Integrating Contextual Phonological Rules in a Large Vocabulary Decoder

Guillaume Gravier[1], *François Yvon*[2], *Bruno Jacob*[1,3], *Frédéric Bimbot*[1]

(1) IRISA, Campus de Beaulieu, F-35042 Rennes Cedex
(2) ENST, 46 rue Barrault, F-75634 Paris Cedex 13
(3) LIUM, Université du Maine, F-72085 LE MANS CEDEX 9

ggravier@irisa.fr, yvon@inf.enst.fr, bruno.jacob@lium.univ-lemans.fr, bimbot@irisa.fr

Abstract

This paper presents an approach to the integratation of contextual phonological rules in the beam-search algorithm of a large vocabulary speech recognition system. The main interest of contextual transcription rules is that they implement constraints on pronunciations sequences which complement the bigram constraints on word sequences. As such, they should help avoiding acoustic confusions and reduce the search space. In our approach, contextual transcription do not incur any augmentation of the lexicon size. This approach is evaluated on a dictation task in French for two different sets of contextual phonological rules. Our results show that, given the current resources, the introduction of contextual rule deteriorates the recognition rate. We discuss the possible factors explaining this surprising result and outline the problems of defining a set of contextual phonological rules and integrating them in the search algorithm.

1. Introduction

Large vocabulary speech recognition systems typically rely on the use of a phonetic description of the vocabulary words, each phonetic unit being modeled by a hidden Markov model (HMM). To be able to cope with different speech styles, it is important that words may have several pronunciations (or pronunciation variants). In most systems, pronunciations variants are introduced in the lexicon and the phonetic description of the vocabulary words is done off-line in a static manner [1]. For each word, a list of pronunciation variants is compiled, possibly with probabilities for each variants. There are several ways of generating the phonetic transcriptions of a word (see *e.g.* [2]) but the use of phonological rules, whether defined by experts [3] or extracted by data-driven techniques [4], is certainly the most popular approach. Phonological rules are typically applied to the isolated words and it is therefore not possible to deal with cross-word rules or variants.

In some cases, the pronunciation of a word highly depends on the context, *i.e.* on the surrounding words. A typical case of this situation is the phenomenon of liaison in French. A liaison is the phonetic realization of a word final consonant in the context of a following word initial vowel or *mute-h*, which can be compulsory, forbidden, or optional. A classical example is the word "*les*" which may be pronounced /le/, or /lEz/ if a liaison occurs. In this particular case, the liaison occurs if the following word is plural and starts with a vocalic sound. Apart from the case of the liaison, there are other situations where the pronunciation depends on the surrounding words.

One solution to introduce contextual information consists in multiplying the number of entries in the lexicon so that each entry corresponds to a word along with its pronunciation. Another solution is to add multi-words to the lexicon. In the first approach, the language model (LM) on the extended lexicon enables to model contextual rules by considering the probability of a (word,pronunciation) pair given the previous (word,pronunciation) pair [5, 6]. However, this approach generates huge pronunciation lexicons and requires large corpora annotated with pronunciation variants to estimate the LM probabilities.

In this paper, we present our approach to introduce contextual transcription rules in a trie-based large vocabulary decoder. The basic principle of our approach is to associate classes to each words, where a class is used to describe a type of orthographic form and intersects information at different level (syntaxic, morphologic, lexical, ...), and to check cross-transcription compatibility (using the word classes) between adjacent pronunciations during the search algorithm.

The motivation for such an approach is that we believe that modeling contextual interactions at the search level should help avoiding (linguistic) confusions and improve the recognition system. Another aspect is that in the current approaches to model pronunciation variants, there are no constraints in the sequence of variants which is not the case in practice. Our approach is a first step toward a model to constrain successive pronunciation variants.

The paper is organized as follows: we first recall the basics of the trie-based search algorithm with a bigram LM as described in [7, 8]. We then present in details the principle of our contextual transcription rules and show how the knowledge of the context can be integrated at the search level before illustrating our method with results on a dictation task in French. We finally discuss the problems of rule generation and of rule set consistency.

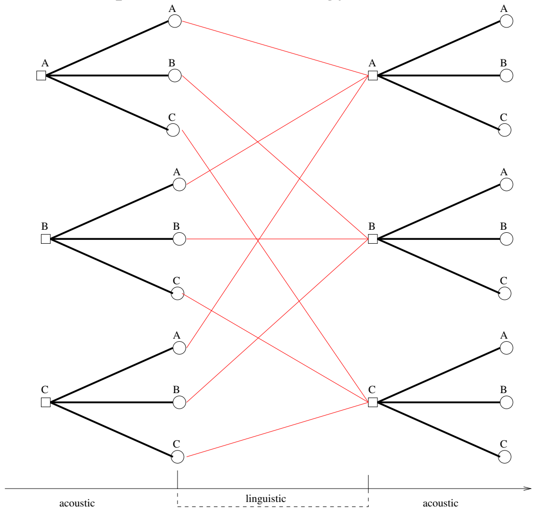
2. Overview of the search algorithm

The *Sirocco* speech recognition system[1] implements a beam-search strategy with bigram language models with a trie-based pronunciation lexicon. The algorithm was fully described in [7] and we recall here the basics of the search algorithm. As only word ends are concerned with contextual transcription rules, we will focus the description of the search algorithm on the processing of word end hypotheses.

With a trie organization of pronunciations, the recognition of the current word is delayed until a leaf node of the trie is reached. This makes it impossible to apply the bigram language model directly since when a word end is found, it is not possible to directly know what is the following word. A solution to this problem consists in introducing a separate trie copy for each predecessor word: when a word end is detected in a trie copy,

[1] See `http://www.enst.fr/~sirocco`

Figure 1: Principle of the search strategy for trie-based lexicons.



the language model can be applied for the current word end given the predecessor word. The principle is illustrated figure 1, where the thin lines correspond to the LM score while the thick ones correspond to the acoustic score.

Inside trie copies, acoustic hypotheses are propagated according to a classical Dynamic Programming (DP) equation. At word ends, *i.e.* when a leaf node of the trie is reached, the bigram score is added to the complete path score and the DP maximization is performed over all previous words. Suppose an acoustic hypothesis is at time t in the leaf node S_w of a trie copy, thus generating the word end hypothesis (w, t) for w at time t. If we denote $h_w(t)$ the log-likelihood of the best word sequence ending with word w at time t, the DP maximization at the word level is given by

$$h_w(t) = \max_{v \in \mathcal{V}} q_v(t, S_w) + \beta \ln P(v, w), \qquad (1)$$

where $q_v(t, S_w)$ denotes the log-likelihood of the best path up to node S_w of the trie copy corresponding to the predecessor word v at time t; $P(v, w)$ denotes the bigram score and β the "fudge" factor. The maximization (1) is performed over the set of all the possible words $\mathcal{V}$. Finally, the (w, t) word end hypothesis is recorded with a back-reference to the word hypothesis (v, t') which maximized (1), where t' is the back-pointer information propagated with the acoustic hypothesis. New acoustic hypotheses are grown with w as their predecessor word with back-pointers to (w, t).

3. Contextual transcription rules

In this section, we first define how contextual transcription rules are specified in the *Sirocco* system before detailing the implementation in the search algorithm.

3.1. Definition of contextual transcription rules

To define contextual phonological rules, a set of lexical classes is first associated with each word[2] of the vocabulary. Contextual phonological rules are then defined for a given left and right context, where a context is a logical combination of classes. In other words, a phonological rule applies to a word only if the predecessor word matches the rule left context and if the following word matches the rule right context. Probabilities may be associated to each rules.

Let us illustrate these concepts with the example of the liaison in "*les*". As mentioned in the introduction, the word "*les*" is pronounced /le/, unless followed by a word starting with a vocalic sound. In the latter case, the pronunciation includes the liaison and is /lEz/. Furthermore, if "*les*" is a noun or an adjective, the pronunciation /lEz/ occurs when the following word starts with a vocalic sound and is plural.

This is a typical example of contextual rules since, for *les taken as a noun or an adjective*, the rule *les* → /lEz/ applies only if the following word matches the condition "is plural and starts with a vocalic sound". In all other cases, the rule *les* → /le/ must be used. The corresponding contextual transcription rules are given table 1 where *Vinit* is the class of words starting with a vocalic sound, *Plural* is the class of plural words and * denotes any word. In this example, probabilities have been associated to the rules.

Table 1: Example of contextual phonological rules for word "*les*".

(*)	*les*	(Vinit & Plural)	→	/lEz/	1.0
(*)	*les*	(!(Vinit & Plural))	→	/le/	1.0

Note that lexical classes were used throughout this section for illustration purposes but the classes associated to a word can virtually be anything. As an example, we are using classes to perform automatic speech alignment: each word is associated to a class representing that particular word at a given position in the sentence and the transcription rule for a word applies only if the left (resp. right) context matches the previous (resp. next) word in the target sentence.

3.2. Integrating rules in the search algorithm

To implement contextual transcription rules at the search level in a trie-based beam search, the information concerning the contexts is attached to the trie leaf nodes along with the corresponding word. In the previous example, the leaf node corresponding to the end of the pronunciation /lEz/ would carry the word "*les*", the left and right contexts (*i.e.* (*) and (Vinit)) and the rule probability. Reaching a trie leaf node means that we have found a word with a given transcription rule (or, in other words, with a given contextual pronunciation variant). Word end hypotheses are therefore characterized by the word, the word end time and the pronunciation rule.

Let us denote the word end hypothesis associated with state S_w by (w, a, t), where a denotes a specific rule for word w. In addition to information on the previous word end time t', the back-pointers associated with the acoustic hypothesis which generated the word end hypothesis (w, a, t) also carry information concerning the previous word transcription rule number a'. When a word end is reached, one therefore has to decide whether the transition $(v, a', t') \rightarrow (w, a, t)$ is valid with respect to the contexts for rules a and a'. The transition is valid if and only if

1. the classes of word w match the right context for the

[2] In this context, and unless otherwise specified, the notion of word refers to an orthographic string. Homographs are therefore considered as a single word.

phonological rule a' of word v, and

2. the classes of word v match the left context for the phonological rule a of word w.

In a more formal manner, a word end hypothesis (w, a, t) is valid if the classes of w match the right context of the previous word end hypothesis (taken along the best path). For all valid word end hypotheses, the word level maximization equation for contextual transcription rules is given by

$$h_w(t) = \max_{v \in \mathcal{V}(w,a)} q_v(t, S_w) + \beta \ln P(v, w) + \ln P(w, a) \ , \tag{2}$$

where $\mathcal{V}(w, a)$ is the set of admissible previous words w.r.t. the left context of transcription rule a for word w. If a word end hypothesis is not valid w.r.t. its predecessor word, it is simply discarded along with the corresponding path.

The control of the right context of a rule is done in the search with a delay of one word. This is due to the trie organization of the pronunciations and is similar to what is done with the language model score. This similarity stresses the fact that contextual rules somehow act like a bigram model at the pronunciation level rather than at the word level.

4. Experiments

The proposed approach for integrating contextual phonological rules in our large vocabulary decoder was tested on a dictation task in French using the BREF corpus [9]. The BREF corpus is made of read sentences extracted from the journal *Le Monde*. A set of 41,000 sentences uttered by 80 speakers was used to estimate the parameters of a set of 40 monophone models with 3 states and 32 Gaussian components per state. The vocabulary contains the 20,000 most frequent words on the LM training corpus. A separate set of 300 sentences uttered by 20 different speakers is used for testing. The test set has a total of about 9,000 words. Three different set of contextual rules were defined.

4.1. Rule sets

The first set actually contains context free transcription rules: all transcription rules have the context (*) as their left and right contexts and always apply.

The second set of contextual rules correspond to the liaison and implements the following basic rule[3]:

> words whose graphic form ends with 'r', 's', 't', 'x', 'd' or 'n' (plus exceptions such as *beaucoup* or *franc*) have a pronunciation variant with a liaison iif followed by a word starting with a vocalic sound; otherwise the liaison is not realized.

Finally, the third set of contextual transcription rules was derived from MHATLex [3]. In MHATLex, a representation of the words at the phonological level is given with contextual phonological groups whose phonetic realization depends on the context. A set of rules has been derived from the MHATLex entries corresponding to our 20k word vocabulary. In MHATLex, lexical units are considered rather than words, *i.e.* homographs with different lexical categories are considered as separate entries, and different rules apply to homograph lexical units. Since

[3]This rule does not pretend to be an exact rule that treats all and solely the possible cases of liaison but it is rather general and accurate and is used as an illustration.

Table 2: Number of contextual pronunciation variants for each set of rules.

rule set	context free	liaison	mhatlex
# of rules	83,215	83,215	116,657

only words, *i.e.* orthographic forms, are considered in the recognition system, all the rules corresponding to the different lexical categories for that word are considered. A total of 20 lexical classes is used to defined the rule contexts, half of the classes carrying lexical information, the second half carrying morphological information.

For the three set of contextual transcription rules, the total number of rules is reported in table 2.

4.2. Results

For the three set of rules defined above, recognition rates, as well as error rates, are reported figure 2 in terms of substitutions, deletions and insertions. Results show that the introduction of contextual information for the transcription at the search level decreases the performances of our system.

For the MHATLex derived rules, this is not surprising since homograph lexical entries are confused into a single entry. This confusion generates conflicts between rules: two different rules apply at the same time on the same pronunciation for the same word and the system does not know which rule must be applied. This problem is discussed in section 5.

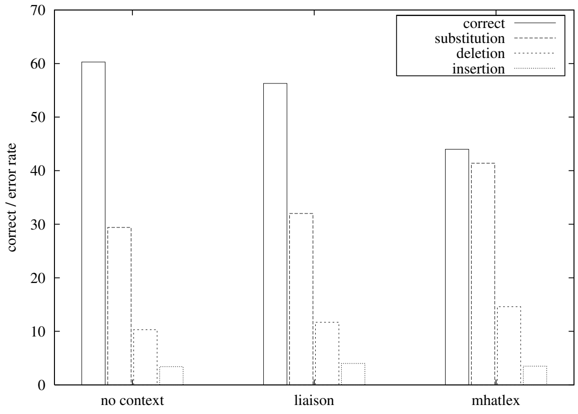
The results are more surprising for the contextual rules, since discarding impossible sequences of pronunciations from the search space should obviously avoid confusions and yield a better recognition rate. For example, the pronunciation /lEz bato/ of the word sequence "*les bateaux*" should no longer be considered during the search. Several factors may explain the decrease of recognition rate in this case. One of them is that the monophone acoustic models are not accurate enough to detect the realization of the liaison thus making the contextual rules inefficient. Another factor maybe the high number of pronunciation variants in our lexicon, which generates too many possible confusions. Finally, all those experiments were performed with a small search beam size (10,000 maximum active acoustic hypotheses) and a larger beam size could yield more positive results. However, we do not really believe that increasing the search space size is a solution since contextual transcription rules should limit by themselves the search space.

5. Discussion

Regardless of the results obtained, several points regarding the definition and use of rules are worth mentioning. The main issue here concerns the potential unconsistencies of the various linguistic resources involved in the search and the incurred loss of efficiency.

A first consistency violation occurs whenever two rules, having partially overlapping left and right contexts, apply to the same word and pronunciation variant. To illustrate the problem, let us assume the pronunciation p of word w is licenced by two different rules a_1 and a_2. During the search, upon reaching the end state of p, while both a_1 and a_2 may simultaneously apply (depending on the previous word v), only one transcription (the best one) will be "blindly" retained, and further developped. Now suppose that a_1 and a_2 express uncompatible restrictions

Figure 2: Recognition and error rates for various contextual phonological rules (context free, contextual liaisons and Mhatlex derived contextual rules).



on their right context, and that a_2 is selected: at the end of the next word, upon matching the right context of w, it might well appear that while a_1 was possible, a_2 is in fact forbidden, causing the deletion of a valid path.

A second type of unconsistency occurs when the various contexts where (the pronunciations of) w can be inserted actually forbid word sequences vw or wx. In this case, the language model will unduly reserve a probability mass for events which are in fact rendered impossible by the rule definition.

Even if it might well be possible to devise explicit sanity checks for those situations, we feel that these checks would not solve the more serious difficulties hiding behind these unconsistencies. One problem is very specific to our current search algorithm, which relies upon the so-called word pair approximation [7] for merging path, when our transcription rules impose second order dependencies between transcriptions: increasing the order of the language model used in the search should remedy to this problem. More serious is the issue of knowledge integration: traditional dictation systems define the notion of a word in terms of an orthographic sequence, which proves very convenient for defining and learning probabilistic models from large corpora. This definition appears overly simplistic for precisely modeling contextual phenomena like liaisons. These phenomena are not rare: for a given speaker, the process of selecting pronunciation variants for successive words is subject to a number of constraints which reflects specific elocution strategies.

The description of such constraints is however rarely attempted, except for local coarticulation effects occurring at word junctures: in all generality, this description would require to integrate a mixture of morpho-syntactic and phonological information at the word level. Our class-based system allows to integrate this kind of information, and more importantly, to effectively use it during the search process.

However, given our preliminary experiments, it is still unclear whether the incurred reduction of the search space would compensate for the induced multiplication of lexical entries. Moreover, given the available resources for French, especially the lack of large phonetically annotated corpora from which to learn and estimate transcription models, the description of large span inter-pronunciation dependencies has to rely on partial and sometimes unconsistent lexical descriptions.

6. Conclusions

We have presented an approach to introduce contextual transcription rules into a large vocabulary speech decoder. This approach was tested on a dictation task in French, demonstrating the effectiveness of our algorithm implementation. The results currently achieved are unconvincing, reflecting the need for more carefully designed resources. We however feel that the description of constraints on pronunciation sequences has to be integrated in the search algorithm. We are currently investigating the automatic induction of such constraints from data.

7. Acknowledgment

The authors express their gratitude to Arnaud Dauchy for his work on the estimation of the acoustic model on the BREF corpus.

This work has been done in the framework of the Sirocco project, partially funded by the Institut National de Recherche en Informatique et Automatique (INRIA) as a Cooperative Reseach Action.

8. References

[1] Helmer Strik and Catia Cucchiarni, “Modeling pronunciation variation for asr: overview and comparison of methods,” in *Modeling Pronunciation Variation for Automatic Speech Recognition Workshop*, 1998, pp. 137–144.

[2] *Workshop on Modeling Pronunciation Variation for Automatic Speech Recognition*, 1998.

[3] Guy Pérennou and Martine De Calmès, “MHATLex: Lexical resources for modelling the french pronunciation,” in *2nd International Conference on Language Resources and Evaluation*, 2000, vol. 1, pp. 257–264.

[4] Ingunn Amdal, Filipp Korkmazskiy, and Arun C. Surendran, “Data-driven pronunciation modelling for non-native speakers using association strength between phones,” in *Workshop on Automatic Speech Recognition – Chalenges for the new millenium*, 2000, pp. 85–90.

[5] N. Cremelie and J.P. Martens, “On the use of pronunciation rules for improved word recognition,” in *Eurospeech*, 1995, vol. 3, pp. 1747–1750.

[6] Florian Schiel, Andreas Kipp, and H. G. Tillman, “Statistical modeling of pronunciation: it’s not the model, it’s the data,” in *Modeling Pronunciation Variation for Automatic Speech Recognition Workshop*, 1998, pp. 131–136.

[7] Stefan Ortmanns and Hermann Ney, “A word graph algorithm for large vocabulary continuous speech recognition,” *Computer Speech and Language*, vol. 11, pp. 43–72, 1997.

[8] N. Deshmukh, A. Ganapathiraju, and J. Picone, “Hierarchical search for large-vocabulary conversational speech recognition,” *IEEE Signal Processing Magazine*, pp. 84–107, September 1999.

[9] Lori F. Lamel, J.-L. Gauvain, and M. Eskénazi, “BREF, a large vocabulary spoken corpus for French,” in *Eurospeech*, 1991, pp. 505–508.