Combining Semantic Word Classes and Sub-Word Unit Speech Recognition for Robust OOV Detection

Axel Horndasch, Anton Batliner, Caroline Kaufhold, Elmar Nöth

Pattern Recognition Lab, Friedrich-Alexander-University Erlangen-Nuremberg, Germany

axel.horndasch@fau.de

Abstract

Out-of-vocabulary words (OOVs) are often the main reason for the failure of tasks like automated voice searches or human-machine dialogs. This is especially true if rare but task-relevant content words, e.g. person or location names, are not in the recognizer's vocabulary. Since applications like spoken dialog systems use the result of the speech recognizer to extract a semantic representation of a user utterance, the detection of OOVs as well as their (semantic) word class can support to manage a dialog successfully. In this paper we suggest to combine two well-known approaches in the context of OOV detection: semantic word classes and OOV models based on sub-word units. With our system, which builds upon the widely used Kaldi speech recognition toolkit, we show on two different data sets that – compared to other methods – such a combination improves OOV detection performance for open word classes at a given false alarm rate. Another result of our approach is a reduction of the word error rate (WER).

Index Terms: OOV detection and classification, semantic word classes, sub-word unit speech recognition, Kaldi

1. Introduction

The out-of-vocabulary word (OOV) problem has always been a serious issue for speech recognition systems and downstream applications. Not only are OOVs the source of recognition errors, but they are also often the main reason for the failure of automated voice searches or human-machine dialogs. Especially content words like person or location names, which are likely to be out-of-vocabulary if they occur rarely, can be of crucial importance. For instance just recently the question *"Give me information about the Argentinian soccer player Jorge Burruchaga!"* was perfectly recognized by a state-of-the-art voice-search system – except for the name of the person in question[1]. The inquiry, which was carried out several times, was never (and probably could not be) recognized correctly by the system; most of the times the result was *portable charger*. A similar question in human-to-human communication could of course also lead to a misrecognition. But the result *portable charger* could be ruled out by a human being because it is not the name of a person and, given the context (*Argentinian*, *soccer player*) and the acoustic information, it would be relatively easy for the listener to get a perfect transliteration of the unfamiliar name in a short clarification dialog.

In the ideal case an automatic system would also make use of the additional information which is contained in such inquiries to come to a better recognition result. In the paper at hand an approach is suggested which is a step in that direction: we describe a way how to configure a speech recognizer which, for correctly hypothesized unknown words, preserves the acoustic information by producing suitable sub-word units and provides the semantic context by assigning a word class.

The rest of this paper will be structured in the following way: in section 2 we will give an overview over the work that has been published in the context of OOV word detection and recovery. Our approach and the implementation based on the Kaldi speech recognition toolkit will be introduced in section 3. After a description of the two speech corpora in section 4 we present the results of our experiments in section 5 and close the paper with a conclusion. (section 6).

2. The OOV Problem – An Overview

The OOV problem still exists in spite of the fact that it has been a very important research topic in the field of automatic speech recognition for a long time and a large number of different approaches have been proposed throughout the years. A good overview can be obtained by studying the following books: [1, 2, 3] and [4].

In this paper we will focus on the use of sub-word units (SWUs) and (semantic) word classes to tackle the OOV problem by creating a one-stage speech recognizer that can be used for a spoken dialog system. We think this is an intuitive solution for the dialog scenario, also because humans seem to apply such strategies in similar situations. For other (off-line) tasks it makes sense of course to include more semantic context when retrieving out-of-vocabulary words like proper names [5].

2.1. Sub-Word Units for OOV Detection

One fundamental concept which emerged relatively early to deal with out-of-vocabulary words was to make use of the simple fact that it takes less sub-word units (SWUs) than words to cover the (virtually unlimited) vocabulary of a language. Tables 3 and 5 provide examples of how the OOV rate goes down if smaller-sized sub-word units are used. Of course this comes at the cost of losing context which is provided by words or larger-size SWUs and which tends to make speech recognition more robust. Apart from linguistically motivated candidates like syllables or phones, a lot of work has gone into finding other units, e.g. data-driven phonetic SWUs (see e.g. [6, 7, 8]) or graphones (see [9, 10, 11]). To use SWUs for OOV detection they are usually added to the vocabulary of the speech recognizer. Regarding the integration of sub-word units into the language model (LM) there are two main approaches (see [3] and [4]):

1. Flat hybrid language models
2. Hierarchical hybrid language models

In the case of *flat hybrid language models*, there is a single language model which includes words and sub-word units. Training text data is usually created by substituting rare words in conventional training data with the corresponding SWUs. This

[1]The Argentinian soccer player Jorge Burruchaga played a major role in the 1986 World-Cup that took place in Mexico scoring the winning goal against Germany in the final.

has a number of advantages, e.g. the dependencies of words and SWUs are modeled automatically. On the other hand it is harder to determine the beginning and the end of an OOV word, especially if the system is limited to one-stage recognition and no complex post-processing is possible.

If *hierarchical hybrid language models* are employed for OOV detection, it means that out-of-vocabulary words get a separate sub-language model which is embedded into the word-level LM. This has the disadvantage that extra training data is needed to train the OOV sub-language model and often an SWU-specific penalty has to be introduced to balance false alarm rate (FAR) and OOV detection rate. For one-stage decoding, however, hierarchical modeling can be useful to get a segmentation of OOV regions, for example if there is an enumeration of unknown names like we observed in the SMARTWEB corpus (see section 4). Also, if there is specific knowledge regarding OOV words to be expected, it can be incorporated when creating the sub-language models. For example if the word class `PERSON_NAME` is of relevance for a task and there is a-priori knowledge about the ethnic affiliation of the names which have to be recognized, the sub-language models can be trained accordingly (e.g. with language-specific SWU sequences).

2.2. Semantic Word Classes

It is a well-known fact that class-based n-gram language models can help reduce test set perplexity as well as the word error rate (WER) in automatic speech recognition [12, 13]. But word classes can also be used to support the detection of out-of-vocabulary words by embedding OOV models to create hierarchical LMs. For example in [14] Part-of-Speech (POS) and automatically derived OOV classes are introduced. The authors of [15] and [16] focus on semantically motivated word classes which are combined with generic or more generalized word models. The idea is to focus on open word classes like person or location names for which OOVs are very common. Approaches which are similar to the one described in this paper – a sub-word unit language model embedded within a word-based language model for covering out-of-vocabulary words combined with a word class for named entities – have been suggested in [17] and recently in [18]. Unfortunately these publications do not focus on the aspect of OOV detection, e.g. no detection and false alarm rates are provided and the number of OOV word classes is limited to one in both cases.

3. Word Class-Based Hierarchical OOV Detection

The motivation for the solution described in this paper was that the out-of-vocabulary problem for both of the data sets we experimented with (see section 4) was a major source of errors for speech recognition. While in many cases the class or category of the unknown words was obvious from the sentence structure (*"I want to go to ..."*, *"Did ... play the title role in the movie ..."*) the speech recognition systems we used produced typical OOV errors, e.g. *"Montag erst" (only Monday)* instead of *"Budapest"* or *"Hamburg okay"* instead of *"Jean-Paul Gaultier"*.

However, in both cases we had – apart from the resources usually needed to train a speech recognition system (speech recordings, transcribed texts, pronunciation dictionary) – extra knowledge regarding the corpora: manually annotated word classes. The categorized words mostly belonged to semantically relevant classes for the task, e.g. city names for the train information system EVAR [19] and the names of celebrities, movie titles etc. for the open domain question-answering system SMARTWEB [20]. To make use of this knowledge, we integrated word classes into the speech recognizer.

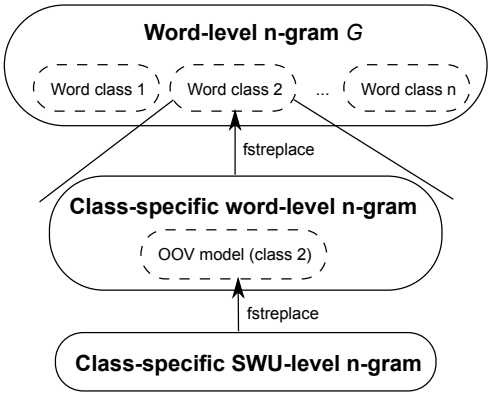

Figure 1: *Hierarchical OOV modeling in Kaldi: For each word class embedded into the language model transducer G an SWU-based OOV model can be created; the OOV model itself is also embedded into the word class using* `fstreplace`.

Formally, non-overlapping word classes can be defined as a mapping $C : \mathcal{W} \rightarrow \mathcal{C}$ which determines a sequence of word classes $\mathbf{c}$ given a sequence of words $\mathbf{w}$ (definition taken from [21]):

$$\mathbf{w} = w_1 \dots w_n \rightsquigarrow C(w_1) \dots C(w_n) = c_1 \dots c_n = \mathbf{c} \quad (1)$$

When using word classes the probabilities of the language model for word sequences $\mathbf{w}$ have to be adjusted. In the case of bigram modeling, the formula

$$P(\mathbf{w}) = P(w_1) \prod_{i=2}^{n} P(w_i|w_{i-1}) \quad (2)$$

needs to be rewritten as

$$P(\mathbf{w}) = P(w_1|c_1)P(c_1) \prod_{i=2}^{n} P(w_i|c_i)P(c_i|c_{i-1}) \quad (3)$$

Following the maximum likelihood principle, the class-related probability of a word $P(w_i|c_i)$ can simply be estimated by counting the number of occurrences of the word divided by the number of all words in the word class (unigram modeling). The (conditional) class probabilities $P(c_i|\dots)$ can be determined in the same way normal n-gram probabilities are computed. The only difference is that the word sequences used for training must be converted to word class sequences.

For our experiments, we had to integrate our word classes as well as the OOV model into the WFST-based (Weighted Finite State Transducers [22]) architecture of the speech recognition toolkit Kaldi [23]. Because Kaldi does not support word classes out of the box, we created all necessary (sub-)language models (see Figure 1) with the SRI language modeling toolkit [24], converted them to WFSTs and used `fstreplace` from the OpenFST library [25] to create a hierarchically structured language model transducer G. The only other change we had to make to the Kaldi recipes was to add self-loops in the lexicon transducer L for the class- and OOV-model-specific disambiguation symbols. The disambiguation symbols are needed to keep the WFSTs determinizable and they have to be added to the L transducer so the word classes and OOV models are not eliminated when composing L and G.

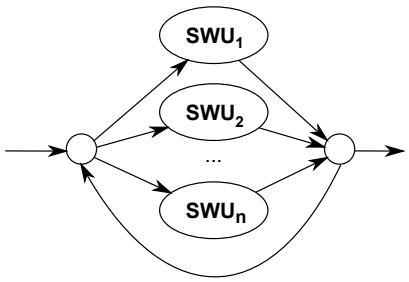

Figure 2: *A simple SWU-based OOV model to be embedded into a word class.*

As indicated in formula 3 we used unigrams for the word class sub-language models but in general any n-gram language model can be incorporated into the word-level LM. To train the SWU-based OOV models embedded in the word classes, we did not introduce any additional data – e.g. other pronunciation lexica – but reused the sub-word units which were already available from the lexica of each corpus. Based on these SWUs (phones, syllables, variable-length phone sequences), we modeled the OOVs by creating simple sub-word unit zerograms with a self-loop (see Figure 2). Again, it is possible to integrate other SWU n-gram language models, but they require additional data and make the overall system more complex.

4. Data Sets

4.1. EVAR train information system

The first data set we studied for this paper was collected during the development of the automatic spoken dialog telephone system EVAR [19]. Just as in [26] we use the subset of 12,500 utterances (10,000 for training and 2,500 for testing) which are the recordings of users talking to the live system over a phone line as opposed to read data which was initially acquired to train the speech recognition module.

Set	Utt.	Words		Syllables	
		Types	Tokens	Types	Tokens
Training	10,000	1,423	34,934	1,118	55,874
Test	2,500	712	8,286	669	13,154
All	12,500	1,603	43,220	1,221	69,028

Table 1: *Statistics regarding words and syllables in the training and test set of the* EVAR *corpus*

Table 1 shows that the vocabulary of the EVAR corpus is limited to 1,603 words or 1,118 syllables; it also indicates the low number of words per utterance on average (less than 3.5). But even though the data set is small, it is interesting for our approach because it contains the open word class CITYNAME. Given the functionality of the overall system – providing callers with the schedule of express trains within Germany – it is not surprising that CITYNAME is the dominating word class and that the 7.05% OOV rate[2] is way above the average rate of 2.98% for the whole corpus (see Table 3).

[2]The definition of *out-of-vocabulary word* for our experiments is simply "a word in the test set which is not in the training set". But for the real EVAR system OOVs occurred very often too, because originally the system's vocabulary only included the names of German cities with express train stations.

Set	Utt.	Data-Driven		Phones	
		Types	Tokens	Types	Tokens
Training	10,000	502	63,468	45	158,671
Test	2,500	400	15,264	43	37,610
All	12,500	502	78,732	45	196,281

Table 2: *Statistics regarding data-driven sub-word units and phones in the training and test set of the* EVAR *corpus*

Apart from syllables and phones, we also created a set of variable phone sequences (labeled *"data-driven"* in Tables 2 and 3) as sub-word units for OOV detection. These data-driven units are based on a mutual-information (MI) measure which is iteratively applied to a set of phones/phone sequences to find the next pair to be merged until a certain number of units has been formed. It has been reported in [2] that such MI units are better suited for OOV detection than smaller-sized units like phones.

SWU / Word Class	Types		Tokens		
	#OOVs	#all	#OOVs	#all	%
Words	180	712	247	8,286	2.98
Syllables	103	669	156	13,160	1.19
Data-driven	0	400	0	15,264	0.00
Phones	0	43	0	37,610	0.00
CITYNAME	41	115	78	1,106	7.05

Table 3: *OOV statistics regarding words, syllables, data-driven SWUs, phones as well as entries of the word class* CITYNAME *in the test set of the* EVAR *corpus.*

4.2. SMARTWEB Handheld Corpus

The *SmartWeb Handheld Corpus*, the second data set we used for our experiments, was collected during the SMARTWEB project. It lasted from 2004 to 2007 and was carried out by a consortium of academic an industrial partners [20]. The recordings of the handheld corpus were made using cell phones and the signals underwent *a complex chain of speech transmissions including Bluetooth and UMTS* [27]. The resulting data was given one of three labels: *normal*, *bad* or *unusable*. For the experiments in this paper only the *normal* recordings were used.

Task	Utt.	Words		Syllables	
		Types	Tokens	Types	Tokens
Training	8,760	5,464	83,776	2,679	151,375
Test	983	1,311	8,887	1,157	17,225
All	9,743	5,787	92,663	2,754	168,600

Table 4: *Statistics regarding words and syllables in the training and test set of the* SMARTWEB *corpus.*

This corpus is interesting because SMARTWEB was designed to be an open-domain question-answering system. As a consequence there are many more open word classes than in EVAR and the vocabulary is much larger as well (5,787 word types, see also Table 4). The out-of-vocabulary rates for the test set for words, SWUs and selected word classes are shown in Table 5: overall the OOV rate on the word level is 4,86% (432 out of 8,887 test tokens); on the syllable level, however, it is only 0,60% (104 out of 17,225 test tokens). The word class for which most out-of-vocabulary words are encountered is CELEBRITY (59 tokens). Due to data sparsity there is the effect that more MOVIETITLEs are unknown (15 out of 18 test tokens) than CITYNAMEs (8 out of 197 tokens).

SWU / Word Class	Test Tokens			Tr. Tokens
	#OOVs	#all	%	#all
Words	432	8,887	4.86	83,776
Syllables	104	17,225	0.60	151,375
Phones	0	46,124	0.00	405,332
CITYNAME	8	197	4.06	1,569
CELEBRITY	59	197	29.95	1,383
MOVIETITLE	15	18	83.33	141

Table 5: *OOV statistics regarding words, syllables, phones as well as entries of the categories* CITYNAME, CELEBRITY *and* MOVIETITLE *in the test set of the* SMARTWEB *corpus.*

5. Experimental Results

To compare our approach using hierarchical hybrid language models in combination with word classes, we made experiments using two other methods to hypothesize OOV words:

1. A flat hybrid language model (words and SWUs)
2. Comparing the results of a word and an SWU recognizer

For the flat hybrid language model (1.) we simply tagged the sub-word units in the training text and the vocabulary. If there were SWUs in the recognition result, they were classified as out-of-vocabulary words. Sequences of adjacent sub-word units were interpreted as one single OOV.

For 2. we combined the output of a word and an SWU recognizer and compared the recognition results (on the SWU level): If the lexical sequence of SWUs for a hypothesized word differed from the sequence of the corresponding hypothesized SWUs, an OOV was assumed. Another way to look at this approach is to think of the combined result as a confusion network.

To asses the performance of our system, we used the standard measures word error rate (WER), detection rate / recall (RCL), precision (PRC), and false alarm rate (FAR) as they are defined for example in [4].

Tables 6 and 7 show that, compared to a baseline speech recognizer, our hierarchical approach with data-driven MI units or syllables improves the word error rate on both data sets. At the same time the method is well-suited to robustly detect out-of-vocabulary words (low false alarm rates ranging from 0.22 to 0.34). Since the hierarchical approach only generates OOVs for certain word classes, the overall recall is relatively low. But if the focus is on OOVs from a specific word class, the detection rates are quite high (see for example the detection rates for CITYNAME on the EVAR corpus in figure 3).

As outlined earlier, we introduced more than one word

Model	WER	OOV Results		
		RCL	PRC	FAR
Baseline	14.7	0.0	0.0	0.00
Baseline + word class CITYNAME	14.5	0.0	0.0	0.00
Flat hybr. OOV syllables	15.5	27.0	50.0	0.83
Combined word / syllable recognition	16.8	**64.0**	27.0	5.24
Hier. OOV: data-driven + word class CITYNAME	**14.0**	21.0	65.0	0.34
Hier. OOV: syllables + word class CITYNAME	14.1	21.0	**75.0**	**0.22**

Table 6: *Word error rates and OOV detection results (recall, precision, false alarm rate) for all OOVs in the test set of the* EVAR *corpus.*

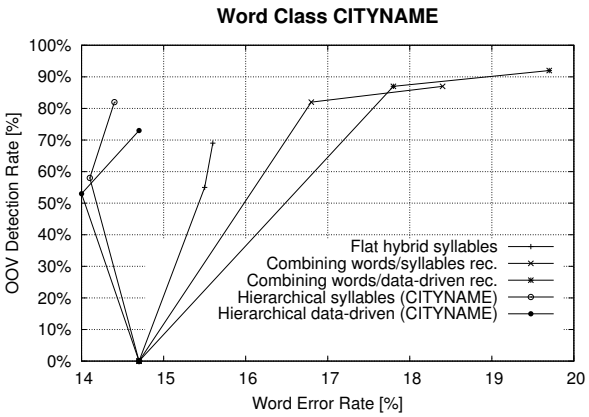

Figure 3: *Comparing OOV detection results for the word class* CITYNAME *and word error rates for different approaches on the* Evar *test set.*

class with embedded OOV models in the speech recognizer for SMARTWEB. But even though this led to correctly detected OOVs with the wrong class label, we were still able to get good results for each word class:

- CELEBRITY: 34 (recall), 63 (precision)
- CITYNAME: 50 (recall), 13 (precision)
- MOVIETITLE: 47 (recall), 54 (precision)

Only the precision for CITYNAMEs was lower than expected. The main reason for this were non-categorized out-of-vocabulary words which were often hypothesized as city names.

Model	WER	OOV Results		
		RCL	PRC	FAR
Baseline	19.7	0.0	0.0	0.00
Baseline + multiple word classes	18.8	0.0	0.0	0.00
Flat hybr. OOV syllables	21.0	42.0	56.0	1.67
Combined word / syllable recognition	21.3	**69.0**	34.0	6.98
Hier. OOV: syllables + multiple word classes	**18.4**	26.0	**80.0**	**0.34**

Table 7: *Word error rates and OOV detection results (recall, precision, false alarm rate) for all OOVs in the test set of the* SMARTWEB *corpus.*

6. Conclusion

In this paper we presented a one-stage speech recognition system with a hierarchical hybrid OOV model that is based on a combination of word classes and sub-word units. With this system, which is an extension of the well-known Kaldi speech recognition toolkit, we showed that out-of-vocabulary words from multiple word classes can be detected with good precision and that the resulting word error rate is reduced compared to a baseline recognizer. The classification of out-of-vocabulary words makes it possible for a spoken dialog system to react in a more appropriate way, especially when an OOV is detected which belongs to a word class that is essential for the success of the dialog. Highlighting an OOV by allowing the user to clarify, for example, can reduce misunderstandings and prevent frustration in spoken human-machine interaction.

7. References

[1] P. Fetter, "Detection and transcription of out-of-vocabulary words in continuous-speech recognition," Ph.D. dissertation, Fachbereich 12 (Elektrotechnik) der Technischen Universtität Berlin, Germany, 1998.

[2] I. Bazzi, "Modelling out-of-vocabulary words for robust speech recognition," Ph.D. dissertation, MIT Department of Electrical Engineering and Computer Science, Cambridge, Massachusetts, USA, 2002.

[3] M. C. Parada, "Learning sub-word units and exploiting contextual information for open vocabulary speech recognition," Ph.D. dissertation, Baltimore, Maryland, USA, 2011.

[4] L. Qin, "Learning out-of-vocabulary words in automatic speech recognition," Ph.D. dissertation, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA, 2013.

[5] I. Sheikh, I. Illina, D. Fohr, and G. Linarès, "Document level semantic context for retrieving OOV proper names," in *Proc. of ICASSP 2016*, Pudong, Shanghai, China, 2016, pp. 6050–6054.

[6] T. Kemp and A. Jusek, "Modelling unknown words in spontaneous speech," in *Proc. of ICASSP 1996*, vol. 1, Atlanta, Georgia, USA, 1996, pp. 530–533.

[7] D. Klakow, G. Rose, and X. Aubert, "OOV-detection in large vocabulary system using automatically defined word-fragments as fillers," in *Proc. of EUROSPEECH 1999*, Budapest, Hungary, 1999, pp. 49–52.

[8] I. Bazzi and J. Glass, "Modeling out-of-vocabulary words for robust speech recognition," in *Proc. of INTERSPEECH 2000*, vol. 1, Beijing, China, 2000, pp. 401–404.

[9] L. Galescu and J. F. Allen, "Bi-directional conversion between graphemes and phonemes using a joint n-gram model," in *4th ITRW on Speech Synthesis, Perthshire, Scotland, UK*, 2001, p. 131.

[10] A. Horndasch, E. Nöth, A. Batliner, and V. Warnke, "Phoneme-to-grapheme mapping for spoken inquiries to the semantic web," in *Proc. of INTERSPEECH 2006*, Pittsburgh, Pennsylvania, USA, 2006, pp. 13–16.

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

[12] P. F. Brown, P. V. deSouza, R. L. Mercer, V. J. D. Pietra, and J. C. Lai, "Class-based N-gram models of natural language," *Comput. Linguist.*, vol. 18, no. 4, pp. 467–479, Dec. 1992.

[13] R. Kneser and H. Ney, "Improved clustering techniques for class-based statistical language modelling." in *Proc. of EUROSPEECH 1993*, Berlin, Germany, 1993, pp. 973–976.

[14] I. Bazzi and J. R. Glass, "A multi-class approach for modelling out-of-vocabulary words," in *Proc. of INTERSPEECH 2002*, Denver, Colorado, USA, 2002.

[15] T. Schaaf, "Detection of OOV words using generalized word models and a semantic class language model." in *INTERSPEECH*, 2001, pp. 2581–2584.

[16] F. Gallwitz, "Integrated Stochastic Models for Spontaneous Speech Recognition," Ph.D. dissertation, Pattern Recognition Lab, Computer Science Department 5, University of Erlangen-Nuremberg, Berlin, Germany, 2002.

[17] S. Seneff, C. Wang, I. L. Hetherington, and G. Chung, "A dynamic vocabulary spoken dialogue interface," in *Proc. of INTERSPEECH 2004*, Jeju Island, Korea, 2004.

[18] P. S. Aleksic, C. Allauzen, D. Elson, A. Kracun, D. M. Casado, and P. J. Moreno, "Improved recognition of contact names in voice commands," in *Proc. of ICASSP 2015*, South Brisbane, Queensland, Australia, 2015, pp. 5172–5175.

[19] W. Eckert, T. Kuhn, H. Niemann, S. Rieck, A. Scheuer, and E. Schukat-Talamazzini, "A Spoken Dialogue System for German Intercity Train Timetable Inquiries," in *Proc. of EUROSPEECH 1993*, Berlin, Germany, 1993, pp. 1871–1874.

[20] W. Wahlster, "SmartWeb: Mobile Applications of the Semantic Web." in *GI Jahrestagung (1)*, 2004, pp. 26–27.

[21] E. Schukat-Talamazzini, *Automatische Spracherkennung – Grundlagen, statistische Modelle und effiziente Algorithmen*. Braunschweig, Germany: Vieweg, 1995.

[22] M. Mohri, F. Pereira, and M. Riley, "Weighted finite-state transducers in speech recognition," *Computer Speech & Language*, vol. 16, no. 1, pp. 69–88, 2002.

[23] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The kaldi speech recognition toolkit," in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. Waikoloa, Hawaii, USA: IEEE Signal Processing Society, 2011.

[24] A. Stolcke, "SRILM - an extensible language modeling toolkit," in *Proc. of INTERSPEECH 2002*, Denver, Colorado, USA, 2002, pp. 901–904.

[25] C. Allauzen, M. Riley, J. Schalkwyk, W. Skut, and M. Mohri, "Openfst: A general and efficient weighted finite-state transducer library," in *Implementation and Application of Automata, 12th International Conference, CIAA 2007*, Prague, Czech Republic, 2007, pp. 11–23.

[26] G. Stemmer, "Modeling variability in speech recognition," Ph.D. dissertation, Pattern Recognition Lab, Computer Science Department 5, University of Erlangen-Nuremberg, Berlin, Germany, 2005.

[27] H. Mögele, M. Kaiser, and F. Schiel, "SmartWeb UMTS speech data collection: The SmartWeb Handheld Corpus," in *Proc. of LREC 2006*, Genova, Italy, 2006, pp. 2106–2111.