Available online at www.sciencedirect.com

SciVerse ScienceDirect

Computer Speech and Language 28 (2014) 141–162

COMPUTER SPEECH AND LANGUAGE

www.elsevier.com/locate/csl

An improved two-stage mixed language model approach for handling out-of-vocabulary words in large vocabulary continuous speech recognition☆

Bert Réveil*, Kris Demuynck, Jean-Pierre Martens

Ghent University – iMinds, ELIS Multimedia Lab, Sint-Pietersnieuwstraat 41, 9000 Gent, Belgium

Received 17 May 2012; received in revised form 23 March 2013; accepted 3 April 2013
Available online 17 April 2013

Abstract

This paper presents a two-stage mixed language model technique for detecting and recognizing words that are not included in the vocabulary of a large vocabulary continuous speech recognition system. The main idea is to spot the out-of-vocabulary words and to produce a transcription for these words in terms of subword units with the help of a mixed word/subword language model in the first stage, and to convert the subword transcriptions to word hypotheses by means of a look-up table in the second stage. The performance of the proposed approach is compared to that of the state-of-the-art hybrid method reported in the literature, both on in-domain and on out-of-domain Dutch spoken material, where the term 'domain' refers to the ensemble of topics that were covered in the material from which the lexicon and language model were retrieved. It turns out that the proposed approach is at least equally effective as a hybrid approach when it comes to recognizing in-domain material, and significantly more effective when applied to out-of-domain data. This proves that the proposed approach is easily adaptable to new domains and to new words (e.g. proper names) in the same domain. On the out-of-domain recognition task, the word error rate could be reduced by 12% relative over a baseline system incorporating a 100k word vocabulary and a basic garbage OOV word model.

Keywords: Out-of-vocabulary words; OOV detection; OOV modeling; Phoneme-to-grapheme conversion

1. Introduction

The performance of any large vocabulary continuous speech recognizer (LVCSR) is bounded by the finity of its vocabulary. Words outside that vocabulary are called out-of-vocabulary (OOV) words and cannot be recognized without additional processing steps. If such steps are not pursued, it is generally acknowledged that each OOV word occurring in the speech signal translates to multiple errors in the recognition output: an early measurement for English (Hetheringthon, 1995) counted 1.22 errors per OOV word, but currently, one typically anticipates an average of 1.5 to more than 2 errors per OOV word (Adda-Decker and Lamel, 2000; Bisani and Ney, 2005).

Given that computer resources (memory and computation speed) are nowadays less of a concern, there is a tendency to reduce the word error rate (WER) by simply including more words in the vocabulary. It has been argued (Rastrow

☆ This paper has been recommended for acceptance by A. Potamianos.
* Corresponding author. Tel.: +32 09 264 33 97; fax: +32 09 264 35 94.
E-mail addresses: breveil@elis.ugent.be (B. Réveil), kris.demuynck@elis.ugent.be (K. Demuynck), martens@elis.ugent.be (J.-P. Martens).

0885-2308/$ – see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.csl.2013.04.003

et al., 2009; Parada et al., 2010a) that this approach has a down-side, namely that it can raise the confusability among the in-vocabulary (IV) words, and as such end up in raising the WER. Moreover, it is impossible to create a vocabulary that contains all words that will ever be spoken to the recognizer during its deployment. This is especially unfortunate in case of unforeseen content words. In Demuynck et al. (2009), it has been observed that for a vocabulary size of more than 100k words, 90% of the OOV words in a large Dutch newspaper text corpus are content words such as compound nouns and proper names. It is clear that an incorrect recognition of these content words will hamper the comprehension or the further processing of the recognition output in applications such as automatic translation or spoken term detection. That is why pursuing the lowest possible OOV word rates is always important.

A first step in that pursuit is choosing an optimal set of lexical units. This choice is language-specific. English for instance is known to exhibit only small amounts of inflection and compounding, which implies that regular words make good lexical units. In highly agglutinative languages such as Finnish or Turkish on the other hand, morphemes are typically preferred over words. The usage of subwords as lexical units of course implies that the language model needs to be adjusted to be able to join subwords back to words. A complete description of such a system, including the automatic derivation of morphemes, can be found in Hirsimäki et al. (2006). Other languages such as Mandarin Chinese may even benefit from combining syllabic character-based lexical units with the more traditional word-based units (Hieronymus et al., 2009).

For a mildly generative language such as Dutch, it has been shown (Demuynck et al., 2009) that for vocabulary sizes of 100k entries or more, regular words make good lexical units: word inflections are effectively covered by vocabularies of such sizes, while OOV compound words can to a large extent be recovered by means of a post-processing step that merges individual successive words to form compounds.

However, as mentioned above, a large portion of the OOV words (in Dutch as well as in other languages) comprise proper names and other word categories such as foreign words that do not follow the language-specific morphological rules. Hence, OOV words remain a problem, irrespective of the lexical units that are chosen. In this paper, we therefore try to go beyond the specific optimization of the recognition vocabulary and conceive a methodology that aims to recover all types of OOV words.

Conceptually, most of such complete OOV word handling methods consist of the following three steps: (1) the detection (spotting) of regions in the speech signal that contain an OOV word, (2) the generation of a subword representation of the detected region, and (3) the generation of an orthographic representation for the detected region. From a practical viewpoint, the different steps may be inter-twinned. In their simplest form for instance, the second and third step together may just generate a filler symbol like "<OOV>" or an in-vocabulary word between brackets to alert the user that the word at that position is probably an unknown word. Using this approach, one could already obtain WER improvements: e.g. if the recognizer is forced to ignore presumed OOV word segments in its decoding of the surrounding speech, a recognition error caused by an OOV word might not propagate.

If the user is a human being, like a hearing-impaired person reading subtitles, it may also be acceptable to generate a pseudo-orthographic transcription of the OOV word, provided that this word obeys the pronunciation rules of the language and reads as the OOV word. Such a system could for instance output '[Perris]' for the capital of France, with the brackets indicating that it is a pseudo-transcription. In that respect, we refer to the work of Decadt et al. (2001) where alternative orthographic transcriptions were created for words in the recognition output that had a low word confidence score (most of these words were OOV words). Their system uses a phoneme recognizer incorporating a 5-gram phoneme LM to generate a phonemic transcription for these words, after which a phoneme-to-grapheme (P2G) converter is employed to generate a graphemic transcription. Although the obtained graphemic strings were usually not fully correct, the resulting speech transcriptions were easier to read than the original ones.

The ultimate goal of an OOV word handling method is however to produce a full and correct orthographic representation of every OOV word, and to recover most, if not all, of the errors in non-OOV regions emanating from the presence of OOV words.

In the rest of this paper, we first review some previously proposed OOV word handling methods that aim at detecting OOV word regions in the speech and at producing an orthographic transcription for these regions (Section 2). Our own OOV word detection and recovery method is introduced in Section 3. The approach combines the knowledge of an open-vocabulary word-level language model (LM) with that of a subword-level generative model of OOV words and a background pronunciation dictionary that was created in a fully automatic way. Section 4 discusses our experimental set-up. In Section 5, we conduct an experimental study (for the Dutch language) to assess the performance of the most popular existing method and our proposed method as a function of the control variables they embed for optimizing their

behavior. From this analysis we draw some general conclusions and identify possibly promising further extensions of this work in Section 6.

2. Formerly proposed OOV word handling methods

There exists a large body of literature on dealing with OOV words in automatic speech recognition. Not all work falls under the conceptual description of a complete OOV word handling method that we have presented in Section 1. In Geutner et al. (1998) for example, one gets around the fact that the recognition lexicon is limited in size by dynamically adapting the recognition vocabulary in a two-step recognition process. In the first stage, a regular word-based recognizer is used to create a word recognition lattice per sentence. In the second stage, a new recognition system is built per sentence, of which the vocabulary contains all words found in the lattice that was generated in the first stage, supplemented with the most likely words in the LM training text corpus that are acoustically similar to the lattice words. This reduces the OOV rate in the second stage, resulting in a 9% relative raise in word accuracy.

Although the above example proves that other approaches to OOV word recovery can be successful too, the subsequent literature overview focusses on two specific approaches: *mixed* and *hybrid* OOV word modeling.

2.1. Mixed OOV word modeling

So-called mixed approaches are all characterized by the fact that they attempt to detect and phonemically transcribe OOV words by including an explicit OOV word model into the vocabulary. Furthermore, an open-vocabulary word-level LM is typically used in which the occurrence of an OOV word (or by extension the occurrence of multiple classes of OOV words) is explicitly modeled. The mixed approach has been elaborated in e.g. Asadi et al. (1990), Kemp and Jusek (1996), Chung (2000) and Bazzi and Glass (2000, 2001, 2002). All cited papers focus on small vocabulary recognition tasks. Most of them do not provide a P2G step to convert the phonemic transcriptions of OOV word into orthography.

The work of the respective authors mainly differs in the way the OOV word model is conceived. In Asadi et al. (1990), that model allows for any sequence of context-independent phonemes of at least two phonemes long to represent an OOV word. Using an open-vocabulary class grammar and an optimized bias against OOV words to reduce the false alarm rates, 74% of the non-vocabulary words occurring in sentences of the 1k word DARPA naval resource management task can be detected at a false alarm rate of 3.4%.

Kemp and Jusek (1996) introduce a more complex subword-based generative model. It consists of a syllable loop containing 11k syllables that are derived from the CELEX dictionary containing phonemic word transcriptions with syllable boundary marks. Each syllable is conceived as a sequence of phoneme states. The syllables are divided into two classes, *reduced* and *non-reduced* syllables, and phonotactic rules concerning the occurrence of syllable class sequences are employed to weigh the syllable transitions in the loop. Within each syllable, phonemic bigram estimates retrieved from the phonemic transcriptions of a list of infrequent words occurring in a text corpus are used to determine the phoneme state transitions. Using a system with a vocabulary of 3k words and a bigram LM, the authors can spot 18.8% of the OOV words occurring in spontaneous continuous speech at a false alarm rate of 31%, without affecting the recognition of the in-vocabulary words. Since the false alarms typically occur at places where the baseline system already made a recognition error, the word error rate (WER) does not change significantly.

Bazzi and Glass (2000) elaborate a similar generative OOV word model incorporating a phoneme recognizer using bigram phoneme LM restrictions. It is used to detect OOV words in weather information queries. The model is added to a 2k word vocabulary recognizer employing an open-vocabulary bigram word LM. The bigram phoneme LM is trained on the phonemic transcriptions of 88k training utterances, which implies that the LM also models IV word transcriptions and cross-word bigrams, and that the word frequencies have a direct influence on the estimated bigrams. The authors employ a penalty for entering the OOV word model. For a zero penalty, they find that 47% of the OOV words can be spotted, without severely affecting the WER for the IV words. Positive penalties lead to higher detection rates, but also to higher false alarm rates and higher WERs. The authors refine their approach in subsequent work. In Bazzi and Glass (2001), the OOV word model is trained on the 90k LDC PRONLEX word dictionary, which discards cross-word phoneme bigrams and eliminates the impact of the word frequency on the bigram estimates. The

set of subword units is also expanded by introducing longer-length phoneme sequences. About 2k eligible phoneme sequences are identified following a bottom-up approach that uses a mutual information criterion to merge phoneme sequences. Two-thirds of the discovered subword units correspond to regular English syllables. The new OOV word model reduces the false alarm rate by 60% relative at a constant OOV detection rate of 70%. In Bazzi and Glass (2002), OOV words are grouped into 8 classes according to phonotactic similarities. With respect to a single OOV word model trained on PRONLEX and only using phonemes as subword units, the new approach yields a 45% relative decrease of the false alarm rate for a constant OOV detection rate of 70%.

Chung (2000) devises a three-stage recognition system to decode similar weather information queries containing out-of-vocabulary city names. In the first stage, her system relies on a domain-independent sub-lexical recognizer to generate acoustic-phonetic recognition lattices. The sub-lexical recognizer integrates linguistic knowledge from several subword levels (phonological level, phonemic level, syllabic level and morphological level) into a single FST grammar representation to aid its search. In the second stage, an open-vocabulary word recognizer incorporating a bigram language model that accounts for the occurrence of an OOV city name is used to generate initial word hypotheses. The second recognizer furthermore includes a subword level in the grammar that generates likely letter-phoneme (a subword unit existing of a grapheme and a corresponding phoneme) sequences for presumed OOV regions. The letter-phoneme combination ensures that an orthography can be generated for presumed OOV word regions. An empirically optimized OOV word penalty is used to control the amount of detected OOV words, while an OOV word length penalty is employed to make sure that the detected OOV word zones do not overlap with in-vocabulary sentence parts. In a third stage, an *N*-best word hypothesis list is rescored with an open-vocabulary 3-gram word LM in order to find the final hypotheses. Testing the three-stage system with respect to a baseline all-in-one word recognizer, the WER can be reduced by 29.3% relative.

2.2. *Hybrid OOV word modeling*

A nowadays popular approach to OOV word modeling employs a so-called hybrid LM that models sentences simultaneously in terms of words and subword units. Its *N*-gram probabilities are estimated from a text in which the OOV words are replaced by their constituent subword unit sequences. The main advantage of a hybrid LM is that it takes nothing more than a standard search method to generate a sequence of words and subword units, with the subword sequences representing the OOV words. This makes the mixed approach easily transferable towards large vocabulary recognition tasks, which is probably why it is nowadays the state-of-the-art approach. The hybrid approach has been elaborated in e.g. Klakow et al. (1999), Bisani and Ney (2005), Rastrow et al. (2009) and Parada et al. (2010a,b).

Klakow et al. (1999) were among the first to use the hybrid approach. Their subword units are individual phonemes and multiphonemic fragments (diphones, triphones, etc.) that are discovered automatically as frequently occurring fragments in the phonemic transcriptions of infrequent words. An optimal number of subword units are identified by minimizing the perplexity (with respect to the hybrid LM) of training text parts containing an OOV word in any one of the positions covered by the *N*-gram span. Three penalties are introduced to control the OOV detection rate: a word insertion penalty (between two words), a garbage insertion penalty (between a word and a subword unit) and an OOV length penalty (between two subword units). For a vocabulary of 5.5k words, a reduction of the WER for IV words of 16% relative is achieved with a trigram LM (11% with a bigram model). This means that the suggested approach leads to an increased IV word accuracy in the regions surrounding the correctly identified OOV words. However, following a similar methodology with baseline vocabularies up to 20k, Yazgan and Saraclar (2004) report that the actual phonemic transcriptions that are generated for the OOV words are seldom correct: only 7.5% of the OOV words get a correct phonemic transcription.

Bisani and Ney (2005) also attempt to recover the spelling of OOV words. They introduce the graphone (similar to the letter-phoneme of Chung, 2000) as a new type of subword unit. A graphone is defined as a grapheme sequence and its corresponding phoneme sequence. Whenever a graphone sequence appears in the recognition output, it yields, by definition, an orthographic representation. For vocabulary sizes of 5k, 20k and 64k words, graphones are inferred from the pronunciation dictionary of the IV words. All graphones occurring in these words and counting no more than 6 graphemes are added to the baseline vocabulary. On a dictation task, the authors obtain relative WER improvements of 30% (5k lexicon), 15% (20k lexicon) and 0.5% (64k lexicon). The low value for the latter vocabulary size (the absolute

improvement is only 0.05%) is somewhat in line with the fact that the OOV rate for that vocabulary is as small as 0.5%, and therefore only 0.75–1% of the errors are anticipated to be related to OOV words.

In Rastrow et al. (2009), OOV word spellings are recovered by finite state transducer operations on the recognition lattice generated by a recognizer with a hybrid LM. The subword unit set is derived in a data-driven manner from the phonemic transcriptions of the in-vocabulary words occurring in a large broadcast news training text. The procedure starts with the creation of a 5-gram phoneme LM that is pruned on the basis of relative entropy until it contains no more than 20k subword *N*-grams. The latter are then added as additional entries to the standard word vocabulary incorporating 10k, 20k, 40k, 60k and 84k words. The detection of OOV segments is further aided by a Maximum Entropy classifier that uses the posterior probabilities of the words and subwords occurring in the recognition output lattice as features. To recover the OOV words, the lattice generated by the recognizer is first expanded to a phonemic version by using the phonemic transcriptions available in the recognition lexicon. Next, the phoneme lattices are rescored using the lexicon and the LM built for the largest available vocabulary of 84k words. This way, a baseline 10k vocabulary system can be improved by 9.4% relative, but the improvements of systems with a baseline vocabulary of 40k words and more are negligible.

In Parada et al. (2010a,b), the approach of Rastrow et al. (2009) is further refined. The OOV detection is improved through the combination of a more sophisticated classifying technique (Conditional Random Fields) and more appropriate contextual features. The recovery of detected OOV words is improved by employing web harvesting techniques that search the Internet for words that co-occur with the words that are recognized in the vicinity of detected OOV regions. These words are then added to an OOV term list. By doing so, 29k new terms covering 53% of the actual OOV tokens in the test data could be discovered. In a second step, each new term is phonemically transcribed by means of a G2P converter. A detected OOV region is decoded as one of the newly gathered terms as soon as a *term-region specific threshold* (TRST) is exceeded (for details, we refer to the paper). In a series of experiments, the authors find an OOV recovery recall of 15.2%, while the final WER decreases from 17% to 16.8% (statistically significant reduction). Using an oracle OOV detection, the OOV recovery recall is 40.7% leading to a WER of 16.4%. When the OOV recall would have been 100%, the WER would have been 14.6%.

3. Proposed approach

The approach that we propose here consists of two parts: (1) the detection and phonemic subword transcription of the OOV word regions and (2) the P2G conversion of these phonemic transcriptions.

Our aim was to devise a modular system that adheres to the mixed approaches, but now for a large vocabulary recognition task. Furthermore, we wanted to compare the performance of such an approach to that of the state-of-the-art hybrid approach.

Unlike in the methods that were discussed in Section 2.1, we do not worry that much about IV words being mistakenly detected as OOV words, because we expect that, if this happens, these words will get a correct subword transcription, and will therefore be recovered in the P2G stage. Furthermore, we argue that good subword transcriptions in OOV regions can only be generated by very restrictive (e.g. high-order) subword LMs that are well armed to compete with the regular word-based LM. Consequently, our method employs such restrictive subword LMs and it does not need an OOV word penalty to suppress the number of false alarms. In Section 4.3, we detail how we can cope with such high-order LMs in our recognizer. Other minor differences with the cited work are that we train the OOV subword LMs on OOV word transcriptions only and that we take account of the OOV word frequencies in constructing these models.

Our approach to the second part relies on a straightforward look-up dictionary to map the phonemic subword transcriptions to word orthographies.

3.1. Detection and transcription of OOV words

We argue that a state-of-the-art hybrid LM is in a sense suboptimal for the detection of OOV words because (1) it models the inter-dependencies between IV and OOV words through OOV subword units and (2) a single LM is used to model two distinct phenomena, namely sequences of short subword units in OOV word regions and sequences of (much) longer words in IV word regions. As an example, we have expanded the LM probabilities of a 4-word sequence

$w_1\ w_2\ OOV\ w_4$ containing an OOV word that is pronounced as $s_1 \ldots s_M$ (=M subword units). In case of a trigram hybrid LM one obtains that

$$P(w_1\ w_2\ OOV\ w_4) = P(w_1)P(w_2|w_1)P(s_1|w_1w_2)P(s_2|w_2s_1)(\prod_{k=3}^{M} P(s_k|s_{k-2}s_{k-1}))P(w_4|s_{M-1}s_M)$$

By conditioning the probability distribution of the initial OOV subword units s_1 and s_2 on the preceding IV words w_1 and w_2, the amount of LM training material for modeling the OOV subword sequences is subdivided, raising the risk of data scarcity. Furthermore, conditioning the probability distribution of the IV word w_4 on the preceding OOV subwords (s_{M-1}, s_M) masks the effect that the IV words w_1 and w_2 have on the prediction of w_4. A potential benefit of the hybrid approach is of course that it can capture predictable (beginnings of) OOV subword sequences. One may wonder however whether an OOV word whose subword sequence can be accurately predicted in this way should not have been an IV word in the first place.

The use of a single all-in-one LM also implies that the modeled context length in the LM can strongly vary from $N-1$ phonemes in OOV word areas to $N-1$ words (representing far more than $N-1$ phonemes) in IV word zones. Given that infrequent words tend to be rather long, especially in compounding languages like German or Dutch, one might thus expect that a hybrid LM is more powerful if it works with longer span subword units such as phoneme trigrams or syllables. When working with short subwords such as phonemes, we conjecture that a longer context length is needed in an OOV word region (modeled as a subword sequence) than in other regions where normal word sequences need to be modeled.

In this work we therefore propose to work with a mixed LM composed of two components. The first component is a traditional word level LM that can help to detect OOV regions. The second component is a subword level LM that can generate subword transcriptions for the detected OOV regions. At the level of the words, every OOV word is mapped to a single generic OOV word class. A straightforward extension would be the introduction of multiple OOV word classes. However, to be beneficial, this would require that the OOV words can be clustered in groups that are both acoustically and grammatically disparate. During LM training, the OOV tokens are allowed to occur in any position of an N-gram (i.e. also in the context) so that the linguistic context can be maximally exploited and the recognition performance is maximally maintained. In order to transcribe a candidate OOV region, a generative OOV subword model is employed that relies on an N-gram subword LM to help identify the subword unit sequence best matching the acoustics. The subword LM is trained on the phonemic transcriptions (generated by a G2P transcriber) of the OOV words appearing in the LM training text. IV words are not included during training. To give more frequent OOV words more weight than less frequent OOV words, the word frequencies measured in the text corpus are taken into account as well. An OOV word consisting of subword units $s_1 \ldots s_M$ can be recognized correctly if the acoustics of the time segment covered by that word make the sequence $s_1 \ldots s_M$ more likely than any other IV word in that segment, given the preceding words.

If 'lc' represents the linguistic left context of the OOV word, we try to achieve this objective by putting the following mixed LM

$$P(s_1 \ldots s_M|lc) = P(s_1 \ldots s_M|OOV, lc)P(OOV|lc) \approx P(s_1 \ldots s_M|OOV)P(OOV|lc)$$

in competition with the traditional word-level LM composed of probabilities of the form $\log\ P(w|lc)$ where w is an IV word.

The conditional probability of the subword unit sequence $P(s_1 \ldots s_M|OOV)$ is calculated as follows:

$$\prod_{m=1}^{M+1} P(s_m|s_{\max(m-N+1,0)} \ldots s_{m-1})$$

with s_o and s_{M+1} representing predefined word-start and word-end tags respectively. Note that by preventing the left subword context in the above equation to cross the word boundary and by adding the probability of the word-end tag, the OOV probabilities are independent of any effects induced by the IV words. Consequently, if new OOV words have to be accommodated or if a better grapheme-to-phoneme (G2P) converter becomes available to transcribe the list of OOV words, it suffices to retrain the OOV subword model, meaning that the large word-based LM must not be re-estimated.

For the exemplary word sequence $w_1\ w_2\ OOV = (s_1 \ldots s_M)\ w_4$, one now gets the following LM probability (in case of a trigram word LM):

$$P(w_1\ w_2\ OOV = (s_1 \ldots s_M)\ w_4) = P(w_1)P(w_2|w_1)P(OOV|w_1 w_2)P(s_1 \ldots s_M|OOV)P(w_4|w_2\ OOV)$$

Clearly, the IV word context is better preserved in the conditional probability of word w_4 than in the case of a hybrid LM. Thanks to the clear separation between the two language model components one can search more or less independently for the best subword unit set and for the best *N*-gram orders for both the word-level and the subword level component. Our approach can for example combine any high-order LM at the subword level with a standard trigram LM at the word-level.

In that respect, our approach is bound to be robust against changes in the domain. In fact, in many cases the probability of a word sequence containing an OOV is mainly determined by the (unknown) word class of that OOV (e.g. a person name) rather than by the full identity of that OOV (it does not matter what the person's name really was). Therefore, the word-level LM part that deals with OOV words in the training domain is expected to generalize well to other domains. Furthermore, as the subword *N*-gram model is taking word frequencies into account, and since it is anticipated that the domain-specific OOV words represent only a small part of the OOV probability mass, the subword *N*-gram models (taking the larger frequencies of the non-domain-specific OOV words into account) will not be affected much by switching between domains either. Unseen OOV words are therefore hypothesized to fit the (general) phonotactic rules embedded in the subword LM.

Although the selection of the subword units could be optimized in a data driven way, we did not pursue such an optimization here, mainly because the literature is inconclusive regarding the type of subword units that should be preferred. Furthermore, we anticipate that any type of subword unit should work fine in our approach, given that the subword LM is of an appropriate order. That is why we have chosen to work with two standard generic unit sets, namely the phonemes and the syllables that we retrieved from the phonemic transcriptions (including syllabification) of all the OOV words that occurred in the LM training text.

3.2. Phoneme-to-grapheme conversion

For the conversion of a subword unit sequence $s_1 \ldots s_M$ to an orthographic representation we propose a straightforward table look-up strategy. We argue that most of the recoverable OOV words can be listed in a background OOV word list that can for one part be retrieved a priori from a large text corpus (e.g. the material that was used for the LM training) and for another part be obtained by harvesting additional words that are identified as being important in the context of the application. The phonemic transcriptions of the background words can for instance be generated by an available grapheme-to-phoneme converter.

The straightforward table look-up approach searches for a transcription in the background pronunciation dictionary that perfectly matches the subword sequence generated by the recognizer. If such a match is found, the subword sequence is replaced by the corresponding word in the dictionary. Otherwise, the sequence is marked as OOV in the final hypothesis. It is clear that the table look-up could be supplemented with approximate matching strategies or with a generic P2G converter as proposed in Decadt et al. (2001). However, we have restricted ourselves to the look-up approach in this work in order to not further complicate the comparison between the mixed LM approach and the hybrid LM approach with yet another technique that must be fine-tuned.

Apart from word pronunciations, the look-up dictionary also includes word frequencies that are derived from the LM training text. In case multiple word pronunciations match the recognized phoneme sequence (=homonyms), the most frequent word is returned. New words added at a later stage can be assigned a low frequency or frequencies that are derived from new sources and that are normalized so as to match with the frequencies that were retrieved for the original words. As was shown in Parada et al. (2010b), the Internet forms a good source to detect new content words and their frequencies.

Since the recognizer may also employ its OOV model to explain IV word regions that are difficult to decode (e.g. due to a bad fit of the main word-level LM), the IV words are also included in the background dictionary.

4. Experimental set-up

This section describes the databases, the toolkits and the evaluation metrics that we have used to conduct the experimental validation of our proposed OOV handling approach. All evaluations are performed on (Flemish) Dutch speech material.

4.1. Lexicon

The word pronunciations for the training material, the test vocabulary and the large word list for the P2G conversion were all generated with the commercially available Flemish Dutch G2P converter of Nuance.[1] This converter is the one that has been incorporated in all Flemish Dutch Nuance speech synthesizers that were commercialized over the last decade. Per word, one single G2P transcription is generated.

Next to word forms, all lexicons contain symbols for pauses (silence), speaker noises (e.g. breathing, laughing, coughing) and filled pauses. These symbols can be inserted anywhere in the hypothesized word sequence at the cost of a small insertion penalty. Their insertion does not have any effect on the language model whatsoever. The transcriptions[2] for the filled pause are /9:/, /9:m/, /9$\tilde{}$/ and /9$\tilde{m}$/.

The lexicons of our baseline recognizers, i.e. the systems without dedicated OOV word modeling, also contain a generic OOV word token that is described as a repetition of 1–5 garbage acoustic units (see Section 4.3 for more details). The lexicons of the recognizers with dedicated OOV word modeling are augmented with the subword units (phonemes or syllables). These subwords are described using the same acoustic units, i.e. cross-word context-dependent tied-state triphones that are used to describe the regular vocabulary words.

4.2. Language model training

All LMs were trained using the SRI language modeling toolkit (Stolcke, 2002; Stolcke et al., 2011) with modified Kneser–Ney discounting.

The training text for the word-level LM is a collection of Flemish newspaper and magazine articles covering the period from 1999 to 2004, called the Mediargus corpus. After text normalization (e.g. removing duplicate sentences from the corpus, writing out numbers and correcting misspellings), the data set contains about 1.2 billion words (=tokens) and 5M unique words (=types).

In order to create an initial baseline recognition system, we selected the 100k most frequent words in the LM training corpus as the word vocabulary. This 100k vocabulary is used in the majority of our experiments, because larger vocabulary sizes reduce the amount of OOV words to a point where a detailed analysis of the impact of the OOV word modeling approaches would require unwieldy large amounts of test material. Note however that the final systems are also evaluated in combination with word vocabularies containing the 200k and 400k most frequent words. For the 100k, 200k and 400k vocabulary, the OOV rate in the training text was 3%, 1.8% and 1% respectively.

After having mapped all OOV words to one OOV type, baseline open-vocabulary 3-gram LMs were created from the training material. The 3-gram LMs include all *N*-grams that were seen during training (no cut-offs), as we empirically established that this LM led to the best baseline recognition results. We also created open-vocabulary 4-gram, 5-gram, 6-gram and 7-gram LMs for additional experiments that are presented in Section 5.9. During the training of these LMs, we used a cut-off value of 2 for the 4-grams, 3 for the 5-grams, 4 for the 6-grams and 5 for the 7-grams.

In order to construct the subword-level LM component, we first generated a phonemic transcription for all the words that occur in the LM training corpus. As the transcriptions comprise syllable boundaries it was easy to identify the phonemic syllable types that are needed to construct a subword LM based on syllables. No less than 67.6k syllables are required to explain all OOV word transcriptions. However, the 10k most frequent syllables already cover 99% of the OOV word transcriptions. The subword-level LMs include all seen subword *N*-grams (no cut-offs).

[1] http://www.nuance.com.

[2] All phonemic transcriptions in this paper follow the SAMPA notation (http://www.phon.ucl.ac.uk/home/sampa/).

4.3. Recognition system

The acoustic modeling and primary decoding of the speech is achieved by means of the state-of-the-art SPRAAK toolkit (Demuynck et al., 2008). SPRAAK uses a time-synchronous decoder designed to integrate complex knowledge sources including high-order *N*-grams and cross-word context-dependent phones in a single pass. This is achieved by means of a token passing algorithm (Demuynck et al., 2000; Demuynck, 2001) that combines the pronunciation information with the LM on-the-fly. Having a clear separation between the two major components – pronunciation information and LM – allows the use of dedicated and optimal representations for each component and minimizes the constraints put on the components. The pronunciation information comprises the pronunciation dictionary composed with the context-dependent tied states from the acoustic models and is stored as a static state-emission finite state transducer (FST). The only limitations the decoder puts on the LM is that it must have a start node and that it must be able to list all transitions (destination state and corresponding probability) when a new lexical token is presented. The large *N*-grams used in this work are stored as compressed trees, reducing the memory load to 5.7 bytes per *N*-gram entry on average. The composition of the word *N*-gram and the subword *N*-gram is done on-the-fly and is equivalent to replacing every OOV arc in the word *N*-gram with a copy of the OOV subword *N*-gram. Insertion of (filled) pauses and speaker noises is also handled on-the-fly by the LM module and is equivalent to adding pause/noise loops on every node in the word *N*-gram. The overall effect of this set-up is that the decoder can treat subwords, (filled) pauses, speaker noises and regular words uniformly, as all the specifics are handled by the LM module. The decoding is accelerated by smearing unigram probabilities, LM forwarding and caching, and adaptive beam pruning (Demuynck, 2001). Note that equivalent configurations are attainable with decoders based on weighted finite state transducers, e.g. using the techniques presented in Hori et al. (2007).

The acoustic models were trained on 40 h of Flemish Broadcast News data from the NBEST benchmark (Kessens and van Leeuwen, 2007; van Leeuwen et al., 2009). 49 three-state acoustic units (46 phonemes, silence, garbage and speaker noise) are modeled using SPRAAK's default tied Gaussian approach. The automatic training produced 3853 cross-word context-dependent tied triphone states that share a pool of 49,636 Gaussians. The word boundary was not marked in the lexicon so no distinction was made between within-word and across-word triphones (position independent triphones). On average, each state uses 181 of these Gaussians. The acoustic models were trained using the maximum likelihood criterion (no discriminative training).

The models for silence, speaker noises and garbage are context-independent but share the same Gaussians. The garbage model is trained on all segments that are labeled as "unclear speech" in the training corpus. Aside from a maximum of 5 repetitions per segment, the garbage acoustic unit is free to choose what it models. Based on forced alignment of the train data, the garbage acoustic unit seems to loosely follow the syllable structure of the unclear speech.

The acoustic features consist of 22 mean-normalized log MEL spectra and their first and second order time derivatives. These 66 features are reduced to 36 by means of a discriminative linear transformation and decorrelation (Demuynck, 2001).

4.4. Test data

Recognition experiments were conducted on two test sets. The first set is the Flemish NBEST Broadcast News development set. It consists of about 1 h of speech (10k words) spoken by 18 different speakers and its content covers more or less the same time period and topic areas as the training data. It is therefore a typical example of an in-domain task. The second test set is read speech selected from component 'o' of the Spoken Dutch Corpus (SDC, Oostdijk, 2000). It contains 1 h and 20 min of speech (12.5k words) spoken by 40 different speakers. Each speaker reads a 2 min paragraph from one of the five novels. The novels cover five diverse topics, ranging from a review of some philosophic theory over a story on colonial history to plain crime fiction. The data are thus typical for an out-of-domain task. In the following, we will refer to both test sets as the NBEST set and the SDC set respectively.

4.5. Other elements

In all recognition experiments we adopt the traditional word error rate (WER) as our primary evaluation criterion. This is in line with the idea that the final goal of OOV word recovery is to obtain the lowest possible WER. Statistical significance of WER differences is determined using the Wilcoxon signed ranks test (Conover, 1999).

Table 1
Baseline WER (%) on the NBEST and SDC test sets. Also added are the test set OOV rates (%) w.r.t. the 100k word vocabulary and w.r.t. the complete background word list.

Set	WER	OOV_{100k}	OOV_{All}
NBEST	14.1	1.6	0.2
SDC	26.6	4.6	0.4

In case syllables are used as the subword units, we consider all 67.6k syllables that are needed to generate all transcriptions in our OOV word pronunciation dictionary. This way we rule out any effect of not working with a closed-vocabulary subword set. However, we did verify that virtually the same results as the ones presented below can be obtained with smaller systems incorporating only the 10k most frequent syllables.

Note that the set of 67.6k syllables also covered all syllables occurring in the G2P transcription of the test set OOV words that are not included in the background dictionary.

5. Experimental validation

In our experimental validation, we first establish baseline recognition results on both test sets. Next, we investigate the effect of the subword LM order on the predictions of that model in the OOV regions. Then, we compare the test set perplexities for the proposed mixed model and for a hybrid model, and we assess the capability of the newly proposed approach to achieve a lower WER. We also evaluate the effectiveness of our approach in combination with larger word vocabularies and higher-order LMs. In the end, we compare our method to an approach designed specifically to cope with Dutch compounds and to the straightforward approach of extending the number of IV words.

5.1. Baseline system

Table 1 presents the recognition results that were obtained on the two datasets with a baseline system that includes a simple garbage OOV word model in combination with a 100k vocabulary and a 3-gram LM. The table also lists the test set OOV rates with respect to the 100k word vocabulary and with respect to the complete background word list extracted from the LM training material. As could be expected, the recognizer performs significantly better on the in-domain task than on the out-of-domain task. The differences in performance are primarily caused by differences in the test set perplexity and not as much by differences in the OOV rate w.r.t. the 100k vocabulary (see Fig. 2 for perplexities). Nevertheless, the OOV rates are interesting because they show that the full word list covers 87.5% and 91.3% of the OOV words appearing in the NBEST and SDC test sets respectively. This demonstrates that a large background vocabulary, regardless of its origin, can provide a high coverage of words from different domains. This claim is also supported by OOV rate measurements that were conducted on other parts of the Spoken Dutch Corpus (spontaneous speech data, lectures, discussions, etc.).

When we evaluate the performance of our baseline OOV word model, we find that it does not affect the recognition. Virtually the same recognition results (absolute WER differences <0.1%) are obtained when we do not include such a model.

5.2. Effect of the subword LM order

The aim of the OOV subword model is to predict the probability of the subword sequences corresponding to OOV words. A good way to measure the quality of that model (L_{sw}) is to consider the set of all OOV words ($\mathcal{S}_{OOV}$) and to measure the cross-entropy between the OOV word probabilities that the model predicts and the OOV word probabilities that were derived from the context-independent word counts $c(w)$ in the LM training text:

$$CE = -\sum_{w \in \mathcal{S}_{OOV}} \frac{c(w)}{\sum_{w \in \mathcal{S}_{OOV}} c(w)} \log \frac{P[s_{1_w} \dots s_{M_w} | L_{sw}]}{\sum_{w \in \mathcal{S}_{OOV}} P[s_{1_w} \dots s_{M_w} | L_{sw}]}$$

Fig. 1 shows the cross-entropies for N-gram models using phonemes and syllables as the subword units. One can see that for both model types the cross-entropy decreases with the model order until it saturates. Moreover, they seem to

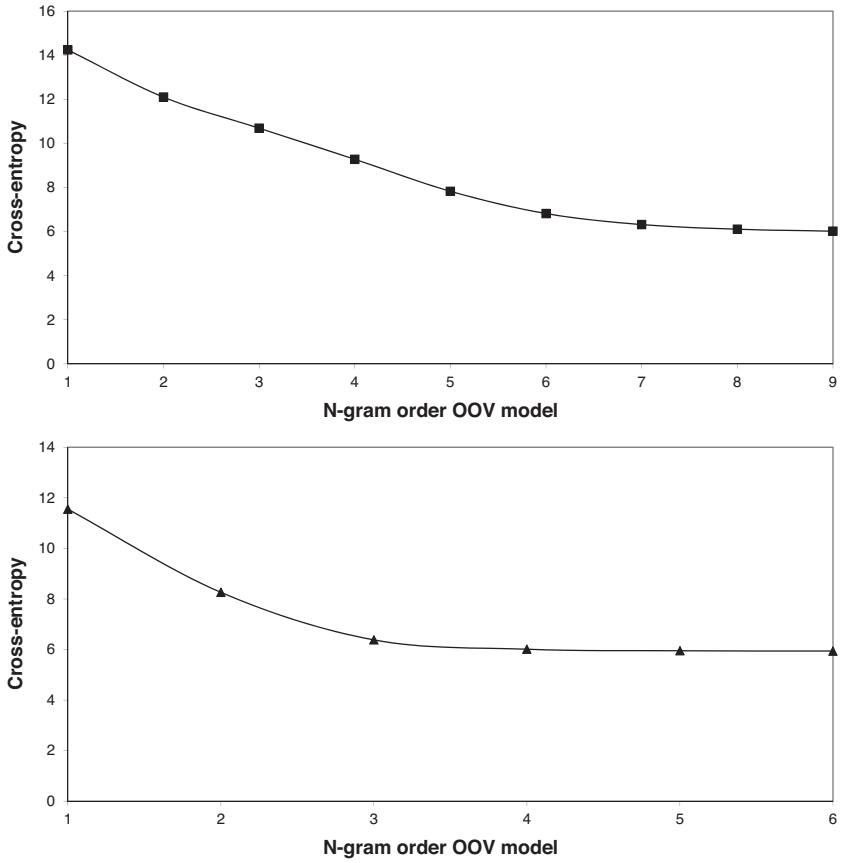

Fig. 1. Cross-entropy between the relative OOV word frequency and the OOV subword sequence probability provided by phoneme-level (top) or syllable-level (bottom) *N*-gram models of different orders.

saturate at the same value. For syllables, the saturation is reached at an order of 4 whereas for phonemes it takes an order of 9 to reach saturation. This is not surprising since the average context length (in phonemes) that is modeled by a syllable LM of a certain order is larger than that of a phoneme LM of the same order.

The figures definitely suggest that the low-order subword models that were tested in earlier work on mixed OOV word handling exhibit suboptimal performance.

5.3. Mixed versus hybrid language model

Before conducting recognition experiments, we first investigate whether the newly proposed mixed LM is able to compete with the commonly used hybrid LM. Two such hybrid language models were trained on hybrid text versions of the LM training data. In one version, we replaced all the OOV words in the sentences by their constituent phonemes. In a second version, we replaced them by their constituent syllables. Trigram hybrid language models were derived from each of these text versions.

The hybrid models as well as the various mixed models comprising a subword component of a variable order were then assessed on the test sets in terms of their perplexity. To that end, we employed the different models to calculate the total LM probability of the reference test set texts (appropriate hybrid text versions were used to accomplish this) and divided this probability by the total number of word tokens in the texts. All results can be found in Fig. 2.

The plots confirm that the NBEST sentences fit the grammar of the language models much better than the SDC sentences. The effects of the model order are also in line with the results discussed in the previous section.

An expected result is that a hybrid approach with syllables is substantially better than one with phonemes, as a trigram phonemic context is clearly insufficient to adequately cover the OOV word zones. In contrast, the figures prove

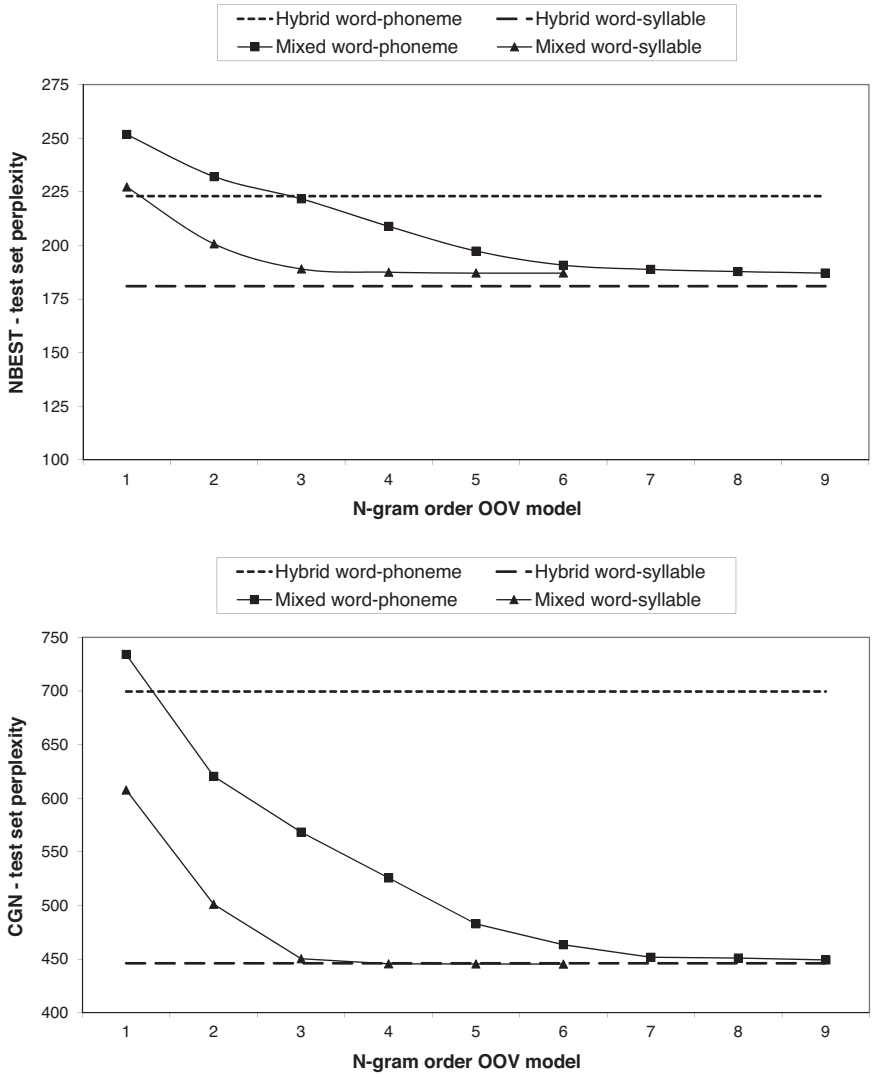

Fig. 2. Test set perplexity for hybrid or mixed word-phoneme and word-syllable language models on the NBEST development set (top) and the SDC read speech set (bottom).

that the mixed approach leads to equal performances for both choices of the subword units, be it that the required model order is much smaller in the case of syllables.

Somewhat disappointing for us is that a hybrid 3-gram LM with syllabic subword units outperforms all mixed language models on the NBEST data, while it matches their performance on the SDC data. On the other hand, the differences in perplexity become small as soon as the subword model order is sufficiently large.

In a more detailed calculation, the perplexities of the IV words and of the OOV words were determined separately. This showed that a mixed LM obtains slightly lower perplexities for the IV words in both sets, while the hybrid model yields the lowest perplexity for OOV words. The differences are small however.

In the next section we make a systematic comparison of two types of hybrid and mixed language models when incorporated in a large vocabulary speech recognizer.

5.4. Recognition with OOV word handling

Here, we assess systems incorporating a hybrid or a mixed LM for OOV word modeling in combination with the table look-up method for phoneme-to-grapheme conversion. The background pronunciation dictionary consists of the

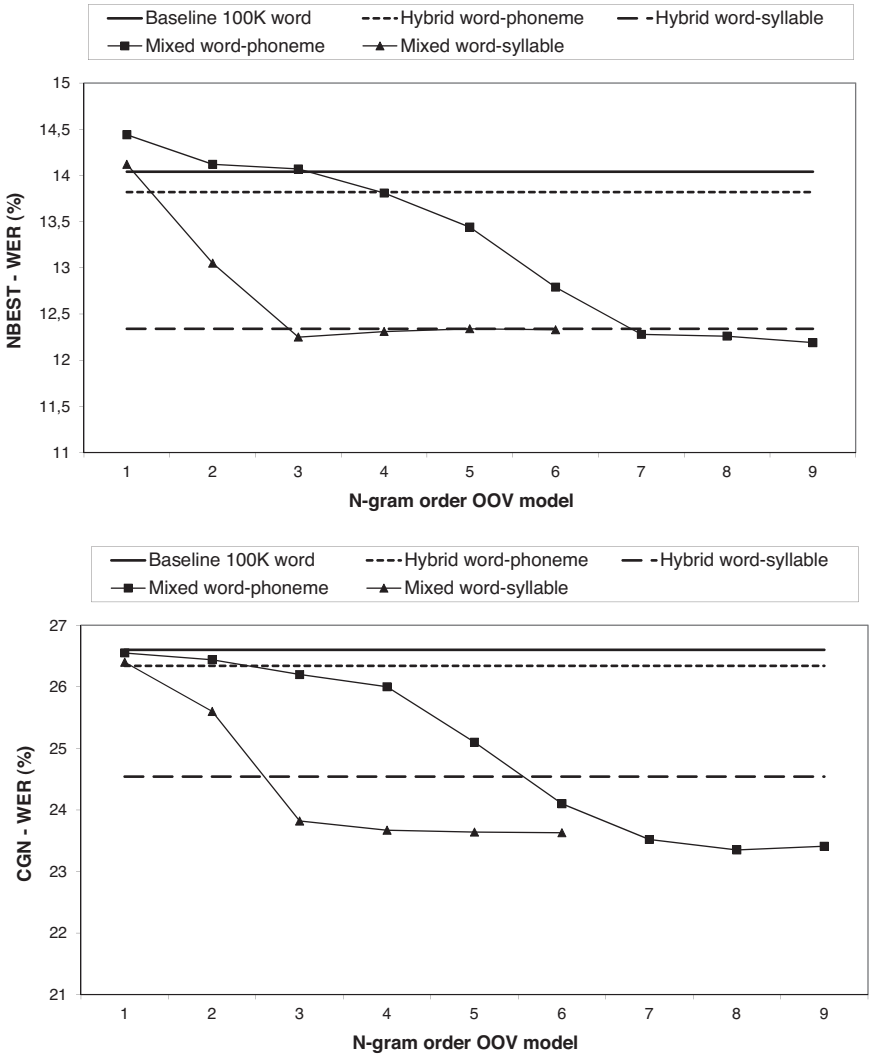

Fig. 3. WERs (%) on the NBEST development set (top) and the SDC read speech set (bottom) for a baseline 100k word vocabulary recognizer and for recognizers incorporating hybrid or mixed word-phoneme and word-syllable language models.

5M words retrieved from the word-level LM training corpus and their automatically computed G2P pronunciation (one pronunciation per word).

Fig. 3 shows the WERs obtained with the different systems. On the in-domain task (NBEST), the two mixed systems perform equally well as the hybrid system with syllabic subword units, provided of course that the subword LM order is sufficiently large. This means that the slightly higher perplexity of the mixed models (see previous section) does not cause a higher WER. All three systems significantly ($p < 0.01$) improve the recognition performance w.r.t. the baseline recognizers. The substantially higher perplexity of the hybrid word-phoneme LM does translate into a higher WER. In fact, in the latter case there is no significant gain due to OOV handling anymore, on neither of the test sets.

On the out-of-domain task (SDC), the hybrid word-syllable system and the two mixed systems significantly ($p < 0.01$) improve the recognition as well. More important however is that the mixed systems significantly ($p < 0.01$) outperform the hybrid systems as soon as the subword LM order is sufficiently large. This result confirms the hypothesis that mixed language models are more robust against changes of the domain than hybrid models.

For clarity reasons, we will – from here on – fix the subword LM order and focus the analysis of mixed LM performance on two specific systems: the mixed word-syllable model that uses a 4-gram subword model, and the mixed word-phoneme model incorporating a 9-gram subword model. Furthermore, we will no longer consider the

Table 2
WER (%) on the NBEST development set and the SDC read speech set for a baseline 100k word vocabulary recognizer, for a hybrid word-syllable recognizer, for a mixed word-phoneme recognizer using a 9-gram subword LM and for a mixed word-syllable recognizer using a 4-gram subword LM.

Set	Baseline	Hyb_{syl}	Mix_{syl}	Mix_{phon}
NBEST	14.1	12.3	12.3	12.2
SDC	26.6	24.5	23.6	23.4

hybrid system that incorporates phonemic subword units. Under these conditions, one finds that the hybrid model (with syllabic subword units) yields a relative improvement of 12.1% for the NBEST data and 7.8% for the SDC data. The mixed models obtain relative improvements of 12.3% (syllabic subwords) and 13.2% (phonemic subwords) on the NBEST set, and 11.0% (syllabic subwords) and 12.0% (phonemic subwords) on the SDC set. Table 2 summarizes the obtained WERs with baseline, hybrid and mixed recognizers on both test sets.

5.5. Result analysis

In a first result analysis, we wanted to establish how much each resource and operation contributes to the overall WER improvement for both approaches. Therefore, we investigated how much of the improvement is owed to the phoneme-to-grapheme conversion. Comparing the obtained WERs with and without the P2G table look-up shows that a hybrid system without P2G conversion can already attain 66% (NBEST set) and 77% (SDC set) of the total WER improvement attainable with P2G conversion. The mixed language model without P2G conversion offered only 57% (true for both subword types on the NBEST set) and 66% (true for both subword types on the SDC set) of the gain attainable with P2G conversion. Combining these results with some of the forthcoming analysis results suggests that (1) the subword sequences produced by the mixed language models are more often correct, in the sense that they enable a correct P2G conversion, and (2) the mixed language models transcribe more speech segments in terms of subwords, which implies that more recognition errors (deletions) occur if these subword sequences are not converted to an orthography.

In another analysis, we have detailed how many reference words of the test sets were misrecognized for each of the following three word categories: (1) OOV words, (2) all IV words and (3) IV words adjacent to an OOV word.

The percentages listed in Table 3 show that the mixed systems cause a stronger reduction of the misrecognitions of OOV words and of IV words surrounding an OOV word than the hybrid system does, both for the in-domain and the out-of-domain task. For the latter task, the difference in performance is even significant. The mixed word-phoneme system slightly outperforms the mixed system with syllabic subwords. This seems to indicate that correctly decoding the OOV subword sequences is harder when the large set of 67.6k confusable syllables is used.

The misrecognition rates for the IV words in general are virtually the same for all methods. With respect to the baseline system, this mainly implies that the total number of misrecognized IV words does not increase. One can

Table 3
Misrecognition rates (%) for OOV words, all IV words and IV words adjacent to an OOV word (IV_{OOV}). The percentages are given for a baseline 100k word vocabulary recognizer and for hybrid and mixed systems incorporating 100k regular words. The mixed word-phoneme system uses a 9-gram subword LM, the mixed word-syllable system a 4-gram subword LM. Results are presented for the NBEST development set and the SDC read speech set.

Category	Baseline	Hyb_{syl}	Mix_{syl}	Mix_{phon}
NBEST development set				
OOV	100.0	57.6	51.5	48.5
IV	10.5	10.3	10.5	10.4
IV_{OOV}	51.0	24.6	24.0	23.8
SDC read speech set				
OOV	100.0	79.5	73.0	70.0
IV	18.7	18.7	18.7	18.7
IV_{OOV}	56.4	46.3	42.1	40.6

Table 4
Misrecognition rates (%) on the NBEST development set for OOV words, all IV words and IV words adjacent to an OOV word (IV_{OOV}). The percentages are given for a baseline 200k word vocabulary recognizer and for hybrid and mixed systems incorporating 200k regular words. The mixed word-phoneme system uses a 9-gram subword LM, the mixed word-syllable system a 4-gram subword LM.

Category	$Baseline_{200k}$	Hyb_{syl}	Mix_{phon}
OOV	100.0	71.1	53.0
IV	10.6	10.4	10.5
IV_{OOV}	48.1	29.5	25.3

notice however that for the in-domain task, the hybrid system yields a slightly better IV word recognition than the mixed systems. We first hypothesized that this is due to the fact that, for a baseline word vocabulary of 100k words, the in-domain test set still contains a lot of predictable OOV subword sequences, giving an advantage to the hybrid modeling approach. To verify this hypothesis, an additional set of experiments was performed in which the baseline word vocabulary was raised to 200k words. In that setting, the baseline recognition result on the in-domain task improves from 14.1% (for a 100k system) to 12.9%, which proves that increasing the IV word vocabulary for the in-domain task is definitely a viable option. The result for a hybrid system is 12.1%, while a mixed system using a 9-gram phoneme subword LM obtains a WER of 11.9%. For the examined systems, we then performed a similar analysis as in Table 3, of which the results are presented in Table 4. It shows that the recognition rate for IV words is still slightly better in case of a hybrid approach.

In order to find the true reason for the performance difference on the IV words we broadened our analysis. We did not just look at WERs anymore, but we also examined the faulty word hypotheses generated by the hybrid and the mixed systems. This examination indicated that the mixed approach potentially has a larger advantage over the hybrid system than the WERs suggest. One phenomenon that is responsible for masking this advantage in the WERs is the fact that a subword sequence generated by the recognizer can correspond to a word tuple (mostly of IV words) and that this transcription, although sometimes correct (see examples listed in Table 5), is not replaced by a word hypothesis from the background dictionary because there is no single word in this dictionary that matches this transcription.

For the NBEST benchmark, we observed that around 12.5% of the subword sequences in the output of both mixed systems correspond to word tuples, while this is only true for 6.7% of the subword sequences generated by the hybrid LM. For the SDC benchmark, 19.5% (mixed word-syllable) and 22% (mixed word-phoneme) of the subword sequences in the output of the mixed recognizers correspond to word tuples, against 14% for the system with the hybrid LM. In other words, the above phenomenon is more apparent when a mixed LM is used. In a preliminary attempt to exploit some of these correctly recognized word tuples, we considered a P2G mapping approach in which a recognized phonemic sequence can be replaced by a word pair in case (1) the phonemic sequence does not match the transcription of a single word and (2) both words of the word pair are IV words. However, our attempt was not successful because some of the erroneous phonemic sequences could be explained as combinations of two words as well, which introduced additional insertion errors. A more intelligent P2G mapping approach is probably required for these cases. Such an approach could be devised in future work.

A second phenomenon that deserves a citation is the fact that some of the in-vocabulary names appear with multiple spellings in the background dictionary. Consequently, if such a name is recognized as an OOV word, it is not necessarily mapped to the spelling that is used in the reference transcription at that time. Some examples of these phenomena are listed in Table 6. This phenomenon too occurs slightly more often for the mixed language models, which explains the slightly higher IV word misrecognition rates.

Table 5
Examples of correctly transcribed in-vocabulary word tuples that occur as single transcriptions in the output of the recognizer.

Word pair	Subword sequence
Club Brugge	/klYbbrYG@/
heen brachten	/he:nbrAxt@n/
zei de koopman	/zEid@ko:pmAn/
de vrouw van nonkel	/d@vrAuvAnnONk@l/

Table 6
Examples of recognized names that are considered incorrect, although they are valid alternatives of the name in the reference transcription.

Reference transcription	Recovered orthography
van Praetlaan	VanPraetlaan
Van de Looverbosch	Vandelooverbosch
Casarotto	Cassarotto
Hvastija	Hvastia
Rijckaert	Rijkaard

5.6. Detection of OOV words

In addition to the WER result analysis, we have also investigated how the OOV word detection as such improves in the first stage of the recognition. Table 7 lists the OOV word detection rate (the recall) and the false alarm rate (complement of the precision) per system.

The mixed systems clearly detect more OOV words and this is especially true for the out-of-domain task. On the other hand, they have a larger false alarm rate for the in-domain task. However, due to the inclusion of the IV words and their word frequencies in the background dictionary, the extra false alarms do not cause an increase in WER. We only noticed very small increases of the word insertion errors. Apparently, an IV word that is recognized as a subword sequence is very frequently recovered correctly from the dictionary.

To find out how well the presented methodologies generalize towards new OOV words, we also measured the OOV detection rates for the words that do not belong to the background dictionary. For the in-domain task, all systems detect 5 of the 19 OOV words that do not appear in the LM training corpus (and hence are not listed in the 5M background dictionary). The detection rate is thus lower than average for these words. For the out-of-domain task, the detection rate is closer to the one observed for the OOV words that do occur in the LM training corpus. The hybrid and mixed word-syllable systems both detect 27 of the 49 OOV words that are not in the background dictionary. The mixed word-phoneme system detects one more.

5.7. Phonemic transcriptions of OOV words

In a last result analysis the recognized phonemic sequences for the OOV words were compared to typical and automatically generated transcriptions of these words. The typical transcriptions for the circa six hundred unique OOV words (731 OOV word tokens) that occurred all-together in the two test sets were created by the first author of this paper. He made them without having listened to any recording and without having looked at the phonemic sequences produced by the recognizers.

A comparison with the typical transcription yields a good measure of how accurate the recognized output is. A comparison with the automatically generated dictionary transcription yields a measure of how likely the recognized transcription will match to the transcription in the background dictionary.

All transcriptions were first deprived of their syllable boundaries and the Levenshtein distance between transcriptions was used as the evaluation metric. The phoneme error rate (PER), which is defined like the WER, but then for phoneme sequences, was adopted as the global evaluation measure. Table 8 shows the PERs for the three systems.

Table 7
OOV detection rate (%) and false alarm (FA) rate (%) for a hybrid word-syllable recognizer, a mixed word-phoneme recognizer incorporating a 9-gram subword LM and a mixed word-syllable recognizer incorporating a 4-gram subword LM.

Measure	Hyb_{syl}	Mix_{syl}	Mix_{phon}
NBEST development set			
OOV_{det}	66.7	67.9	68.5
FA	33.3	38.1	43.5
SDC read speech set			
OOV_{det}	53.0	58.5	61.0
FA	29.9	28.2	30.4

Table 8
PERs (%) for the phonemic sequences returned by the hybrid word-syllable recognizer, the mixed word-phoneme recognizer using a 9-gram subword LM and the mixed word-syllable recognizer using a 4-gram subword LM. The PERs are computed w.r.t. the typical (TY) transcriptions and the dictionary transcriptions (G2P).

Reference	Hyb_{syl}	Mix_{syl}	Mix_{phon}
NBEST development set			
TY	13.5	12.7	13.0
G2P	14.0	14.4	15.5
SDC read speech set			
TY	26.7	22.0	21.1
G2P	28.4	23.8	22.7

The main conclusion is that the newly proposed approach yields transcriptions that are better (closer to the typical transcriptions) than the ones generated by the hybrid approach. However, for the in-domain task the transcriptions created by means of a mixed approach with a phonemic subword LM are (on average) further away from the dictionary transcriptions. This means that the risk of not recovering the OOV words is somewhat raised for that case. For the SDC task, all odds are in favor of the new approach.

The transcriptions generated for the OOV words that do not even belong to the background dictionary are very accurate for the in-domain task (lower PERs than average, yet only measured on 5 words), but less accurate than average for the out-of-domain task (PERs raised by a factor of 2). In the latter case, most errors are found in the transcriptions of very uncommon proper names, such as *Gecukiro* or *Lubusi*.

The numbers in Table 8 also demonstrate that the automatically generated dictionary transcriptions are suboptimal. A large part of this suboptimality stems from the fact that the G2P converter has a lot of difficulties with the transcription of proper names, which often comprise parts of a foreign origin. To give a few examples, the names SCHUMACHER, HAWKING, DELPHI, MADOU and HVASTIJA have the following typical transcriptions (in terms of the Dutch phonemes only): /SumAx@r/, /hOkIN/, /dElfi/, /ma:du/ and /vAstia:/. However, the G2P converter generates the transcriptions /sxymAx@r/, /hawkIN/, /dElphi/, /ma:dAu/ and /vAstEia:/. Consequently, optimizing the pronunciation variants for proper names (Réveil et al., 2012) in the background dictionary might induce further progress.

5.8. OOV subword models trained on other data

As described in Section 3.1, we have chosen to train the OOV subword LMs on G2P transcriptions of OOV words only. Furthermore, we argued that OOV word frequencies should be taken into account to weigh the N-gram estimates. In this section, we want to verify whether these intuitively obvious decisions effectively lead to the best performance. To that end, we have considered recognition experiments with three other OOV subword models that all incorporate a 9-gram phoneme LM. The three LMs are respectively trained on (1) OOV words only, but without taking word frequencies into account, (2) all words (IV and OOV words), without taking word frequencies into account, (3) all words, taking word frequencies into account. In Table 9, the WERs obtained with these models are depicted, as well as the WER obtained with the original 9-gram LM phoneme-based OOV word model.

The table shows that our original approach yields the best WER results. In case of the out-of-domain test set, it even yields a significantly ($p < 0.01$) better performance than the other OOV word models.

As was argued in Bazzi and Glass (2001), including IV word frequencies in the LM estimation process is definitely not a good option, because then the LM is strongly biased towards the transcriptions of the IV words. However,

Table 9
WERs (%) for mixed word-phoneme recognizers incorporating different 9-gram subword LMs. The LMs are respectively trained on (1) OOV words only, taking word frequencies into account (OOV+), (2) OOV words only, without taking word frequencies into account (OOV), (3) IV and OOV words, taking word frequencies into account (ALL+) and (4) IV and OOV words, without taking word frequencies into account (ALL).

Set	OOV+	OOV	ALL+	ALL
NBEST	12.2	12.3	12.4	12.3
SDC	23.4	23.9	24.7	23.9

Table 10
WERs (%) with varying word vocabulary sizes (V_{size}) and word-level *N*-gram orders for baseline recognizers, for hybrid word-syllable systems, for mixed word-syllable systems using a 4-gram subword LM and for mixed word-phoneme systems using a 9-gram subword LM.

V_{size}	LM_{order}	Base	Hyb_{syl}	Mix_{syl}	Mix_{phon}
NBEST development set					
	3-gram	14.1	12.3	12.3	12.2
	4-gram	13.5	12.1	12.0	12.0
100k	5-gram	13.8	12.0	11.9	11.8
	6-gram	13.4	11.8	11.7	11.6
	7-gram	13.4	11.8	11.8	11.7
	3-gram	12.9	12.1	12.0	11.9
	4-gram	12.3	11.9	11.6	11.7
200k	5-gram	12.5	11.8	11.8	11.6
	6-gram	12.3	11.8	11.6	11.4
	7-gram	12.3	11.7	11.5	11.4
	3-gram	12.4	12.1	12.1	12.1
	4-gram	11.9	11.9	11.6	11.7
400k	5-gram	12.0	11.9	11.9	11.7
	6-gram	11.9	11.7	11.6	11.5
	7-gram	11.8	11.7	11.6	11.5
SDC read speech set					
	3-gram	26.6	24.5	23.6	23.4
	4-gram	25.7	23.4	23.1	22.6
100k	5-gram	25.6	23.0	22.8	22.3
	6-gram	25.6	22.9	22.7	22.2
	7-gram	25.6	22.9	22.7	22.2
	3-gram	24.9	23.9	23.3	22.9
	4-gram	24.1	23.0	22.6	22.3
200k	5-gram	24.0	22.8	22.4	22.0
	6-gram	23.9	22.6	22.4	22.0
	7-gram	23.9	22.6	22.4	21.9
	3-gram	24.0	23.5	23.2	23.0
	4-gram	23.2	22.9	22.5	22.3
400k	5-gram	23.0	22.5	22.2	21.9
	6-gram	22.9	22.3	22.1	21.9
	7-gram	22.9	22.3	22.1	21.9

our experiments suggest that the training procedure followed in Bazzi and Glass (2000, 2001, 2002), i.e. training a subword LM on pronunciations of IV and OOV words, without taking word frequencies into account, is not optimal either.

5.9. *Effectiveness in combination with larger baseline vocabularies and higher order word-level LMs*

In some additional experiments, we have verified the effectiveness of the mixed approach when the included baseline word vocabulary is extended and when the order of the word-level LM is increased. Table 10 shows WERs obtained with standard word-based recognizers, hybrid word-syllable recognizers, mixed word-syllable recognizers incorporating a 4-gram subword LM and mixed word-phoneme recognizers incorporating a 9-gram subword LM. The vocabularies contain either 100k, 200k or 400k words. The word-level LMs are 3-grams, 4-grams, 5-grams, 6-grams or 7-grams. Corresponding hybrid LMs were trained for comparison. During the training of these LMs, we used the same cut-off values that were mentioned in Section 4.2.

For a 3-gram word-level LM, with respect to the baseline system performance, both the hybrid and the mixed systems still lead to significant ($p < 0.01$) recognition improvements when the word vocabulary is increased. When the

order of the word-level LM is increased, the improvements remain significant, except in the case of 5-gram, 6-gram and 7-gram LMs using a 400k word vocabulary. There, the improvement brought by the hybrid approach on the in-domain NBEST set is no longer significant. The same holds for the mixed word-syllable approach in case of a 5-gram LM.

On the out-of-domain data, the mixed approach always significantly outperforms the hybrid approach ($p < 0.01$ for the mixed word-phoneme system, at least $p < 0.05$ for the mixed word-syllable system).

On the in-domain data, the mixed approach performs at least as good as the hybrid approach. In a few cases, the mixed systems significantly ($p < 0.05$, sometimes even $p < 0.01$) outperform the hybrid word-syllable system. An interesting comparison is the one between a hybrid word-syllable recognizer using a 4-gram LM and a mixed word-syllable recognizer using a 4-gram word-level and subword-level LM. The latter significantly ($p < 0.05$) outperforms the hybrid system when the vocabulary contains 200k or 400k words. Together with the results obtained on the out-of-domain data, this proves that separately modeling word N-grams and subword N-grams is beneficial.

Furthermore, the figures seem to indicate that both the mixed and the hybrid systems saturate once the word-level LM is a 6-gram. In an extra series of experiments, we verified that this is also the case for hybrid word-phoneme systems. The results obtained with these systems are not included in Table 10 as they are significantly worse than the ones obtained with a hybrid word-syllable system.

A final interesting observation is that the mixed systems with phonemic subwords mostly perform better (sometimes significantly better) than the mixed systems with syllabic subwords. This confirms that decoding the OOV subword sequences using the large set of 67.6k confusable syllabic subword units is harder than decoding them using the set of phonemes. Perhaps, there might be an intermediate set of subwords that leads to an optimal recognition performance, but we did not examine this so far. This could be done in future work.

5.10. Computational efficiency of mixed and hybrid approaches

As we have indicated before, an advantage of the hybrid approach is that it requires nothing more than a standard search algorithm to generate a sequence of words and subword units. The mixed approach is more complex. To get an idea of the difference in computational efficiency, we have determined the overhead of both approaches with respect to the performance of a baseline recognition system.

On the 100k 3-gram set-up, the overhead introduced by the OOV word modeling is 25%, 75% and 50% for the hybrid word-syllable model, the mixed word-phoneme model and the mixed word-syllable model respectively. This overhead drops to 10%, 40% and 25% respectively when using 400k words and a 6-gram word-level LM.

Most of the overhead in the mixed systems is owed to the parallel decoding of the audio at the frame level: each possible sequence of acoustic units is evaluated both in terms of regular words and in terms of subword units. Decoupling the low-level acoustic decoding from the higher-level linguistic processing, as done in Chung et al. (1999), Chung (2000) and Demuynck et al. (2003), is expected to decimate the overhead. An additional reduction in overhead could be attained by introducing dedicated pruning thresholds for the IV and the OOV words, or dedicated search structures to avoid duplication of work (e.g. handling the same OOV subword sequence in different word N-gram contexts). We plan to look into this matter in future work.

5.11. Comparison with an OOV word modeling approach for Dutch compounds

In Demuynck et al. (2009), a post-processing of the output of a regular word-based recognizer (no prior text decompositions) was used to recover compound OOV words in very large vocabulary Dutch continuous speech recognition. In essence, two successive words were merged whenever the unigram count of the compound (measured on the LM training corpus) was larger than the bigram count of the hypothesized word pair. Some specifications and exceptions, that are fully described in the paper, were added to tune the merging process.

In this section, we compare our mixed word-phoneme approach to this post-processing approach for systems incorporating 100k words and a 3-gram LM (Table 11).

We find that the mixed word-phoneme system leads to significantly larger ($p < 0.01$) WER reductions. This is of course owed to the fact that OOV words other than compounds can now be recovered too. It also implies (verified manually) that compounds are successfully recovered with our approach as well.

Table 11
WERs (%) for a baseline recognizer, for a system incorporating the post-processing approach to recover compound OOV words (Demuynck et al., 2009) and for a mixed word-phoneme recognizer using a 9-gram subword LM. The word vocabulary size is 100k, the word-level LM contains 3-grams.

Set	Baseline	Post-process	Mix_{phon}
NBEST	14.1	13.4	12.2
SDC	26.6	26.2	23.4

Table 12
Word error rates (%) for a 100k baseline recognizer with a garbage OOV word model, for a mixed word-phoneme LM system using an 800k background dictionary and for a system with an 800k lexicon and a 100k open-vocabulary word language model.

Task	Word_{100k}	$\text{Mix}_{phon800k}$	Word_{800k}
NBEST	14.0	12.3	12.5
SDC	26.6	23.4	23.8

5.12. Comparison with a lexical expansion approach

In this final section, we further position our method based on usage scenarios of speech technology such as the transcription of broadcast news shows or the indexing of multimedia archives. These tasks require a general-purpose speech recognizer that can rapidly be adjusted to handle new words. Retraining the language model is not an option, since either the large text corpus used to design the LM is not available to the end user, or since there is no text material available that uses the new words. It is realistic though that the end user can provide a list of domain-specific terms that he expects to occur in the speech. In the case of multimedia indexing, these words can even be supplied after the actual speech recognition was performed. It is clear that our approach (as well as a hybrid approach) fits well in such a scenario because it can simply incorporate the domain-specific terms and their corresponding transcriptions (generated automatically by the G2P converter) in the background lexicon. As this background lexicon is only consulted after the actual speech recognition phase, the process of adding new words can even be made interactive based on the automatically detected OOV word regions (one could even use the transcriptions generated by the recognizer in the background dictionary) or, in the case of multimedia indexing, can be performed incrementally based on search terms the users are inputting.

A more traditional way of including domain-specific terms is to add them to the lexicon of the baseline recognizer. However, if w_{OOV} is such a term, we need to find a good estimate of

$$P(w_{OOV}|lc) = P(w_{OOV}|OOV)P(OOV|lc)$$

with $P(w_{OOV}|OOV)$ being the chance that an OOV word is actually w_{OOV}. The most logical thing to do is to make the latter probability equal to the mean unigram probability of the OOV words (=mean over all unique OOV words) in the LM training text.[3] We evaluated this approach on both test sets, using a recognition lexicon of 800k words, the baseline lexicon plus the 700k most frequent OOV words from the LM training text.

Table 12 shows the recognition results that were obtained with the approach as well as with our approach (mixed word-phoneme LM) using the same 800k words in a background lexicon.[4] Our more modular approach outperforms the straightforward technique on both test sets. However, the better performance is not statistically significant (significance at the level of $p < 0.1$). The figures in Table 12 also reveal that an 800k background dictionary leads to virtually the same performance as a very large dictionary comprising 5M words.

Apart from the slightly better recognition performance, our method is more informative and flexible (as is the hybrid approach), in the sense that presumed OOV word regions are phonemically transcribed rather than in terms of provided

[3] Note that better average probabilities for OOV words could probably be estimated if multiple OOV word classes would be considered as in Gallwitz et al. (1996), but since we do not use word classes in our mixed or hybrid approaches, we do not incorporate word classes in this method either.

[4] We do not present results with a larger lexicon extension, as our preprocessing script fails to compile the static FST for a lexicon size of 1.6M words.

in-vocabulary words. This alerts a user that certain segments are likely to contain an OOV word, while it also allows him to optimize the P2G conversion step.

6. Conclusions and future work

In the above sections, we have proposed our own implementation of the mixed language model approach for the detection and phonemic transcription of OOV word regions, as well as a simple table look-up method for converting these transcriptions to word spellings. With this two-stage approach it was possible to achieve a significant reduction of the word error rate with respect to the error rate obtained with baseline systems incorporating a vocabulary of 100k, 200k or 400k words. The proposed OOV recovery approach yields at least the same performance as the popular hybrid approach on an in-domain test, and outperforms the hybrid approach on an out-of-domain test. It was also demonstrated that the better performance of the mixed LMs transfers to higher-order word-level LMs (up to 7-grams). Furthermore, it was shown that the mixed approach outperforms the traditional approach of extending the lexicon (without retraining the LM) of a standard single-stage recognizer. Achieving tangible improvements with the mixed LM approach on (very) large vocabulary tasks required (1) the use of high-order subword *N*-grams (e.g. a phoneme 9-gram), and (2) judicious attention to how the subword LM is trained (train on OOV words only, take their relative frequencies into account, incorporate the non-emitting word end symbol in the *N*-gram, etc.).

As was already indicated in the text, several aspects of the system are open to further improvements. We expect that the most gain can be achieved with a more sophisticated phoneme-to-grapheme conversion technique. The straightforward dictionary-based P2G approach correctly recognizes about 50% of the OOV words in an in-domain task and 30% in an out-of-domain task, but evidently it should be possible to further raise that percentage. One potential avenue for improvement is to allow that phonemic hypotheses can be matched to word sequences rather than to single words only. Such a P2G mapping strategy should definitely consider *N*-gram statistics to figure out if and which orthographic word sequences should be matched to the recognized phonemic transcription. A second potential approach is to improve the phonemic transcription of the proper names that appear in the background lexicon. The so-called *P2P converters* proposed in Réveil et al. (2012) could be useful to reduce the mismatch between the recognized phoneme sequences and the canonical forms found in the background lexicon. A third improvement could be to devise an approximate matching strategy that searches for the best possible match of a phonemic transcription in the background lexicon. That search might be guided using additional information about e.g. the identities of the words that were recognized in the vicinity of the phonemic transcription, similar to what was done in Parada et al. (2010b).

Furthermore, we believe that our proposed mixed approach could become even more effective when using position-dependent triphone acoustic models. This should help in detecting word boundaries, improving the transition towards and from the OOV subword LM in case an OOV word occurs.

Introducing multiple OOV word classes as proposed in Bazzi and Glass (2002) may further enhance the system. Ideally, the OOV words should be clustered in groups that are both acoustically and grammatically disparate. As stated in Demuynck et al. (2009), most OOV words in large vocabulary Dutch speech recognition are proper names and (compound) nouns. Therefore, three possible initial OOV word classes could be “proper names”, “nouns” and “other words”.

Something that is also worth investigating is the creation of optimal OOV word models. The subword *N*-gram model can be fine-tuned in several aspects. Preliminary experiments in that respect for example showed that a more coarse smoothing is beneficial when a large mismatch between the OOV words in the training data and the test data occurs. WER reductions of up to 0.3% absolute can be obtained for the out-of-domain test set with a mixed word-phoneme system that incorporates a 9-gram phoneme LM in which not all seen *N*-grams are included. On the in-domain NBEST set, the performance remains more or less the same.

Finally, we would like to assess our approach for other languages than Dutch, in order to see whether its merits are transferable.

References

Adda-Decker, M., Lamel, L., 2000]. The use of lexica in automatic speech recognition. Lexicon Development for Speech and Language Processing, 235–266.

Asadi, A., Schwartz, R., Makhoul, J., 1990]. Automatic detection of new words in a large vocabulary continuous speech recognition system. In: Proceedings of ICASS, pp. 125–128.

Bazzi, I., Glass, J., 2000, September. Modeling out-of-vocabulary words for robust speech recognition. In: Proceedings of ICSL, pp. 401–404.

Bazzi, I., Glass, J., 2001, September. Learning units for domain-independent out-of-vocabulary word modelling. In: Proceedings of Eurospeech, pp. 61–64.

Bazzi, I., Glass, J., 2002]. A multi-class approach for modelling out-of-vocabulary words. In: Proceedings of ICSL, pp. 1613–1616.

Bisani, M., Ney, H., 2005]. Open vocabulary speech recognition with hybrid flat models. In: Proceedings of European Conference on Speech Communication and Technology, pp. 725–728.

Chung, G., 2000]. Automatically incorporating unknown words in Jupiter. In: Proceedings of ICSL, pp. 3520–3523.

Chung, G., Seneff, S., Hetherington, L., 1999]. Towards multi-domain speech understanding using a two-stage recognizer. In: Proceedings of Eurospeech, pp. 2655–2658.

Conover, W., 1999]. Practical Nonparametric Statistics, vol. 3. John Wiley & Sons, Inc., New York.

Decadt, B., Duchateau, J., Daelemans, W., Wambacq, P., 2001, December. Phoneme-to-grapheme conversion for out-of-vocabulary words in large vocabulary speech recognition. In: Proceedings of ASRU, pp. 413–416.

Demuynck, K., 2001. Extracting, modelling and combining information in speech recognition. Ph.D. thesis. KU Leuven, ESAT.

Demuynck, K., Duchateau, J., Van Compernolle, D., Wambacq, P., 2000]. An efficient search space representation for large vocabulary continuous speech recognition. Speech Communication 30 (January (1)), 37–53.

Demuynck, K., Laureys, T., Compernolle, D.V., Hamme, H.V., 2003]. FLaVoR: a flexible architecture for LVCSR. In: Proceedings of European Conference on Speech Communication and Technology, pp. 1973–1976.

Demuynck, K., Puurula, A., Compernolle, D.V., Wambacq, P., 2009]. The ESAT 2008 system for N-Best Dutch speech recognition benchmark. In: Proceedings of ASRU, pp. 339–344.

Demuynck, K., Roelens, J., Compernolle, D.V., Wambacq, P., 2008]. SPRAAK: an open source speech recognition and automatic annotation kit. In: Proceedings of ICSL, pp. 495–498.

Gallwitz, F., Nöth, E., Niemann, H., 1996]. A category based approach for recognition of out-of-vocabulary words. In: Proceedings of ICSL, pp. 228–231.

Geutner, P., Finke, M., Scheytt, P., 1998]. Adaptive vocabularies for transcribing multilingual broadcast news. In: Proceedings of ICASS, pp. 925–928.

Hetheringthon, I.L., 1995]. A Characterization of the Problem of New, Out-of-Vocabulary Words in Continuous-Speech Recognition and Understanding. Massachusetts Institute of Technology, Boston.

Hieronymus, J., Liu, X., Gales, M., Woodland, P., 2009]. Exploiting Chinese character models to improve speech recognition performance. In: Proceedings of Interspeech, pp. 364–367.

Hirsimäki, T., Creutz, M., Siivola, V., Kurimo, M., Virpioja, S., Pylkkönen, J., 2006]. Unlimited vocabulary speech recognition with morph language models applied to Finnish. Computer Speech and Language 20 (October (4)), 515–541.

Hori, T., Hori, C., Minami, Y., Nakamura, A., 2007]. Efficient WFST-based one-pass decoding with on-the-fly hypothesis rescoring in extremely large vocabulary continuous speech recognition. IEEE Transactions on Audio, Speech, and Language Processing 15 (May (4)), 1352–1365.

Kemp, T., Jusek, A., 1996]. Modelling unknown words in spontaneous speech. In: Proceedings of ICASS, pp. 530–533.

Kessens, J., van Leeuwen, D.A., 2007]. N-best: the northern and southern Dutch benchmark evaluation of speech recognition technology. In: Proceedings of Interspeech, pp. 1354–1357.

Klakow, D., Rose, G., Aubert, X., 1999]. OOV-detection in large vocabulary system using automatically defined word-fragments as fillers. In: Proceedings of Eurospeech, pp. 49–52.

Oostdijk, N., 2000]. Het Corpus Gesproken Nederlands. Nederlandse Taalkunde 5 (3), 280–284.

Parada, C., Dredze, M., Filimonov, D., Jelinek, F., 2010a]. Contextual information improves OOV detection in speech. In: Proceedings of HLT, pp. 216–224.

Parada, C., Sethy, A., Dredze, M., Jelinek, F., 2010b]. A spoken term detection framework for recovering out-of-vocabulary words using the web. In: Proceedings of Interspeech, pp. 1269–1272.

Rastrow, A., Sethy, A., Ramabhadran, B., Jelinek, F., 2009]. Towards using hybrid word and fragment units for vocabulary independent LVCSR systems. In: Proceedings of Interspeech, pp. 1931–1934.

Réveil, B., Martens, J.-P., van den Heuvel, H., 2012]. Improving proper name recognition by means of automatically learned pronunciation variants. Speech Communication 54 (3), 321–340.

Stolcke, A., 2002]. SRILM – an extensible language modeling toolkit. In: Proceedings of International Conference on Spoken Language Processing, pp. 901–904.

Stolcke, A., Zheng, J., Wang, W., Abrash, V., 2011]. SRILM at sixteen: update and outlook. In: Proceedings of IEEE Automatic Speech Recognition and Understanding Workshop, p. 5.

van Leeuwen, D.A., Kessens, J., Sanders, E., van den Heuvel, H., 2009]. Results of the N-Best 2008 Dutch speech recognition evaluation. In: Proceedings of Interspeech, pp. 2571–2574.

Yazgan, A., Saraclar, M., 2004]. Hybrid language models for out of vocabulary word detection in large vocabulary conversational speech recognition. In: Proceedings of ICASS, pp. 745–748.