Hybrid Language Models Using Mixed Types of Sub-lexical Units for Open Vocabulary German LVCSR

M. Ali Basha Shaik, Amr El-Desoky Mousa, Ralf Schlüter, Hermann Ney

Human Language Technology and Pattern Recognition – Computer Science Department
RWTH Aachen University, 52056 Aachen, Germany
{shaik,desoky,schlueter,ney}@cs.rwth-aachen.de

Abstract

German is a highly inflected language with a large number of words derived from the same root. It makes use of a high degree of word compounding leading to high Out-of-vocabulary (OOV) rates, and Language Model (LM) perplexities. For such languages the use of sub-lexical units for Large Vocabulary Continuous Speech Recognition (LVCSR) becomes a natural choice. In this paper, we investigate the use of mixed types of sub-lexical units in the same recognition lexicon. Namely, morphemic or syllabic units combined with pronunciations called *graphones*, normal graphemic morphemes or syllables along with full-words. This mixture of units is used for building hybrid LMs suitable for open vocabulary LVCSR where the system operates over an open, constantly changing vocabulary like in broadcast news, political debates, etc. A relative reduction of around 5.0% in Word Error Rate (WER) is obtained compared to a traditional full-words system. Moreover, around 40% of the OOVs are recognized.

Index Terms: open vocabulary, morpheme, syllable, graphone

1. Introduction

German language is characterized by high lexical variety as a large number of distinct lexical forms can be derived from the same root due to different factors like inflection, derivation, and compounding. This morphological richness leads to high OOV rates and causes data sparsity problems, and high LM perplexities. On the other hand, the open vocabulary LVCSR tasks require the number of recognizable words to almost be infinite like in the open domain dictation, broadcast news transcription, political debates translation, etc. Therefore, the recognition of OOV words is a major challenge for such systems. To improve the OOV recognition rate, sub-lexical LMs are good candidates. Where, the sub-lexical units can be properly combined to produce a wide range of words achieving better lexical coverage, and thus fitting the task of open vocabulary speech recognition.

One of the main issues of sub-lexical language modeling is the proper choice of the sub-word type. A non-careful choice of the sub-word type could increase the WER. A possible type of sub-word is the *morpheme* which is the smallest linguistic component of the word that has a semantic meaning. Normally, morphemes are generated from the full-words by applying word decomposition based on supervised or unsupervised approaches. The supervised approaches make use of linguistic knowledge like in [1], where a set of manual rules is developed for German word decomposition. However, in [2], a manually decomposed lexicon is used for recognition. Other supervised methods rely on carefully built morphological analyzers based on lexical and syntactic knowledge like in [3, 4, 5]. Although the supervised decomposition is normally optimized for high performance, it requires labor-intensive work and still suffers from the so-called *unknown word problem*, that is, words that are not explicitly coded into the system. On the other hand, the unsupervised approaches are statistical based data driven approaches like in [6, 7, 8]. In [9], an algorithm is proposed that decomposes words according to the statistical relevance of the resulting constituents. Other unsupervised methods are based on the Minimum Description Length principle (MDL) like in [10, 11]. On the contrary, the unsupervised approaches are language independent as they do not require any language specific knowledge and can be applied to any language.

Another type of sub-word is the *syllable* which is a phonological building block of words. A German syllable consists at least of a nucleus that can either be a vowel or a diphthong. Consonant clusters can enclose the nucleus and must fulfil the phonotactic restrictions to form a valid syllable [12, 13]. Although syllables are used as sub-lexical units for various languages, little attention is paid to use them for the German language. For example, syllable based LMs are successfully used for languages like Chinese [14], Polish [15], and English [16].

A different type of unit is the *graphone* which is a combination of the graphemic sub-word with its context dependent pronunciation forming one joint unit. Graphones are mainly used to model OOV words. In [17], a set of graphones is used for OOV words in an English ASR task, where the graphones are constructed based on fixed-length sub-words without any linguistic considerations. While, in [18], a set of graphones based on morphemes derived from data driven segmentation is used to model OOV words in a German LVCSR system. Normally, Graphones are automatically derived from Grapheme-to-phoneme (G2P) conversion as illustrated in [19].

So far, extensive research is conducted to improve the OOV recognition rate by using sub-lexical LMs. Therein, no attempt is made to include more than two types of units in the same lexicon and LM, for example: full-words/morphemes or full-words/graphones. In this work, we investigate the use of hybrid lexicons and LMs based on three mixed types of sub-lexical units for building an open vocabulary LVCSR system for German language. For the most frequent in-vocabulary words, normal full-words are used. While, for less frequent in-vocabulary words, graphemic morphemes or syllables are used. Moreover, for OOV words, a set of graphones based on morphemic or syllabic sub-words is added. This mixture of units is hypothesized as a more reliable methodology to achieve better lexical coverage and experimented for an open vocabulary LVCSR system. The experimental results show significant improvements in OOV recognition rates and WERs. To the best of our knowledge, the proposed methodology was not explored before.

2. Methodology

2.1. Morpheme based sub-words

We perform word decomposition using a data driven tool called *Morfessor* [20]. It is a statistical tool that can automatically discover the optimal decomposition for words of a text corpus based on the MDL principle. It is considered a general model for unsupervised induction of morphology from raw text. It is mainly designed to cope with languages having rich morphology, where the number of morphemes per word is varying so much and not known in advance [11]. In our previous publication [18], Morfessor was successfully used to model some fraction of in-vocabulary words leading to significant improvement in WER for a German LVCSR task.

We train our decomposition model using a list of unique words that occur more than 5 times in the LM training data; this gives about 0.5 Million words. We do not include other words in order to avoid irregular words that are harmful to the training process. Nevertheless, the model is still capable of decomposing unseen words. In addition, the resulting decompositions are adapted so as to produce a cleaner set of morphemes and to avoid very short morphemes which are usually difficult to recognize. This is found to be helpful to improve the final WER.

2.2. Syllable based sub-words

The general structure of a German syllable consists of the onset, nucleus and coda. The nucleus is usually a vowel or a diphthong. while the onset and coda are usually optional consonants. The onset is the sound occurring before the nucleus, and the coda is the sound following the nucleus. A syllable without a coda is called a free syllable, while a syllable with a coda is called a closed syllable [21]. Normally, a syllable is not considered valid unless it fulfils certain phonotactic restrictions [12, 13]. We perform syllabification of German words using a phonological rule based tool called *KombiKor v.8.0* [22]. For the same reasons as in case of morphemes, we modify the syllabification output so as to avoid very short syllables. For example, free syllables are merged to the adjacent ones.

2.3. Graphone based sub-words

For words whose pronunciations are unknown, we use a statistical based language independent G2P approach to obtain the missing pronunciations. Our approach is based on joint-sequence grapheme-phoneme models as shown in [19]. Therein, the objective is to find the most likely pronunciation $\varphi \in \Phi^*$ for a given orthographic form $g \in G^*$, where Φ and G are the sets of phonemes and letters respectively:

$$\varphi(g) = \arg\max_{\acute{\varphi} \in \Phi^*} p(\acute{\varphi}, g) \tag{1}$$

We refer to the joint probability distribution $p(\varphi, g)$ as a "graphonemic" joint sequence model. We assume that for each word, its orthographic form and its pronunciation are generated by a common sequence of graphonemic units called *graphones*. Each graphone is a pair $q = (g, \varphi) \in Q \subseteq G^* \times \Phi^*$ of a letter sequence and a phoneme sequence of possibly different lengths. The joint probability distribution $p(\varphi, g)$ is reduced to a probability distribution over graphone sequences $p(q)$ which are modeled by a standard M-gram:

$$p(q_1^N) = \prod_{i=1}^{N+1} p(q_i | q_{i-1}, ..., q_{i-M+1}) \tag{2}$$

This model has two parameters: the order of the M-gram model, and the allowed size of graphones. The number of letters and phonemes are allowed to vary between zero and an upper limit L. Such a model can be trained using Maximum Likelihood (ML) training via the Expectation Maximization (EM) algorithm as presented in [19]. To produce a pronunciation for a given word, we use the maximum approximation over the set $S(g, \varphi)$ of all joint segmentations of g and φ:

$$p(\varphi, g) \approx \max_{q \in S(g,\varphi)} p(q_1, ..., q_L) \tag{3}$$

In the above model, the inventory of graphones Q is automatically inferred from the training data. The letter and phoneme sequences are grouped into an equal number of segments. The number of letters in each segment depends on the parameter L. Normally, we choose the optimum L which gives the minimum Phoneme Error Rate (PER) over a held-out test dictionary. This guarantees the best possible pronunciations for given letter sequences. Here, it is worth noting that the letter sequences do not represent any type of linguistic units; rather, they are groups of letters of almost fixed length.

The set of graphones inferred during G2P training constitutes a graphone model that can be integrated with the normal word model. Thus, it is possible to combine vocabulary entries with sub-lexical graphones to form a unified set of recognition units. In our experiments, we estimate a normal graphone model as described above, then we modify the letter sequences such that they represent morphemes or syllables of the underlying words. To modify the initial graphones, we need to do letter-phoneme alignment. For that, we follow the approach described in [23] based on Dynamic Programming (DP), and Expectation Maximization (EM) algorithms.

2.4. Mixed unit types

As shown in our previous work [18, 3], a sub-lexical language model can perform significantly better than a full-word based language model. The recognition vocabulary is divided into two parts: the N most frequent words are kept as full-words without decomposition, while the rest of the vocabulary consists of sub-words. This prevents the most frequent words from being mixed-up with other sub-words in the search space. In this paper, we follow a similar but improved threefold approach. The value of N is optimized for each type of sub-word (morphemes: $N = 5k$; syllables: $N = 10k$) over the development corpus to obtain the best WER. In addition, we use M graphones ($M = 200k$), where the graphemic sub-word component of the graphone represents either a morpheme or a syllable. To compare this threefold approach to the best conventional twofold sub-lexical approach (full-words + sub-words), we replace graphones with normal morphemes or syllables.

2.5. Experimental considerations

For easy recovery of full-words from sub-words in the recognition output, we attach a '+' sign to the end of each non-boundary sub-word. For example: (*adventswochenenden* → *advents+ wochenenden*; *adventssamstagen* → *advents+ samstagen*). The sequences of graphones are marked differently, For example:(*adventswochenenden* → **advents:a_t_v_E_n_**ts*** **wochenenden:v_O_x_@_n_@_n_d_=n**; *adventssamstagen* → **advents:a_t_v_E_n_t_**z*** **samstagen:z_a_m_s_t_a:_g_=n**). In the case of graphones, we note that the same sub-word could have different pronunciations in different contexts.

On the other hand, we compute the OOV rate of any corpus such that a word is considered an OOV if and only if it is not found in the vocabulary and it is not possible to compose it using in-vocabulary sub-words. We call this *effective OOV rate*.

3. Experimental setup

Our acoustic models are triphone models trained using about 343h of audio material taken from German Broadcast News (BN), European Parliament Plenary Sessions (EPPS), read articles, dialogs, and some web data. The acoustic models are trained based on Maximum Likelihood (ML) method.

Our LM training corpus consists of around 188 Million running full-words including the official data provided for the Quaero project (mainly news data). The text corpus is used for vocabulary selection (M most frequent words) and to estimate back-off N-gram LMs using modified Kneser-Ney smoothing by the SRILM toolkit [24].

Our speech recognizer works in 2 passes. In the first pass, across-word acoustic models are used without speaker adaptation. The second pass performs speaker adaptation based on both Constrained Maximum Likelihood Linear Regression (CMLLR), and Maximum Likelihood Linear Regression (MLLR). In each pass, a 3-gram LM is used to construct the search space and to produce recognition lattices. The lattices are then rescored by a 4-gram LM.

To evaluate the recognition performance, we use the Quaero 2009 development and evaluation corpora (dev09: 7.5h; eval09: 3.8h). Each corpus consists of audio material from EPPS sessions and web sources. Additionally, eval09 has some BN data.

4. Experiments

In this section, we explain our recognition experiments. First, we introduce our baseline experiments. Then, we present results using hybrid sub-lexical language models based on mixed types of units as discussed in Sections 2. At the end, we analyze the advantages and disadvantages of our approaches.

4.1. Baseline recognition

In Table 1, we show the results of our baseline recognition experiments using traditional full-word LMs. We consider the system of 100k full-words as a reference baseline, while the other baselines are listed for comparison purposes.

Table 1: *Baseline recognition results using LMs based on full-words (voc: vocabulary).*

voc size	**Dev09**		**Eval09**	
	OOV [%]	**WER** [%]	**OOV** [%]	**WER** [%]
100k	4.6	32.8	4.5	28.4
300k	2.9	31.2	2.6	27.3
500k	2.4	30.9	2.1	27.1

4.2. LMs based on mixed unit types

In Table 2, we summarize the results of our recognition experiments using LMs based on mixed types of units. We distinguish two main types of experiments: the ones where the basic sub-lexical unit is the morpheme, and the one where the basic sub-lexical unit is the syllable. The vocabulary size is fixed to 300k. Out of that, some initial part consists of full-words (5k in the case of morphemes, and 10k in the case of syllables). This is decided after a series of optimization experiments over dev09 corpus as previously discussed in Section 2.4. the rest of the 300k entries are either fully given as graphemic sub-words (morphemes or syllables) or as a mixture of graphemic sub-words and graphones. In all our experiments, the first 100k entries cover the in-vocabulary words as well as some part of the OOV words of the original 100k full-words vocabulary, and the following 200k entries are augmented to model additional OOV words. The reason behind choosing 200k entries is to achieve nearly similar OOV rate as in the case of 500k baseline (for further OOV analysis refer to Section 4.3). The detailed experiments are given in Table 2. The baseline systems are re-tabulated for easy comparison.

Table 2: *Recognition results using LMs based on mixed unit types along with baseline systems (sbws: sub-words, wrds: words, grfs: graphones, morf: morpheme, slb: syllable).*

sys	**sbws type**	**#full wrds**	**# sbws**	**# grfs**	**Dev09 WER [%]**	**Eval09 WER [%]**
s1	morf	5k	295k	-	31.0	27.1
s2	morf	5k	95k	200k	**31.0**	**27.0**
s3	slb	10k	290k	-	32.5	28.7
s4	slb	10k	90k	200k	32.1	28.5
b1	-	100k	-	-	32.8	28.4
b2	-	300k	-	-	31.2	27.3
b3	-	500k	-	-	30.9	27.1

Checking the results of Table 2, we see that, the use of morpheme based units (s1, s2) is better than using syllable based units (s3, s4). Specifically, the morphemic graphones (s2) perform much better than the syllabic graphones (s4). In addition, the graphone approach (s2, s4) outperforms both the normal morpheme (s1) or syllable (s3) approaches. The best results are achieved for system 's2' using a mixture of units: *5k full-words + 295k morphemes + 200k graphones*. This gives WER reductions of [dev09: 5.5% relative (1.8% absolute); eval09: 5.0% relative (1.4% absolute)] compared to the 100k baseline system (b1). Moreover, the WERs are almost equal to the 500k baseline system (b3), and slightly better for eval09 corpus.

4.3. Experimental analysis

In Table 3, for each system, we record the following values *with respect to the 100k full-words vocabulary*:

- OOV : effective OOV rate.
- COOV: percent of correctly recognized OOVs.
- MIV: percent of mis-recognized in-vocabulary words (negative effect).

Table 3: *Analysis of recognition results using mixed units (OOV : effective OOV rate, COOV: percent of correctly recognized OOVs, MIV: percent of mis-recognized in-vocabulary words).*

sys	**Dev09**			**Eval09**		
	OOV [%]	**COOV [%]**	**MIV [%]**	**OOV [%]**	**COOV [%]**	**MIV [%]**
s1	2.6	34.2	22.9	2.3	39.4	21.8
s2	2.6	34.0	22.8	2.3	39.1	21.7
s3	2.5	30.6	23.6	2.2	36.3	22.9
s4	2.5	33.2	23.7	2.2	37.9	22.9
b1	4.6	-	21.6	4.5	-	21.0
b2	2.9	32.3	22.8	2.6	37.6	21.8
b3	2.4	35.7	22.8	2.1	40.6	21.8

From Table 3, using the proposed mixture of units in system 's2', we achieve a reduction in OOV rate of more than 2.0% compared to the 100k full-words vocabulary. We also see that the effective OOV rate is comparable to the OOV rate of the

500k full-words baseline (b3). Thus, we are able to get a similar OOV rate using only 60% of the vocabulary size (300k vs. 500k). In addition, using the proposed approach, we recognized around 40% of OOV words. On the other hand, the negative effect of mis-recognizing the in-vocabulary words (MIV) due to lexical confusion is kept to minimum having around 0.7% increase for eval09 over the 100k full-words baseline (b3). By comparing 's1' to 's2' for eval09 corpus, we see that the COOV is higher for 's1'. While, the negative effect of MIV is relatively less for 's2', thus leading towards a better WER for 's2' than 's1'. It is also clear that the relative performance of morpheme based units (s1, s2) is better than the syllable based units (s3, s4) in terms of COOV, MIV as well as WER.

5. Conclusions

We investigated the use of mixed types of sub-lexical units for building an open vocabulary LVCSR system for German language. The best results are obtained using a threefold mixture of: *5k full-words + 95k morphemic sub-words + 200k graphones*. We achieved a significant improvement in WER of 5.0% relative (1.4% absolute) for eval09 corpus compared to a 100k full-words baseline. In addition, we achieved a WER reduction of 0.4% relative (0.1% absolute) for eval09 corpus compared to the conventional twofold sub-lexical approach: *5k full-words + 295k morphemic sub-words*. Moreover, we recognized around 40% of the OOVs with respect to the baseline 100k vocabulary. At the same time, the percent of mis-recognized in-vocabulary words is limited to 0.7% for eval09 corpus. The obtained results are almost equal to the 500k baseline experiment. This emphasizes the effectiveness of this open vocabulary approach. For German language, it appears that the use of morphemic sub-words outperforms the use of syllabic sub-words. One of the reasons behind this is that the number of syllables per single word is relatively much higher than the number of morphemes due to the high degree of compounding in German.

6. Acknowledgements

This work was partly funded by the European Community's 7^{th} Framework Programme under the project SCALE (FP7-213850), and partly realized under the Quaero Programme, funded by OSEO, French State agency for innovation.

7. References

[1] M. Adda-Decker and G. Adda, "Morphological decomposition for ASR in German," in *Workshop on Phonetics and Phonology in ASR*, Saarbrücken, Germany, Mar. 2000, pp. 129 – 143.

[2] A. Berton, P. Fetter, and P. Regal-Brietzmann, "Compound words in large-vocabulary German speech recognition systems," in *Proc. Int. Conf. on Spoken Language Processing*, vol. 2, Philadelphia, PA, USA, Oct. 1996, pp. 1165 – 1168.

[3] A. El-Desoky, C. Gollan, D. Rybach, R. Schlüter, and H. Ney, "Investigating the use of morphological decomposition and diacritization for improving Arabic LVCSR," in *Interspeech*, Brighton, UK, Sep. 2009, pp. 2679 – 2682.

[4] W. Byrne, J. Hajič, P. Ircing, P. Krbec, and J. Psutka, "Morpheme based language models for speech recognition of Czech," in *Text, Speech and Dialogue*, ser. Lecture Notes in Computer Science, 2000, vol. 1902, pp. 139 –162.

[5] J. Kneissler and D. Klakow, "Speech recognition for huge vocabularies by using optimized sub-word units," in *Proc. European Conf. on Speech Communication and Technology*, vol. 1, Aalborg, Denmark, Sep. 2001, pp. 69 – 72.

[6] M. Adda-Decker, "A corpus-based decompounding algorithm for German lexical modeling in LVCSR," in *Proc. European Conf. on Speech Communication and Technology*, Geneva, Switzerland, Sep. 2003, pp. 257 – 260.

[7] R. Ordelman, A. V. Hassen, and F. D. Jong, "Compound decomposition in Dutch large vocabulary speech recognition," in *Proc. European Conf. on Speech Communication and Technology*, Geneva, Switzerland, Sep. 2003, pp. 225 – 228.

[8] T. Rotovnik, M. S. Maučec, and Z. Kačič, "Large vocabulary continuous speech recognition of an inflected language using stems and endings," *Speech Communication*, vol. 49, no. 6, pp. 537 – 452, Jun. 2007.

[9] M. Larson, D. Willett, J. Köhler, and R. Rigoll, "Compound splitting and lexical unit recombination for improved performance of a speech recognition system for German parliamentary speeches," in *Proc. Int. Conf. on Spoken Language Processing*, Beijing, China, Oct. 2000.

[10] M. Creutz, T. Hirsimäki, M. Kurimo, A. Puurula, J. Pylkkönen, V. Siivola, M. Varjokallio, E. Arisoy, M. Saraclar, and A. Stolcke, "Morph-based speech recognition and modeling of out-of-vocabulary words across languages," *ACM Transactions on Speech and Language Processing*, vol. 5, no. 1, Dec. 2007.

[11] M. Creutz, "Induction of the morphology of natural language: Unsupervised morpheme segmentation with application to automatic speech recognition," Ph.D. dissertation, Helsinki University of Technology, Finland, 2006.

[12] T. Kemp and A. Jusek, "Modelling unknown words in spontaneous speech," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, vol. 1, Atlanta, GA, USA, 1996, pp. 530 – 533.

[13] B. Möbius, "Word and syllable models for German text-to-speech synthesis," in *Proc. of the Third ESCA Workshop on Speech Synthesis*, NSW, Australia, Nov. 1998, pp. 59 – 64.

[14] B. Xu, B. Ma, S. Zhang, F. Qu, and T. Huang, "Speaker-independent dictation of Chinese speech with 32K vocabulary," vol. 4, Philadelphia, PA , USA, Oct. 1996, pp. 2320 – 2323.

[15] M. Piotr, "Syllable based language model for large vocabulary continuous speech recognition of polish," in *Text, Speech and Dialogue*, ser. Lecture Notes in Computer Science, 2008, vol. 5246, pp. 397 – 401.

[16] C. Schrumpf, M. Larson, and S. Eickeler, "Syllable-based language models in speech recognition for English spoken document retrieval," in *Proc. of the 7th International Workshop of the EU Network of Excellence DELOS on AVIVDiLib*, Cortona, Italy, May 2005, pp. 196 – 205.

[17] M. Bisani and H. Ney, "Open vocabulary speech recognition with flat hybrid models," in *Interspeech*, Lisbon, Portugal, Sep. 2005, pp. 725 – 728.

[18] A. El-Desoky, M. Shaik, R. Schlüter, and H. Ney, "Sub-lexical language models for German LVCSR," in *IEEE Workshop on Spoken Language Technology*, Berkeley, CA, USA, Dec. 2010, pp. 159 – 164.

[19] M. Bisani and H. Ney, "Joint-sequence models for grapheme-to-phoneme conversion," *Speech Communication*, vol. 50, no. 5, pp. 434 – 451, May 2008.

[20] M. Creutz and K. Lagus, "Unsupervised morpheme segmentation and morphology induction from text corpora using Morfessor 1.0," Computer and Information Science Helsinki University of Technology, Finland, Tech. Rep., Mar. 2005.

[21] K. Kohler, "Einführung in die phonetik des deutschen." Berlin, Germany: Erich Schmidt Verlag, 1995.

[22] "Kombikor v.8.0," http://www.3n.com.pl/kombi.php.

[23] R. I. Damper, Y. Marchand, J. D. Marsters, and A. Bazin, "Aligning letters and phonemes for speech synthesis," in *5th ISCA Speech Synthesis Workshop*, Pittsburg, PA, USA, Jun. 2004, pp. 209 – 214.

[24] A. Stolcke, "SRILM - an extensible language modeling toolkit," in *Proc. Int. Conf. on Spoken Language Processing*, vol. 2, Denver, Colorado, USA, Sep. 2002, pp. 901 – 904.