ACCURATE KEYWORD SPOTTING USING STRICTLY LEXICAL FILLERS

Rachida El Méliani *Douglas O'Shaughnessy*

INRS-Télécommunications
16 Place du Commerce, Verdun (Île-des-Sœurs), H3E 1H6, Québec, Canada
meliani@inrs-telecom.uquebec.ca

ABSTRACT

Our goal is to design an accurate keyword spotter that can deal with any size of keyword set, since the size actually required in a wide range of applications is large (number of airports, number of names in a directory, etc.). This justifies the choice of an architecture based on a large-vocabulary continuous-speech recognizer. In a previous paper [1] we introduced the use of strictly-lexical subword fillers for keyword spotting based on the INRS large-vocabulary continuous-speech recognizer [2] showing that they are, when compared to acoustic fillers, a good compromise between memory and time consumption, keyword choice freedom and task-independence training on one hand and accuracy on the other hand. We propose here two new high-performance designs of individual strictly-lexical subword fillers that perform, this time, better than their acoustic counterparts while still keeping the mentioned advantages.

1. INTRODUCTION

Some continuous-speech-recognizer-based keyword spotters [4] use acoustic fillers to make the distinction between keywords and out-of-vocabulary words: they define two sets of subword models, one trained on the occurrences of all keywords in the training corpus while the other learns on all out-of-vocabulary speech. However in such an architecture the training is dependent on the keyword vocabulary. Moreover, to be efficient in terms of acoustic discrimination between the two kinds of words (keywords and extraneous words), this representation architecture needs:

1- That enough occurrences of the subwords composing keywords be available in the training corpus in order to get well-trained keyword subword models; that means a limitation in keyword choice.

2- That the intersection between the two sets of subwords respectively used for keywords and extraneous-speech models is small enough. For instance, short keywords may often be part of out-of-vocabulary word sequences, thus their associated models will represent those parts of extraneous speech too, leading to false alarms or deletions. This adds another restriction to keyword choice.

For all these reasons we introduced the use of strictly lexical fillers [1]: the two kinds of words are this time represented by a unique set of context-dependent phoneme models trained on the whole corpus. The discrimination between them is performed through the lexical graph as well as the language model. Thus the training-part of the keyword spotter is task-independent, while the detection-part consumes less memory and time for model-score determination than when the acoustic fillers were used for discrimination.

The use of individual strictly-lexical subword fillers with an adequate language model instead of a background word model [6] is motivated by the importance of the language-specific lexical constraint brought by subword unigram or bigram frequencies. We present here two high-performance individual strictly-lexical subword filler architectures differing in the orthography of the fillers in the lexicon: the first one is phonemic-based while the second one is syllabic-based.

2. KEYWORD SPOTTER DESCRIPTION

2.1. The INRS Continuous-Speech Recognizer

Our keyword spotter is based on the INRS continuous-speech recognizer [2] which is an HMM-based real-time very-large-vocabulary continuous speech recognizer. An overview of this recognizer is necessary to the understanding of the final system.

This recognizer processes the input speech block after block, the output beam of a block becoming the input beam of the following one. The lexicon presents for each word orthography all the different corresponding pronunciations. The system transforms the lexicon into an ordered lexical tree; only phoneme sequences belonging to this graph will be recognized. From this lexical tree, with the use of the computed table of context-dependent phonemes scores (B*), phonetic transcriptions are scored through the two passes; then with the use of the given language models, the most probable word strings are derived.

The INRS recognizer used here computes language models based on the deterministic back-off form from bigram distributions P($w_i|w_N$), and unigram distributions P(w_i), where w_i is the considered word and w_N the preceding one in its history. The language model score contribution to the final score is given through the formula:

$$score = logP_{HMM} + \alpha logP_{LM} + \beta \qquad (1)$$

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.

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 number of keywords while q is the number of fillers.

The main parts that will be modified in the design of our spotter are the training part, the lexical tree and the language models.

2.2. Acoustic fillers

We first tried acoustic fillers, guided by previous work [4]. We thus defined two sets of context-dependent phoneme HMMs trained as described in the introduction. In addition, those fillers, to be usable in the INRS recognizer, must have their definition completed by adding to the lexicon orthographic fillers representing all out-of-vocabulary words (table 1).

2.2.1. Phonemic fillers

Our first idea, inspired from the use of acoustic context-dependent-phoneme models has been to construct the orthographic fillers using only isolated phonemes as phonetic transcriptions. We propose here an improvement of the unique phonemic filler described in [1] by defining a set of 40 orthographic fillers: one filler for each phoneme as shown in table 2. We will refer to them as *"individual acoustic phonemic fillers"* (IAP).

2.2.2. Syllabic fillers

However, due to the strong lexical constraint the syllable imposes on phoneme strings and to the fact that a minimum-sized word is a one-syllable word, we designed a new set of orthographic fillers: one filler for each syllable. We will call them, in this paper, *"individual acoustic syllabic fillers"* (IAS).

Thus, phoneme sequences will be ruled in a deterministic way, related to the language structure, and dealing with one-phoneme-long phoneme sequences as "e" to six-phoneme-long ones like "franks" or "dwardz" , instead of following a statistical criterion (language models) that may allow unrealistic phoneme sequences as "kslp" for example.

word	phon.transc.
filler1	phoneme1
filler2	phoneme2
.	.
.	.
filler40	phoneme40

Table 2. Individual phonemic filler

As a matter of fact, [3] recalls that the history of writing enhanced the importance of syllables in the transcription of major world languages, while Segui and al. [5] demonstrated that the syllable is a fundamental unit of speech perception and processing.

2.3. Strictly-Lexical Subword Fillers

The training of acoustic fillers is dependent on the keyword set, since each time this set changes the training has to be performed again. To avoid the important loss of time required by retraining, and obtain a system more flexible to keyword changes, as well as to get total freedom of keyword choice, we propose to represent the kinds of words by a unique set of context-dependent phoneme models trained, this time, on the whole training corpus.

We thus withdraw the acoustic discrimination between the two kinds of words while keeping only the lexical and language model discriminations. Such a design leads to a new system using less memory for model and lexical-tree computation, while being less time consuming because there are fewer models to compute and score.

We thus define as before *"individual strictly-lexical phonemic fillers"* (ISLP) as well as *"individual strictly-lexical syllabic fillers"* (ISLS).

2.4. Language Models

The bigram and unigram distributions of the fillers are computed on the occurrences of out-of-vocabulary words in the training corpus.

3. EXPERIMENTS

3.1. System parameters

The system samples speech at 16 kHz using a block size of 25 ms as well as a block shift of 10 ms. The coefficients used are 15 static and dynamic MFCC. The two passes use the same acoustic models: three-state right-context phoneme HMMs, with all distributions sharing the same covariance matrix as well as a set of 256 means.

The terms α and β are set to different values for each one of the two passes. For memory limitation reasons, the INRS recognizer used here has been simplified to allow our spotter to be usable for all the proposed fillers, especially the acoustic ones. This results in a decrease of the recognition rate (80% for Wall Street Journal), and obviously affects the detection rate.

The four terms, α's and β's, as well as the two beamwidths, one for each pass, are taken as open parameters of our keyword spotter, that have to be set for each vocabulary.

We gathered from the dictionary database 10536 syllables to be used in fillers.

3.2. Database and Vocabularies

Test results are reported for the Wall Street Journal database already described in [1]. The training set is 172.6 minutes long; it contains 4131 different words. As for the test set, it is 21 minutes long and contains 984 different words, from which 260 are not in the training set.

As no specific task is targeted here, and in order to keep our results as general as possible, we define six different vocabularies, the size of which range from 10 to 99 words of

Name	IAP (%)	fa rate	ISLP (%)	fa rate
DIGI	75	8	83	4.2
NBRE	92.1	3.2	89.3	2.6
ONBR	92	7.6	90.4	4.6
FWOR	88.4	1.7	94	2.9
VFWO	89.8	2.8	94.2	4.4
VFW+	92.5	1.3	92.1	5.8

Table 3. Results for individual phonemic fillers for less than 10 fa/h/kw.

variable frequencies in the training corpus, to perform our experiments on:

- DIGI includes the ten digits. Their frequencies in the training set range from 8 (word "zero") to 154 (word "one") with an average of 90. The total number of their occurrences in the test set is 128. Most of those words have a one-syllable length.
- NBRE includes all the 51 ordinal and cardinal numbers available in the database; their frequencies vary from 1 to 154. They occur 299 times in the test set.
- ONBR is the subset of NBRE containing 32 ordinal numbers. They are found 284 times in the test set. Cardinal numbers are among the closest derived forms (i.e., words accepting keywords as subwords: genetive forms, plurals, etc.) of the ordinal numbers.

	IAS	ISLS	IAP	ISLP
score (%)	91.4	94.3	88.3	90.5

Table 4. Average results.

- FWOR contains 99 words of frequency greater than 10. They are present 345 times in the test set.
- VFWO is a list of 23 very frequent (more than 30 times) words that occur 239 times in the test set.
- VFW+ is an extension of VFWO where the derived forms of its keywords are added. The 56 words have frequencies ranging from 1 to 191 (word "dollars") and are present a total of 234 times in the test set.

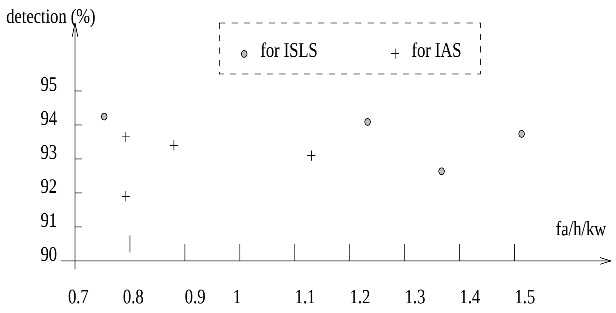

Figure 1. Detection score variation for syllabic fillers for VFW+.

Name	IAS (%)	fa rate	ISLS (%)	fa rate
DIGI	87.8	2.8	91.5	3.8
NBRE	91.5	1.2	92.3	1.4
ONBR	91.4	1.5	95.8	2.1
FWOR	90.9	1	95.7	1.5
VFWO	93	2.3	96.1	2.8
VFW+	93.8	.7	94.3	.74

Table 5. Results for individual phonemic fillers for less than 10 fa/h/kw.

4. EXPERIMENTAL RESULTS

4.1. False Alarm Rate

The results of our experiments are reported in tables 3, 4 and 5 for a false-alarm rate lower than 10 false alarms per hour per keyword (fa/h/kw). However, in fact the false-alarm rate is much lower for the syllabic fillers (less than 4) than for the phonemic ones.

The best detection values correspond to nearly the same beamwidth values independently of the choice of the vocabulary or the kind of parameters. Moreover, it is interesting to note that the lowest false-alarm rate for phonemic strictly-lexical fillers is obtained for a null flat distribution term β of the language models of both passes. In fact, the false-alarm rate increases drastically with β while the detection score increases slightly (see figure 2). However this term has very slight influence on detection scores obtained for syllabic fillers (see figure 1), which shows that the syllabic fillers have a more stable behavior than the phonemic ones.

In fact, the detection scores of our keyword spotter are not proportional to the false-alarm rate. The range of false-alarm rate is different for each kind of filler and each vocabulary. Thus detection rates in tables 3, 4 and 5 are given for the best corresponding false-alarm rates.

4.2. Filler Comparison

We see, at first glance, that the results highlight the obvious superiority of individual strictly-lexical syllabic fillers over all the others studied, as well as the pertinence of syllabic fillers when compared to phonemic ones. Furthermore, individual strictly-lexical phonemic fillers are found mostly to

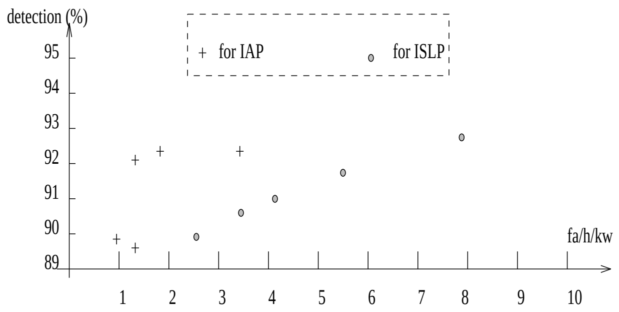

Figure 2. Detection score variation for phonemic fillers for VFW+.

be slightly more performing and less time and memory consuming than their acoustic counterparts while leading to a higher amount of false alarms: we can then conclude that they are a fair compromise between memory and time consumption, keyword choice freedom and task independence training on one hand and accuracy on the other hand.

4.3. Analysis of the results

4.3.1. Effect of the Keyword Frequency

A deeper analysis of our results shows that the ISLS fillers perform better on vocabularies where words are frequent; it is mostly due to the language model effect that will favor frequent words rather than syllabic fillers.

4.3.2. Effect of the Keyword Length

The lowest results are obtained for DIGI, the vocabulary of digits. They can be first explained by the presence of homonyms of words "two" and "four" in the extraneous speech (words "to" and "for"), which are twice more frequent than the original keywords, thus leading to an important increase in the false alarm rate.

It is secondly due to the fact that, as the words of DIGI are mostly very short, it often happens that they are parts of the out-of-vocabulary word sequences (for instance nine appears in ninety) and are more likely to be misdetected (commonly false alarms include insertion of keywords as well as keyword substitutions).

When digits are combined with all other numbers (NBRE, ONBR), some of the previous misdetections are corrected by the new bigram repartition.

Here again, we notice the relevance of the individual syllabic design and especially the higher performance of the individual strictly-lexical syllabic fillers for such a difficult task as detection of digits (DIGI).

4.3.3. Effect of the Derived Words

A close look at the output word sequence obtained shows that false alarms of keywords may occur whenever a derived word is found; that remains a classical problem for all kinds of keyword spotters.

Our system used with individual strictly-lexical syllabic fillers thus detected in the same file, for VFWO :

- "yesterday" instead of "yesterday's" in its only occurrence,
- "company" instead of "company's" three times among five,
- "market" instead of "marketing" three times among four,
- "stock" instead of "stocks" once among two occurrences,

which leads to an average of 8 among 12 (66%) false alarms related to derived words.

In VFW+ those derived words are added to the keyword set. The results reveal that some (4) of the previous false alarms are completely corrected, while others (3) are found as substitutions of keywords, that are still false alarms; however they can be verified in an additional process. Nevertheless some new insertions of VFWO words happened.

4.3.4. Effect of Keyword Number

The number of words does not seem to have an effect on the performance of the system. The shorter vocabulary is DIGI (10 words), and its low performance has already been explained by means of word length effects, and the presence of derived words and homonyms. Moreover, the vocabulary VFWO, which is the second shortest one (26 words), is among the most performing.

However, the necessity, for any kind of spotter, of adding derived words to the keyword set in order to decrease the number of false alarms, implies that very small vocabularies are not the best deal.

5. CONCLUSION AND PERSPECTIVES

This paper describes the use of two new efficient individual strictly-lexical fillers in keyword spotting. Our results show that, while individual strictly-lexical phonemic fillers often give better scores that their acoustic counterparts, individual strictly-lexical syllabic fillers always perform far better than the corresponding acoustic fillers.

We thus designed an accurate keyword spotter combining task-independence for training, reasonable memory and time consumption as well as total keyword choice freedom. Therefore the superiority of the lexical filler architecture to the acoustic design is clearly demonstrated.

Moreover, we found the syllabic design more performing in both topologies than the phonemic one, thus highlighting the importance of the syllable for English in that case.

We investigated the effects of keyword frequency, keyword length and the presence of derived words on our system and proposed some solutions to overcome bad effects.

Further work is concerned by the use of those fillers to improve our unknown-word detector [1].

REFERENCES

[1] 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.

[2] P. Kenny, G. Boulianne, H. Garudadri, S. Trudelle, R. Hollan, Y.M. Cheng, 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.

[3] P. Ladefoged, "A course in phonetics", Hartcourt Brace Jovanovich Inc., chapter 10.

[4] R.C. Rose, "Keyword Detection in Conversational Speech Utterances Using Hidden Markov Model Based Continuous Speech Recognition", *Computer Speech and Language*, volume 9 (1995), pp. 309-333.

[5] J. Segui,E. Dupoux and J. Mehler, "The Role of Syllable in Speech Segmentation, Phoneme Identification, and Lexical Access", in *Cognitive Models of Speech Processing*, Altmann, editor, chapter 12, pp. 263-280.

[6] M. Weintraub, "Keyword-Spotting Using SRI's $DECIPHER^{TM}$ Large-Vocabulary Speech-Recognition System", *Proceedings IEEE ICASSP*, 1993, pp. II-463-466.