6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

LANGUAGE MODEL LEVEL VS. LEXICAL LEVEL FOR MODELING PRONUNCIATION VARIATION IN A FRENCH CSR

Laure BRIEUSSEL-POUSSE *Guy PERENNOU*

IRIT
University Paul Sabatier - 118, Route de Narbonne
F-31062 TOULOUSE CEDEX - FRANCE
{pousse, perennou}@irit.fr

ABSTRACT

In this paper, we present a markovian system, which takes into account pronunciation variation of French. After doing a brief overview of different methods allowing to deal with pronunciation variation in ASR, we describe our approach (which is based on the MHAT (Markovian Harmonic Adaptation and Transduction) model), as well as the lexical and phonological materials defined in order to implement MHAT into a classic ASR system based on HMM models. We finally compare two approaches (both issue from the MHAT model) that differ each other by the level of pronunciation modeling : at the lexicon level and at the language model level (by introducing an intermediate level of words representations depending on the context of words in the sentence). Results show an improvement of French continuous speech recognition when taking into account the context of words in the sentence within the language model.

INTRODUCTION

Automatic speech recognizers can now deal with continuous speech in the framework of very large vocabulary applications. In order to refine the recognition, many researches have been done over the last past years to make the recognizers efficient independently from users. Each user has its own way to pronounce words and even a same user may not pronounce twice the same way the same utterance. That is why efforts are made to model pronunciation variations within the ASR systems. In [1], Strik and Cucchiarini make an overview of the most important characteristics that distinguish the various studies on pronunciation variation modeling, depending on how they answer these questions :

- What type of pronunciation variation is modeled (intra-word variation and/or variation between words) ?
- Where should the information on variation come from (data-driven methods vs. knowledge-based studies) and should the information be formalized or not ?
- In which component of the ASR should variation be modeled ?

In this paper, we will try to answer those questions by presenting our ASR system for French. Its fundaments lie on our theoretical model MHAT (Markovian Harmonic Adaptation and Transduction) which has been largely presented in [2] or [3].
The experimentation will try to answer the questions 2 and 3 for handling variations within and between words.

1. PRONUNCIATION VARIATIONS

Usually, studies on pronunciation variation consider that the variation is internal to a word. Then, the way to model the variation is to introduce in the lexicon several entries per orthographic word (one entry per pronunciation) [4] or a single meta-representation reflecting the word variation [5]. It is probably the best way to model intra-word variation, which is the main problem for a lot of languages. French exhibits also important cross word variations (Sandhi rules). Among them the most important are:

1. Liaison (e.g. *deux heures* -> /døʒœrə/)
2. Schwa deletion at the ending of a word (e.g. *me rendre à* -> /mərɑ̃dʀa/)
3. Liquid deletion (e.g. *votre billet* -> /vɔtbije/)
4. Nasal assimilation of stop consonants

These phenomena are all induced by the context of the words in the sentence. Pronunciation variation between words for French should therefore be considered in a continuous ASR.

2. SOURCE OF KNOWLEDGE

As far as how the variation information is obtained, it is clear that the choice depends principally from the amount of transcribed speech data that we can use or not. On one hand, if a very large database is available, a data-driven method will be chosen. Pronunciation lexicons are obtained by measuring the frequencies of pronunciations in the corpus and by applying a threshold to select only the most frequent pronunciations of each word. On the other hand, if phonological rules exist, why not using them to generate pronunciation materials ? We decided to develop our pronunciation lexicon by applying phonological rules on the vocabulary of the application. These French rules were obtained in the framework of BDLEX and MHATLex [6].

3. LEVEL OF PRONUNCIATION REPRESENTATION

Three components of an ASR may be involved in the process of pronunciation handling : the acoustic level, the lexicon level and the language model level.

3.1 Acoustic level

One way to improve ASR is to generate better acoustic-phonetic units during the training process. Using only one pronunciation representation of each word will lead to a bad acoustic-phonetic alignment if the word is not pronounced in the standard way in the training speech corpus. But handling multiple pronunciation during the training process involves the phonetic transcription of all the training material, which can be a long task. As many other researches teams, we decided to generate automatically the phonetic transcription of our training corpus, by using the TAPES system.

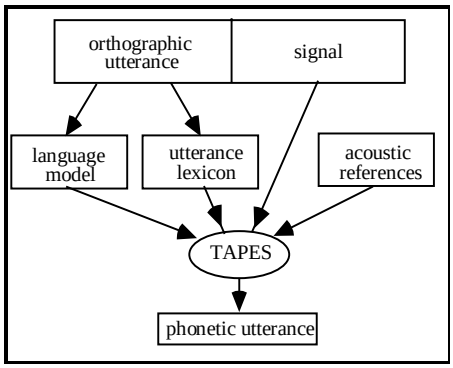

Fig. 1 - The TAPES system

At the beginning, TAPES was first used with acoustic references obtained on a bootstrapping corpus manually phonetically transcribed.

From the orthographic utterance, TAPES generates a language model and a lexicon containing all the possible pronunciations for each word (the pronunciations present in MHATLex).

Then a recognition module is used: giving a very high weight to the language model, the recognition system is obliged to choose the correct orthographic form, and the alignment with the signal forces the recognition system to choose the closest pronunciation. An option is added to take into account (within the lexicon and in the language model) extra-linguistic realizations like hesitations, breath of the speaker or room noise.

3.2. Lexicon level

In the classical approach of ASR, inflected words are represented through a dictionary Dic(W,P) where a lexical entry (at the word level W of the MHAT model) is associated to one or several pronunciations represented at the phonetic level P. HMM recognizers perform a Viterbi search.
Let $W=w_1...w_L$, $U=u_1...u_L$ and $Y=y_1...y_T$ be respectively a word string at level W, its realization at level P and the input speech signal of the recognizer.
Consequently, the recognizer decides in favor of

$$W^*=argmax_W\{P(W,Y)\}$$

where

$$P(W,Y)=P(W)\ Max_U\ \{ß(W,U)P(Y/U)\}$$

and $ß(W,U)$ $=1$ if U is a pronunciation of W,
$=0$ else.

3.3 Language model level

When using the (W,P) model, i.e. when introducing pronunciation variation at the lexicon level, we do not take into account the context of words in the sentence. A solution consists in using multi-words [7] at the lexicon level. Another solution is to model pronunciation variation at the language model level either by using directly the variants themselves to calculate the N-grams of the language model, or by introducing an intermediate level of representation of words depending on the context.
Due to practical considerations, we chose to develop the latest possibility. We have introduced a "phonotypical" level W'. Each word at the W level

may have one or more W' representation depending on its context.
For example, the word "les" has two W'-entries:
- les$_1$ pronounced [lɛʒ] in bigrams as (les$_1$ amis)
- les$_2$ pronounced [lɛ] in bigrams as (les$_2$ frères).

One W'-representation may have more than one P-representation. For exempla "notre" has two entries at the level W':
- notre$_1$ pronounced [nɔtʀ] in bigrams as (notre$_1$ ami)
- notre$_2$ pronounced [nɔt(ʀə)] in bigrams as (notre$_2$ frère).

The W'-representation notre$_2$ stands for 2 P-representations : either [nɔt] or [nɔtʀ6].

The language model of a (W',P) recognizer applies on W' representations. Therefore, in a classical HMM recognizer, the Viterbi search decides in favor of:

$$W'^* = argmax_{W'}\{P(W',Y)\}$$

where

$$P(W',Y) = P(W') \; Max_U \{ß(W',U)P(Y/U)\}$$

and $ß(W',U)=1$ if U is a pronunciation of W',
$=0$ else.

4. LEXICAL vs. LANGUAGE MODEL LEVEL

4.1 The Experimental Framework

4.1.1 Acoustic features
The following parameter setting has been adopted:
- sampling rate : 8kHz
- Windowing: frame-width 25 ms, shift: 10 ms
- either Filter bank: 14C+14ΔC+E+ΔE+ΔΔE
  or: 12 MFCC, 10ΔMFCC

4.1.2. Dictionary of phonetic units
The dictionary of phonetic units includes 38 monophones, modeled as Bakis HMM including three double-states with mixture gaussian densities.

4.1.3. Speech data
We use telephonic spontaneous speech recorded through the dialogue demonstrator developed in IRIT within ARISE project [8].

The Apr97 corpus used includes:
- 530 calls;
- amount of speech 4h;
- number of words: 23261 (665 distinct words);

The May98 corpus used includes:
- 1007 calls;
- amount of speech: 7h30;
- number of words: 42930 (924 distinct words);

The corpora have been transcribed manually at standard orthographic level and automatically (TAPES) at phonetic level.

4.1.4. Training corpus for LM
31000 words / 7200 sentences from Apr97 corpus.

4.1.5. Test corpus
The first test corpus used in the reported experiences includes 1079 words and 273 sentences from Apr97 corpus.
The test2 corpus includes 662 words and 283 sentences from the last corpus we collected during the first semester of 1999.

4.1.6. Lexicon
We have adapted our lexical materials (among them BDLEX and MHATLex which correspond to the (W',P) level for 500000 inflected words) and create a new lexicon and its environment.
From MHATLex completed by the vocabulary of the timetable task of our application within the ARISE project, we have derived:
- (W,P) model: 895 W-entries / 1914 P-variants
- (W',P) model: 2512 W'-entries / 4834 P-variants

4.2. Results

The goal of this experiment was to determine whether the use of context-free pronunciations was better than the use of context-dependent pronunciations.
Results are presented in the following table as word error rates using the following formula :
*WER=((#insertions+#deletions+#substitutions) *100/#words).*

Corpus	Apr97 (4h)				May98 (7h30)	
Training →	filter bank		MFCC		MFCC	
Test ↓	16 mixt. of gauss.	32 mixt. of gauss.	16 mixt. of gauss.	32 mixt. of gauss.	16 mixt. of gauss.	32 mixt. of gauss.
(W,P)	17.6%	15.8%	15.2%	13.3%	14.1%	13.2%
(W',P)	15.8%	15.1%	14.1%	12.8%	12.7%	12.2%
gain	10.2%	4.4%	7.2%	3.7%	9.9%	7.6%

Table 1 - Comparison between context-free (W,P) and context-dependent (W',P) recognizer using the test1 corpus

Training corpus	May98 (7h30)
Parameters	32 mixture of gaussians ; MFCC
(W,P)	**15.8%**
(W',P)	**13.9%**
gain	**12%**

Table 2 - Comparison between context-free (W,P) and context-dependent (W',P) recognizer using the test2 corpus

The results suggest that the use of context-dependent pronunciations gives better results than context-free pronunciations for French. As expected, the longest the training corpus, the best recognition ; and the longest the training corpus, the most important gain due to context-dependent pronunciations.
We could fear that the gain obtained with the use of context-dependent pronunciations within the language model was a lucky consequence of the way the test corpus was built. In table 2, we can see that the use of another test set yields to the same improvement.

CONCLUSION

In French, crossword pronunciation variations are very frequent. The comparison between a recognizer taking into account phonological phenomena at the language model level and a recognizer taking into account only intra-word variations at the lexicon level shows that it is necessary to handle the context of words during the recognition process.
It would be interesting to see whether our method is still efficient with other languages in which the juncture phenomena are less important than in French.

ACKNOWLEDGMENTS

The authors acknowledge the invaluable contribution of Martine de Calmès for the preparation of training environment and *Philips Research Labs Aachen* for providing the Dialogue workstation used during ARISE Project and for the recognition tests.

REFERENCES

[1] H. Strik, C. Cucchiarini, *"Modeling Pronunciation Variation for ASR : Overview and Comparison of Methods"* in Proc. of the Workshop on Modeling Pronunciation Variation for Automatic Speech Recognition, Rolduc, pp. 137-144, 1998

[2] G. Pérennou, L. Brieussel-Pousse, *"Phonological component in automatic speech recognition"*, in Proc. of the Workshop on Modeling Pronunciation Variation for Automatic Speech Recognition, Rolduc, pp. 91-96, 1998

[3] G. Pérennou, *"Les règles et les niveaux en phonologie: du générativisme aux modèles markoviens"* in Fondements et Perspectives en traitement automatique de la parole (Ed. H. Méloni), Universités Francophones, HACHETTE or ELLIPSE, pp. 185-204, 1996

[4] J. Ferreiros, J. Mucias-Guarasa, J.M. Pardo, L. Villarubia, *"Introducing multiple pronunciations in Spanish speech recognition systems"* in Proc. of the Workshop on Modeling Pronunciation Variation for Automatic Speech Recognition, Rolduc, pp. 29-34, 1998

[5] L. Lamel, G. Adda, *"On designing pronunciation lexicons for large vocabulary, continuous speech recognition"*, in Proc. ICSLP'96, Philadelphia, 1996.

[6] de Calmès M, Pérennou G, *"BDLEX: A Lexicon for Spoken and Written French"*, 1st Int. Conf. on Language Resources & Evaluation, pp.1129-36, May1998

[7] M. Wester, J. Kessens, H. Strik, *"Improving the performance of a Dutch CSR by modeling pronunciation variation"*, in Proc. of the Workshop on Modeling Pronunciation Variation for Automatic Speech Recognition, Rolduc, pp. 145-150, 1998

[8] G. Pérennou, M. de Calmès, C.A.Lavelle, "*Documentation and motivation for the final technological and operational improvements for DEMON"* Deliverable D42221, ARISE LE3-4229, Dec. 1998