DEALING WITH PRONUNCIATION VARIANTS AT THE LANGUAGE MODEL LEVEL FOR THE CONTINUOUS AUTOMATIC SPEECH RECOGNITION OF FRENCH

L. Pousse and G. Pérennou
IRIT
University Paul Sabatier
118 Route de Narbonne, 31062 TOULOUSE, France.
Tel. 33 561 55 61 73, FAX: 33 561 55 62 58, E-mail: pousse@irit.fr, perennou@irit.fr

ABSTRACT

In this paper, we describe three approaches of continuous speech recognition. Two of them (referred to as (W,P) and (W',P) models) take into account pronunciation variants of words. They allow to handle (very common) phonological french phenomena like liaisons or mute-e elision. The (W',P) model introduces the phonotypical level as defined in the MHAT Model [4,5]. Comparing (W,P) and (W',P) models show a significant improvement in recognition accuracy when a contextual language model is introduced at this phonotypical level.

1. INTRODUCTION

At the present time, automatic speech recognition systems pay attention to pronunciation variants [1], especially for French: many words could be affected by liaisons (insertion of a consonant at the end of a word) mute-e and consonant cluster reduction. Taking these phenomena into account during the training and the recognition processes is a crucial matter. Much researches has been done to introduce various pronunciations of a single word in the lexicon [1,2,3] and results have shown a significant reduction of error rate (by relatively approx. 20%) for French, but this solution is not optimal as it does not handle the context of words within the sentence.

In previous years our main focus was to model French phonological phenomena. This has led us to propose the MHAT model (Markovian Harmonic Adaptation and Transduction) [4,5]. Words are represented both through contextual phonological groups (cpg's) and through multiple pronunciation groups (mpg's) reflecting the variants owed to word context within the sentence and all the pronunciation variants of one, or more, contiguous phonemes realized by a group of speakers. Linguistic material has been developed (lexicons, phonological rules, boundary rules, pronunciation rules) allowing us to implement our phonological model. In order to deal with speech variation for French, we decided to integrate phonological contextual constraints at the level of the language model. And this led us to work out new methodologies for language model training and to adapt recognition strategies.

2. METHOD

2.1. (W) and (W,P) models

2.1.1 (W) model

The classical approach to speech recognition consists in deciding in favor of inflected word string $M^* = m^*_1...m^*_N$ satisfying:

$$M^* = argmax_M \{p(M).q(Y|M)\},$$

where $M=m_1...m_N$ denotes an inflected word string (the W level of MHAT). The probabilities $p(M)$ and $q(Y|M)$ are the contributions of respectively the language model and the acoustic model.

2.1.2 (W,P) model

In order to take into account the pronunciation variants we can introduce a model that includes a phonetic level P involving complex representations (M,U) = ($m_1...m_N$, $u_1...u_N$) where u_N is the pronunciation, at the P level, of m_N at the W level. Such a model is referred to as the (W,P) model.

The classical approach is modified as follows. We decide for $(M^*,U^*) = (m^*_1..m^*_N, u^*_1...u^*_N)$ satisfying:

$$(M^*, U^*) = argmax_{M,U} \{ p(M).r(U|M).q(Y|U)\}$$

(assuming that the acoustic representation Y depends only on its phonetic representation).
In this model the probabilities $p(M)$, $r(U|M)$ and $q(Y|U)$ are the contributions of respectively the language model, the phonetics model and acoustics one. The usual recognition algorithms of HMM-based systems can be applied if we assume that:

- a k-gram language. In the case of a bigram model:

$$p(M) = \prod_k b(m_{k-1}m_k)$$

where $b(m_{k-1}\, m_k) = p(m_k \mid m_{k-1})$

- a non-contextual phonetic model where

$$r(U \mid M) = \prod_k r(u_k \mid m_k)$$

The (W,P) model is illustrated in Figure 1. The inflected word «grande» (feminine form of French adjective for «great») in this example has three phonetic variants [gRãdə] , [gRãd] and [gRãn].

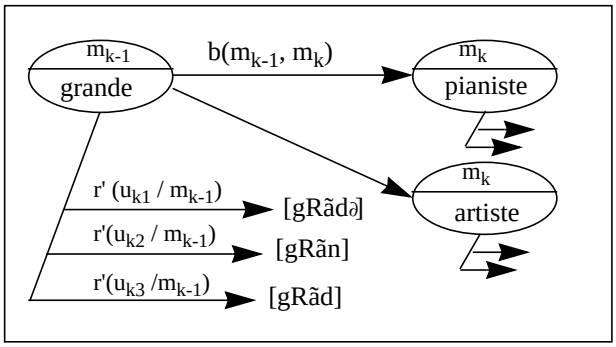

Figure 1: Illustration of an example of (W,P) model.

Previous work has proved that using a (W',P) model yields better recognition rates than using a (W) model [2,3,6].

2.2 Example of Sandhi rules for French

The main problem resulting from the (W, P) model is that it does not take into account word context.

As an example reflecting this problem for French, let us describe two very common phonological phenomena: latent consonant and mute-e.

2.2.1. An example for latent consonant

The orthographic word «grand» (masculine form for «great») can be realised either [gRã] or [gRãt] at the phonetic level. At the phonological level, it is assumed that «grand» cannot be pronounced [gRãt] when the liaison is forbidden. It means that «grand» has two contextual variants (called phonotypical variants):

- grand_1: «grand» / _ -liaison
- grand_2: «grand» / _ +liaison

«grand_1» (respectively «grand_2») is a phonotypical variant of «grand» in forbidden (respectively authorized) liaison context. In a more representative way, and using mpg's, «grand_1» could be written /gRã(t")/ with the following pronunciation rule:

t" --> ϕ | t

while «grand_2» would be /gRã/.

2.2.2. An example for mute-e

The inflected word «grande» (feminine form) is pronounced [gRãd] in liaison context, otherwise [gRãdə] or [gRãn]. It means that «grande» has two phonotypical variants:

grande_1: «grande» / _ -liaison

grande_2: «grande» / _ +liaison

«grande_1» (respectively «grande_2») is a phonotypical variant of «grande» in a forbidden liaison (respectively authorized liaison) right-hand context. In a more representative way, and using mpg's, «grande_1» could be written /gRã(~də)/ with the following pronounciation rule:

(də) --> [də] | [n]

while «grande_2» has only one pronunciation [gRãd].

These two examples show specific difficulties raised by Sandhi rules applicable to French, as these rules lead to major context-dependent modifications.

2.2.3 (W',P) model

We have proposed to take into account the effect of the context by introducing a W' level (called phonotypical level) of inflected contextual words. The decoding problem is now as follows:

Decide in favor of $(M'^*, U^*) = (m'^*_1 ... m'^*_N, u^*_1 ... u^*_N)$ satisfying:

$(M'^*, U^*) = \text{argmax}_{M',U} \{ p'(M') \,.\, r'(U|M') \,.\, q'(Y|U) \}$,

where $(M', U) = (m'_1 ... m'_N, u_1 ... u_N)$ consists of a string of phonotypical words and a string of pronunciations at level P. In this (W',P) model we assume that:

- $p'(M') = \prod_k b'(m'_{k-1} m'_k)$

(in the case of a bigram model of phonotypical words) where:

$$b'(m'_{k-1}\, m'_k) = p'(m'_k \mid m'_{k-1})$$

- $r'(U|M') = \prod_k r'(u_k|m'_k)$

(non-contextual phonetic model).

In this model the effect of context is taken into account by the language model defined by the phonotypical bigrams $b'(m'_{k-1}m'_k)$.

Figure 2 shows the modifications introduced by the (W',P) model on the example of Figure 1.

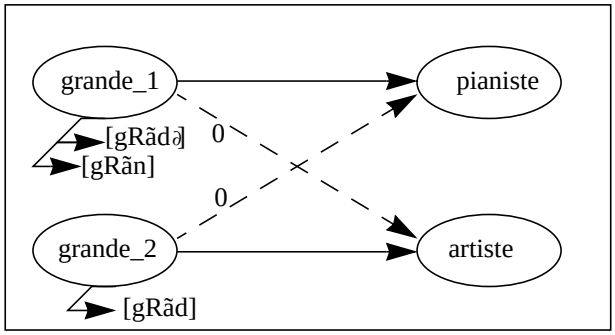

Figure 2: Example of phonotypical bigrams.

3. BUILDING A PHONOTYPICAL LANGUAGE MODEL

As a first step, we chose the method presented below (limited for the moment to bigram probabilities).
We can translate an orthographic corpus into phonotypical representations, and then compute phonotypical bigrams probabilities by taking into account the occurrence of each phonotypical bigram in the following formula:

$$b^*(m'_i\, m_k) = count\,(m'_i\, m_k) / count(m'_i).$$

The phonotypical translation was carried out successively applying two kinds of rules:

- boundary rules between two words (optional, obligatory or forbidden liaison) depending on their syntactic categories,
- contextual adaptation: giving a word its phonotypical representation, depending on boundaries as well as on its phonological context in the sentence.

In addition we worked out a lexicon of phonotypical representations of words and of their phonetic realisations found out by applying pronunciation rules.

4. EXPERIMENTAL RESULTS

The investigation of the different models was possible whithin the framework of the ARISE Project [7]. We used the Philips continuous speech recognizer [8] that we localized to French train-schedule information task.

4.1. Characteristics chosen for the recognizer

4.1.1 Acoustic parameters

Two different options were compared:

- either filter banks (14 channels (C) in the range of 350 to 3400 Hz + Δ C + energy (E) + Δ E + $\Delta\Delta$E)
- or 12 MFCC + the first 10 Δ MFCC.

The number of gaussians per state was set either to 16 or to 32 gaussians.

4.1.2 Number of phones

Because of the shortness of our corpus, we did not use triphones. On the one hand, we use only monophones (38 of them, including extralinguistic sounds like F, oral expiration, or x, something produced by the speaker but unrecognized as a phoneme) and, on the other hand, we used the 39 phones making a distinction between two contextual allophones of [R] --one of them been realized in the unvoiced, obstruent left-hand context.

4.2. Corpora

The corpus used in our investigation was collected as part of the European ARISE Project [7].
It is 4 hours long and contains 4539 utterances. It was divided into a training set of 4266 sentences and a test corpus containing 273 utterances.
For the all these tests we were careful to select only sentences that were both linguistically and pragmatically relevant to the task.

4.3. Lexica

All our lexica were built from the BDLEX lexicon [9].
During the training of acoustic parameters, the lexicon covers all the words contained in the training corpus (no out-of-vocabulary word). Words may have different pronunciations (average of 2.2 variants per inflected word). It contains 2048 entries (i.e., 921 different inflected words).
During the recognition process, the entries in the lexicon depend on the representation of the language model:

- with a (W, P) level language model, the lexicon contains 1914 entries (i.e., 895 different orthographic words).
- with a language model at the (W',P) level , the lexicon contains 4834 entries (i.e., 2512 different phonotypical words or 895 different inflected words).

The same inflected words are present in both the training lexicon and the recognition lexica. But in the recognition lexica, we compacted some expressions into a single inflected entry (for example: je_voudrais (*I would like)*).

5. RESULTS AND DISCUSSION

Results are presented in Table 1.
The word error rates presented in Table 1 are given by the sum of the number of insertions, substitutions and deletions (100%-accuracy rate).

It can be seen that the best results still show a 20.4% error rate which, of course, seems high. However, this unfavorable impression should be dispelled for we are dealing with speaker-independent recognition of spontaneous telephonic speech.

Furthermore, the amount of training data used is not such as to allow for a triphone-based model.

inflected words' language model (W,P)				phonotypical language model (W',P)				
# gaussian	# phones	MFCC	word error rate	# gaussian	# phones	MFCC	word error rate	gain
16	38	no	28.2	16	38	no	24.8	-12.0%
16	38	yes	23.6	16	38	yes	21.6	-8.5%
16	39	no	26.7	16	39	no	26.0	-2.6%
16	39	yes	23.4	16	39	yes	22.5	-3.9%
32	38	no	23.6	32	38	no	21.7	-8.1%
32	38	yes	23.5	32	38	yes	21.8	-7.2%
32	39	no	25.4	32	39	no	24.6	-3.2%
32	39	yes	22.8	32	39	yes	20.4	-10.5%

Table 1: comparative results using an orthographic language model and a phonotypical language model.

It can be seen, that independently from the acoustic and phonetic options, successively tried, results turned out better with the (W',P) model (relatively 7% better).
As could be expected, the best results were obtained when using 32 gaussians per state. Similarly, using MFCCs makes for better results than when using filter banks.

Finally, the use of 39 monophones (with 2 contextual allophones of [R]) generally yields better results than 38.
To sum it up, the best results are obtained by using 32 gaussians per state, MFCCs, 39 phones and the (W',P) model (20.4% error rate with words).

6. CONLUSION

We have investigated the contribution to ASR made by a phonotypical language model in opposition to a standard orthographic language model.

The experiments reported in this article confirm for French the importance of the phonological context of words which is the essential feature of this model. We note that making use of the MHAT model with its formalism --which is compatible with HMM-based systems-- has enabled us to experiment under conditions of real oral dialogues.

There remains, however, the drawback that this context renders both lexicon and language model more complex. Therefore, we are at present striving to find means of simplifying the (W',P) model so as to make it easier to handle.

ACKNOWLEDGEMENT

The authors wish to thank Prof. J.F. Malet, CSU Sacramento, for assisting in the translation of their French manuscript, and Dipl.-Ing. F. Seide, Philips, for his help and his suggestions.

REFERENCES

[1] E.P. Giachin, A.E. Rosenberg, L. Chin-Hui, «Word Juncture Modelling Using Phonological Rules for HMM-Based Continuous Speech Recognition», in *Computer Speech and Language*, pp155-168, 1991.

[2] X. Aubert, Ch. Dugast, «Improved Acoustic-Modelling in Philips' Dictation System by Handling Liaisons and Multiple Pronunciations», *Proc. EUROSPEECH' 95*, pp.767-770, Madrid, 1995.

[3] L.F. Lamel, G. Adda, «On designing Pronunciation Lexicons for Large Vocabulary, Continuous Speech Recognition», *Proc. ICSLP'96*, Philadelphia, 1996.

[4] G. Pérennou, «Phonological Component In Automatic Speech Recognition. The Case Of Liaison Processing», in *Speech Communication: Relations and Interactions*, C. Sorin et al. (Editors), pp. 211-223, 1995.

[5] G. Pérennou, «Les règles et les niveaux en phonologie: du générativisme aux modèles markoviens», in *Fondements et Perspectives en Traitement Automatique de la Parole*, AUPELF-UREF, pp. 185-203, 1996.

[6] L. Pousse, «Introduction d'une composante phonologique à la reconnaissance automatique de la parole continue», in *Proc. JEP' 96*, pp.70-74, Avignon, 1996.

[7] ARISE: «Automatic Railway Information Systems for Europe», LE3 - 4229.

[8] H. Aust and M. Oerder, «A realtime prototype of an automatic inquiry system», in *International Conference on Spoken Language Processing*, Vol.2, pp. 703-706, 1994.

[9] G. Pérennou, M. de Calmès, D. Cotto, I. Ferrané, J.M. Pécatte, «Le projet BDLEX de Base de Données Lexicales», *Actes du Séminaire Lexique*, CHM-Pôle Parole et Language Naturel, pp. 153-71, Toulouse, 1992.