IMPROVED STRATEGIES FOR A ZERO OOV RATE LVCSR SYSTEM

M. Ali Basha Shaik[1], Amr El-Desoky Mousa[1,3], Stefan Hahn[1,4], Ralf Schlüter[1], Hermann Ney[1,2]

[1]Human Language Technology and Pattern Recognition – Computer Science Department
RWTH Aachen University, 52056 Aachen, Germany
[2]Spoken Language Processing Group, LIMSI CNRS, Paris, France
{ shaik, desoky, hahn, schlueter, ney }@cs.rwth-aachen.de

ABSTRACT

In this work, multiple hierarchical language modeling strategies for a zero OOV rate large vocabulary continuous speech recognition system are investigated. In our previously proposed hierarchical approach, a full-word language model and a context independent character-level LM (CLM) are directly used during search. The novelty of this work is to jointly model the character-level prior and the pronunciation probabilities, to introduce across-word context into the character-level LM, and to properly normalize the character-level LM using prefix-tree based normalization for the hierarchical approach. Significant reductions in-terms of word error rates (WER) on the best full-word Quaero Polish LVCSR system are reported.

Index Terms— OOV, hierarchical, prefix-tree, LVCSR

1. INTRODUCTION

Although modern state-of-the-art decoders handle large vocabularies, LVCSRs still suffer from significant OOV rates and data sparsity [1, 2]. A straight-forward approach would be to increase the full-words to decrease the OOV rate, with the cost of huge search space. But this might not guarantee a decrease in WER. Alternatively, even with the inclusion of sub-lexical vocabularies, a zero OOV rate is not guaranteed unless chosen sub-lexical units cover all possible words [2, 3, 4]. Thus, any non-zero OOV rate LVCSR system, independent of the size or the type of the vocabulary fails to recognize a fraction of the OOVs as their phoneme sequence to word mapping is not present in the lexicon.

One of the solutions to address this problem could be to use a single hybrid LM (HLM) containing a mixture of full-words and characters to satisfy the condition of zero OOV rate. However, the recognition system most likely tends to produce higher recognition errors due to more acoustic confusions introduced by characters [5, 6]. In the literature, character-only LMs (CLMs) using different smoothing methods, are applied in [7]. Different hierarchical models are described in [8, 9, 10]. Multiple class-based LMs are used along with full-word LM to detect OOVs [8]. Similarly, a full-word LM and a CLM are used for a OOV recognition for LVCSR in [9] and later applied for hand-writing recognition [10].

In a hierarchical approach, a full-word LM and a CLM are directly used in search for recognizing in-vocabulary (IV) words and OOV words separately within a WFST framework, unlike the hybrid approach [9]. Thus, the recognition system should be able to output sequences of characters in places of OOVs. The hierarchical approach also provides a configurable framework to use multiple level CLMs, as we demonstrate in this paper. However, a composite network consisting of both the full-word LM and CLM could also be viewed as a single composite LM network during search. This implies that the composite LM network could be normalized at sequence level. Precisely, if a CLM provides zero-probability mass for the IV character-level sequences and non-zero probabilities for the OOV character-level sequences, then the composite network satisfies the condition for sequence normalization. In the literature, a prefix-tree based approach has been addressed to solve this type of problem for a hand-writing recognition task and only perplexities are reported in [11]. In this work, we propose a similar prefix tree approach in conjunction with our previously proposed approach [9] for LVCSR and report word error rates.

On the other hand, it is observed that the counts of the OOV words are generally lower than the counts of the IV words for large corpora. Thus in [9], the probability estimates of the character-level sequences of the OOVs are not robustly estimated in the character-level LM and also the across word context is not taken into account.

2. CHARACTER-LEVEL OOV MODELING

Here, we make an attempt to improve our previous approach by taking across-word context into account for the character sequences to be recognized in the OOV region. Normally, introducing longer across-word context in the model will bring the data sparsity problem into picture. Therefore, we use only one previous word as context. Even though, this could not overcome the data sparsity problem. For this reason, we cluster the context into multiple classes. This is also motivated by

[3]Amr Mousa is now with Technische Universität München, München.
e-mail: amr.mousa@tum.de
[4]Stefan Hahn is now with Nuance Communications, Inc., Aachen.
e-mail: stefan.hahn@nuance.com

the fact that some words are similar to other words based on their semantic meaning and thus can be grouped into classes [12]. In this paper, two different data-driven word clustering methods are used to cluster the context. We train a separate CLM for every class. Each LM is trained using the character-level sequences (of the words) that follow the corresponding class in the LM data. Thus, the resulting CLM for every class is assumed to be richer in context compared to the single class alone, but sparse. This sparsity is reduced by interpolating the resulting LMs with the background CLM.

Alternatively, approaches to open vocabulary recognition necessitate automatic pronunciation generation using grapheme-to-phoneme (G2P) model trained using a generative approach [13, 14]. In principle, G2P, also called a pronunciation model is an alignment model. If high correlation between the pronunciation distributions of the lexical entries and the OOV words is assumed, then an M-gram P2G model (ie., inverted G2P model) could be directly used as a language model in the hierarchical framework. Here the pronunciations of the OOVs are hidden variables necessitating a corresponding joint character and phoneme sequence distribution. Ideally, during recognition, the sum over all possible pronunciations of each OOV word hypothesized would have to be performed, which can nevertheless be approximated by a maximum. As an outline, the following hierarchical strategies are investigated using the best full-word large vocabulary for the Quaero Polish LVCSR task:

a. context independent character-level LM (*cf.* [9])
b. pronunciation model as a language model
c. across-word multi-class context dependent CLMs
d. sequence level prefix-tree normalization for (a,b or c)

The results are compared in-terms of WER. According to our best knowledge, investigations conducted in this paper have never been made in literature for LVCSR tasks.

3. HIERARCHICAL APPROACH

In the first approach, the formulation of the open-vocabulary decision rule as described in our previous work [9] is used. We represent $\mathcal{W}$ and $\mathcal{C}$ as the full-word and the character-level model inventory size respectively. Consider a word sequence of length K: $w_1^K = w_1 ... w_K$ with $w_i \in \mathcal{W} \; \forall \; i = 1, \ldots, K$. Each word $w_i \in \mathcal{W}$ is represented as a character sequence $C_i = c_{i,1}^{|C_i|} \in \mathcal{C}^*$, where $c_{i,l} \in \mathcal{C} \; \forall \; i = 1, \ldots, K \; \wedge \; l = 1, \ldots, |c_i|$. This model is created with a pre-assumption that it can represent all the OOV words. The function C:$\mathcal{W} \rightarrow \mathcal{C}^*$ maps words w to their respective sequences C(w). The acoustic model distribution is $p(x_1^T | w_1^K, C_1^K)$ for an acoustic observation sequence $x_1^T = x_1, \ldots, x_T$ given both a word and the corresponding M-gram sequence. Let L and M represent the context lengths of full-word and character models respectively. The decision rule in search, $r(x_1^T)$, is computed as :

$$r(x_1^T) = \underset{K, C_1^K}{\operatorname{argmax}} \max_{w_1^K} p(x_1^T | w_1^K, C_1^K) \prod_{l=1}^{K} \left[p(w_l | w_{l-L+1}^{l-1}) \cdot \left\{ \begin{array}{ll} \prod_{m=1}^{|C_l|} p(c_{l,m} | c_{l,m-M+1}^{m-1}) & \text{iff } w_l = w_{oov} \\ 1 & \text{iff } w_l \neq w_{oov} \; \wedge \; \text{C}(w_l) = C_l \\ 0 & \text{otherwise} \end{array} \right\} \right] \quad (1)$$

4. INCLUDING ACROSS-WORD CONTEXT

In word clustering, words are grouped into subsets using some similarity measure. Table 1 shows several examples of bigrams listed under three different sets. In each set, all the *italicized words* belong to the same class. For simplicity, it is assumed that some of the non-italicized words are OOVs. It can be seen that the OOV words like (aachen, aeons, office) have a strong relationship with the class of its preceding word. Following this observation, we want to recognize the characters of the OOV word based on the class of its preceding word. In general, OOVs are the least frequent words or unseen words in the training text. Thus, clustering either OOVs or its preceding context is a challenging task as count statistics become weaker as the OOV rate decreases. Nevertheless, OOVs like frequent words follow specific set of grammar rules. In this approach, words are clustered using two data-driven methods namely, (loose) semantic clustering using Singular Value Decomposition (SVD) [15, 16] and Maximum-Likelihood criterion [17]. In SVD clustering, all the words in the text (bigrams) are converted into real valued vectors using word-pair co-occurrence matrix and then SVD is applied. The generated vectors are grouped into ϱ subsets/clusters using standard k-means algorithm. Alternatively, all the words in the text (bigrams) are also grouped into required number of clusters (ϱ) using the *mkcls* tool [17].

Table 1. *Examples showing OOV word dependencies in context with its preceding word*

set-1	set-2	set-3
two meters	*outside* aachen	*my* birthday
twenty aeons	*inside* berlin	*your* office
square meters	*near* köln	*his* wallet

In the first approach (*cf.* Section 3), the probability mass for the OOV character-level sequences is not reliable due to relatively low count estimates compared to the count estimates of the IV character-level sequences in the CLM. Thus, to overcome this problem, all the OOV classes are interpolated with the background CLM. In principle, the parameter $p(c_{l,m} | c_{l,m-M+1}^{m-1})$ could be changed to $p(c_{l,m} | c_{l,m-M+1}^{m-1}, \varrho(w_h))$ in Eq. 1, where $\varrho(w_h)$ represents the class-level sequence of the words in the history. Due to complexity reasons, it is reduced to $p(c_{l,m} | c_{l,m-M+1}^{m-1}, \varrho(w_{l-1}))$. Thus, the latter part of the Eq. 1 within the braces changes to:

$$\left\{ \begin{array}{ll} \prod_{m=1}^{|C_l|} p(c_{l,m} | c_{l,m-M+1}^{m-1}, \varrho(w_{l-1})) & \text{iff } w_l = w_{oov} \\ 1 & \text{iff } w_l \neq w_{oov} \; \wedge \; \text{C}(w_l) = C_l \\ 0 & \text{otherwise} \end{array} \right\} \quad (2)$$

In principle, the conditions are applied in Eqs. 1 or 2 in such way that character-level sequences are enabled only

when w_{oov} class is observed. In practice, the character-level sequences from the CLMs could also represent the words present in the full-word vocabulary. This leads to: $\sum_{w\in\mathcal{W}} p(w|w_h) + p(w_{oov}|w_h)\sum_{w\notin\mathcal{W}} p(C(w)) < 1$, ie., although both the full-word and CLMs are independently normalized, the overall hierarchical LM sums always less than one. But, if the CLM satisfies the constraint $\sum_{w\notin\mathcal{W}} p(C(w)) = 1$ and $\sum_{w\in\mathcal{W}} p(C(w)) = 0$, then $\sum_{w\in\mathcal{W}} p(w|w_h) + p(w_{oov}|w_h)\sum_{w\notin\mathcal{W}} p(C(w)) = 1$. This problem is solved using prefix-tree normalization as described in Section 6.

5. JOINT CHARACTER AND PRONUNCIATION DISTRIBUTION

The joint-sequence M-gram G2P model is converted into a finite state automaton [13]. Here, the input labels are *graphemes* and output labels are *phonemes*. P2G model is generated by interchanging the input and output symbols.

6. LEXICAL PREFIX-TREE NORMALIZATION

Let $\mathcal{W}$ be the set of words present in the vocabulary excluding unknown word symbol U and sentence-end symbol S. Let $\mathcal{C}$ be the set of all characters excluding the word end symbol #, but including the word begin symbol $. If h represents word-level history, then the word-level N-gram model can be represented as $p(w|h)\ \forall\ w \in \mathcal{W} \cup \{\mathtt{U}, \mathtt{S}\},\ h \in \{\mathcal{W} \cup \mathtt{U}\}^*$. If h_c represents character-level history then, a character-level M-gram model can be represented as $p(c|h_c)\ \forall\ c \in C \cup \{\#\},\ h_c \in C^*$. Thus, using the above notations, a character-level prefix tree can be constructed with inner arcs representing characters, leaves representing word ends as indicated by the word end symbol # and sharing a common word-begin node. As shown in Fig. 1, nodes attached directly to characters from words of the full-word lexicon ($\mathcal{W}$) might have outgoing arcs, i.e., successor characters from $\mathcal{W}$ and OOV words. Therefore, the leaves of the tree necessarily denote word-ends with the corresponding word-end symbol, since characters from the alphabet can always be continued to form further, possibly new OOV words. For example, an OOV word character sequence 'TAXON' shares the path of the IV character sequence 'TAX' as its prefix. But, from the prefix-tree it can be seen that the word-ends leaves (#) are exclusive. In other words, all the words, ie. either IV or OOV word have a unique word-end leaf (#). Now, we define the prefix subsets of the words present in the lexicon in the prefix-tree. Let $\mathcal{W}(c_1^N)$ denote a set of all the character-level sequences

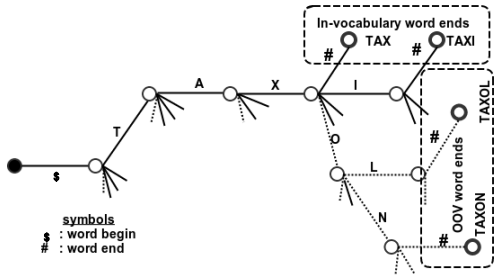

Fig. 1. *An illustration of a lexical prefix-tree*

of the words from the fixed lexicon. A character sequence might also end with the word-end symbol (#). Let $w(c_1^N)$ be a function that returns the word represented by a character sequence c_1^N, and $\mathcal{W}(c_1^N)$ is the set of all words from the fixed vocabulary $\mathcal{W}$ with prefix c_1^N (prefix subsets). Then,

$$\mathcal{W}(c_1^N\#) = \left\{ \begin{array}{cc} w(c_1^N) & \text{iff } w(c_1^N) \in \mathcal{W} \\ \varnothing & \text{iff } N > 0 \wedge w(c_1^N) \notin \mathcal{W} \\ \mathcal{W} \wedge \{\mathtt{S}\} & \text{iff } N = 0 \end{array} \right\} \quad (3)$$

Using Bayes rule, the character sequence probabilities are computed as $p(c_1^N) = \prod_{i=1}^{N} p(c_1|c_1^{n-1})$, which satisfy the following constraints: $\sum_{c\in\mathcal{C}\cup\{\#\}} p(c|h_c) = 1$ and $\sum_{c_*\in\mathcal{C}^*} p(c_*\#) = 1$. Thus, the requirement for sequence-level normalization is the joint probability of character sequences should exclude all the words from the fixed lexicon $\mathcal{W}$, and only represent OOV words. Therefore, we strictly distinguish all the character sequences which end with a word-end symbol (#) and also all the character sequences which do not end with a word-end symbol (#). Thereby, the modified CLM only needs to exclude complete character sequences that represent words from the fixed lexicon $\mathcal{W}$. For example, the modified CLM generated using a prefix-tree approach should provide 'non-zero' probabilities for the word-end terminals of the character sequences of an OOV words 'TAXON' and 'TAXOL', as shown in Fig. 1. It should also provide 'zero' probabilities for the word-end terminals of the IV character sequences 'TAX' and 'TAXI'. Thus prefix-tree based LM can be generated as follows, assuming that $f(h_c)$ is returning the first character in h_c:

$$\bar{p}(c|h_c) = \left\{ \begin{array}{cc} 0 & \text{iff } c = \# \wedge w(h_c) \in \mathcal{W} \wedge f(h_c) = \$ \\ \frac{p(c|h_c)}{1-\sum_{(w\in\mathcal{W}: w=w(h_c)\wedge f(h_c)=\$)} p(\#|h_c)} & \text{otherwise} \end{array} \right\} \quad (4)$$

7. EXPERIMENTAL SETUP

Maximum Likelihood based triphone across-word acoustic models are generated using 110 hours of Polish audio data. The source of audio data is mainly European Parliament Planery Sessions (EPPS) and Broadcast News (BN) articles. For the domain adapted LM training, around 600 Million running full-words from EPPS, news articles, pod-cast, audio-data are used. For the transducer operations, the Openfst toolkit is used [18]. For the experiments, the N-gram backoff LMs (N=3) with modified Kneser-Ney smoothing are estimated using vocabulary sizes : 100k, 200k, 300k and 600k based on word frequency using srilm tools [19]. A background 11-gram CLM is trained on the character sequences of all the words in the corpus using Witten-Bell smoothing. Similarly, multiple 11-gram CLMs are trained dependent on the class of the immediate previous word. They are linearly interpolated with the background CLM. Likewise, a 32-gram prefix-tree normalized LM is generated. A hybrid LM is generated using a mixture of 200k full-words and characters. For the realization of Eq. 1 and 2, the WFST based dynamic decoding is used due to memory constraints. The recognition performance is investigated using the development cor-

pus (Dev10: 3.2h) and the evaluation corpus (Eval10: 3.5h) from the Quaero project.

8. RESULTS

As shown in Table 2, both the development and evaluation corpora have similar OOV rates across different vocabularies. For the full-word systems, although an increase in word perplexity is observed, character perplexity is least affected. The relationship between PPL and WER is not straight forward.

It is observed that the addition of full-words to the existing vocabulary always does not guarantee an optimal WER. It is found that the 100k system has a reasonable OOV rate. 200k system has a lower OOV rate and is the optimal system in terms of WER. Thus, both of these systems are chosen for the multi-class hierarchical experiments to investigate the effect of classes across different vocabularies. The 200k system is chosen for further hierarchical experiments.

Table 2. Selection of an optimal full-word baseline (ppl_w: word level PPL, ppl_c: character-level PPL)

corpus	metric	Vocabulary size			
		100k	**200k**	300k	600k
Dev10	*WER/OOV*	27.6/3.9	**25.5**/2.0	25.9/1.4	26.0/0.7
	ppl_w/ppl_c	432/2.73	485/2.78	609/2.89	686/2.95
Eval10	*WER/OOV*	31.1/4.5	**28.5**/2.3	28.7/1.6	28.7/0.9
	ppl_w/ppl_c	457/2.74	526/2.80	664/2.92	748/2.98

Multi-class hierarchical approach results are shown in Table 3. Here, it is observed that for the 100k vocabulary size, using two or five clusters is beneficial in-terms of the WER. For the 200k vocabulary size, as the OOV rate drops further, clustering did not help, as it is difficult to cluster the OOVs having poor count statistics. Alternatively, due the effect of linear interpolation of the background CLM, better WER is achieved compared to our first hierarchical approach (Section 3). WERs using ML clustering method are not shown in the Table 3, as they are very similar to the WERs obtained using SVD based clustering method. All the experimental results are shown in Table 4 for better comparative analysis. As hypothesized earlier, the system using a hybrid LM produced higher recognition errors due to more acoustic confusions introduced by characters. Limited and consistent improvements in-terms of WER are obtained using the pronunciation model (P2G) or the character-level LM in the hierarchical framework. The reason is generating proper pronunciations for real-world OOVs using a G2P model is complicated [13, 20]. Without prefix tree normalization, best results are obtained using an improved hierarchical model. Alternatively, the prefix tree normalized improved hierarchical model outperformed all other approaches. Using our best system, the reductions in WERs are reported as: [Dev : $\approx$ abs: 0.7%, rel: 2.7%], [Eval: $\approx$ abs: 0.6%, rel: 2.1%]. and, OOV recognition rates are reported as [Dev: 25%, Eval: 23%].

Table 3. *WERs Vs. ML/SVD classes (***FW***: fullword baseline)*

vocab.	corpus	WER[%]				
		FW	No. of classes			
			1	2	5	10
100k	Dev10	27.6	26.5	**26.4**	**26.4**	26.4
	Eval10	31.1	30.1	**29.8**	**29.8**	29.9
200k	Dev10	25.5	**25.0**	25.9	–	–
	Eval10	28.5	**28.0**	–	–	–

Table 4. Detailed Results (**200k+char** : hybrid LM containing full-words and characters, **char** : hierar. approach *cf.* Section 3, **P2G** : hierar. approach *cf.* Section 5, **multi**: improved hierar. approach *cf.* Section 4, **PPL** : character-level perplexity, **CER**: character error rate)

Hierarchical LMs		prefix	PPL	WER	CER
Level-1	Level-2	norm	(char)	[%]	[%]
200k+char	-	-	2.60	25.8	14.3
200k	-	-	2.78	**25.5**	14.0
(Dev10)	P2G	-	2.99	25.3	13.8
	char		2.95	25.3	13.7
	multi		2.93	25.0	13.7
		yes	2.92	**24.8**	13.7
200k	-	-	2.80	**28.5**	20.9
(Eval10)	P2G	-	3.02	28.4	20.9
	char		2.99	28.3	20.9
	multi		2.98	28.0	20.7
		yes	2.98	**27.9**	20.7

9. CONCLUSIONS

We made an attempt to recognize OOVs as a sequence of characters for a zero OOV rate LVCSR system using multiple strategies. Various strategies include incorporation of a pronunciation model or a multi class character level model or an interpolated character level model using the prefix-tree normalization within the hierarchical framework. Limited, yet consistent improvements are obtained using the pronunciation model. In the multi class based approach, the characters of the OOVs are recognized based on the class of its preceding word during decoding. Here, each class-level LM is linearly interpolated with the background character-level LM and then prefix-tree normalization is applied, leading to better probability estimates in the OOV regions compared to our previously proposed approach [9]. Furthermore, our proposed prefix-tree based sequence normalization approach helped in further reducing the word error rate when applied to the multi-class hierarchical LMs. Significant number of OOVs are recognized. The experimental results are found to be statistically significant (under 10% significant level, p-value ≤ 0.1) [21].

10. ACKNOWLEDGEMENTS

This research work was funded by the European Community's 7^{th} Framework Programme under the projects EU-Bridge (FP7-287658) and transLectures (FP7-287755). Hermann Ney was partially supported by a senior chair award from DIGITEO, a French research cluster in Ile-de-France.

11. REFERENCES

[1] Takaaki Hori, Chiori Hori, and Yasuhiro Minami, "Fast on-the-fly composition for weighted finite-state transducers in 1.8 million-word vocabulary continuous speech recognition," in *Interspeech*, Jeju Island, Korea, Oct. 2004, pp. 289 – 292.

[2] M. Shaik, A. El-Desoky, R. Schlüter, and H. Ney, "Hybrid Language Models Using Mixed Types of Sublexical Units for Open Vocabulary German LVCSR," in *Interspeech*, Florence, Italy, Aug. 2011, pp. 1441 – 1444.

[3] Teemu Hirsimäki, *Advances in Unlimited-vocabulary Speech Recognition for Morphologically Rich Languages*, Ph.D. thesis, Helsinki University of Technology, 2009.

[4] M. Shaik, A. El-Desoky, R. Schlüter, and H. Ney, "Using Morpheme and Syllable Based Sub-words for Polish LVCSR," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Prague, Czech Republic, May 2011, pp. 4680 – 4683.

[5] Karmele López de Ipiña, Inés Torres, Lourdes Oñederra, Amparo Varona, and Luis Javier Rodríguez, "Selection of sublexical units for continuous speech recognition of Basque," in *Interspeech*, Beijing, China, oct 2000, pp. 544 – 547.

[6] Karmele López de Ipiña, N Ezeiza, G. Bordel, and M. Graña, "Morphological segmentation for speech processing in Basque," in *Proceedings of IEEE Workshop on Speech Synthesis*, Santa Monica, USA, sept 2002, pp. 187 – 190.

[7] Gerasimos Potamianos and Frederick Jelinek, "A study of n-gram and decision tree letter language modeling methods," *Speech Communication*, vol. 24, no. 3, pp. 171–192, June 1998.

[8] Issam Bazzi, *Modelling Out-of-Vocabulary Words for Robust Speech Recognition*, Ph.D. thesis, Massachusetts Institute of Technology, 2002.

[9] M. Shaik, D. Rybach, S. Hahn, R. Schlüter, and H. Ney, "Hierarchical Hybrid Language Models for Open Vocabulary Continuous Speech Recognition using WFST," in *Workshop on Statistical And Perceptual Audition*, Portland, OR, USA, Sept. 2012, pp. 46 – 51.

[10] M. Kozielski, D. Rybach, S. Hahn, R. Schlüter, and H. Ney, "Open Vocabulary Handwriting Recognition Using Combined Word-Level and Character-Level Language Models," in *ICASSP*, Vancouver, Canada, May 2013, pp. 8257–8261.

[11] M. Kozielski, M. Matysiak, P. Doetsch, R. Schlüter, and H. Ney, "Open-lexicon Language Modeling Combining Word and Character Levels," in *ICFHR*, Crete, Greece, 2014.

[12] P. Brown, P. deSouza, R. Mercer, V. Della Pietra, and J. Lai, "Class-based N-gram models of natural language," *Computational linguistics*, vol. 18, pp. 467 – 479, 1992.

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

[14] Josef R. Novak, Paul R. Dixon, Nobuaki Minematsu, Keikichi Hirose, Chiori Hori, and Hideki Kashioka, "Improving WFST-based G2P Conversion with Alignment Constraints and RNNLM N-best Rescoring," in *Interspeech*, Portland, OR, USA, Sept. 2012.

[15] 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, June 2004, pp. 209 – 214.

[16] A. El-Desoky, M. Shaik, R. Schlüter, and H. Ney, "Morpheme Level Feature-based Language Models for German LVCSR," in *Interspeech*, Portland, OR, USA, Sept. 2012.

[17] Franz Josef Och, "An Efficient Method for Determining Bilingual Word Classes," in *Proc. the Conf. of the European Chapter of the ACL*, Bergen,Norway, June 1999, pp. 71 – 76.

[18] Cyril Allauzen, Michael Riley, Johan Schalkwyk, Wojciech Skut, and Mehryar Mohri, "OpenFst: a general and efficient weighted finite-state transducer library," in *Proc. the International Conference on Implementation and Application of Automata*, Prague, Czech Republic, July 2007, pp. 11 – 23.

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

[20] Panagiota Karanasou and Lori Lamel, "Pronunciation variants generation using SMT-inspired approaches," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Prague, Czech Republic, May 2011, pp. 4908 – 4911.

[21] N. Parihar and J. Picone, "DSR Front End LVCSR Evaluation," Tech. Rep. AU/384/02, Aurora Working Group, European Telecommunications Standards Institute, France, Dec. 2002.