A METHOD FOR MODELING LIAISON IN A SPEECH RECOGNITION SYSTEM FOR FRENCH

*L. Bahl, S. De Gennaro, P. de Souza, E. Epstein, J.M. Le Roux, B. Lewis, C. Waast**

T.J. Watson Research Center,

P.O. Box 218, Yorktown Heights, NY 10598, USA

*IBM Speech Systems, European Speech Research,

68, 76 Quai de la rapée 75012 Paris, France

ABSTRACT

In French the pronunciations of many words change dramatically depending on the word immediately preceding it. The result of this phenomenon, known as “liaison”, in an ASR system that does not model “liaison” is the requirement of unnatural pronunciation and much user dissatisfaction. We present, in this paper, the development of an acoustic model which takes into account the wide variability of word pronunciations caused by the liaison, the integration of this model into a French continuous speech recognition system and decoding results.

1. INTRODUCTION

In some languages, such as French, a phenomenon occurs at word junctions that can cause significant changes in pronunciation. This phenomenon, generally referred to as “liaison”, allows the optional insertion of a phone between some pairs of words [10]. For example, in the French phrase “les deux à la fois", a Z phone can be optionally inserted between “deux”' and “à”. The word sequence “deux à” can be pronounced as either /D AX AA/ or /D AX Z AA/ (The phones are represented in the widely used ARPABET). In our terminology we say that “deux” generates a Z-liaison and that “à” accepts a liaison. In many contexts, the speaker is free to either pronounce the liaison or not and is also free to either pause between liaisonable words or not. There is, however, a natural tendency on the part of most speakers to carry the liaison over into the next word even if a pause is made between the words. Thus, a speaker is more likely to pronounce “deux à” as /D AX silence Z AA/ rather than /D AX Z silence AA/, i.e. insert the /Z/ after the silence rather than before the silence. In our early discrete speech recognition systems for French, no liaison was allowed and speakers were expected to pronounce all words without any liaison. The inconvenience of being forced to omit liaison was one of the most frequent complaints from users. For continuous speech recognition in French, modeling liaison is a necessity.

2. METHOD FOR MODELING LIAISON

As shown in Table 1, French has 6 different phones that can be generated when a liaison occurs. The word endings causing these phones to be inserted are also shown in Table 1. The P and G-liaisons are rare. Liaison strongly depends on a complex interaction involving orthography, syntax, semantics and other factor [7][8][9]. Constructing a complete set of formal rules for liaison is very difficult as there are many exceptions and dialectal variations and quite often the usage contradicts the formal rules. However, as a first approximation we can say that:

If a word's spelling ends with a non-pronounced consonant c, where c ∈ {(s, x, z), n, r, (t, d), p, g}, and the following word starts with a vowel-like sound then a liaison phone L can be inserted in front of the second word, where L ∈ {Z, N, R, T, P, G}

Ending consonant	Generated phones	Example
s, x, z	/Z/	Mes (Z) amis
n	/N/	Un (N) ami
r	/R/	Premier (R) ami
t, d	/T/	Petit (T) ami
p	/P/	trop (P) amis
g	/G/	Long(G) et difficile

Table 1: Phones generated when a liaison occurs.

Even to this rather simple rule, we must add some additional rules to handle words beginning with the letter “h”. Word initial letter “h” is always silent, and all words beginning with “h” have a vowel sound at the beginning. However, some words starting with “h” accept liaison, others don't. There are no simple rules for this and dictionaries explicitly identify whether a word accepts liaison or not. Liaison can be handled by adding extra phonetic transcriptions in the pronunciation dictionary. There are two obvious ways to do this. The first method would be to create extra phonetic transcriptions for the words that accept liaison by inserting each liaison phone in front of the normal phonetic transcription. In this case, for each liaison acceptor word, we have to create 6 new phonetic transcriptions. This substantially increases the size of the pronunciation dictionary. A second method would be to create extra phonetic transcriptions for liaison generator words by inserting the generated liaison phone at the end of the current phonetic transcription. In this case, for each liaison generator word, we have to create one new phonetic transcription. This solution increases the size of the dictionary much less but doesn't take into account the natural tendency to carry the liaison over into the next word. Our solution consists of encoding liaison information into the pronunciation dictionary without adding new phonetic transcriptions. Two types of liaison

information are identified for each word. Consequently, two flags are set for each word indicating the presence or absence of the two liaison characteristics. The first is a “generating liaison” flag, which specifies whether or not the word generates a liaison, and if so, which of the 6 phones is generated. The second is an “accepting liaison” flag, which specifies whether or not the word accepts a liaison at its beginning. The “accepting liaison” flag allows the handling of exception words which start with a vowel-like sound but do not accept liaison.

3. SYSTEM DESCRIPTION

Essential aspects of the system used in the experiments here have been described earlier [2][3]. We summarize below the important elements of the system and point out the enhancements that have been introduced for French. The feature extraction technique uses differences in the cepstral vectors between frames to model the dynamics of speech as described in [6]. The system uses different acoustic models for sub-phonetic units in different contexts. The context dependent classes are identified by growing a decision tree from available training data [3] and the acoustic vectors that characterize the training data at the leaves are modeled by a mixture of gaussian densities with diagonal covariance matrices. The HMM's used to model the leaves are simple 1-state models, with a self-loop and forward transition. The training procedure assumes that we have an initial speaker independent training that can be used to bootstrap the procedure. We begin by making, for each training utterance, the particular pronunciation of each word and also the presence of silence between words. The training data is then aligned against these scripts using a Viterbi algorithm, giving us the class label (transition within the phonetic HMM to which the vector is aligned) to each feature vector and the phonetic context. Decision networks are then constructed using the training data, one for each class label. The training data at each terminal node is then used to determine a mixture of gaussian densities with diagonal covariance. This is done by first clustering the feature vectors and refining the models using the forward-backward training algorithm. The number of mixture components is variable and an upper bound is often imposed. The language model is a mixture of a word trigram model and a class based trigram model [5]. The classes are only related to nouns, adjectives, verbs and adverbs. The words belonging to two different classes, like “adresses” (address) that can be either a noun or a verb, are merges into a same class: noun-verb. Each spelling is associated to only one class and consequently no tagging is needed when computing class sequence probabilities. For our continuous speech recognition system, we currently use 610 classes. The mixture of word and class based trigram models provides not only a better accuracy but also more “acceptable errors”: the class based trigram model frequently corrects grammatical errors especially on poorly trained words (most of singular and plural are homophones in French, number of conjugated verbs too).

4. LIAISONS IN THE RECOGNITION PROCESS

Liaison cannot be imposed as a requirement upon the speaker, since the use of liaison in the language is optional according to the individual's speaking [1][4]. Thus, in a speech recognition system, liaison must be optional and the recognizer must be guided by the acoustic signal to decide whether liaison was or was not used in each instance. In our current system two acoustic matches are performed: a fast match and a detailed match. The fast match filters the total vocabulary down to a short list of reasonable candidate words based on a simplified acoustic model. The language model is used to further prune this list. Detailed matches are performed for the words in the pruned list. In order to reduce the computational cost and to obtain a fast response time, we use a “megaphone” to model liaisons at the fast match level. The “megaphone” is defined to be the phones /Z/ /N/ /R/ /T/ /P/ /G/ in parallel. Since liaison is optional a null transition is also added in parallel Figure 1. When computing fast match score, every liaison-accepting word is prefixed with the “megaphone”. The “megaphone” probability is computed as follows:

$$\text{Prob (megaphone)} = \max_{L \in \{Z, N, R, T, P, G, \text{null}\}} \text{prob}(L)$$

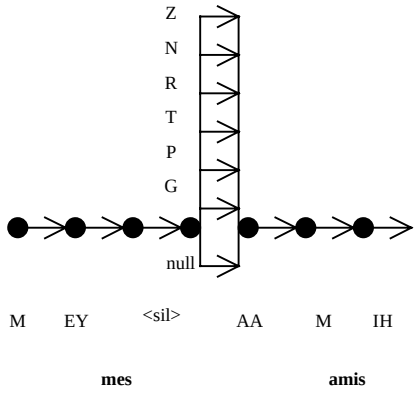

Figure 1: The “megaphone” for the fast match.

Thus, during the fast match, it is assumed that all liaison phones can occur, independent of what liaison phones can actually be generated by the previous word. This situation is remedied in the detailed match, where the exact acoustic match for a word sequence is computed. As shown in Figure 2, in the detailed match, if the current candidate word accepts liaison we determine whether the preceding word is a liaison generator. If so, the “megaphone” is replaced with only the appropriate liaison phone in parallel with a null transition. If the current word does not accept liaison, then the “megaphone” is removed.

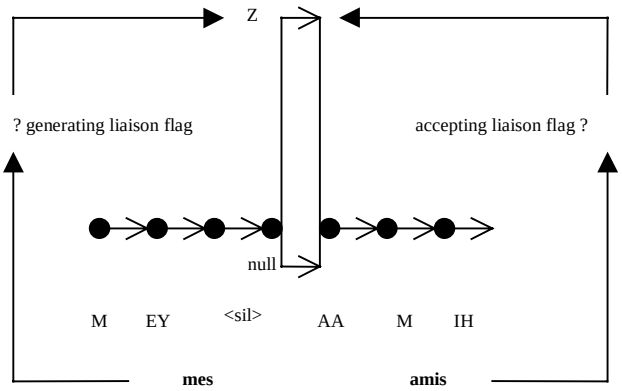

Figure 2: Liaison modeling during the detailed match.

5. EXPERIMENTAL RESULTS

The flags have been determined by the use of a set of 300 ordered rules which involve conditions on the left and right spelling context, grammatical condition (Gram) and phonetic condition (Pho). The first row of the Table 2 describes a rule for the adverbs ending by the letter "z". In this case the phonetic condition, the right and the left contexts are empty, the grammatical condition is ADV for adverb. The phonetic condition is used when we don't want to generate a liaison if the phonetic transcription of the current word already ends by a liaison sound. For example, the second row of the Table 2 describes a rule for the words ending by the letters "er". This rule generates an R-liaison-flag except when the phonetic transcription of the word ends by an R sound. According to this rule the word "inter" doesn't generate a liaison.

Ending consonant	Gram	Pho	Flag	Example
z	ADV			Assez
er		/R/	/R/	Premier
er	SUBS			Papier

Table 2: Example of generating-liaison-flags rules.

To handle the liaisons during training, first we tagged our training data to determine where a liaison can be pronounced. This was done automatically using the set of 300 ordered rules. Each word that can accept a liaison has been replaced into the training scripts by a liaisonable word, which is the current word annotated by the left context liaison. For example, the word sequence "mes amis" (my friends) is replaced by "mes Z-amis". The liaisonable words have two different sets of phonetic transcriptions, one without liaison and one with the appropriate liaison phone on the front of its phonetic transcriptions. "Z-amis" can be pronounce either /AA M IH/ or /Z AA M IH/. Then, at the beginning of the training procedure, when making for each utterance, the particular pronunciation of each word, the appropriate pronunciation for these liaisonable words will be selected according to what the speaker really said. When the selected pronunciation contains a liaison phone one can say that the word has been liaisoned. The Table 3 shows the average number (in percent) of liaisonable and liaisoned words found into two different corpora. The corpus "PhoR" contains 6,500 sentences of phonetically rich sentences extracted from the SpeechDat data base [11]. The "PhoR" sentences have been recorded by a set of 990 different speakers. The corpus "SonrP" contains 20,000 sentences randomly extracted from a large corpus of general business domain. No constraint on phone coverage has been applied to select the 20,000 sentences. The "SonrP" sentences have been recorded by a set of 1,130 different speakers. The Table 4 shows the average number (in percent) of sentences of those corpora containing 0, 1, 2, ... liaisonable (and liaisoned) words.

Corpora	SonrP	PhoR
% Li-able words	6.77	6.80
% Li-ed words	2.40	2.55
% Li-ed / Li-able	35.55	37.60

Table 3: Liaisonable and liaisoned words distribution.

	SonrP		PhoR	
	Li-able	Li-ed	Li-able	Li-ed
X	M	M	M	M
0	33.90	66.29	39.62	69.04
1	32.77	26.54	33.69	24.73
2	20.49	6.11	16.99	5.35
3	8.40	0.95	6.24	0.70
4	3.18	0.08	2.37	0.15
5	0.94	0.01	0.70	0.01
6	0.24	0.00	0.30	0.00
7	0.06	0.00	0.04	0.00
8	0.00	0.00	0.01	0.00

Table 4: liaison (able/ed) words frequency.

X = number of liaison (able/ed) words per sentence, M = percentage of sentences containing x liaison (able/ed) words over the total number of sentences.

According to Table 3 and Table 4, there is no big differences on liaison frequency distribution over the two corpora even if they have been build in a very different way. Only a few words are liaisonable, and even fewer are liaisoned. But almost one sentence out of three contains at least one liaisoned word. This number is highly speaker-dependent. The Table 5 shows the more frequent and the more frequently pronounced liaisons. According to the corpus, 36 to 38%, of the predicted liaisons are pronounced. The most frequent liaison is Z; followed by N, T, and R. The best predicted liaison is the N-liaison: 61 to 64% of the N-predicted-liaisons are pronounced when only 34 to 37% of the Z-predicted-liaisons are pronounced. One could improve the set of rules in order to make a better prediction of the liaisons by adding more specific rules. But as the liaison phenomenon is speaker dependent, the best could be to adapt liaison prediction to each speaker.

	SonrP		PhoR	
Liaison	**Li-able**	**Li-ed**	**Li-able**	**Li-ed**
Z	55.89	20.50	54.19	18.51
N	8.55	5.21	10.98	7.05
R	6.01	0.49	5.55	0.60
T	29.45	9.30	28.97	11.35

Table 5: Percentage of each type of liaison over the total number of liaisons.

We carried out the experiments on a large vocabulary task (65k words general business task vocabulary). The training data consisted of 160,000 sentences from 900 French speakers. A total of 30,000 gaussians were used. Each context dependent HMMs was modeled with a maximum of 14 gaussians. During decoding, the liaisons are handled as described in the previous sections. The test data consisted of 921 words per speaker from 10 speakers. 0.33% of the test words are out of the vocabulary. Ignoring the out of vocabulary words, the test perplexity was about 300. No instructions were given to the speakers. They were free to either pronounce the liaison or not. 8.4% of the test script words are liaisonable. The performance of the French continuous speech recognition system was evaluated both with and without liaison modeling. We used 100 sentences to adapt the acoustic models to each test speaker.

Handled liaison	**WER**	**Gain**	**RT**
No liaison modeling	9.51	00.00	100.00
Z liaison modeling	6.27	34.06	99.33
Z, N liaison modeling	6.08	36.06	100.11
Z, N, R liaison modeling	6.03	36.59	100.22
Z, N, R, T liaison modeling	5.90	37.96	100.99

Table 6: Percentage of WER, gain, normalized real time ratio (RT) with and without liaison modeling.

The Table 6 shows the word error rate (WER) when the test script is decoded using no liaison modeling, just a Z-liaison modeling, a Z- and N-liaison modeling, etc… There is a significant gain in performance if the acoustic models are adjusted to properly model liaison without any impact in speed. The test script contains in average 28 liaisoned words. The gain in error rate when liaisons are modeled is 38%, meaning 34 corrections of word errors. Consequently, modeling liaisons not only improves the recognition rate on liaisoned words, but avoids also errors on the neighborhood of liaisoned words. This method was first designed for French liaison, but could be apply to other strong co-articulation phenomenon, in French (for example, to handle multiple pronunciations of the word "y") but also in other languages. In English, "you or I" can be pronounced /Y UW W AO R AY/ (W-liaison and R-liaison), "the other" can be pronounced /DH IY Y AH DH AX/ (Y-liaison), "an apple" can be pronounced /AE silence N AE P AX L/ (silence shifts to non-word boundary). However, as shown in [5] for discrete speech, this method introduces phonetic ambiguities that can increase the number of errors if the speaker doesn't pronounce any liaison (which is almost impossible in French). A tradeoff needs to be found between the gain due to the strong co-articulation modeling and the loss due to the introduction of phonetic ambiguity.

6. REFERENCES

1. X. Aubert, C. Dugast "Improved Acoustic-Phonetic Modeling in Philip's Dictation System by Handling Liaisons and Multiple" Proceedings of the EUROSPEECH, pp. 767-770, 1995.
2. L. Bahl, S. Balakrishnan-Aiyer, J. Bellgarda, M. Franz, P. Gopalakrishnan, D. Nahamoo, M. Novak, M. Padmanabhan, M. Picheny, S. Roukos "Performance of the IBM large vocabulary continuous speech recognition system on the ARPA Wall Street Journal task", Proceedings of the ICASSP, pp. 41-44, 1995.
3. L. Bahl, P. de Souza, P. Gopalakrishnan, D. Nahamoo, M. Picheny "Context-dependent vector quantization for continuous speech recognition", Proceedings of the ICASSP, 1993.
4. J. Brousseau, C. Drouin, G. Foster, P. Isabelle, R. Kuhn, Y. Normandin, P. Plamondon "French Speech Recognition in an Automatic Dictation System for Translators: the TransTalk Project", Proceedings of the EUROSPEECH, pp. 193-196, 1995.
5. H. Crépy, J.C. Marcadet, C. Waast "Dictée à grand vocabulaire en français: IBM VoiceType 3.0, un produit de la recherche", 1ères JST FRANCIL, pp. 19-23, 1997.
6. J.L. Gauvain, L. Lamel, M. Adda-Decker "Developments in Continuous Speech Dictation using the ARPA WSJ Task", Proceedings of the ICASSP, pp. 65-68, 1995.
7. M. Grevisse "Le bon usage, grammaire française", (c) Duculot, ISBN 2-8011-0588-0, 1986.
8. A. Lerond "Dictionnaire de la prononciation", (c) Librairie Larousse. ISBN 2-03-340101-4, 1980.
9. C. Waast "Contribution a l'élaboration d'un système de reconnaissance de parole continue à grand vocabulaire", These E.N.S.T, Jan. 1991.
10. S.J. Young, M. Adda-Dekker, X. Aubert, C. Dugast, J.-L. Gauvain, D.J. Kershaw, L. Lamel, D.A. Leeuwen, D. Pye, A.J. Robinson, H.J.M. Steeneken, P.C. Woodland "Multilingual large vocabulary speech recognition: the European SQUALE project", Computer Speech and Language, pp. 73-89, Vol. 11, 1997
11. French fixed network speech corpus. Version 1.0. Phonetically rich sentences. The SpeechDat project is a CEC-funded initiative (LRE-63314).