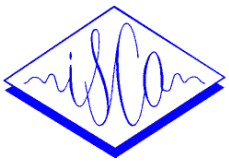

4th European Conference on
Speech Communication and Technology
EUROSPEECH '95
Madrid, Spain, September 18-21, 1995

LEXICAL FILLERS FOR TASK-INDEPENDENT-TRAINING BASED KEYWORD SPOTTING AND DETECTION OF NEW WORDS

R. El meliani & D. O'Shaughnessy
e-mail: meliani@inrs-telecom.uquebec.ca
INRS telecommunication
16 place du commerce
Verdun (Ile des Soeurs)
H3E1H6 CANADA.

ABSTRACT

In this paper we describe a preliminary investigation of the use of fillers at the lexical level rather than the modelling level of a hidden Markov model-based keyword spotter and a detector of new words. In our last system, keywords and out-of-vocabulary speech shared the same context-dependent phoneme models with no explicit modelling of extraneous speech. Thus a task-independent training is performed for all models while the scoring method uses a two-pass Viterbi-type algorithm based on a lexical tree constructed with transcriptions of keywords and fillers using the same set of 40 English phonemes. The distinction between keywords and extraneous speech is performed during the search by using the lexical tree and language models.
Thus using a simple method, we perform a faster training and allow easier modifications for the word-spotting task. On the other hand this kind of architecture allows our system to be used for both keyword spotting and new word detection tasks.

1. INTRODUCTION

It is well known that users of speech recognition systems usually do not limit their speech to the pre-defined vocabulary: they use spontaneous speech. *Depending on the application*, two different ways of dealing with out-of-vocabulary speech can be chosen:
- Detecting only the vocabulary words when the aim is to detect the *information* brought by the input speech, as for operator assisted calls, for instance. It is called keyword spotting.
- Or recognizing the whole text, new words included, when only an *orthographic transcription* is needed (dictation machine, unlimited-vocabulary continuous speech recognition...), which is referred to as new word detection. Only a few new word detectors have been reported [9,10].

In the past those two kinds of systems have been addressed separately [6,7,9,10,11], even though the main difficulty is the same for both: *how to make a sufficient distinction between keywords and out-of vocabulary words.*
Moreover, most continuous speech recognizer based keyword spotters as well as new word detectors use acoustic-phonetic fillers to make this distinction, which usually requires a large amount of out-of-vocabulary speech for a correct training of fillers. However, it seems that for human listener the distinction between keywords and out-of-vocabulary words is simply lexical: the keyword is spotted whenever its phonemic or syllabic sequence is detected. On the other hand, the history of writing showed the importance of syllables in the transcription of main world major languages [8].

This paper presents, in this view, a preliminary investigation of the performances of lexical fillers for task-independent-training keyword spotting as well as for new word detection.

2. SYSTEM DESCRIPTION

2.1. The INRS Continuous Speech Recognizer: Motivations of its Use

The HMM-based word spotter described in this work uses the INRS large-vocabulary continuous speech recognizer [1,2,3,4]. Using such a recognizer provides our spotter with a lot of benefits. Thus, as this system uses context-dependent phoneme hidden Markov models, the spotter models will be independent from the size of the keyword dictionary.
Moreover in the case where acoustic filler models are chosen to be the same as acoustic keyword models, the spotter will become task-independent which will lead to a gain of computing time and flexibility for changes of the keyword dictionary. With such an architecture our system can then be used as a keyword spotter as well as be taken as a new word detector.
On the other hand, as the recognizer uses a large dictionary, the size of the keyword set can be very large as it is suitable for many applications like phone directory consultation for instance. Finally, one more advantage of using this recognizer, and not the least, is that it is nearly real-time (one block-time delay). All those advantages lead to a larger choice of applications.

2.2. The INRS Continuous Speech Recognizer: Description

The block diagram of the INRS continuous speech recognizer is given in figure 1. This system processes the input speech block after block: the output beam of the word search processor of one block is the input beam of the following block. This temporal block processing is a fair compromise between the use of a forward-backward algorithm and the real time issue, that allows, at the same time, important memory reduction. Each input speech block is represented by a set of static and dynamic Mel-frequency cepstrum coefficients (MFCC).
The word search processor starts with a backward Viterbi search on the phonetic graph derived from the lexical tree using a coarse acoustic phonetic HMM and produces the B* table of the estimates of all phone segment scores as well as phone segment end times. The phonetic graph is an ordered tree of phonetic transcriptions of all the dictionary words. Only phoneme sequences belonging to this graph will be recognized.

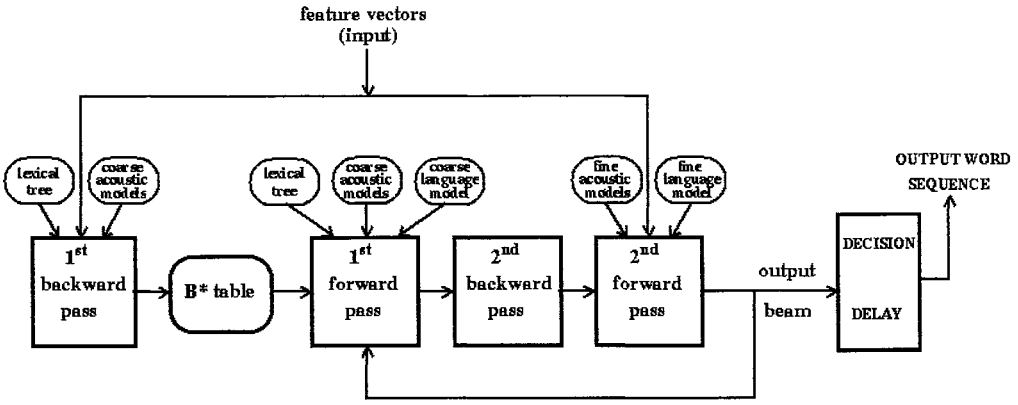

Figure 1: The two-pass word search processor of the INRS system.

This lexical constraint allows a gain in computing time and accuracy.

The second step of the search processor is an extremely fast processor that can deliver a small set of word hypotheses (a word graph) highly guaranteed to include the correct hypothesis. It consists of an A* forward search [3,5] which uses the B* table to produce an overestimated probability of future scores in order to reduce the beam search efficiently. This first pass uses acoustic and language models that are usually coarser than the ones to be used in the second pass. The language model computes the probability of a word knowing its predecessors using unigram, bigram and trigram frequencies.

Finally during the second pass the pruned word graph is re-evaluated usually using finer acoustic phonetic HMMs and language models to give the list of the most likely sequences of words with, for each hypothesis, a multiple segmentation based on the unique segmentation provided by the first pass.

This scheme is used for both training (segmentation phase) and recognition. During the training the lexical graph is the sequence of phonemes included in the incoming phonemic segmentation of the input file. As for recognition, its lexical graph is a parallel of all phonemic transcriptions of the dictionary words.

2.3. The Keyword Spotter

2.3.1 Motivations

In our first keyword spotter all extraneous words were modelled using a filler model consisting of the parallel arrangement of all English phones and trained on extraneous speech whereas keyword phone models were trained on all occurrences of keywords. This filler model has been shown [6,7] to be the best choice for a continuous-speech-based spotter concerning the trade-off between time computing and memory needs on one hand, and accuracy on the other hand.

The difference between keywords and out-of-vocabulary words being first lexical, we thought of trying lexical fillers instead of the classical acoustic-phonetic fillers to represent extraneous speech. Thus, we lowered significantly the acoustic model number, removed the need of a large amount of extraneous speech to train efficiently filler models, and reached a task-independent training. The fillers will be detected in the word graph whenever the score of one of their phonetic transcriptions is more likely than that of keywords, using only the lexical graph constraint and language model constraints. The only part of the system that is not task-independent is the word frequency determination for the language model.

Thus, in order to more reduce the computing time and memory needs, we first experimented with lexical phone fillers: the models to train being the same for keywords and out-of-vocabulary words, their number has thus been reduced in half. Unfortunately the first results given by those fillers showed a noticeable increase in the false alarm rate, the design of this filler making it maybe too general.

The idea of using syllabic fillers instead is based on the importance of this lexical constraint which accounts for most coarticulation effects. This will allow us to reduce the number of confusions between keywords and fillers. Lleida et al [11] have already used syllabic fillers but at the acoustic phonetic level. Our choice has been directed as well by remarks on the length of most hesitations, cut words and the average size of words that compose the extraneous speech. The least extraneous word size is one syllable, the average size is two syllables.

2.3.2 Description of Syllabic Fillers

The frequencies of all possible syllables have been computed on the used databases. The set of all those syllables has then been divided between syllabic fillers, thus becoming possible phonemic transcriptions of those fillers. Every filler gathered only syllables of the same frequency in order to get a frequency homogeneous filler: the reason is that in the language models frequencies and probabilities correspond only to orthographic transcriptions due to the lack of phonetic transcription frequencies; mistakes are then possible whenever a phonetic transcription frequency is very different from the orthographic transcription frequency. The increase of the lexical set is thus negligible.

2.4. The New Word Detector

As for the new word detector, it is simply an application of our lexical-filler-based keyword spotter with a few out-of-vocabulary words, the keyword set covering nearly all the word set. This design is possible because keywords and extraneous speech share the same acoustic phonetic models as required in a classical continuous speech recognition task. Moreover the pronounced phonetic transcription is easily recovered, thus making

Name	Type	Speakers	Number of words	Noise environment
ATIS	Travel database	285 (mixed)	1030	quiet
White Fang (whfng)	Book on tape	1 male	3600	35 db
Wall Street Journal (WSJ)	Read newspaper	1 male	4872	normal

Table 1: Experimental Corpus.

this system more suitable for a dictation machine than a limited-vocabulary continuous speech recognizer.

2.5. The Language Model

Usually, in speech recognition the information given by bigram frequencies is stronger than that given by unigram frequencies and using both of them may lead to errors if the bigram is not frequent while the considered unigram is very frequent.
Nevertheless, in our case, due to the fact of the presence of fillers in the output word sequence and to the impossibility to evaluate correctly the frequencies of bigrams of the form {keyword filler}, using a complete bigram frequency set will be inefficient. So the language model of our keyword spotter uses the unigram frequency set mostly. The frequency of a filler is set to the common frequency of the syllables it includes.
As for our new word detector language model, it uses a complete unigram set as well as known-word bigram frequency sets.

3. EXPERIMENTAL ENVIRONMENT

3.1. The Experimental Corpus

Test results are reported for three different databases: Wall Street Journal, ATIS (Air Travel Information System) and White Fang. Their main differences are summarized in table 1.
Wall Street Journal has been recorded near a work station in an office, contains 4872 different words and consists of the reading of the Wall Street Journal newspaper by one male speaker.
ATIS has been recorded in a very quiet environment, contains nearly 1030 different words and consists of spontaneous speech uttered by 285 male and female speakers. The training is performed for the 285 speakers with a total of 9269 sentences. As for the test set, it includes 802 sentences uttered by 19 speakers.
White Fang is a book on tape read by a male speaker in a 35 dB noise level. It uses 3600 different words.

Vocabulary set	Number of keywords	Word number ratio (%)	Word frequencies	Keyword training time (min.)	Filler training time (min.)	Time ratio (kw/filler) (%)
WSJ1	4844	99.4	various	165.91	6.62	2506
WSJ2	30	.62	> 10	7.58	165.05	4.59
ATIS	102	9.9	mostly > 5	31.2	128	24.38
WHFNG	101	5.56	mostly > 5	17.13	60	28.55

Table 2: Characteristics of vocabularies

Vocabulary set	Detection rate (%)	False alarm number (/kw/h)
WSJ1	80	4.4
WSJ2	77	0.8
ATIS	81	2.7
WHFNG	70	2.4

Table 3: Acoustic-Phonetic Filler Keyword Spotting

We experimented with different sizes of keyword sets varying from 30 to 4700. There was no limit on the number of keywords per sentence. Four kinds of vocabularies were used in these first experiments to analyse the influence of various parameters. Their characteristics are shown in table 2.

3.2. Experimental Setup

Speech has been sampled at 16 kHz with a block size of 25 ms and a block advance of 10ms;15 static and dynamic MFCC were used. Our first results having been run on a Dec-station, we were confronted with memory problems which forced us to simplify the recognizer. Thus our continuous speech recognition rate is lower than that of the complete INRS system.
The acoustic models were the same in the two passes: three-state right-context phone HMMs with all distributions sharing the same covariance matrix and a set of 256 means.
As no exhaustive list of possible syllables was available, we gathered possible syllables from the syllabic transcriptions provided in our complete dictionary. Moreover, in new word detection the keyword frequencies have been doubled to take in account the ratio between syllables and the average-syllabic size of words.

4. EXPERIMENTAL RESULTS

4.1. Acoustic-Phonetic Filler

The results obtained for the acoustic-phonetic filler are shown in table 3. The scores reflect the noise level showing the necessity of noise filtering. However for the comparison between ATIS and the Wall Street Journal, results are opposite to those found for continuous speech recognition, maybe because of the keyword energy level.

4.2. Lexical Phonemic Filler

We experimented with the lexical phonemic filler only for Wall Street Journal and White Fang. The results were very dependent on the language model because we used only one filler gathering all phonemes. The best results were found with no language models, but presented a false alarm rate at least 5 times higher than the preceding acoustic phonetic filler results and a loss of nearly 5% on the detection rate.
A set of phonemic frequency-homogeneous fillers, as well as the definition of a relevant language model, are under investigation.

4.3. Lexical Syllabic Filler

Vocabulary set	Detection rate (%)	False alarm number (/kw/h)
WSJ1	72	10
WSJ2	85	0.5
ATIS	70	3

Table 4: Lexical Filler Keyword Spotting

	Word Recognition (%)	New word Detection (%)	Transcription Detection (%)
WSJ1	72	82	85

Table 5: Lexical Filler New Word Detection

4.3.1. In Keyword Spotting

The results of our first experiments of syllabic lexical fillers are given in table 4. Comparison to the acoustic phonetic filler shows that this syllabic filler choice is a fair trade-off between task-independent training, computing time and memory needs on one hand, and accuracy on the other hand. Moreover these results may be improved by a better definition of the language model.

4.3.2 In New Word Detection

Table 5 gives the first results of the new word detection for the Wall Street Journal large-vocabulary set. The recognition rate decreased by 10% compared to that of continuous speech recognition while the new word detection rate as well as the phonetic transcription detection are relevant. Using the complete INRS system instead of this simplified one may improve significantly these results.
Other vocabulary sets are still to be tested to confirm those first results.

5- CONCLUSION

In this paper we presented lexical fillers of two kinds: phonemic fillers and syllabic fillers. We tested them on different databases for different sizes of vocabularies and compared them to the acoustic-phonemic phoneme-based filler.
Our first result shows the relevance of lexical syllabic fillers when used in a task-independent-training keyword spotter as well as when tested in a new word detector. The loss in performance is compensated by the time computing gain in training (fewer models) and during keyword modifications. However, the conception of those fillers as well as the definition of their language models are being refined to achieve better scores.

6. REFERENCES

[1] P. Kenny, P. Labute, Z. Li and D. O'Shaughnessy, "New Graph Search Techniques for Speech Recognition", Procs. of Int. Conf. on Acous. Speech and Sig. Proces.1994, pp. I-553-556.

[2] P. Kenny, R. Hollan, G. Boulianne, H. Garudadri, Y. M. Cheng, M. Lennig and D. O'Shaughnessy, "Experiments in Continuous Speech Recognition with a 60,000 Word Vocabulary", Proc. of Int. Conf. on Spoken Language Processing, Banff, Canada, 1992, pp. 225-228.

[3] P. Kenny, R. Hollan, P. Kenny H. Garudadri, M. Lennig and D. O'Shaughnessy, "An A* Algorithm for Very Large Vocabulary Continuous Speech Recognition", DARPA Workshop, 1992.

[4] P. Kenny, G. Boulianne, H. Garudadri, S. Trudelle, R. Hollan, M. Lennig and D. O'Shaughnessy, "Experiments in Continuous Speech Recognition Using Books on Tape", Speech Communication, V. 14 n. 1, pp. 49-60, Feb. 1994.

[5] N.J. Nilsson, "Principles of Artificial Intelligence", Palo Alto: Tioga, 1980.

[6] R.C. Rose and D.B. Paul, "A Hidden Markov Model Based Keyword Recognition System", Proc. of Int. Conf. on Acous. Speech and Sig. Proces., 1990.

[7] R.C. Rose and E.M.Hofstetter, "Task Independent Words potting Using Decision Tree Based Allophone Clustering", Proc.of Int. Conf. on Acoust., Speech and Sig. Proces., 1993.

[8]P. Ladefoged, "A Course in Phonetics", Hartcourt Brace Jovanovich Inc., Chapter 10.

[9] A. Asadi, R.Schwartz and J. Makhoul, "Automatic Mode ling for Adding New Words to a Large-vocabulary Continuous Speech Recognition System", Proc. .of Int. Conf. on Acous. Speech and Sig. Proces.1991, pp. 305-308.

[10] A. O. Asadi and H.C. Leung, "New-Word Addition and Adaptation in a Stochastic Explicit-segment Speech Recognition System", Proc. of Int. Conf. on Acous. Speech and Sig. Proces.1993, pp. V-642-645.

[11] E. Lleida, J. B. Mariño, J. Salavedra, A. Bonafonte, E. Monte and A. Martinez, "Out-of-Vocabulary Word Modelling and Rejection for Keyword Spotting", Proc. of Eurospeech, 1993, pp. 1265-1268.