SPECIFIC LANGUAGE MODELLING FOR NEW-WORD DETECTION IN CONTINUOUS-SPEECH RECOGNITION

Rachida El Méliani and Douglas O'Shaughnessy

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

ABSTRACT

The objective of this work is to allow the INRS continuous-speech recognizer to process accurately new words and incorporate them into the vocabulary. Until now only a few new-word detectors have been reported, all of them defining an acoustic filler model different from the models used to represent vocabulary words. In this paper, we define several designs using, unlike other researchers, strictly-lexical fillers and a unique process to perform speech recognition, new-word detection and new-word phonetic transcription. Moreover, we propose here four different types of language models differing in the way they use the limited information we gathered on new words. The best combinations are found to be different from the ones we obtained for keyword spotting.

1. INTRODUCTION

Speech recognizers are submitted to many constraints; one of the most important ones in use with spontaneous speech is vocabulary limitation since it is largely admitted that users of such systems do not restrict their speech to the selected vocabulary. Even increasing the vocabulary is not an efficient solution: interrupted words, mispronounced ones and the continuing mass of new names will not be covered [1].

However the problem of distinguishing between vocabulary words and out-of vocabulary words is more difficult for new-word detection than for keyword spotting, since in keyword spotting most out-of-vocabulary words are already present in the training corpus, and may thus be used to obtain task-related models to represent those words, while for new-word detection, on the contrary, the training corpus includes only vocabulary words, but no information on new words.

Until now, only a few new-word detectors have been reported. In 1990 Asadi and his coauthors [2] introduced first designs using an explicit HMM for the filler modeling the new words. Young [3] used the same model adding definitions of confidence measures. Then Fetter [4] used 15 word models to represent new words while vocabulary words are modeled with diphone models. As for Jusek [5], he preferred representing vocabulary words with context-dependent phoneme models while reserving context-independent phoneme models to define the fillers. So, all of those new-word detection systems define an acoustic filler model different from the models used to represent vocabulary words with still modest detection performance, and without new-word phonetic transcription.

The aim of this work is to extend INRS continuous-speech recognizer applications [6] by adding the processing of new words and their incorporation into the vocabulary. In this paper, we define several designs using, at the opposite of the other researchers, a unique process to perform accurate speech recognition, new-word detection and new-word phonetic transcription, and no acoustic discrimination.

The fillers will be defined at the lexical and language-model levels only, not at the acoustic level, since the set of context-dependent phonemes gathered from vocabulary words may be very close to the set of phonemes that may occur in potential new words because of the large number and the wide dispersion of the former, as well as of the unknown character of the latter and of their potential links with vocabulary words (derived words like subwords of vocabulary words or words accepting vocabulary words as subwords). The two architectures that have led to efficient keyword spotting [7] will be compared to other filler designs. We propose here four different types of language models differing in the way they use the limited information we have on new words. The best combinations will be shown to be different from those we obtained for keyword spotting.

2. SYSTEM DESCRIPTION

Our system is based on the INRS continuous-speech recognizer [6], which is an HMM-based real-time very-large-vocabulary continuous speech recognizer that processes the input speech block after block, in two passes based on Viterbi and $A^\star$ algorithms and using acoustic (context-dependent phoneme HMM) as well as language models.

2.1. Filler design

The acoustic models used to represent vocabulary words as well as unknown words are trained *on the whole training set*. Thus, only strictly lexical fillers are used, that is, fillers defined at the lexical and language-model levels only (see the lexicon general format in table 1). In [7] we showed the superiority in keyword spotting of this type of filler compared to the acoustic filler designs. Moreover, in new-word detection the number of vocabulary words is far larger and thus includes most of the context-dependent phonemes that are possible in the language. On the other hand, new words may contain any of those phonemes. Thus, there is no specificity of the acoustic models to any of the two kinds of words (in and out-of vocabulary). Moreover, new words are unavailable in the

keyword$_1$	phtr$_{k_1 1}$	...	phtr$_{k_1 n_1}$
$\vdots$	$\vdots$		$\vdots$
keyword$_p$	phtr$_{k_p 1}$	...	phtr$_{k_p n_p}$
filler$_1$	phtr$_{f_1 1}$	...	phtr$_{f_1 m_1}$
$\vdots$	$\vdots$		$\vdots$
filler$_q$	phtr$_{f_q 1}$	...	phtr$_{f_q m_q}$

Table 1: Lexicon general format. p is the number of vocabulary words while q is the number of fillers. "Phtr" specifies phonetic transcriptions associated with words.

training corpus. We thus believe that in that case there is no real improvement in a separate acoustic modeling and that the discrimination between the two types of words may be performed at the lexical and language model levels only.

The successful architectures defined in [7] construct the orthographic fillers using only one transcription for each filler, that is one phoneme for each filler in the *"individual phonemic fillers"*, or one syllable for each filler in the *"individual syllabic fillers"*. Their performances are compared to those obtained for the *"unique phonemic filler"* (see table 2) where the forty English phonemes are the phonetic transcriptions of this filler, as well as to those obtained for the *"syllabic fillers with multiple transcriptions"* where the set of syllables has been divided between all fillers according to the frequencies in the database: each filler accepts as phonetic transcriptions only syllables occuring with the same frequency.

filler	phoneme1	...	phoneme40

Table 2: Form of the unique phonemic filler in the lexicon.

This last type is motivated by a desire to stay compatible with the design of the chosen language models which uses unigram and bigram frequencies. We thus avoid a classical mistake that occurs each time two phonetic transcriptions corresponding to the same lexical word have very different frequencies, one high, the other low. In that case, the frequency attributed to the lexical word by the language model is the sum of the two frequencies, and will be thus given to both phonetic transcriptions in the score evaluation instead of their true frequencies.

2.2. Language Models

Because of the lack of information on new words as well as of their absence in the training corpus, it is rather difficult to find an efficient language modelling for the fillers. We propose here four different types of language models differing in the way they use the limited information we gathered on new words.

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.

The second kind of language model (LM2) is drawn from the previous one where, for fillers, *only unigram frequencies are kept*. Their bigram frequencies are dropped because they bring mostly false lexical information for new words since they are specific to vocabulary word composition.

In the third type of language model (LM3) we consider that *the words of frequency equal to one* in the training corpus are suitable to represent the behaviour of new words. Thus the frequencies corresponding to fillers are computed on a modified training set where those low frequency words are replaced by fillers.

Finally the last language model (LM4) uses the *results of the new-word analysis* performed by Hetherington and Zue [1]. The latter state that for any database the curve representing the logarithm of the vocabulary size V is an increasing function of the number of training sentences s which becomes linear after several hundred sentences, thus:

$$V \ \alpha \ s^{\gamma} \tag{1}$$

They state as well that the curve representing the logarithm of the new-word rate r is a decreasing function of s becoming, too, linear after several hundred sentences. That is:

$$r \ \alpha \ s^{\beta} \tag{2}$$

From these two equations we obtain:

$$r \ \alpha \ V^{\delta} \tag{3}$$

with

$$\delta \ = \ \frac{\beta}{\gamma}. \tag{4}$$

Thus the knowledge of V and δ leads to the evaluation of r. This factor may then be used to update in LM1 the filler frequencies.

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 experiments have been performed with the simplified INRS recognizer already used in [7], on the same Wall Street Journal database. The 10536 syllables gathered are divided between 165 syllabic fillers with multiple transcriptions. The tests use the whole vocabulary from which 218 words 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. Thus V equals 4656, and then r is about 13%.

3.2. Scoring methods

We found in the papers reporting new-word detection [2, 3, 4, 5] no precise or common definition of detection or the false alarm rate. 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.

In fact, those last ones are relevant in cases where parts of the new word are already present in the vocabulary, like when new words are derived forms of some vocabulary words for instance (e.g.: “decliners” and “decline”). However, in that case, the phonetic transcription is less easy to obtain than in the case of total detection: a multiple hypothesis verification taking in account all possibilities must be proposed to the user to select the correct new word. 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.

In this work the false-alarm rate, FA, is defined as the ratio in % between the number of false alarms in the file and the number of vocabulary words in the same file. False alarms due to pronunciations not available in the dictionary, though correct, are frequent (e.g.: “and” and its shortened pronunciation /n/).

As for the phonetic transcription rate, PT, it is the ratio in % between the number of phonemes detected correctly and the total number of phonemes in the chosen phonetic transcription. This definition may, nevertheless, be improved by allowing more freedom on the vowels that can be pronunced differently or even vanish, as well as by taking in account classical transformations due to coarticulation effects, like for example unvoicing of voiced sounds in unvoiced contexts. The extended phonetic transcription rate thus obtained is, here, nearly 15% higher than PT.

The detection rate of vocabulary words, Det, as well as their recognition rate, Rec, are given too. They are compared to the values obtained for the vocabulary words with the recognizer used with the whole dictionnary, that is, Rec equals 75% and Det 83%.

4. RESULTS AND DISCUSSION

4.1. Evaluation

The results are reported in tables 3, 4, 5, 6. The best values are highlighted in bold font. For comparison of those results we look first for the best Det and Rec since the system is judged acceptable if it loses only a little accuracy on vocabulary words. In fact, complete evaluation is based on the total recognition rate, TR, corresponding to the total vocabulary, and which is the average of D and Det, while the total recognition rate is drawn from TR and FA. Then a compromise between D and FA has to be found. PT is finally taken in account when TD or at least D is relevant enough.

4.2. The unique phonemic filler

An analysis of the results in table 3 shows that in the case of the unique phonemic filler the language model LM1 is the most efficent since it has some of the best vocabulary word detection and recognition rates, the best detection D of new words, and one of the best phonetic transcription rates with the least false alarm rates. LM2 follows with, however, the highest FA, then LM4 is next. We

	Det	Rec	D	TD	PT	FA
LM1	72	66	**86**	0	64	**12**
LM2	**74**	**68**	78	**14**	64	22
LM3	70	63	57	0	61	14
LM4	71	65	71	**14**	**67**	16

Table 3: Test results for the unique phonemic filler.

can conclude that, for that type of filler, information on new words does not improve the results because of the complete lack of specificity of this filler. Moreover, for all the proposed language model designs the total detection rate is very low. However LM1 with this filler forms a simple system with interesting performances.

4.3. The individual phonemic fillers

	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**
LM4	65	62	84	**65**	64	40

Table 4: Test results for the individual phonemic fillers.

The results in table 4 correspond to quite high false alarm rates. However they show good detection rates. Thus, except for TD, the results are lower than those obtained for the more general previous filler. It is obvious from the results that LM3 performs quite better than the others, followed by LM1. The high false alarm rate is related to the too small size of the phonetic transcription of fillers (phoneme), as we already found in the tests on keyword spotting [7].

4.4. The syllabic fillers with multiple transcriptions

For the syllabic fillers with multiple transcriptions (see table 5), detection rates D and DT are better and phonetic transcriptions rates are higher than for the previous ones. Moreover LM2 and LM3 obtain the least false alarm rates. These FA values can be considered interesting. However, Rec and Det are lower than for phonemic fillers. Their best values are obtained for LM2 and LM3. Therefore we can conclude that LM2 and LM3 are, with this kind

	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**
LM4	47	46	**90**	**90**	67	24

Table 5: Test results for the syllabic fillers with multiple transcriptions.

of filler, the most efficent for new-word detection but with a loss

in vocabulary word detection and recognition. Moreover the extended phonetic transcription is then around 85%, a quite relevant figure. A combination of LM2 and LM3 does not improve either the result of LM2 or that of LM3. The detection rates obtained with syllabic fillers with multiple transcriptions combined with LM3 are higher than those reported by other researchers with different architectures [2], [4], [5].

4.5. The individual syllabic fillers

We see in table 5 that the results of the individual syllabic fillers are mostly worse than for the previous filler. However LM3 shows here a nicer behaviour than the other language models. It even obtains the best Det and PT of all the designs proposed in this paper.

	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**
LM4	76	**70**	75	45	65	17

Table 6: Test results for the individual syllabic fillers.

4.6. General comparison

At the opposite of the results obtained for keyword spotting [7], the best performances have been obtained with the unique phonemic filler and the syllabic fillers with multiple transcriptions. The best compromise is obtained for the latter used with LM2 or LM3 with quite satisfying values, followed with the choice joining the unique phonemic filler to LM1, as well as the one combining individual phonemic fillers with LM3, or when individual syllabic fillers are used with LM3.

LM3 seems then to bring noticeable improvement with all fillers except with the unique phonemic one. The information brought by vocabulary words with frequency equal to one in the training corpus is thus demonstrated to be relevant enough in terms of language modeling of new words. The detection rates obtained with syllabic fillers with multiple transcriptions combined with LM3 are higher than those reported by other designs [2], [4], [5]. The extended phonetic transcription reached with the same combination is quite good.

As for LM4, it seems that even if there is a proportionality between r and the logarithm of V, this does not mostly apply to subword frequencies because vocabulary word subunits and new word ones have unrelated dispersions.

5. CONCLUSION AND PERSPECTIVES

In this paper we proposed several designs to convert the INRS continuous speech recognizer into a system performing in a single process speech recognition, new-word detection and new-word phonetic transcription. Four different architectures of fillers are combined with four variations of language modelling based on new-word specific information.

The best fillers are shown to be different from those obtained for keyword spotting [7]. The use of syllables allows here again a lower false alarm rate.

As for the language model, we conclude that LM3, the language model using words of frequency equal to one in the training corpus to represent the behaviour of new words, shows a satisfying behaviour for all the fillers except with the unique phonemic one, and especially for syllabic fillers with multiple transcriptions. An improvement of this model is under investigation.

6. REFERENCES

[1] I.L. Hetherington and V.W. Zue, "New Words: Implications for Continuous Speech Recognition", *Proc. EUROSPEECH* 1993, pp. 2121-2124.

[2] 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.

[3] S.R. Young, W. Ward, "Learning New Words from Spontaneous Speech", *Proc. ICASSP* 1993, pp. II 590-991.

[4] P. Fetter, A. Kaltenmeier, T. Kuhn, P. Regel-Brietzmann, "Improved Modeling of OOV Words in Spontaneous Speech", *Proc. ICASSP* 1996, pp. 534-537.

[5] A. Jusek, G.A. Fink, F. Kummert, H. Rautenstrauch, G. Sagerer, "Detection of Unknown Words and its Evaluation", *Proc. EUROSPEECH* 1995, pp. 2107-2111.

[6] 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.

[7] R. El méliani and D. O'Shaughnessy, "Accurate Keyword Spotting Using Stricly Lexical Fillers", *Proc. ICASSP* 1997, Vol.2 pp.907-910.