DYNAMIC LEXICON FOR A VERY LARGE VOCABULARY VOCAL DICTATION

Marie-José Caraty, Claude Montacié and Fabrice Lefèvre

LIP6 - Université Pierre et Marie Curie - CNRS
4, place Jussieu - 75252 Paris Cedex 5 - France
Tel : (33/0) 1 44 27 26 74, Fax : (33/0) 1 44 27 70 00, E-mail : caraty@laforia.ibp.fr

ABSTRACT

For very large vocabulary vocal dictation systems, we present a decoding strategy useful to reduce the lexical decoding cost. For each test-utterance, a sub-lexicon is selected from a very large recognition vocabulary. Such a recognition sub-lexicon is called Dynamic Lexicon (DL). Various algorithms of DL selection are developed and tested in terms of coverage rate of textual corpus. From these experiments, we describe a DL constitution we choose to use in D-DAL, our HMM-based recognizer competing for the first campaign of french vocal dictation supported by AUPELF. The contribution made by this original DL is a posteriori confirmed through the AUPELF-B1 test-dictation.

1. INTRODUCTION

Whatever the automatic recognition system is, we have to solve the major problem of the memory requested by the language models to process very large vocabulary recognition. In the case of HMM-based recognition systems using a one pass algorithm for the lexical decoding, several resources (e.g., acoustic, lexical, language) have to be simultaneously allocated in memory. With a very large vocabulary, it becomes important to determine a short list of candidate words (i.e., a recognition sub-lexicon) [1, 2, 3, 4] before computing slow and detailed acoustic, lexical and language matches [5]. To use efficiently such a sub-lexicon, we propose a 4-step test-sentence decoding (gender recognition, phonetic decoding, word hypothesizer, lexical decoding) described in the figure 1. Thus, the computational cost is unchanged when the output probabilities are stored and the memory request is lower. Of course, the performance of this approach depends on the coverage of the test-sentence by the recognition sub-lexicon. Experiments are carried out on the development and test corpus of the campaign AUPELF-B1.

2. D-DAL: A SPEECH RECOGNIZER

D-DAL (Dictaphone-Dactylographe Automatique du LIP6) is a HMM-based vocal dictation system developed from HTK (Hidden Markov model ToolKit) [6]. The continuous speech recognition system is speaker-independent and without any constraint of vocabulary. The system is designed from three main distinctive parts, the recognition core (free-language component), the acoustic representation (language-dependent component) and the language of the application (task-dependent component).

2.1. Recognition Core

In D-DAL, the test-sentence is decoded from the connected models of the words using the token passing algorithm and the information of the n-gram and n-class [5, 6]. The lexical decoding cost of a test-sentence is estimated in terms of computation cost and memory cost. These costs are split up into an acoustic part and a linguistic part. Let N_A be the number of acoustic models, let N_L be the recognition vocabulary size and let N_T be the number of test-sentence frames. When the bigrams are used, the computation cost is $N_T [A_1 N_A + L_1 N_L{}^2]$, the memory cost is $A_2 N_A N_L + L_2 N_L{}^2$, with A_1, A_2 representing the acoustic parts and L_1, L_2 the linguistic parts of the two costs. For a very large vocabulary recognition, the linguistic cost is major. A strategy to decrease the linguistic cost is to use only a part of the recognition vocabulary for the decoding.

2.2. Acoustic of the Language

For the acoustic representation of the words making up the language, we have developed the phonetic dictionary of the LIP6 (DP-L). The DP-L is constituted of the phonetic transcriptions of the 80k-words of *Le Grand Robert* dictionary. Numerous modifications and additions to the phonetic transcriptions are processed, -the missing elisions, -the grammatical flexions, -the transcriptions sounding the liaisons and -the transcriptions of proper names. Actually, the DP-L is a 500k-words vocabulary, with 2.15 phonetic variants (mainly due to the liaisons) per word, that is to say more than 1M of patterns (i.e., phonetic transcriptions). The analysis vectors are 15 Mel frequency cepstrum coefficients, the energy and their delta-coefficients. The phonetic models are 3-states Bakis models. Gaussian mixtures represent the HMM states output distributions. BREF is the acoustic corpus we use, a database of 100

speech hours, including 65585 sentences uttered by 120 speakers (i.e., 65 female and 55 male speakers). Four types of acoustic models are trained from the BREF database, 37 monophone models including the silence, 1,104 left biphone models, 1,104 right biphone models, 10,005 triphones models. For a gender adaptation, the contextual phonetic models are specialized on the acoustic corpus of the female and male speakers.

2.3. Language of the Application

A standard application is the newspapers recognition task. The archives of *Le Monde* newspaper (from 01/01/87 to 11/02/96) constitute the textual corpus. The corpus has been corrected (e.g., accent mistakes, typing errors) using the DP-L. A 813k-words vocabulary covers the whole corpus. Let L-Wk be the vocabulary corresponding to the list of the Wk-words sorted by decreasing order of occurrence frequency observed on the textual corpus. Taking the *Le Monde* training text corpus from january the 1st of 1987 to december the 31st of 1995, a preliminary study has consisted in computing the percentage of words covered by various L-Wk vocabularies for the articles published in the first quarter of 1996 (Table 1). The percentage of out-of-vocabulary words becomes insignificant over a 100k-words vocabulary. The statistical language models (e.g., n-gram and n-class) are computed on the textual corpus from a recognition vocabulary. In consequence of the DL-based decoding, the n-gram language models are dynamic language models.

L-1k	L-2k	L-5k	L-10k	L-20k
70.2%	76.6%	84.8%	90.1%	94.2%
L-64k	**L-100k**	**L-200k**	**L-500k**	**L-700k**
98.1%	98.8%	99.4%	99.7%	99.7%

Table 1. Coverage rate of a quater of "Le Monde" by the L-Wk vocabulary as a function of the W size

3. DYNAMIC LEXICON

For each test-utterance (e.g., test-sentence or test-paragraph), a DL is selected from a recognition vocabulary by a 2-step processing. The first step is the computation of dissimilarities between the phonetic lattice of the test-utterance for each word of the vocabulary. The second step is the selection of the nearest words from the phonetic lattice. Various Dynamic Lexicon Selection algorithms (DLS) have been tested on the development-corpus of AUPELF-B1 before we define an original DL constitution. The interest of this DL is analyzed through the AUPELF-B1 test.

3.1. Acoustic-Lexical Dissimilarities

The DL selection is based on a dissimilarity measure between the phonetic transcription of a word and the acoustic-phonetic decoding of a test-utterance. This measure is computed using the Wagner and Fisher algorithm [7]. Originally designed to compare strings, the algorithm has been modified in order to find the optimal location of a sub-string in a string. Three operations are defined to match two symbols : the substitution, the omission and the insertion. Each operation has a cost. The goal of the algorithm of dynamic programming is to minimize the summation of the costs, corresponding to the sequence of elementary operations, involved in the comparison of the strings. Two cost functions are studied. The first one consists in taking a fixed value for the operation costs. In the second one, the costs are trained to take into account the errors of the phonetic decoding. The substitution cost between two phonemes is an inverse function of the percentage given by the confusion matrix computed from the acoustic-phonetic decoding of a training set. The N-choices phonetic lattice is the result of the acoustic-phonetic decoding of the test-utterance. It consists in storing the N-best phonemes hypothesis at the end of each phoneme of the optimal solution. To integrate the N-choices lattice in the dynamic programming, a substitution cost is computed as a function of the minimum of the N substitution costs. The minimum is wheigthed to take into account its corresponding rank in the phonetic lattice. To use the occurrence frequency of the word, the dissimilarity computed by dynamic programming is wheighted by an inverse function of the corresponding probability (e.g., -log(p)).

3.2. DL Selection Algorithms

Five DLS algorithms combining three variants are experimented. The variants concern the depth of the phonetic lattice, the choice of the cost functions and the use of the occurrence frequency of the words. The notations introduced for the variants are the following : $\overline{T}$ the first choice lattice, T the 3-choices lattice, $\overline{C}$ the Levenshtein cost function, C the cost function based on the confusion matrix, $\overline{P}$ the equiprobability of the words occurrence, P the probability of the words occurrence obtained on the textual database. For a test-sentence, the table 2 gives the index (I) of the words in the L-813k list, the standard phonetic transcription (Ph) and the three choices of the phonetic lattice (T1, T2, T3).

S	Le	Luxembourg	lui-même	pourrait	verser
I	3	3384	81	68 259	4545
Ph	l «	l y ɑ̃ k b œ r	l ɥ i m ɛ m	p u r ɛ	v
T1	l «	l « ɑ̃ k œ s u	y m ɛ m	p w a	l e r s e
T2	l «	l « ɑ̃ k œ s n	n m ɛ m	p r i	l e r s
T3	l e	l « ɑ̃ k œ s n	n m l n	p r i	l ɛ r

S	une	quote-part	de	cinq cent	millions.
I	13	27255	1	55 23	117
Ph	y n	k ɔ t p a ʁ	ɛ̃	s ɑ̃	d « m ɔ̃ j
T1	# i n e	k a # p a b	d «	ɛ̃ s ɑ̃	m ɔ̃ #
T2	# i n d	k a # p r	d «	ɛ̃ s ɑ̃	m ɔ̃ k
T3	# i n d	k a # # r	d s	ɑ̃ s ɑ̃	i ɔ̃ # l

Table 2. Test-sentence (S) - Indices (I) of the words in the L-813k list - Standard phonetic transcription (Ph) 3-choices phonetic lattice (T1, T2, T3)

For each one of the five DLS, the table 3 gives the index of each word of the test-sentence in the corresponding DL. These indices can be compared to the indices (I) in the L-813k list given in the table 2.

Sentence	$\overline{T}\overline{C}\overline{P}$	$\overline{T}C\overline{P}$	$T\overline{C}\overline{P}$	$TC\overline{P}$	TCP
Le	159	159	141	141	141
Luxembourg	1589	1394	802	762	519
lui	47795	24907	72634	40054	1381
même	164	164	146	146	146
pourrait	81954	65412	23623	16240	1514
verser	2926	1408	509	609	439
une	12052	3595	18783	7252	322
quote-part	10973	9000	2605	2386	3092
de	136	136	115	115	115
cinq	998	777	1864	1304	309
cents	133	133	112	112	112
millions	2470	1393	4820	2276	416

Table 3. Indices of the words in the DL as a function of the DLS.

3.3. Development Experiments

For each DLS, we compute the coverage rate of development-sentences by the 20k-words corresponding DL. The experiments are carried out on the 96 first sentences, covered by the DP-L, of the development-corpus of AUPELF-B1. The sentences (2334 words) are uttered by 3 speakers (i.e., 2 female and 1 male speakers). The recognition vocabulary corresponds to the L-100k covered by the DP-L. The training set of the substitution costs uses 279 sentences of development. The results are given in the table 4, the coverage is expressed in words and in percentage. To get the same coverage performance than the TCP-based DL, the recognition vocabulary L should have a 43448-words size involving a linguistic computational cost five times bigger.

	$\overline{T}\overline{C}\overline{P}$	$\overline{T}C\overline{P}$	$T\overline{C}\overline{P}$	$TC\overline{P}$	TCP	L-20k
Words	2112	2180	2172	2211	2319	2255
%	90.5	93.4	93.1	94.7	99.3	96.6

Table 4. Coverage rate in words and in percentage by a 20k-words DL as a function of the DLS

3.4. Test Experiments

For the first campaign AUPELF-B1, D-DAL has used the dynamic lexicon principle. In the category "Vocal Dictation with Unconstraint Vocabulary", the test has consisted to the speaker-independent recognition of 655 sentences (i.e., 19241 words) structured by paragraphs. At first, we present the DL constitution we have chosen at the end of the development step. The interest of the DL is analyzed through the AUPELF-B1 test-dictation.
During the decoding (i.e., synchronous and asynchronous processings), D-DAL uses the statistical n-gram and n-class language models. The textual corpus is automatically tagged from 72 grammatical labels. The language models are DL-dependent while the DL is test-dependent. The dynamic lexicon is more precisely a 40k-patterns vocabulary. The patterns are lexically or grammatically different. The chosen composition of the 40k-patterns vocabulary is the following : -20k-words are selected according to the decreasing grammatical-lexical occurrence frequencies, -10k-words are $TC\overline{P}$-based DL and -10k-words are TCP-based DL. The occurrence frequency is interesting in case of bad decoding, whereas the $TC\overline{P}$ algorithm is interesting when the decoding is accurate for an unprobable word. The DL is selected from a recognition vocabulary : a LG-354k patterns vocabulary (i.e., lexically or grammatically different patterns) corresponding to the 162k-words lexicon constituted of the words of L-813k belonging to the DP-L. The performance of the dynamic lexicon is estimated by its coverage rate of the test-corpus of AUPELF-B1. The table 5 gives the coverage rate of this corpus for various LG vocabularies including the DL-40k recognition vocabulary. On the test-corpus, the number of out-of-vocabulary words is twice higher in the case of the simple LG-40k recognition vocabulary than in the case of the DL-40k one. To get the same coverage rate than the DL-40k, a LG-100k is to be considered. On the textual corpus, a 1.6 size factor is observed from a L to a LG vocabulary. Therefore, a 20k-words recognition vocabulary constituted with the DL principle is as powerful as a L-50k recognition vocabulary.

LG-40k	LG-64k	DL-40k	LG-128k	LG-354k
95.7 %	97.2 %	97.9 %	98.4 %	98.8 %

Table 5. Coverage rate as a function of the LG vocabulary

3.5. Dynamic Language Models

The dynamic lexicon involves the development of a dynamic language models algorithm. The n-class are lexicon-independent as they are grammatically-based. For n-gram, we have developed a two-pass algorithm. The L-64k counts are previously computed and a filtering of the textual corpus by the 64k-words vocabulary is processed. The following sentence is given as an example of filtering.

Original text :
{ contre les excès de la finance ,# il appelle à un renforcement des systèmes de surveillance et des règlementations ,# faute de quoi "# de graves accidents surviendront "# }

Filtered text :
{ accidents surviendront }

In the first step of the algorithm, the L-64k n-gram counts are filtered by the intersection between the L-64k and the DL-40k words. In the second step, the filtered corpus is used to compute the remaining n-gram counts. The speed of this algorithm depends on the size of the filtered textual corpus. In our recognition task, the

filtered corpus has a 5 % size of the original textual corpus.

4. CONCLUSION

The dynamic lexicon is a posteriori shown to be very efficient. Once its constitution has been decided, the DL has been assessed in terms of coverage rate on the test-corpus of the AUPELF-B1 vocal dictation. The results show a DL-based 20k-words recognition vocabulary is as powerful as a L-50k recognition vocabulary. The DL is an efficient strategy for very large vocabulary speech recognizers. More, its constitution allows a priori to process for the best what is not taken into account in a L recognition vocabulary : the "rare" words and the "bad" decoded words. On a Pentium-Pro 200MHz, for a 11 s duration sentence (about 30 words), the DL-selection we described takes 2 mn and the corresponding dynamic bigram computation takes about 5 mn. A strategy to speed up the DL time processing is to select it from clusters of the original vocabulary and to code the filtered text for the dynamic language models computation.

5. REFERENCES

[1] L. Fissore, P. Laface, G. Micca & R. Pieraccini. "Lexical Access to Large Vocabularies for Speech Recognition". IEEE TASSP, vol. 37, n° 8, pp. 1197-1213, 1989.

[2] C. Waast, L. Bahl & M. El-Bèze, "Fast Match on Decision Tree", Eurospeech, pp. 909-912, 1995.

[3] J. Macias-Guarasa, A. Gallardo, J. Ferreiros, J. M. Pardo, & L. Villarrubia, "Initial Evaluation of a Preselection Module for a Flexible Large Vocabulary Speech Recognition System in Telephone Environment", ICSLP 96, pp. 30-33, 1996.

[4] C. Montacié, M.-J. Caraty & C. Barras, "Mixture Splitting Technic and Temporal Control in a HMM-Based Recognition System", ICSLP 96, pp. 977-980, 1996.

[5] M.-J. Caraty, C. Barras, F. Lefèvre & C. Montacié. "D-DAL : un système de dictée vocal developpé sous l'environnement HTK". 21èmes JEP, pp. 289-292, 1996.

[6] S.J. Young, "HTK Version 1.4: Reference Manual and User Manual", Cambridge University Engineering Department - Speech Group, 1992.

[7] R. M. J. Fisher, "The String to String Correction Problem", JACM, 1974.

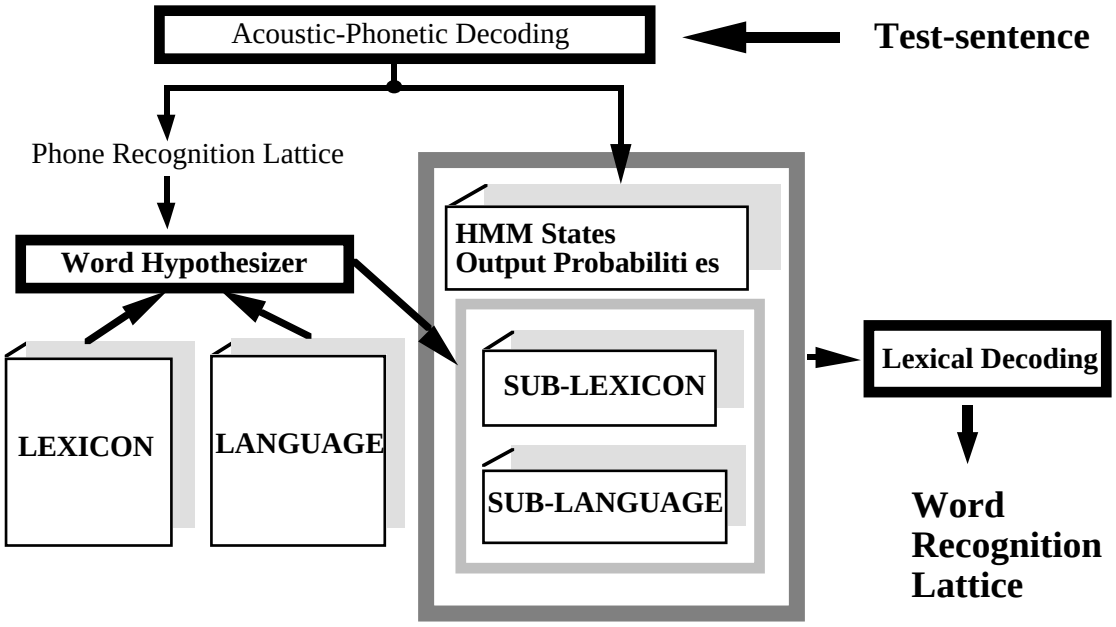

Figure 1. Principle of lexical decoding of a test-sentence using a dynamic lexicon