Hierarchical Hybrid Language models for Open Vocabulary Continuous Speech Recognition using WFST

M. Ali Basha Shaik, David Rybach, Stefan Hahn, Ralf Schlüter, Hermann Ney

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

Abstract

One of the main challenges in automatic speech recognition is recognizing an open, partly unseen vocabulary. To implicitly reduce the out-of-vocabulary (OOV) rate, hybrid vocabularies consisting of full-words and sub-words are used. Nevertheless, when using sub-words, OOV rates are not necessarily zero. In this work, we propose the use of separate character level graphones (orthography and phoneme sequence pair) as sub-words to effectively obtain zero OOV rate. To minimize negative effects on the core vocabulary of the most frequent words, a hierarchical language modeling approach is proposed. We augment the first level hybrid language model with an OOV word class, which is replaced by character level graphone sequences using a second-level graphone based character language and acoustic model during search. This approach is realized on-the-fly using weighted finite state transducers. We recognize a significant fraction of OOVs on the Wall Street Journal corpus, compared to the full-word and former hybrid language model based approaches.

Index Terms: open vocabulary, OOV, language model, filler models

1. Introduction

In automatic speech recognition (ASR) system, the words which are not present in the recognition vocabulary are OOV words. Hence, using full-word vocabulary, we cannot recognize them. Currently, most of the state-of-the-art open vocabulary ASRs operate with a hybrid vocabulary containing full-words and sub-words. Though, they would correctly recognize most of the OOVs using sub-words, they still fail to recognize constantly changing words due to limited sub-word vocabularies, high lexical variety e.t.c.,. Moreover, during recognition, an OOV word could be substituted by some in-vocabulary word, leading to neighboring word errors.

In the literature, some of the recent methods focus on OOV detection using confidence scores [1, 2, 3]. Other methods explicitly model OOVs using *filler models* [4, 5, 6]. The combination of these two methods is analyzed in [7]. In general, confidence based methods are commonly used to detect the correctness of hypothesized words. The filler models focus on explicit modeling of the OOVs using hybrid language models.

In [7] it is also demonstrated that the explicit OOV modeling approach is better for OOV detection, and confidence scoring methods perform better in detecting mis-recognitions. Though filler models perform better in recognizing OOVs, OOV rates are not necessarily reduced to zero. Some of the main reasons for non-zero OOV rates are limitations on vocabulary size, and huge lexical varieties leading to data sparseness. Moreover, if the OOVs are directly modeled using characters as sub-words in the language model (LM), this would lead to additional in-vocabulary mis-recognitions.

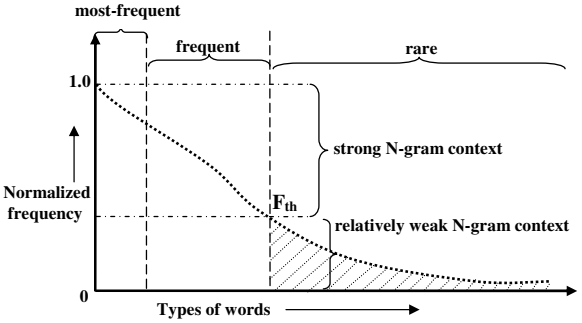

Figure 1: *Word categorization in a text corpus*

On the other hand, according to Zipf's law [8], in a given natural language text, the frequency of a word is inversely proportional to its rank in the frequency table. Thus, we categorize words as most-frequent, frequent and rare words. Examples of most-frequent words are functional words, verbs and adjectives. Frequent words are content or lexical words. Rare words are typical OOVs like proper names, foreign words e.t.c.,. As shown in Figure 1, rare words could be distinguished from frequent words using an experimentally derived cut-off frequency F_{th}. Normally, when unigram counts are computed for the corpus, rare words are found as a long tail (least frequent words). For example, in any LM training corpus, usually a major part of the rare words ($\approx 40\%$) are singletons.

Thus, when N-gram full-word LMs are created, lower probability masses are observed for the rare words (weak N-gram context regions) whose frequency is low due to data sparsity. Similarly, N-gram hybrid LMs also suffer from data sparsity in rare word regions though they can recognize OOVs as sequences of sub-words. Therefore, rare words cannot be effectively utilized in conventional full-word/hybrid LMs. In addition, in the rare word regions, the ASR depends highly on the quality of the acoustic model, leading to phoneme errors.

2. Proposed Method

In this paper, as we focus on a zero OOV rate single pass system, we effectively try to cover all words, i.e., including words unseen in training. Our general concept is similar to [9], who used part-of-speech based multi-class modeling for OOV recognition. In the conventional hybrid LM system, the hybrid vocabulary consists of a selected number of full words (most-frequent) and sub-words from segmented frequent and rare words. Nevertheless, even hybrid vocabularies containing sub-words to not guarantee zero OOV rate in any case, unless the set of all individual characters, or, alternatively all possible combinations thereof for longer sub-word unit lengths are covered in the vocabulary.

Rare words can be effectively represented by sequences of single characters. Nevertheless, we still want to take advantage of modeling words and sub-words observed frequently during LM training. To enable modeling of *any* unknown word, which cannot be represented by the existing hybrid vocabulary of full-words and sub-words, we introduce a character-based LM in a hierarchical way. A first-level hybrid LM comprising frequent full-words and graphone sub-words is augmented with an unknown-word class, which can be hypothesized by the recognizer. Every time the recognizer starts a hypothesis for the unknown-word class, a second-level LM is applied, which hypothesizes on a character by character level using a separate, graphone based character-only LM and corresponding acoustic models. Thus, we refer to our proposed approach as *Hierarchical Hybrid LM*.

3. Methodology

3.1. Graphone Generation

A *graphone* is a sub-word unit derived from aligning a grapheme with its corresponding phoneme sequence. To generate sub-word graphones, we train a statistical joint-sequence grapheme to phoneme (G2P) model. We compute 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 characters respectively as in Eq. 1.

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

A graphone is represented as a pair of phoneme and grapheme sequences $q = (g, \varphi) \in Q \subseteq G^* \times \Phi^*$. The joint phoneme and grapheme sequence probability distribution $p(\varphi, g)$ is reduced to a probability distribution (M-gram) over graphone (sub)sequences q_1^R as:

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

If the number of characters and phonemes are in the range zero and an upper limit *'L'*, the M-gram model is trained using Maximum Likelihood (ML) training using the Expectation Maximization (EM) algorithm as :

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

where $S(g,\varphi)$ is the set of co-segmentations of g and φ. In our experiments, we use an open-source G2P tool [10] for training G2P models. we use M-gram length, M=3. If L is the maximum length of the sub-word grapheme, we generate graphone inventories for 5k, 20k and 64k vocabularies using $L = 4, 4, 3$ respectively. In addition, we generate single character graphone inventory (using L=1) for the aforementioned vocabularies.

3.2. Decision Rule

We distinguish the hybrid word/sub-word vocabulary $\mathcal{W}$ and the separate character vocabulary $\mathcal{C}$. Consider a word sequence of length n: $w_1^n = w_1 ... w_n$ with $w_i \in \mathcal{W} \; \forall \; i = 1, \ldots, n$. Each word or sub-word $w_i \in \mathcal{W}$ is represented by a character sequence $C_i = c_{i,1}^{|C_i|} \in \mathcal{C}^*$ with characters $c_{i,l} \in \mathcal{C} \; \forall \; i = 1, \ldots, n \; \wedge \; l = 1, \ldots, |C_i|$. The function C:$\mathcal{W} \rightarrow \mathcal{C}^*$ maps words/sub-words w to their respective character sequences C(w). We represent N and M as the length of the history for the hybrid LM and the second-level character based LM respectively.

Further, we define the acoustic model distribution $p(x_1^T | w_1^n, C_1^n)$ for an acoustic observation sequence $x_1^T = x_1, \ldots, x_T$ given both a word/sub-word and corresponding character sequence. Formally, the character sequence is added here, to enable modeling words from the unknown-word class w_{oov} by character sequences. For words/sub-words from the hybrid vocabulary the corresponding character sequence aligned to the same word position can be ignored. We then define a hierarchical decision rule to enable recognition with zero OOV, i.e. the recognition of arbitrary character sequences, as given in

$$r(x_1^T) = \underset{n, C_1^n}{\operatorname{argmax}} \max_{w_1^n} \; p(x_1^T | w_1^n, C_1^n) \prod_{l=1}^{n} p(w_l | w_{l-N+1}^{l-1}) \begin{cases} \prod_{m=1}^{|C_l|} p(c_{l,m} | c_{l,m-M+1}^{m-1}) & \text{iff } w_l = w_{oov} \\ 1 & \text{iff } w_l \neq w_{oov} \; \wedge \; \mathrm{C}(w_l) = C_l \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

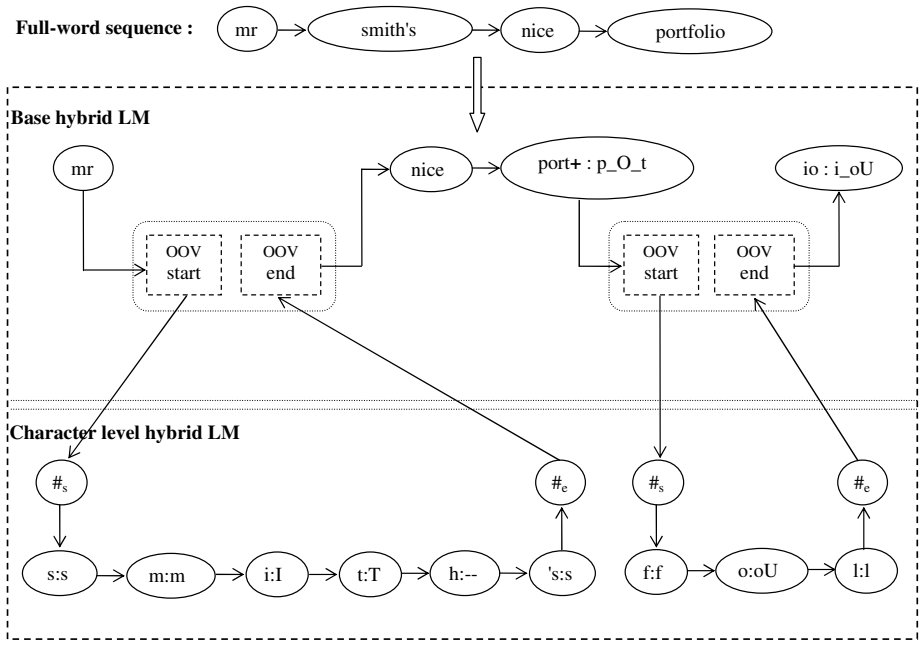

Figure 2: *An example of hierarchical hybrid LM (SAMPA phoneme set notation is used for pronunciations, ':' is a separator between orthography and phoneme sequence in a graphone)*

Eq. 4. The condition in Eq. 4 makes sure, that words/sub-words within the vocabulary are only aligned to their respective character sequences, whereas unknown words can be represented by arbitrary character sequences.

3.3. Decoding using WFST

We use a WFST-based dynamic network decoder, which integrates the LM dynamically as needed during recognition. The composition of the language model transducer $\mathtt{G}$ and the expanded lexicon transducer $\mathtt{C} \circ \mathtt{L}$ ($\mathtt{L}$ is derived from the pronunciation dictionary, $\mathtt{C}$ encodes the context dependency of the acoustic model) is computed on demand using composition filters for on-the-fly pushing of labels and weights as shown by Allauzen in [11]. A detailed description of the decoder can be found in [12] as shown by Rybach.

Here we need to integrate two LMs: the hybrid language model as the base LM and the graphone based character language model for OOVs. Each of these LMs is compiled into a separate weighted automaton. The base LM contains arcs labeled with "unknown" word class. These arcs are replaced by the second-level sub-LM automaton. The fully expanded LM automaton would be too large to be kept in memory. Therefore, we expand these arcs dynamically using OpenFst's `ReplaceFst` [13].

3.4. Basic Example

Using Eq. 4, as shown in Figure 2, we show a basic idea behind hierarchical hybrid LM decoding. In the recognition vocabulary, the most-frequent words like {mr, nice} are preserved as full-words. The sub-word graphones like {port+, io} are used as sub-word vocabulary. The rare-word OOVs are {alexander's, fol+}. The symbols '$\#_s$ and $\#_e$' represents character sequence start or end states respectively. During search, the weighted transition to the character LM is hypothesized for the "unknown (OOV)" word class.

For easy recovery of full-words from sub-words, we mark a '+' sign at the end of the non-boundary sub-words in the base hybrid LM. An example is the word 'portfolio', which is decomposed into 'port+ fol+ io'. Similarly, we also mark the non-boundary characters in the character based LM. In the final step, to form full words, all the recognized sub-word (or character) sequences are post-processed by combining the non-boundary sub-words until the boundary of the sub-word is observed.

In rare cases, half-words could be recognized as a sub-words during recognition. For example, the word like '(ev-) every' could be recognized as 'ev+ every', leading to a word error after post-processing.

4. Experimental Setup

In this work, we try to evaluate our proposed hierarchical hybrid LM systems by comparing to one of the state-of-the-art hybrid LMs in the literature, as shown by Bisani in [5].

Our speech recognizer works in a single pass. We train across-word, triphone based acoustic models using Minimum Classification Error (MCE) method as described by Macherey in [14] using about 80 hours of Nov. '94 North American Business (NAB) audio training

corpus. To create LMs, we use the Wall Street Journal (WSJ) corpus consisting of around 10 Million running full-words. We select 5k, 20k and 64k full-word vocabularies based on word frequency to estimate back-off N-gram LMs using modified Kneser-Ney smoothing by the SRILM toolkit [15].

For recognition evaluation, we select the WSJ corpus comprising of ARPA 1993 and Hub-1 development data, combinedly referred to as "dev 93+94" (812 sentences). The results are also compared on a selected subset (406 sentences) corpus called "dev rare", where the sentences contain rarest words. We use the "dev 93+94" and "dev rare" corpus as test data. Alternatively, We use WSJ evaluation 1995 (300 sentences) corpus as development corpus to differentiate rare words.

5. Differentiating Rare Words

We use a word-frequency based method to empirically identify rare words in the LM training corpora. We also utilize a conventional hybrid LM to determine the rare word cut-off frequency, F_{th} for the full-words as shown in Figure 1. For all the unique words in the full-word LM training corpus, we preserve the top most N words and convert *all* the remaining OOVs into sub-word graphone sequences to create hybrid LM corpus. We select a hybrid vocabulary (U_v: $5k$ full-words and $4k$ graphones) based on word frequencies. We report an OOV Rate of 12.1 [%] for the WSJ 1995 corpus, as the selected vocabulary, U_v do not represent all the words.

We create a hybrid LM (experiment 1) using vocabulary U_v. In the next step, we exclude the singletons ($counts = 1$) from the full-word LM training corpus by explicitly mapping them to 'unknown' and convert the remaining OOVs into sub-word graphone sequences to create an another hybrid LM (experiment 2) using the same vocabulary U_v. Similarly, we exclude the words with $counts \leq 2$ from the full-word LM training corpus and convert the remaining OOVs into sub-word graphone sequences to create the other hybrid LM (experiment 3) using vocabulary U_v. Using these various hybrid LMs, we run the recognition over WSJ 1995 corpus.

Table 1: *Initial recognitions to determine F_{th} using vocabulary [U_v: $5k$ full-words and $4k$ graphones] (expt: Experiment, counts: Unigram counts, WER: Word error rate [%])*

expt	**OOVs converted into graphones in full-word LM corpus to create hybrid LM**	**WER [%]**
1	all	18.6
2	$counts > 1$	18.7
3	$counts > 2$	18.8

As shown in the Table 1, it is observed that all the OOV words converted to graphones in the hybrid LM as in experiment 1 are not much useful. Exclusion of full-word singletons in experiment 2 produced a marginally degraded WER. As WER is further degraded for hybrid LM in experiment 3, we hypothesize that *full-word singletons* as *rare words* for our main hierarchical hybrid LM experiments.

6. Experiments

We conduct a baseline recognition experiments for 5k, 20k and 64k full-word vocabularies. Results are shown in Table 2.

Table 2: *Full-word baseline results (sys: system, OOV: out of vocabulary rate [%], WER: word error rate [%])*

sys.	**dev 93+94**		**dev rare**	
	OOV	**WER**	**OOV**	**WER**
5k	11.2	24.2	15.6	32.0
20k	2.6	11.2	4.6	15.1
64k	0.5	8.8	0.8	10.4

We repeat the hybrid LM experiments, as described by Bisani in [5], which are best in terms of WER and refer them as *reference hybrid* LM. On the other hand, we explicitly exclude all the rare words by mapping them to *unknown* token during LM training in the reference hybrid LM and refer it as *Optimized reference hybrid* LM system.

For the hierarchical hybrid LM experiments, we group all the rare words and sub-word OOVs as a single 'unknown' word category. For the first hierarchical LM experiment, we use the separate sub-word level hybrid LM for the 'unknown' tagged words. All the words in the corpus are converted into 'sub-word graphone sequences' to create separate sub-word level hybrid LM. We use the hybrid vocabulary U_v [U_v: N full-words and K graphones] to create base hybrid LM and the partial hybrid vocabulary [K graphones] to create the sub-word level hybrid LM.

For the second hierarchical LM experiment, we use the separate character level hybrid LM for the 'unknown' tagged words. All the words in the corpus are converted into 'single character graphone sequences' to create separate character level hybrid LM. Here, we use the hybrid vocabulary U_v to create base hybrid LM and the separate graphone character level vocabulary [C graphones] to create the character level hybrid LM.

We construct a 3-gram full-word LMs. Similarly, we use 6-grams for constructing the base hybrid LMs, graphone based sub-word LMs and the graphone based character LMs. Empirically N-gram lengths are optimized. We experiment with 5k, 20k and 64k full-word vocabularies as base vocabularies. The recognition results for all the experiments are shown in Table 3. We also report real time factors (RTF) for the hierarchical hybrid LM exper-

iments, as shown in Table 3.

7. Results

In this section, for all the vocabularies, we analyze word error rates, in-vocabulary word error rates followed by OOV recognition accuracy.

7.1. WER Comparison

As shown in Table 3, for the 5k and 20k experiments, the hierarchical hybrid LM system using the characters in the second-level LM is the best in terms of WER. we obtain the relative WER reductions of 6.9% and 6.6% for the dev 93+94 and dev rare corpus respectively compared to the hybrid LM system in which the rare-words are included in the LM training corpora. Moreover, we report the relative WER reductions of around 4.5% for both the dev 93+94 and dev rare corpus compared to the hybrid LM system in which the rare-words are excluded in the LM training corpora. Similarly, for 20k experiment, using the best system, we achieve the relative WER reductions of 6.3% and 5.7% for the dev 93+94 and dev rare corpus respectively compared to the hybrid LM system in which the rare-words are included in the LM training corpora. In addition, we obtain relative WER reductions of around 3.5% for both the dev 93+94 and dev rare corpus compared to the hybrid LM system in which the rare-words are excluded in the LM training corpora.

We notice that hierarchical hybrid LMs performed better for the systems, 5k and 20k having significant OOV rate in the base hybrid LM. We also notice that, the character based second-level LM is more useful than using the sub-word LM, as the rare words can be better represented using the characters, thus minimizing the data sparsity problem in the language model. For the 64k system, as the base hybrid LM has low OOV rate ($\approx 0.5\%$), we could not obtain improvements.

7.2. In-vocabulary Word Error Comparison

A word is considered as an in-vocabulary, if it is found in the baseline full-word vocabulary. We compute the number of in-vocabulary words misrecognized w.r.t. the baseline full-word vocabulary. As shown in Table 3, for the 5k and 20k experiments, the hierarchical hybrid LM system using the characters in the second-level LM is the best in terms of low in-vocabulary error rates. It is worth noting that all the rare words are not useful to model as the graphone sequences. For the rare words, it is often difficult to obtain the correct pronunciations. For the 64k hierarchical hybrid LM systems, we noticed an increase in the number of in-vocabulary word errors due to low OOV rates in the base hybrid LM.

7.3. OOV Word Recognition Accuracy

As shown in Table 3, for the 5k and 20k experiments, the hierarchical hybrid LM systems recognized more number of OOVs compared to hybrid LMs. For the 20k experiment, though we recognize marginally more number of OOVs using the second-level sub-word LM, we select the system using second-level character LM as the best system as the performance is better in terms of WER.

As shown in Table 3, for the 5k experiment, using the best system, we recognize around 40% and 37% of the OOVs (absolute) compared to its full-word system on dev 93+94 and dev rare corpus respectively. Moreover, we recognized around 17% more OOVs (relative) compared to the to the hybrid LM system in which the rare-words are included, in the LM training corpora, for both the dev 93+94 and dev rare corpus. Similarly, for 20k experiment, using the best system, we recognize around 36% of the OOVs (absolute) compared to its full-word system for both the dev 93+94 and dev rare corpus. Also, we recognize approximately 23% more OOVs (relative) compared to the hybrid LM system in which the rare-words are included, for both the dev 93+94 and dev rare corpus.

We notice that, preferring characters over sub-words in second-level LM is more useful to recognize significant number of OOVs, provided the OOV rates in the base hybrid LM are high. As the characters (graphone) cover all the types of words in the LM training corpora, data sparsity is reduced, further increasing the richness in N-gram context in the LM. On the other hand, for the 64k experiments, we could not recognize more number of the OOVs due to low OOV rate in the base hybrid LM.

8. Conclusions

In this paper, we investigated the use of the hierarchical hybrid LMs, utilizing two different hybrid LMs in a single LM structure, effectively to obtain zero OOV rate. We separately handled the rarely observed OOVs by empirically differentiating them from the frequently occurring OOVs in the full-word LM training corpus. From our experiments, we have shown that using a second-level (graphone) character LM is highly useful to recognize the rarely observed OOVs as well as the sub-word OOVs. We recognized a significant more number of OOVs using proposed hierarchical hybrid LM systems compared to the one of the state-of-the-art hybrid LM systems. We also obtain consistent improvements in terms of WER for different vocabularies. In addition, the proposed LM structure is configurable to exploit various types of LMs in a single-pass with different vocabularies for different word categories.

9. 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.

Table 3: *Recognition results in detail (expt: Experiment, FW voc: Full-word vocabulary, frag: Number of fragments in the vocabulary, WER: Word error rate, IER: Fraction of in-vocabulary words mis-recognized, ORA: OOVs recognized w.r.t. full-word baseline system, Hybrid LM: LM containing full-words along with sub-word graphones, hyb: Hybrid LM, Hierarc. LM: Hierarchical hybrid LM, sub-wrd: Graphone based sub-word LM, char: Graphone based character LM, RTF: Approximate real time factors (using WFST decoder), Y/N: Yes/No)*

Base FW voc.	**Hybrid LM**		**Hierarc. LM**		**frag.**	**Dev 93+94**				**Dev rare**			
		exclude rare OOVs (Y/N)	**1st level LM**	**2nd level LM**		**WER** [%]	**IER** [%]	**ORA** [%]	**RTF**	**WER** [%]	**IER** [%]	**ORA** [%]	**RTF**
5k	–	–	–	–	–	24.2	8.1	–	–	32.0	9.8	–	–
	Y	N	–	–	4085	16.0	9.7	33.5	3.6	21.2	12.2	30.4	4.1
	Y	Y	–	–		15.7	9.5	35.1	–	20.7	12.0	32.1	–
	–	–	hyb	sub-wrd		15.4	9.4	36.7	6.4	20.3	11.8	33.6	7.4
	–	–	hyb	char	4234	**14.9**	**9.2**	**39.3**	6.1	**19.8**	**11.7**	**36.4**	6.9
20k	–	–	–	–	–	11.2	6.8	–	–	15.1	7.9	–	–
	Y	N	–	–	11622	9.6	8.0	29.4	2.8	12.2	9.4	29.9	3.0
	Y	Y	–	–		9.4	7.9	32.7	–	11.8	9.1	33.2	–
	–	–	hyb	sub-wrd		9.1	7.7	**36.7**	4.9	11.6	9.1	**37.3**	5.4
	–	–	hyb	char	11827	**9.0**	**7.6**	36.4	4.2	**11.5**	**9.0**	37.1	4.4
64k	–	–	–	–	–	8.8	7.3	–	–	10.4	8.2	–	–
	Y	N	–	–	14346	8.4	8.2	20.8	3.1	10.0	9.4	21.7	3.3
	Y	Y	–	–		**8.3**	**8.1**	**22.2**	–	**9.8**	**9.2**	**23.1**	–
	–	–	hyb	sub-wrd		8.5	8.3	22.2	5.4	10.2	9.6	23.1	6.0
	–	–	hyb	char	14613	8.5	8.3	20.8	4.7	10.1	9.5	23.1	5.0

10. References

[1] L. Burget, P. Schwarz, P. Matejka, M. Hannemann, A. Rastrow, C. M. White, S. Khudanpur, H. Hermansky, and J. Cernocký, "Combination of strongly and weakly constrained recognizers for reliable detection of oovs," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Las Vegas, Nevada, USA, Mar. 2008, pp. 4081–4084.

[2] H. Lin, J. Bilmes, D. Vergyri, and K. Kirchhoff, "OOV detection by joint word/phone lattice alignment," in *Proc. IEEE Automatic Speech Recognition and Understanding Workshop*, vol. 1, Kyoto, Japan, Dec. 2007, pp. 478–483.

[3] F. Wessel, R. Schlüter, K. Macherey, and H. Ney, "Confidence measures for large vocabulary continuous speech recognition," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Salt Lake City, Utah, USA, May 2001, pp. 288–298.

[4] K. Vertanen, "Combining open vocabulary recognition and word confusion networks," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Las Vegas, Nevada, USA, Mar. 2008, pp. 4325–4328.

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

[6] L. Galescu, "Recognition of out-of-vocabulary words with sub-lexical language models," in *Proc. European Conf. on Speech Communication and Technology*, Geneva, Switzerland, Sep. 2003, pp. 249 – 252.

[7] T. J. Hazen and I. Bazzi, "A comparison and combination of methods for oov word detection and word confidence scoring," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, vol. 1, Salt Lake City, Utah, USA, May 2001, pp. 397 – 400.

[8] R. Ferrer-i-Cancho and R. V. Sole, "Zipf's law and random texts," *Advances in Complex Systems*, vol. 5, no. 1, pp. 1–6, 2002.

[9] I. Bazzi, "Modelling Out-of-Vocabulary Words for Robust Speech Recognition," Ph.D. dissertation, Department of Eelectrical Engeneering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA, 2002.

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

[11] C. Allauzen, M. Riley, and J. Schalkwyk, "Filters for efficient composition of weighted finite-state transducers," in *Proc. the International Conference on Implementation and Application of Automata*, Winnipeg, Canada, Aug. 2010.

[12] D. Rybach, R. Schlüter, and H. Ney, "A comparative analysis of dynamic network decoding," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Prague, Czech Republic, May 2011, pp. 5184–5187.

[13] C. Allauzen, M. Riley, J. Schalkwyk, W. Skut, and M. Mohri, "OpenFst: A General and Efficient Weighted Finite-State Transducer Library," in *Proc. the International Conference on Implementation and Application of Automata*, ser. Lecture Notes in Computer Science, vol. 4783, Jul. 2007, pp. 11 – 23, `http://www.openfst.org`.

[14] W. Macherey, L. Haferkamp, R. Schlüter, and H. Ney, "Investigations on error minimizing training criteria for discriminative training in automatic speech recognition," in *Interspeech*, Lisbon, Portugal, September 2005.

[15] 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.