NEW EFFICIENT FILLERS FOR UNLIMITED WORD RECOGNITION AND KEYWORD SPOTTING

Rachida El Méliani and Douglas O'Shaughnessy

INRS-Télécommunications
16 Place du Commerce, Ile-des-Sœurs, H3E 1H6, Québec, Canada
email: meliani@inrs-telecom.uquebec.ca

ABSTRACT

This paper describes our complete results for improved lexical fillers as well as two new kinds of fillers, gives their results in unlimited speech recognition as well as for keyword spotting and compares them to the acoustic-phonetic filler in the case of keyword spotting.

Tests have been conducted on different vocabularies derived from ATIS and the Wall Street Journal database. Results for keyword spotting show the superiority of the independent lexical phonemic filler that combines accuracy (92% for a false alarm rate of 1.2 FA/h/kw) as well as task-independent training. As for new-word detection, the syllabic and the independent lexical fillers perform quite well, and allow relevant detection of the phonetic transcription.

1. INTRODUCTION

There is an obvious similarity between new-word detection and keyword spotting due to the presence, in both, of the duality between vocabulary words on one hand and non-vocabulary words on the other hand. This similarity allows the use of common methods in both cases like the use of the same kind of filler for instance. Nevertheless vocabulary sizes, the nature of non-vocabulary words and the kind of accuracy required for detection are different:

- thus the number of vocabulary words used in keyword spotting is generally low (tens to hundreds of words) for thousands of out-of-vocabulary words, while that used in new word detection is high (thousands of words) with a few new-words (some words to hundreds of words);
- in keyword spotting most out-of-vocabulary words are already present in the training corpus and can thus be used to obtain task-related models to represent vocabulary words, while for new-word detection, on the contrary, the training corpus includes only vocabulary words, but no information on unknown words;
- as for the accuracy, in keyword spotting only a relevant keyword detection with a minimum false alarm rate is required while in new-word detection, mainly used for unlimited-vocabulary speech recognition, the vocabulary-word recognition rate is as important as the new-word detection rate or even their phonetic-transcription detection rate.

Until now, discrimination between vocabulary words and out-of-vocabulary words has been considered separately and differently in keyword spotting [7, 8] and in new-word detection [1, 2, 5]. In a previous paper [6] we gave our first results when using the same kind of lexical fillers for both new-word detection and keyword spotting. This paper describes our complete results for improved lexical fillers as well as two new kinds of fillers, gives their results in unlimited speech recognition as well as in keyword spotting and compares them to the acoustic-phonetic filler in the case of keyword spotting.

2. DESCRIPTION OF THE SYSTEMS

Our systems are based on the INRS real-time very-large-vocabulary continuous speech recognizer [3, 4]. We describe here the lexical tree and the language models, the two parts that are affected by our methods.

2.1. The Lexical Tree

The lexicon examined for each word orthography all the different corresponding pronunciations (see table 1). The program transforms the lexicon into an ordered lexical tree from which, with the use of the computed table of context-dependent phoneme scores (B^*), phonetic transcriptions are scored; then with the use of the given language models, the most probable word strings are derived.

2.2. The Language Models

The INRS recognizer language model used here is based on the deterministic back-off form. It is computed from bigram distributions $P(w_i|w_N)$, and unigram distributions $P(w_i)$, where w_i is the considered word and w_N its preceding word in its history. Its score contribution follows from the formula:

$$logP_{HMM} + \alpha logP_{LM} + \beta \quad (1)$$

keyword 1	phon. transc. 1	...	phon. transc. c_1
⋮	⋮		⋮
keyword p	phon. transc. 1	...	phon. transc. c_p
filler 1	phon. transc. 1	...	phon. transc. g_1
⋮	⋮		⋮
filler q	phon. transc. 1	...	phon. transc. g_q

Table 1. Lexicon general format. p is the keyword number while q is the filler number.

where P_{HMM} is the HMM acoustic score, P_{LM} the language model score, α a weighting coefficient related to the confidence in the language model and β a flat distribution term that allows handling out-of-vocabulary words.

3. DESCRIPTION OF THE FILLERS

3.1. Strictly Lexical Fillers

In this case the distinction between keywords and out-of-vocabulary words is made only at the lexical level: the two kinds of words are represented by a unique set of context-dependent phoneme models trained on the whole training corpus. The discrimination is achieved by computing scores through a lexical tree using adapted language models. These fillers are designed for both keyword spotting and new-word detection. Three kinds of lexical fillers are reported in this study; they differ mostly in their definition of the lexicon and the language models.

In the first one we defined a unique filler considering all phonemes (see table 2) as possible phonetic transcriptions; it is referred as a 'unique lexical phonemic filler' (ULP). This filler frequency sums all phoneme frequencies; thus for the language model all phonemes will be considered as having this common frequency. Consequently the bigram and unigram files are the smallest, but most of the lexical information carried in the out-of-vocabulary speech of the training corpus is ignored.

filler	phoneme1	...	phoneme40

Table 2. The unique lexical phonemic filler.

Then, due to the importance of the syllable, already discussed in [6], and to the strong lexical constraint it imposes on phoneme strings, we divided the set of syllables among fillers we called 'lexical syllabic fillers' (LS).

word	phon.transc.
filler1	phoneme1
⋮	⋮
filler40	phoneme40

Table 3. The multiple lexical phonemic fillers

Finally, we defined a lexical filler for each English phoneme (see table 3) as a solution to weaknesses of the previous kinds. It is referred to as 'multiple lexical phonemic fillers' (MLP). The sizes of bigram and unigram files increased importantly and consequently memory needs too; nevertheless this is compensated by a noticeable result improvement.

3.2. Acoustic-phonetic Fillers

For this case we suppose that the distinction between the two kinds of words must be made at the acoustic level; it means that we define two sets of HMMs context-dependent phonemes, one trained on the occurrences of all keywords in the training corpus while the other is learned on the whole out-of vocabulary speech in this corpus (thus defining out-of-vocabulary phonemes). Because of the architecture of the INRS recognizer, the definition of those fillers must be completed by the addition to the lexicon of orthographic fillers that will take account of the out-of vocabulary words.

The three kinds already described for strictly lexical fillers are applied to this case too, except that this time they obviously use phonemes trained only on out-of vocabulary speech. We refer to them respectively as a 'unique acoustic-phonetic phonemic filler' (UAP), 'acoustic-phonetic syllabic fillers' (AS) and 'multiple acoustic-phonetic phonemic fillers' (MAP). These fillers are not computable for new-word detection because no out-of vocabulary speech is available in the training corpus in that case.

4. EXPERIMENTAL ENVIRONMENT

4.1. Fillers and Language Models

As no list of syllables was available, we created ours by gathering all syllables (10536) present in the transcription of our complete database vocabulary.

In keyword spotting, bigram and unigram distributions for the filler are computed on the occurrences of out-of-vocabulary words in the training corpus, while in new-word detection they are computed on the whole corpus.

Name	KW number	UKW number	UKW ratio (%)
WSJ1	4842	30	.62
WSJ2	4654	218	4.68
ATIS	1018	12	1.18

Table 4. Characteristics of new-word detection vocabularies; KW means known-word and UKW means unknown-word.

4.2. Vocabularies

The tests reported in this paper concern two different databases: Wall Street Journal (noted here WSJ) and ATIS (Air Travel Information System), already described in [6]. All the experiments reported for keyword spotting were using vocabularies extracted from the Wall Street Journal; the vocabulary sizes range from 23 to 100 words of variable frequencies.

The three vocabularies on which new-word detection has been carried out have from 12 to 218 unknown words as shown in Table 4. The UKW ratio r is computed by the formula

$$r = \frac{UKWnumber}{KWnumber} \quad (2)$$

where KW means known-word and UKW means unknown-word.

5. TESTS AND RESULTS

The INRS recognizer has been simplified to fit with the available memory when used with all the proposed fillers; thus the recognition rate of the simplified recognizer used is low (80% for WSJ and 76% for ATIS) and will obviously affect the detection rate.

5.1. Acoustic-Phonetic Fillers

	UAP	AS	MAP
Detection (%)	71	82.6	93.2
F.A.(/KW/h)	.15	.23	.5

Table 5. Keyword spotting results for the acoustic-phonetic fillers.

The results obtained with the acoustic-phonetic filler in keyword spotting were very sensitive to the frequency of the context-dependent phonemes involved in the set of keywords. A frequency rate of at least 30 is necessary to reach a minimum detection rate of 70%. The choice of keywords is thus importantly reduced. The results given in Table 5 respect that constraint. It must be noticed too that this kind of filler is highly memory and time consuming and requires a task-dependent training. The unique acoustic-phonetic phonemic filler perform very well, far better than the other acoustic-phonetic fillers.

5.2. Lexical Fillers

Lexical fillers are very convenient when used, as in this work, with a lexical-tree-based continuous-speech recognizer. Because it uses task-independent training, the system can be turned easily from keyword spotting to new-word detection by simply changing the dictionary and the language models. Independence to the size of the vocabulary added to this flexibility to make our system suitable for many applications.

	ULP	LS	MLP
Detection (%)	67.6	82.9	92
F.A.(/KW/h)	.1	.1	1.2

Table 6. Keyword spotting results for the lexical fillers.

5.2.1. In Keyword Spotting

Lexical fillers are insensitive to the frequency of the context-dependent phonemes involved in the set of keywords, while giving scores slightly lower than the corresponding acoustic-phonetic fillers (see table 6). In this case too, the unique lexical phonemic filler shows a clear superiority to the other lexical fillers, confirming the efficiency of that kind of filler.

5.2.2. In New-Word Detection

Tests have been performed for the three kinds of lexical fillers. Table 7 reports the results on WSJ for both the unique lexical phonemic filler and the multiple lexical phonemic fillers, while Table 8 shows the results obtained for lexical syllabic fillers. The rates given in Table 8 for WSJ are the average rates computed on both the defined vocabularies.

The following notations are used to ease the table reading:
- R is the recognition rate computed for the vocabulary words in %
- NWD is the new-word detection rate in %
- NWT is the new-word transcription in %

fillers	R	NWD	NWT
ULP	66	70	70
MLP	65	68	64

Table 7. Unlimited-vocabulary recognition results for the unique lexical phonemic filler and the multiple lexical phonemic fillers.

The unique lexical phonemic filler scores poorly because of the loss of lexical constraint it implies. Dividing all the

phonemes among the multiple lexical phonemic fillers does not score better because of the difficulty of associating with them their relevant frequencies.

The lexical syllabic fillers results show an important improvement probably related to the linguistic constraint carried by syllables. These results are rather encouraging: we obtain for ATIS a 2% higher performance than the first recognizer, as for WSJ a relevant new-word detection as well as transcription detection are reached despite a decrease of 10% in the recognition rate compared to the first recognizer.

database	R	NWD	NWT
WSJ	72	82	85
ATIS	78	80	65

Table 8. Unlimited-vocabulary recognition results for the lexical syllabic filler.

6. CONCLUSION

This paper describes different designs for improved lexical fillers as well as two new kinds of fillers. It then reports for keyword spotting the study of all different combinations of the use of acoustic-phonetic models specific to out-of-vocabulary words in association with lexical fillers defined, this time, using the out-of-vocabulary phonemes, and determines a more accurate solution.The lexical fillers allow task-independent training and are more flexible than acoustic-phonemic fillers while giving scores only slightly lower.

Despite of the obvious similarity of keyword spotting and new-word detection, the results show a different behavior when used with our lexical fillers. Thus in keyword spotting the multiple lexical phonetic filler is the best in terms of memory and time needs as well as in performance, while the lexical syllabic filler gives the best results for new-word detection.

7. REFERENCES

1. A. Asadi, R. Schwartz and J. Makhoul, 'Automatic modeling for adding new words to a large-vocabulary continuous speech recognition system' ICASSP 91, pp 305-308.

2. A.O. Asadi, H.C. Leung, 'New-word addition and adaptation in a stochastic explicit-segment speech recognition system', ICASSP 93, pp V-642-645.

3. P. Kenny, G. Boulianne, H. Garudadri, S. Trudelle, R. Hollan, M. Lennig, D. O'Shaughnessy, 'Experiments in continuous speech recognition using books on tape', Speech Communication, Vol.14-1, Feb 1994, pp. 49-60.

4. P. Kenny, P. Labute, Z. Li and D. O'Shaughnessy, 'New graph search Techniques for speech recognition', ICASSP 94, pp I-553-556.

5. E. Lleida, J.B. Marino, J. Saravedra, A. Bonafonte, E. Monte and A. Martinez, 'Out-of-vocabulary word modeling and rejection for keyword spotting' Eurospeech 93, pp 1265-1268.

6. R. El Méliani and D. O'Shaughnessy, 'Lexical fillers for task-independent-training based keyword spotting and detection of new words', Eurospeech 95, pp 2129-2132.

7. R.C. Rose and D.B. Paul, 'A hidden Markov model based keyword recognition system', ICASSP 90, pp 129-132.

8. R.C. Rose and E.M. Hofstetter, 'Task independent wordspotting using decision tree based allophone clustering', ICASSP 93, pp II-467-470