Combined systems for automatic phonetic transcription of proper nouns

A. Laurent[1,2], T. Merlin[1], S. Meignier[1], Y. Estève[1], P. Deléglise[1]

[1] Laboratoire d'Informatique de l'Université du Maine
Le Mans, France
firstname.lastname@lium.univ-lemans.fr

[2] Spécinov
Trélazé, France
a.laurent@specinov.fr

Abstract

Large vocabulary automatic speech recognition (ASR) technologies perform well in known, controlled contexts. However recognition of proper nouns is commonly considered as a difficult task. Accurate phonetic transcription of a proper noun is difficult to obtain, although it can be one of the most important resources for a recognition system. In this article, we propose methods of automatic phonetic transcription applied to proper nouns. The methods are based on combinations of the rule-based phonetic transcription generator LIA_PHON and an acoustic-phonetic decoding system. On the ESTER corpus, we observed that the combined systems obtain better results than our reference system (LIA_PHON). The WER (Word Error Rate) decreased on segments of speech containing proper nouns, without affecting negatively the results on the rest of the corpus. On the same corpus, the Proper Noun Error Rate (PNER, which is a WER computed on proper nouns only), decreased with our new system.

1. Introduction

Large vocabulary automatic speech recognition (ASR) technologies perform well in known, controlled contexts. However proper nouns are frequently out of the systems' vocabulary and their recognition is commonly considered as a difficult task.

There are many situations in which we need to transcribe proper nouns correctly. In the context of indexing multimedia contents, recognizing names pronounced during a broadcast news or a show provides interesting clues about the speakers. In the case of meeting transcription, it is important to know who talks about whom.

Although phonetic transcription of proper nouns can be one of the most important resources for a recognition system, accurate phonetic transcription of a proper noun is difficult to obtain. In fact, a proper noun with a given spelling can be pronounced in different ways depending on both the geographic origin of that noun, and the speaker. Pronunciation of proper nouns is less normalized than pronunciation of other words. This is especially the case for nouns foreign to the language of the speaker.

Two common approaches of the problem of automatic phonetic transcription are proposed in the literature: the rule-based approach (Béchet, 2001), and the statistic-based approach, such as classification trees (Damper et al., 1998) or HMM-decoding based methods (Bisani and Ney, 2001; Bahl et al., 1991). For the specific case of proper nouns, a study on dynamic generation of plausible distortions of canonical forms of proper nouns is proposed in Béchet et al. (2002). This study has been carried out for use in the context of a directory assistance application developed by France Télécom R&D. The method consists in re-evaluating of the n best speech recognition hypotheses yielded by a one-pass decoding where distortions depend on the nature of the competing hypotheses.

The method we propose here is based on combinations of a rule-based phonetic transcription generator and an acoustic-phonetic decoding system. With the latter system, phonetic transcriptions for each word are obtained by decoding the parts of the signal containing the word (according to manual transcription of the signal into words). It allows extraction of a high number of phonetic transcriptions for words present in a development corpus, including some unusual pronunciations. The rule-based generator, on the other hand, tends to generate the most "common-sense" phonetic transcriptions for every word, including words not present in the development corpus.

The experiments proposed in this article focus on the automatic phonetic transcription of proper nouns, as in Béchet et al. (2002). New phonetic transcriptions will be evaluated in terms of Word Error Rate (WER) and Proper Noun Error Rate (PNER). These rates will be evaluated using French broadcast news from the ESTER evaluation campaign (Galliano et al., 2005).

First, we will present advantages and drawbacks of the rule-based and acoustic methods. Next, we will explain our combined methods. Finally our results will be presented and commented.

2. Automatic phonetic transcription system

2.1. Rule-based

LIA_PHON, a rule-based phonetic transcription system (Béchet, 2001), uses the spelling of words to determine the corresponding chain of phones. One of the strengths of this system is to perform the transcription without relying on the speech signal.

LIA_PHON participated in the ARC B3 evaluation campaign of French automatic phonetizers, in which phonetic transcriptions generated by the systems were compared with the results of phonetization by human experts. Er-

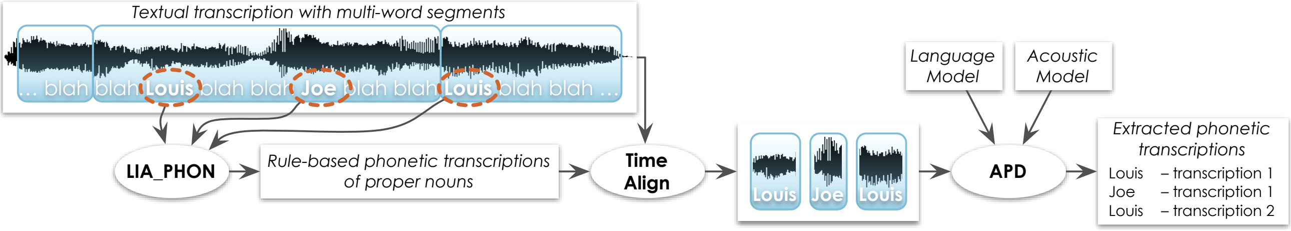
Textual transcription with multi-word segments
... blah blah Louis blah blah Joe blah blah Louis blah blah ...
LIA_PHON
Rule-based phonetic transcriptions of proper nouns
Time Align
Louis
Joe
Louis
Language Model
Acoustic Model
APD
Extracted phonetic transcriptions
Louis – transcription 1
Joe – transcription 1
Louis – transcription 2

Figure 1: Use of the acoustic-phonetic decoding system to extract phonetic transcriptions

ror rate was calculated according to the same principle as for the classical word error rate used in speech recognition. 99.3 % of the phonetic transcriptions generated by LIA_PHON were correct (for a total of 86938 phonemes). However, Béchet (2001) 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 generated by LIA_PHON even though proper nouns only represented 5.8 % of the test corpus, reflecting poorer performance by LIA_PHON on this class of words.

Indeed, phonetic transcription of proper nouns has high and hardly predictable variability. For example in the ESTER development corpus, the first name of the singer “Joey Starr” is pronounced either “`dZoe`”, “`dZoj`”, “`Zoe`”, or “`Zoj`”[1], even though all the speakers involved speak French. It would be very difficult to establish the complete set of rules needed to automatically find all the possible phonetic transcriptions.

In order to do so, an ideal automatic system would 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.

2.2. System based on acoustic-phonetic decoding

The acoustic-phonetic decoding system (APD) 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. Proper nouns which are present several times in the corpus thus potentially get associated with several phonetic transcriptions each.

As is noted in Bisani and Ney (2001), 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 decoding is close to a speech recognition system, 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 (De Calmes and Perennou, 1998) and LIA_PHON. Only the words which were not part of the BDLEX corpus were phonetised automaticaly using LIA_PHON. Words which were identified as proper nouns have been 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 word transcription of the signal. Unfortunately, in the manual transcription we used, words were not aligned with the signal: start and end times of individual word were not available, with only longer segments (composed of several words) having their boundaries annotated. The start and end times of each word of the transcription were 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.

3. Combination

The aim of combining both systems is to get the best out of each, of course without impacting negatively the rest of the speech recognition process.

3.1. Union

The first proposed combination follows the simplest strategy, by building a dictionary as the union of both LIA_PHON and APD phonetic transcriptions. In this dictionary, there is a high number of phonetic transcriptions per word, as can be seen in section 4.1.

3.2. Selection

To eliminate excessive phonetic transcriptions that may generate errors during speech recognition, we propose a way to validate phonetic transcriptions. Selection of valid transcriptions is done by testing each phonetic transcription against the development corpus: only those phonetic transcriptions which allow the corresponding word to be recognized successfully are selected.

[1] Phonetic transcriptions given in Sampa format

For each phonetic transcription variant of each proper noun, a temporary dictionary is built, containing only this phonetic transcription of this proper noun, along with all the non-proper noun words. The speech recognition system is then applied to all the sentences of the development corpus that contain this proper noun, using the temporary dictionary. The tested phonetic transcription for this proper noun is considered as valid only if the proper noun was correctly decoded at least once. In this process, the other words of the temporary dictionary play the role of a rejection model when trying to recognize the proper noun being tested.

4. Experiments

4.1. 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 (Galliano et al., 2005). The ESTER corpus was divided into three parts: training, development and evaluation.

The training corpus is composed of 81 hours of data recorded from four radio stations (France Inter, France Info, RFI, RTM). This corpus was used to train the speech recognition system.

The development corpus is composed of 12.5 hours of data recorded from the same four radio stations. This corpus was used to generate and to validate 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 was recorded 15 months after the development data.

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

4.2. 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 resulting in 1.35 M words. Articles from the French newspaper “Le Monde” were added, resulting in 319 M words.

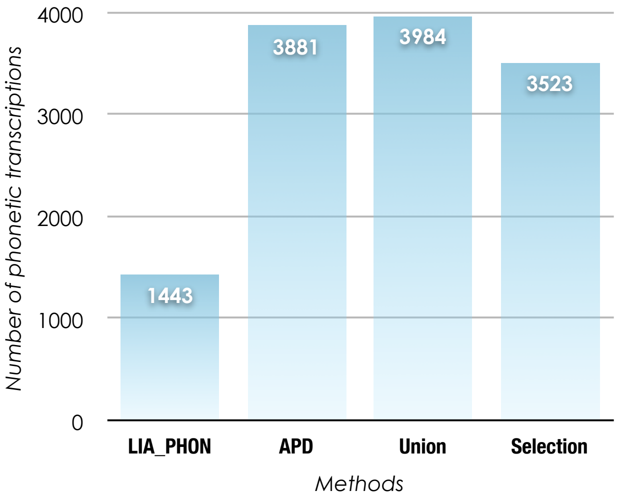

Figure 2: Number of phonetic transcriptions generated by each method

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.

4.2.1. Phonetic transcriptions per proper noun

Figure 2 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 1098 distinct proper nouns, appearing 4791 times.

The rule-based system generates 1443 differents transcriptions, which represents an average of 1.31 phonetic transcriptions per proper noun.

On the same corpus, the APD system generates 3881 phonetic transcriptions, for 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 variant sets generated by both systems represents a total of 3984 transcriptions, *i.e.* an average of 3.64 variants per proper noun.

After filtering with the selection method, which is in charge of eliminating excessive phonetic transcription variants generated by the APD, the number decreases to 3523, *i.e.* an average of 3.21 variants per proper noun.

4.3. 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 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 (in other words, the number of proper nouns which were omitted in the transcription), and N the total number of proper nouns.

The WER permits to evaluate the impact of the dictionaries over the test corpus, whereas the PNER permits to evaluate the quality of the detection of proper nouns.

4.4. Results

Figure 3 presents the PNER obtained when decoding using the various sets of phonetic transcriptions of proper nouns generated by the proposed methods.

Figure 4 presents the WER obtained in the same cases.

The reference system is LIA_PHON, which obtains 26.8 % of WER and 26.0 % of PNER.

The APD system obtains the worst WER and PNER: respectively 27.2 % and 32.3 %.

The union of LIA_PHON phonetic transcriptions and ADP phonetic transcriptions gives the best performance in term of PNER. However, the WER is slightly higher (0.1 point) than for the reference system.

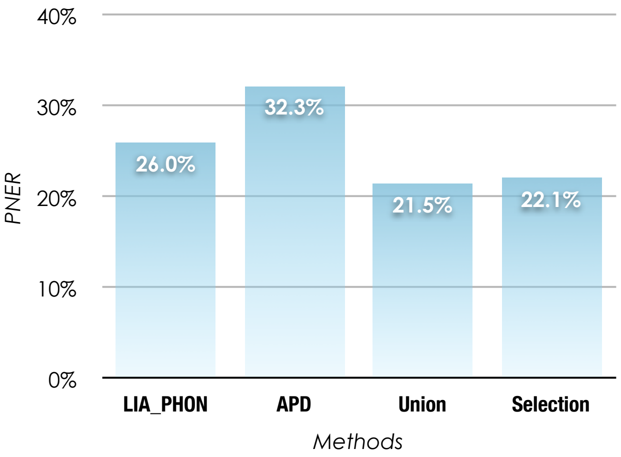

Figure 3: PNER for each method on ESTER test corpus

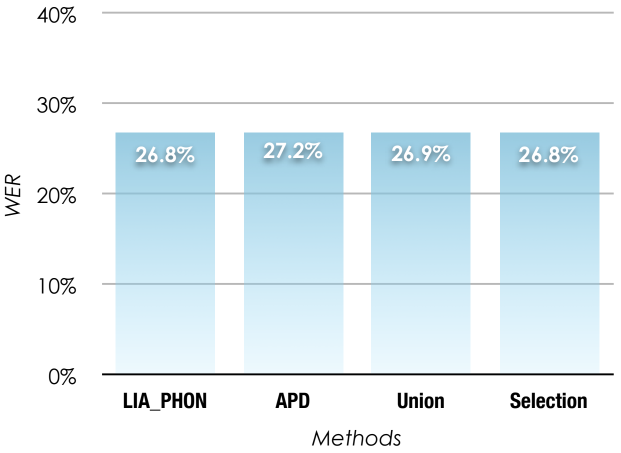

Figure 4: WER for each method on ESTER test corpus

We applied the selection strategy to the phonetic transcriptions generated by the APD system. The union of the filtered phonetic transcriptions and the phonetic transcriptions generated by LIA_PHON is referred to as "Selection" in the figures. For this system, we observed a gain of near 3.9 points of PNER without degrading the WER.

The WER is not widely affected because proper nouns represent only a small part of the words in the corpus: 1840 words out of 113918 words of the test corpus (≈ 1.6 %). To observe the influence of the various proposed methods on the WER, we proposed to evaluate separately the segments that contain proper nouns. Figure 5 shows results for the segments with and without proper nouns.

The most remarkable results are for the "Selection" system: it yields a gain of 0.5 point of WER over LIA_PHON for segments containing proper nouns, without affecting the WER on the other segments.

5. Conclusion

This article presented a method to automatically generate phonetic transcriptions of proper nouns.

We proposed ways of combining a rule-based automatic phonetic transcription generator (LIA_PHON) and an acoustic-phonetic decoding system.

On the ESTER corpus, we observed that the combined systems obtain better results than our reference system (LIA_PHON). With the proposed combination, the WER decreased by 0.5 point on segments of speech containing proper nouns, without affecting negatively the results on the rest of the corpus.

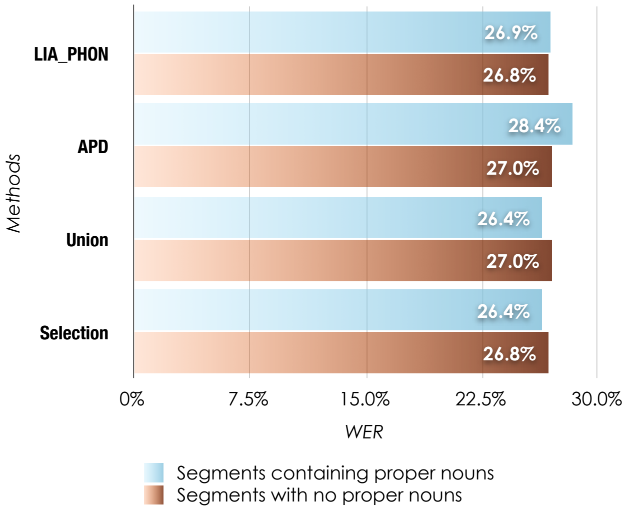

Figure 5: Word Error Rate on ESTER test corpus for segments containing proper nouns and segments with no proper nouns.

An interesting field where the proposed method could be applied is the task of named identification. This task consists in extracting the speaker identities (firstname and lastname) from the transcription (Estève et al., 2007). The new phonetic transcriptions yielded by the proposed method should contribute to make the detection easier by improving the decoding of proper nouns. Preliminary experiments carried out recently at LIUM for a yet unpublished work tend to confirm this hyptohesis.

Pushing further the principle backing the method described in this article, future developments could focus on generalizing the method to other classes of words beyond just proper nouns.

6. References

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

F. Béchet, R. de Mori, and G. Subsol. 2002. Dynamic generation of proper name pronunciations for directory assistance. In *Proc. of ICASSP, International Conference on Acoustics, Speech, and Signal Processing*, pages 745–748.

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

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

R. I. Damper, Y. Marchand, M. J. Adamson, and K. Gustafson. 1998. Automatic phonetic baseform determination. In *Proc. of ESCA International Workshop on Speech Synthesis*, pages 53–58.

M. De Calmes and G. Perennou. 1998. BDLEX: a lexicon for spoken and written French. In *Proc. of LREC, International Conference on Language Resources and Evaluation*, pages 1129–1136.

Y. Estève, S. Meignier, P. Deléglise, and J. Mauclair. 2007. Extracting true speaker identities from transcriptions. In *Proc. of ICSLP, International Conference on Spoken Language Processing*.

S. Galliano, E. Geoffrois, D. Mostefa, K. Choukri, J. F. Bonastre, and G. Gravier. 2005. 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.