonetic transcriptions generated by each method*

G2P method	Generated variants	Extracted variants	After 1 iteration	After 3 iterations
LIA_PHON	4,364	20,218	6,776	6,502
SMT	7,031	20,184	7,065	6,802
JSM	3,626	20,008	6,876	6,708

The APD system extracts an average of 4.34 times the number of variants generated by the different G2P methods. The iterative filtering always keep about 7,000 phonetic transcription variants from the 20,000 variants generated by the APD. The number of variants contained in the final filtered dictionary slightly decreases. For every of our three grapheme to phoneme strategies, the filtering stage completed in 3 iterations.

6.4.2. Results of the first iteration

We want to compare the direct use of the G2P methods with the use of the extraction and filtering of phonetic transcriptions of proper nouns. Figure 4 shows the PNER obtained using the filtering method with each G2P system on the test corpus.

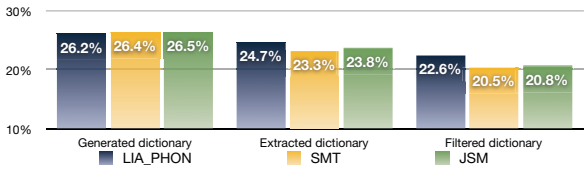

Figure 4: PNER using each G2P method (ESTER test corpus)

These results show that the use of our method allows to have a significant gain in terms of PNER using every phonetic transcription dictionary. As we can see, the APD method supplemented by the SMT-based grapheme to phoneme conversion system is the one that gives the lowest PNER. Figure 5 compares the results of the reference phonetic dictionary (denoted as LIA_PHON) with those of SMT and JSM, in terms of WER computed only over segments that contain proper nouns.

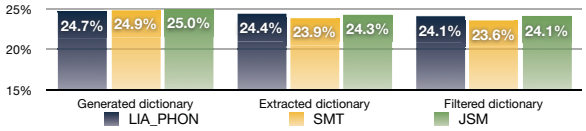

Figure 5: WER on test corpus on segments with proper nouns

Figures 4 and 5 show the interest of filtering: it allows to reduce both PNER and WER on segments with proper nouns.

6.4.3. 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. WER and PNER are computed on segments that contains 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 % to 24.0 % and the PNER decreased from 22.6 % to 22.5 %. SMT allows a gain of 0.2 % in term of WER and a gain of 0.3 % in term of PNER.

Table 2: WER and PNER using the full iterative process

G2P	WER (segments with PN)	PNER
LIA_PHON (ref)	**24.7%**	**26.2%**
SMT	24.9%	26.4%
JSM	25.0%	26.5%
First filtering iteration		
LIA_PHON	24.1%	22.6%
SMT	23.6%	20.5%
JSM	24.1%	20.8%
Second filtering iteration		
LIA_PHON	24.1%	22.6%
SMT	23.5%	20.3%
JSM	24.0%	20.5%
Third filtering iteration		
LIA_PHON	24.0%	22.5%
SMT	23.4%	20.2%
JSM	23.9%	20.5%

Figure 6 shows the WER obtained on the whole ESTER 1 test corpus. It shows that the filtering step does not generate new errors with other word classes.

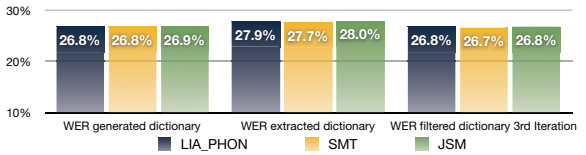

Figure 6: WER on test corpus on every segment

7. Conclusion

In this article, we propose an iterative acoustic-based phonetic transcription generator applied to proper nouns. Our method contains a filtering step used to remove phonetic transcriptions that are the most likely to generate decoding errors. We apply this filtering method to a set of phonetic transcriptions of proper nouns obtained by using various G2P systems to initialize our method, before the extraction of variants from actual audio signals. The use of resulting phonetic dictionaries of proper nouns allows a gain in terms of PNER (Proper Noun Error Rate) and WER on the ESTER corpus. The best results are obtained by using a SMT (Statistical Machine Translation [15]) 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 points. As was expected, the WER on rest of the corpus is unaffected or slightly improved, thanks to the iterative process.

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.

8. References

[1] M. De Calmes and G. Pérennou, "BDLEX: a lexicon for spoken and written French," in *Proc. of LREC 1998*, 1998.

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

[3] J. Tihoni and G. Pérennou, "Phonotypical transcription through the GEPH expert system," in *Proc. of Eurospeech 91*, Genova, Italy, September 1991.

[4] K. Torkkola, "An efficient way to learn English grapheme-to-phoneme rules automatically," in *Proc. of ICASSP*, vol. 2, Minneapolis, USA, April 1993, pp. 199–202.

[5] C. Ma and M. A. Randolph, "An approach to automatic phonetic baseform generation based on bayesian networks," in *Proc. of Interspeech 2001*, Aalborg, Danemark, September 2001.

[6] K. Jensen and S. Riis, "Self-organizing letter code-book for text-to-phoneme neural network model," in *Proc. of ICSLP 2000*, vol. 3, Beijing, China, October 2000, pp. 318–321.

[7] L. Galescu and J. F. Allen, "Bi-directional conversion between graphemes and phonemes using a joint n-gram model," in *Proc. of ISCA 2001*, Perthshire, Scotland, August 2001.

[8] J. R. Bellegarda, "Unsupervised, language-independent grapheme-to-phoneme conversion by latent analogy," in *Speech Communication*, vol. 46, 2005, pp. 140–152.

[9] M. Bisani and H. Ney, "Joint-sequence models for grapheme-to-phoneme conversion," *Speech Communication*, vol. 50, no. 5, pp. 434–451, 2008.

[10] W. Byrne, M. Finke, S. Khudanpur, J. McDonough, H. Nock, M. Riley, M. Saraclar, C. Wooters, and G. Zavaliagkos, "Pronunciation modeling using a hand-labelled corpus for conversational speech recognition," in *Proc. of ICASSP*, Seattle, USA, May 1998.

[11] S. Deligne and L. Mangu, "On the use of lattices for the automatic generation of pronunciations," in *Proc. of ICASSP*, vol. 1, Hong-Kong, China, April 2003, pp. 204–207.

[12] T. Svendsen, F. K. Soong, and H. Purnhagen, "Optimizing baseforms for HMM-based speech recognition," in *Proc. of Eurospeech 1995*, Madrid, Spain, September 1995, pp. 783–786.

[13] A. Laurent, T. Merlin, S. Meignier, Y. Estève, and P. Deléglise, "Iterative filtrering of phonetic transcriptions of proper nouns," in *Proc. of ICASSP*, 2009.

[14] 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 2005*, Lisbon, Portugal, September 2005.

[15] A. Laurent, P. Deléglise, and S. Meignier, "Grapheme to phoneme conversion using an SMT system," in *Proc. of Interspeech*, 2009.

[16] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: a method for automatic evaluation of machine translation," in *Proc. of ACL*, 2002.

[17] M. Bisani and H. Ney, "Breadth-first for finding the optimal phonetic transcription from multiple utterances," in *Proc. of Eurospeech 2001*, 2001.