POWERFUL SYLLABIC FILLERS FOR GENERAL-TASK KEYWORD-SPOTTING AND UNLIMITED-VOCABULARY CONTINUOUS-SPEECH RECOGNITION

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

Since the number of vocabulary words is often very large in both general-task keyword spotting and unlimited-vocabulary continuous-speech recognition, we choose to represent, unlike other teams, vocabulary words and out-vocabulary words with the same set of subword HMMs. Secondly we replace the classical one-phoneme transcription of fillers in the lexicon by a new, more powerful one-syllable transcription. Two different architectures are studied for the two kinds of application and their results are compared to our multiple phonemic fillers. As for the language model, the problem produced, in the case of unlimited-vocabulary continuous-speech recognition, by the lack of information on new words in the training corpus is solved through the use of the limited information we gathered on new words. The results obtained in both applications demonstrate the efficiency of the choice of a one-syllable transcription rather than a one-phoneme one. As for the results in unlimited-vocabulary continuous-speech recognition, the language model using information from words of frequency one is demonstrated to be a new promising method of determination of a language model for new words.

1. INTRODUCTION

In spontaneous speech processing unknown word detection has long been considered independently and differently from keyword spotting. However since the number of vocabulary words is very large in most keyword spotting applications (number of airports, number of names in a directory, etc.), the only differences between the two fields are, first, the kind of words to be detected, keywords in one case, and unknown words and/or vocabulary words in the other, and, secondly, the availibility or not of those words in the training corpus.

Because most known large-vocabulary continuous speech recognizers (SRI, LIMSI, Philips, HTK, Carnegie Mellon University, INRS,...) follow a classical architecture based on the use of subword HMMs as acoustic models, a lexical tree and a language model, we performed the design of our systems on the INRS speech recognizer [7, 8]. Thus, since the number of vocabulary words is often very large in both applications, we choose to represent, unlike other teams [12, 11, 5, 1, 2, 6], vocabulary words and out-of-vocabulary words with the same set of subword HMMs, and discriminate between them mostly at the lexical and language levels. Secondly, we replace the classical one-phoneme-transcription [12, 1] of fillers in the lexicon by a new, more powerful one-syllable-transcription, relying on the strong lexical constraints the syllable imposes on phoneme strings and on its importance in speech. Two different architectures are proposed for the two kinds of application and their results are compared to multiple phonemic fillers [10].

As for the language model, the lack of information on unknown words in the training corpus imposes the use of a different process than the one used for keyword spotting where a number of representatives of keywords as well as out-of-vocabulary words are both available in this corpus. Therefore we designed the unknown-word related language model by using the limited information we gathered on unknown words.

2. DESIGN OF THE SYSTEMS

Our two systems, the keyword spotter as well as the continuous-speech recognizer with unknown-word detector, are constructed using the INRS real-time very-large-vocabulary continuous-speech recognizer [7, 8] because it is designed with a global classical architecture that resembles most known large-vocabulary continuous-speech recognizers. It uses context-dependent phoneme HMMs as acoustic models and language models based on N-gram statistics.

2.1. Filler architecture

The superiority of fillers defined at the lexical and language-model levels only has been clearly demonstrated in the case of keyword spotting [9] when compared to the classical acoustic filler designs, where out-of-vocabulary-word related acoustic models and keyword-related acoustic models are different. This result applies to new-word detection too because of the high coverage of context-dependent phonemes by the large number of vocabulary words. We thus perform the discrimination between vocabulary words and out-of-vocabulary words at the lexical and language model levels only (see the lexicon general format in table 1).

We compare here the performances of two kinds of syllabic fillers to the classical phonemic fillers [10] noted "multiple phonemic fillers" or "phonemic fillers with a

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.

unique transcription" (PFUT) refering to the fact that each filler has only one phonetic transcription represented by a unique phoneme. The 40 English phonemes we used are thus divided among 40 fillers, one for each filler.

The syllabic fillers have each of their phonetic transcriptions constituted by a unique syllable. A strong lexical constraint is thus imposed on phoneme strings whereas for "phonemic fillers with a unique transcription" phoneme strings are following a statistical constraint given by the language model.

The first kind of syllabic fillers ("syllabic fillers with a unique transcription" or SFUT) consists of the division of all syllables among fillers, one syllable for each filler. This allows a better representation of out-of-vocabulary words occuring in the training corpus through quite accurate unigram and bigram frequencies of those fillers. The second kind ("syllabic fillers with multiple transcriptions" or SFMT) divides all syllables between fillers according to their frequencies in the database: each filler gathers as phonetic transcriptions only syllables occuring with the same frequency. This way, part of the language model is already taken in account in the lexicon. This is useful for unknown word detection because of the unavaibility of out-of-vocabulary words in the training corpus.

2.2. Language Models

In a keyword spotting task we use unigram and bigram frequencies computed on the whole training corpus. Filler frequencies are computed from occurrences of out-of-vocabulary words while keyword frequencies are obtained from keyword occurrences in the same corpus. However, in unknown-word detection, out-of-vocabulary words are absent from the training corpus. Nevertheless, we can gather a few cues on unknown words [3, 4]. Our language models are based on simpler ones.

We first defined a simple language model (LM1) by computing the frequencies related to vocabulary words on the whole training corpus, while the ones corresponding to fillers are obtained from a transformation of this corpus where *all words* are converted to fillers. Thus, here, the frequencies of vocabulary-related subunits are supposed simply identical to those of new-word-related subunits. Thus subword (phonemic or syllabic) partition is considered as being the same for vocabulary words as well as out-of-vocabulary words even if, in reality, the last ones have often a different structure than frequent words (new names, new roots, etc). However this method allows a good coverage of classes of out-of-vocabulary words derived from (or with the same roots as) vocabulary words.

In the second kind of language model (LM2) we keep for filler representation *only unigram frequencies* drawn from LM1. No bigram frequencies are defined for fillers. The ones found in LM1 are much more specific to vocabulary-word composition.

Finally, because *words of frequency equal to one* in the training corpus can be viewed as having been potential new words in a previous step, we can consider them suitable to represent more accurately the behaviour of new words. Thus, in this third language model (LM3), the frequencies corresponding to fillers are computed on a modified training set where those low-frequency words are replaced by fillers.

3. EXPERIMENTAL SETTING

3.1. Syllabic Fillers

As no list of syllables was available, we created ours by gathering all syllables present in the transcription of our complete database vocabulary. The 10536 syllables gathered are divided between 165 syllabic fillers with multiple transcriptions.

3.2. Vocabularies

The tests reported in this paper concern Wall Street Journal (noted here WSJ), already described in [9]. All the experiments reported here were using vocabularies extracted from the Wall Street Journal.

As no specific task is targeted here for keyword spotting, and in order to retain for our results as much generality as possible, we define six different vocabularies, the size of which range from 10 to 99 words of variable frequencies in the training corpus, and of variable sizes, more or less confusable, 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 and two of them ("two" and "four") have very frequent homonyms ("too", "to" and "for") in out-of-vocabulary speech. The minimum number of occurrences of these words is lower than for all the five other vocabularies. This vocabulary is the smallest one and the most difficult to detect among the ones studied here.
- 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. This vocabulary includes all words of DIGI.
- ONBR is the subset of NBRE containing 32 ordinal

Name	SFMT	fa	PFUT	fa	SFUT	fa
DIGI	77.3	1.93	83	4.2	91.5	3.8
NBRE	86.63	.76	89.3	2.6	92.3	1.4
ONBR	88.04	1.03	90.4	4.6	95.8	2.1
FWOR	83.19	.91	94	2.9	95.7	1.5
VFWO	86.61	1.92	94.2	4.4	96.1	2.8
VFW+	82.47	.58	92.1	5.8	94.3	.74

Table 2: Results for keyword spotting for less than 10 fa/h/kw.

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.

- 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.

As for unknown-word detection, tests use the whole vocabulary from which 218 words (vocabulary WSJ1) not appearing in the training corpus and the frequencies of which equal one in the test corpus are removed and then considered as new words.

4. EXPERIMENTS

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 (76% for WSJ) and will obviously affect the detection rate.

4.1. Keyword Spotting

The results of the experiments performed on keyword spotting are shown in tables 2 and 3: The detection rate is given in % for a false alarm rate in fa/h/kw. 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 these tables are given for the best corresponding false-alarm rates. In table 3 "score1" is the average score on all vocabularies but DIGI.

These results show that SFUT performs very well with all kinds of vocabularies, even with DIGI. They outperform the two other studied types of fillers. *Using syllables (bigger subunits) leads to a decreased false alarm rate* while using independent (unique) phonetic transcriptions increases the detection rate because the language model is more accurate.

On the other hand DIGI has led, because of the characteristics specified above, to the lowest detection rates

	SFMT	PFUT	SFUT
score (%)	84	90.5	94.3
score1 (%)	85.4	92	94.8

Table 3: Average results for keyword spotting.

for all three kinds of fillers, however the differences with other vocabulary scores is less important for SFUT. The improvement given by a highest frequency of keywords is clear for FWOR and VFWO. As for the occurrence of derived words in the vocabulary, it generally decreases the insertion number while increasing the sustitution rate; therefore its effect on detection rates and false alarm rates is not constant.

4.2. New-Word Detection

We note here a difference between total detection on one hand, when a correct occurrence of a new word is found together with its correct frontiers, and on the other hand partial detection, when the occurrence is correctly detected but with a partial frontier only.

Thus, when parts of the new word are already present in the vocabulary, for instance when new words are derived forms of some vocabulary words, the partial detection will give enough information. However, in that case, phonetic transcription is harder to get than in total detection.

We thus define the total detection rate, TD, as the ratio in % between the number of total detections and the number of new words in the file. D, the partial detection rate, is the ratio in % between the number of partial detections and the number of new words in the file.

The false-alarm rate, FA, is defined here as being the ratio in % between the number of false alarms in the file and the number of vocabulary words in the same file.

The phonetic transcription rate, PT, is evaluated here by the ratio in % between the number of phonemes detected correctly and the total number of phonemes in the chosen phonetic transcription. Moreover, we provide here the detection rate of vocabulary words, Det, as well as their recognition rate, Rec.

Results of the experiments are reported in tables 4, 5 and 6. The highest values are enhanced in bold face. Since our goal is to achieve a continuous-speech recognizer which, in a single pass, takes account of unknown words and includes their phonetic transcription to the dictionary, the protocol followed by the comparison of the results will consider first the best Det and Rec. Then we look for the best compromise between D and FA before considering TD and D.

We can see on tables 4, 5 and 6 that, differently from the results obtained for keyword spotting, the best performances are obtained with SFMT: *it still highlights the importance of the syllable* while showing that, this time, because of the lack of information on unknown words, a partially accurate language model (syllables divided ac-

	Det	Rec	D	TD	PT	FA
LM1	72	**67**	83	35	60	31
LM2	66	61	85	42	**65**	37
LM3	**74**	63	**88**	35	60	**28**

Table 4: Test results for PFUT in unknown-word detection.

	Det	Rec	D	TD	PT	FA
LM1	50	48	**90**	**90**	**71**	29
LM2	**70**	**65**	**90**	70	70	11
LM3	66	64	**90**	83	70	**9**

Table 5: Test results for SFMT in unknown-word detection.

cording to their frequency in the language) performs better than one more detailed (individual syllables).

The best compromise is obtained for SFMT used with LM2 or LM3 with quite satisfying values, followed by the combination of PFUT with LM3, or when SFUT is used with LM3. LM3 seems then to bring noticeable improvement with all fillers. *The information brought by vocabulary words with frequency equal to one in the training corpus is shown to characterize efficiently unknown words in terms of language modeling.*

Finally we notice that the detection rates obtained with SFMT combined with LM3 are higher than the ones reported by other designers [1], [2], [6]. Moreover phonetic transcription given by the same combination is rather good.

5. CONCLUSION

This paper studies three different architectures of fillers for the two major applications of spontaneous-speech processing: general-task keyword spotting and unlimited-vocabulary continuous-speech recognition. Then, in the case of unlimited-vocabulary continuous-speech recognition, because of the lack of information on unknown words in the training corpus, we define three different methods to evaluate the unknown-word related statistics used by the language model with the help of the limited information we gathered on unknown words.

The results obtained in both applications demonstrate the efficiency of the choice of a one-syllable transcription rather than a one-phoneme one. The syllabic fillers with a unique transcription outperform all the other types. As for the results in unlimited-vocabulary continuous-speech recognition, the best performances are reached with the syllabic fillers with multiple transcriptions used with LM3, followed by those with LM2. The best system obtained with LM3 detects new words with an accuracy of 90%, their phonetic transcription with an 83% rate and only a 9% false alarm rate, while keeping a relevant recognition rate. LM3 is thus demonstrated to be a new promising method of determination of a language model for new words.

	Det	Rec	D	TD	PT	FA
LM1	51	50	**83**	56	72	38
LM2	70	67	80	**60**	69	27
LM3	**78**	66	75	35	**75**	**14**

Table 6: Test results for SFUT in unknown-word detection.

6. REFERENCES

1. A. Asadi, R. Shwartz, J. Makhoul, "Automatic Detection of New Words in a Large Vocabulary Continuous Speech Recognition System", Proc. ICASSP 1990, pp.125-128.
2. P. Fetter, A. Kaltenmeier, T. Kuhn, P. Regel-Brietzmann, "Improved Modeling of OOV Words in Spontaneous Speech", Proc. ICASSP 1996, pp. 534-537.
3. I.L. Hetherington et V.W. Zue, "New Words: Implications for Continuous Speech Recognition", *Proc. EUROSPEECH* 1993, p. 2121-2124.
4. I.L. Hetherington, "New Words: Effects on Recognition Performance and Incorporation Issues", *Proc. EUROSPEECH* 1995, p. 1645-148.
5. E.M. Hofstetter, R.C. Rose, "Techniques for Task Independent Word Spotting in Continuous Speech Messages", Proc. ICASSP 1992, pp.II 101-104.
6. A. Jusek, G.A. Fink, F. Kummert, H. Rautenstrauch, G. Sagerer, "Detection of Unknown Words and its Evaluation", Proc. EUROSPEECH 1995, pp.2107-2111.
7. 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.
8. P. Kenny, P. Labute, Z. Li and D. O'Shaughnessy, "New graph search Techniques for speech recognition", ICASSP 94, pp I-553-556.
9. R. El Méliani and D. O'Shaughnessy, "Accurate Keyword Spotting Using Stricly Lexical Fillers", *Proc. ICASSP* 1997, Vol.2 p.907-910.
10. R. El Méliani and D. O'Shaughnessy, "New Efficient Fillers for Unlimited Word Recognition and Keyword Spotting", *Proc. ICSLP* 1996, p.590-593.
11. J.R. Rohlicek, P. Jeanrenaud, K. Ng, H. Gish, B. Musicus, M. Siu, "Phonetic Training and Language Modeling for Word Spotting", Proc. ICASSP 1993, pp. II 459-462.
12. 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.