DEVELOPMENTS IN LARGE VOCABULARY, CONTINUOUS SPEECH RECOGNITION OF GERMAN

M. Adda-Decker, G. Adda, L. Lamel, J.L. Gauvain

LIMSI - CNRS, B.P. 133, 91403 Orsay, France
{madda,gadda,lamel,gauvain}@limsi.fr

ABSTRACT

In this paper we describe our large vocabulary continuous speech recognition system for the German language, the development of which was partly carried out within the context of the European LRE project 62-058 SQALE. The recognition system is the LIMSI recognizer[1] originally developed for French and American English, which has been adapted to German. Specificities of German, as relevant to the recognition system, are presented. These specificities have been accounted for during the recognizer's adaptation process. We present experimental results on a first test set *ger-dev95* to measure progress in system development. Results are given with the final system using different acoustic model sets on two test sets *ger-dev95* and *ger-eval95*. This system achieved a word error rate of 17.3 % (official word error rate of 16.1 % after SQALE adjudication process) on the *ger-eval95* test set.

1. INTRODUCTION

Porting a speech recognizer to a new language consists mainly in the creation of the language specific speech and language models. This can appear as a rather straightforward process, once you have at your disposal sufficient speech and text databases, together with a pronounciation lexicon for the new language. It is nonetheless useful to have prior knowledge about the language characteristics in determining system parameters such as the vocabulary size or the set of phone symbols to be used. When porting our recognition system to German we have mainly considered two language-dependent aspects. The first one is the high lexical variety, or equivalently the low lexical coverage observed in German compared to English or even to French [1]. Low lexical coverage results, for the recognition system, in high out-of-vocabulary (OOV) rates and thus in higher word error rates, as on average one OOV word gives rise to about 1.6 recognition errors. The second language-dependent aspect deals with spoken German, and more precisely concerns the way in which vowel-initial words and morphemes are marked in the spoken language.

In our first German recognition system S1 we used a 40k lexicon represented with a set of 51 phones including silence. No glottal stop symbol was used in the phone symbol set. In a latter version S2 we used a 64k lexicon to reduce the OOV rate. Here the lexicon is represented by a set of 48 phones including a glottal stop and a reduced vowel set. Results are presented for both systems S1 and S2 on a development test set of 200 sentences allowing to measure the impact of lexical coverage on recognition performance. The efficiency of a glottal stop model on the acoustic level has been assessed using the 64k system S2.

In the next section we discuss in some more detail properties of the German written and spoken language and their impact on a speech recognition system. In Section 3 the development of the German system is presented. In Section 4 experimental results are given and in the final section these results are discussed and future directions for further development are outlined.

2. SPECIFICITIES OF WRITTEN AND SPOKEN GERMAN

Concerning the German written language we started looking at the number of distinct lexical items present in the available training data and compared it to English. Using about 35M words of newspaper text for German[1] and English [2], the number of distinct words are 650k and 165k respectively. These figures as well as lexical coverage for different sized lexica are shown in Table 1. Written German has a very large lexical variety resulting in low lexical coverage, compared to English.

Corpus	*Frankfurter Rundschau*	*Wall Street Journal*
language	*German*	*English*
Training text size	31M	37.2M
#distinct words	650k	165k
5k coverage	82.9%	90.6%
20k coverage	90.0%	97.5%
64k coverage	95.1%	99.6%
20k-OOV rate	10.0%	2.5%

Table 1: Comparison of lexical coverage in German and English.

While the OOV rate of a 20k lexicon is 2.5% for English, it is 10% for German. The use of a 64k vocabulary for

THIS WORK IS PARTIALLY FUNDED BY THE LRE PROJECT 62-058 SQALE.

[1] The *Frankfurter Rundschau* training texts were obtained from the ACL-ECI CDROM distributed by Elsnet and LDC. The text material was preprocessed by Philips for use by the SQALE consortium. Philips also provided the 64k lexicon and trigram LM to the consortium.

English essentially eliminates the OOV problem, but for German the OOV rate is still almost 5%.

This low lexical coverage in German mainly stems from word compounding, combined with number and gender agreement and case declension for articles, nouns, adjectives and past participles. The four cases: nominative, dative, genitive and accusative can generate different forms which are acoustically close (differing most often only in the last phoneme). Further lexical variability (semantically equivalent) is observed for male and neutrum genitive (singular): for example, the genitive form of the nominative *der Hof* (meaning: *the yard*) can be written either *des Hofs* or *des Hofes* (meaning: *of the yard*). Acoustically this results in the possibility of adding a schwa-nucleus syllable at the end of the word. In German all nouns or substantives are written with capitalized first letters and most words can be substantivized, thus generating lexical variability and homophones in recognition. But the major reason of the low lexical coverage in German certainly arises from word compounding. Whereas compound words or concepts in English are typically formed by a sequence of words (e.g.: *speech recognition, the speech recognition problem*), in German words are put together to form a new single word (e.g.: *Spracherkennung, das Spracherkennungsproblem*) which in turn include all number, gender and declension agreement variations.

The OOV problem could be reduced by decompounding compound words, as was done for the numbers during text preprocessing. Decompounding is however a non-trivial task requiring a refined morphological analysis and even sometimes semantic information. Many compounds can result in two and more items depending on the degree of morphological analysis carried out. For example consider the following compound word occuring in the training texts: *Bundesbahnoberamtsrat* (approximate translation: *Federal-Rail-Head-Office-Chief*). The following decompositions are possible and semantically correct:
Bundesbahnoberamtsrat → *Bundes Bahn Ober Amts Rat*
Bundesbahnoberamtsrat → *Bundesbahn Ober Amtsrat*
Bundesbahnoberamtsrat → *Bundesbahn Oberamtsrat*
Other decompositions such as:
Bundesbahnoberamtsrat → *Bundes Bahnober Amtsrat*
are possible, but semantically poor. This example clearly illustrates that word compounding in German constitutes an OOV-source, as long the recognition system considers a word to be an item occuring between two spaces.

We are currently investigating effects of decompounding most frequently used words and morphemes. When simply increasing the lexicon size from 64k to 65k by adding the next thousand most frequent words (65k is at present our system's limit) the relative OOV rate reduction so observed on the training material is 1 %. Decompounding semi-automatically all occurrences of 15k compounds in the training material (approximately 3 % of the compounds) results in a new 65k lexicon with a relative OOV rate reduction of 5 % on the training data and of 4 % on the evaluation test set.

In spoken German vowel-initial words and morphemes are often, but not systematically, preceded by a glottal stop or marked by a glottalization (low and irregular pitch rate) of these vowels. This can affect the whole vowel or can result in 2 segments first a glottalized segment, followed by a normal vowel segment. The example shown in Figure 1 is a spectrogram of the compound word *parteiideologisch*. The spectrum of the vowel-vowel sequence (time 4.3s : diphthong /*ai*/ followed by /*i*/ at the boundary) is quite irregular due to glottal stop. While word-initial glottalization also occurs in other languages such as English, its occurrence is less systematic. In German, glottalized segments are good indicators of morpheme boundaries, a characteristic which, when accounted for in the lexicon and the acoustic models, has led to better recognition.

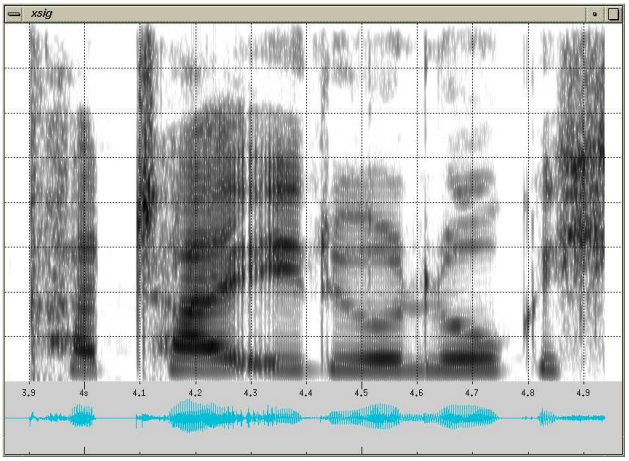

Figure 1: Spectrogram of the compound word *parteiideologisch*. A glottalized initial vowel segment of the morpheme *ideologisch* can be seen around time instant 4.3s (scale: 100ms x 1kHz).

3. SYSTEM DEVELOPMENT

The recognizer makes use of phone-based continuous density HMMs for acoustic modeling and n-gram statistics estimated on newspaper texts for language modeling [1]. The decoding is carried out in multiple passes, where more accurate models are used in successive passes and information is transmitted via word graphs. Thus the first pass which is used to generate the initial word lattice must use accurate enough acoustic models so as to not introduce lattice errors which are clearly unrecoverable with further processing. In practice the graph error rate is small (4.4% on the German SQALE evaluation test set), but poor speakers tend to have higher graph error rates. Note that this rate is necessarily greater than or equal to the observed OOV rate.

Within the SQALE project, speech and text material for training and testing, the phone symbol set and a 64k pronounciation lexicon have been provided to the participants.

At Limsi 40k and 64k bigram language models (LMs) have been trained using the 31 M words of the *Frankfurter Rundschau* text for the S1 and S2 systems respectively. These bigram LMs are used during the first decoding pass to generate word graphs subsequently used with the standard SQALE 64k trigram LM (3 M trigrams, 5 M bigrams).

The acoustic models were trained with the Phondat read speech database, available for research purposes from

the University of Munich. Phondat contains a variety of prompt types including phonetically balanced sentences, a few short stories, isolated letters and train timetable queries. There are a total of 15,000 sentences from 155 speakers. Vocabulary items are rather limited, with only about 1700 different words and the prompt texts are quite different in style from the language model training material (taken from newspaper texts).

During system development, the pronounciation lexicon has been progressively checked and modified. The original lexicon provided within SQALE made use of 52 phone symbols (including glottal stop and silence). For our S1 system, the glottal stop was removed from the transcription lexicon and some minor modifications were carried out on the resulting transcriptions. In the phone symbol set there were 25 vowel symbols, often with 3 different versions for a given vowel type: lax, tense normal and tense long. While a fine transcription can be achieved with the larger phone set (which can be of interest for automatic transcription of the corpus and for explicit modeling of many possible pronounciation variants), it must be ensured that the lexicon is both complete and consistent with regard to such variants. Since it is very difficult to create and maintain a correct and consistent transcription lexicon for large vocabularies, we tried to limit problems due to erroneous or lacking transcriptions by reducing the vowel set. We also manually verified complex consonant clusters to ensure consistency.

For speech recognition purposes, it can be advantageous to allow for pronounciation variation without explicitly adding multiple transcriptions by simply using a smaller symbol set, thus implicitly modeling some of the observed variation. Another argument for reducing the symbol set for the recognizer was that given the relatively small training set lexicon size, many triphones may merely be word-dependent phones. Thus the recognition lexicon of the S2 recognition system was represented with a set of 48 phones (including glottal stop and silence).

Different acoustic model sets were trained after segmentation of the training corpus using the updated phone sets and transcription lexica. In order to model a large number of context-dependent phones we make use of state-tying techniques [3] for the largest models of the S2 system.

The S1 system's acoustic models are derived using 51 phone symbols for training speech transcriptions. All training material, except the longer stories, was used. Sets of 463 and 883 gender-independent context-dependent (CD) phone models were trained. For the S2 system larger acoustic models are trained using a reduced vowel symbol set and an updated transcription lexicon. Here we included the longer stories in the training data, but isolated letters were no longer used. Whitout using the glottal stop symbol 928 gender-independent CD models and 2385 gender-dependent tied-states CD models were trained for the two recognition passes. When using the glottal stop symbol, 986 gender-independent CD models and 2481 gender-dependent tied-states CD models were trained.

4. EXPERIMENTAL RESULTS

Recognition experiments were carried out on two sets of 200 sentences (10 sentences from each of 20 speakers): *Ger-dev95* and *Ger-eval95*. The test data were recorded in the context of the SQALE project by TNO Netherlands, using prompt materials selected from the test portion of the *Frankfurter Rundschau* text data[4]. The test sentences were selected within SQALE so as to limit the OOV rate to be near 2 %. Randomly selecting a set of 400 test sentences from the development texts we observe an OOV rate of more than 7 %.

In Table 2 word error rates of the two systems S1 and S2 on the first test set *Ger-dev95* are compared. The relative error reduction when going from S1 to S2 is 23 %. There are two major reasons for the observed improvement: increased lexical coverage and enhanced acoustic modeling. Extending the lexicon from 40k to 64k words gives about a 10 % relative error reduction, as shown in the lower part of Table 2. To estimate the impact of lexical coverage on the S1 system, without running the costly first decoding pass, we used the 64k-word graphs, produced by the S2 system, as input to the second decoding pass of the S1 system. The error reduction may thus be considered optimistic, but nonetheless indicative. Acoustic modeling improvements account for the remaining 13 % of relative error reduction. These improvements mainly stem from the increased number of CD triphones considered, combined with state tying techniques. But acoustic models have changed along different axes from the S1 to the S2 system: phone symbol set, transcription lexicon, slight changes in acoustic analysis, optional use of glottal stop. The global error reduction achieved here can be estimated to 5 %, as we could measure about 8 % for the increased number of gender-dependent CD phones. A comparative test using an old and new version of the lexicon, without changing anything else in the system showed a relative reduction of 2 %.

System	*configuration*	*%werr*
S2-64k	2500 CDs, gender-dep., opt. glot.	21.8
S1-40k	900 CDs, gender-indep., no glot.	28.4
S1-64k	S2-64k graph, no glott	25.2
S1-64k	S2-64k-graph, opt. glott	24.8

Table 2: The first two lines show recognition results on *ger-dev95* using S2-64k, S1-40k systems. The following lines measure the impact on the S1 system of increased lexical coverage and use of glottal stop symbol.

The S2 system was evaluated within the SQALE project [4], [6], [7], using an ARPA-like procedure and obtained the lowest word error rate, after adjudication, of 18.4 % with the bigram model and 16.1 % with the trigram model on the Ger-eval95 test set.

Beyond the goal of assessing the German recognizer in a task comparable to WSJ in English, we wanted to assess the effectiveness of explicitly modeling the glottal stop in the different system configurations (gender-independent vs. gender-dependent models, bigram vs. trigram LMs). Results are shown for the S2 system in Table 3. The development test data *Ger-dev95* has a slightly higher OOV rate than *Ger-eval95* (2.4 % versus 2.0 %), and a higher word error rate for all conditions.

Using the glottal stop model with the bigram system improves results in most configurations, and its optional use

S2-64k	#acoustic models	gender-dependent	glottal stop	*ger-dev95*	*ger-eval95*
OOV-rate				2.4	2.0
bigram LM	928	no	no	27.6	22.3
bigram LM	986	no	yes	26.2	20.8
bigram LM	928+986	no	optional	26.6	20.8
bigram LM	2385	yes	no	25.3	20.8
bigram LM	2481	yes	yes	25.5	19.6
bigram LM	2385+2481	yes	optional	24.9	19.4
trigram LM	2385	no	no	22.9	19.3
trigram LM	2481	no	yes	23.2	18.5
trigram LM	2385+2481	no	optional	22.1	18.1
trigram LM	2385	yes	no	22.6	18.5
trigram LM	2481	yes	yes	22.9	18.0
trigram LM	2385+2481	yes	optional	21.8	17.3

Table 3: Recognition results (before adjudication) on *ger-dev95* and *ger-eval95* are shown as %Werr (%Werr = %subs+%del+%ins) using the S2-64k system with bigram and trigram LMs. OOV rates are given for the two test sets.

tends to improve the best performance. With the trigram system the use of a glottal stop model only improved the results on the *Ger-eval95* set, giving slightly worse results on the *Ger-dev95* set. However, used optionally, it improves performance on both test sets. The glottal stop model, by absorbing glottalized segments at word boundaries, allows for purer acoustic vowel models. But its use introduces one additional segment in the concerned words' acoustic modeling, and results in a longer duration model, which may or may not be desirable.

5. DISCUSSION

We have presented in this paper our large-vocabulary continuous speech recognizer for German. Concerning the *ger-dev95* test set we measured an improvement of about 20% between the S1 and S2 systems. This error reduction can be mainly attributed to increasing the lexicon from 40k to 64k words, modeling a larger number of context-dependent models using a tied-state approach for parameter sharing, and the use of gender-dependent models. Additional error reduction comes from a sum of modifications, the relative contribution of each being difficult to estimate: a different subset of training sentences was selected (for the S1 system the short stories were not used, whereas isolated letter utterances were included, for the S2 system we did the opposite), the phone set was reduced, glottal stop was optionally used for words starting with a vowel, phonetic transcriptions in the lexicon have been checked for consistency (this is particularly important in German due to the high compound-word rate), alternate transcriptions have been added.

Across the experiments carried out with the S2 system, from the weakest models (bigram LM, small set of gender-independent acoustic models) to the most accurate ones (trigram LM, large set of gender-dependent acoustic models), the relative gain of the optional glottal stop model, increases. An analysis of the errors showed that the glottal stop model often prevented poor word boundary placement and the segmentation of a long word into a sequence of smaller (vowel initial) words, as is often observed for OOV words. The glottal stop model is specific to the German system. This symbol has no distinctive role concerning the phonetic transcription of an isolated word, but in continuous speech its presence often indicates a word or morpheme boundary, and it has proven useful in the recognition system.

Our first steps in decompounding German compound words showed that a significant increase in lexical coverage can be achieved without increasing the system's lexicon. Decompounding seems to be a necessary component of a German recognition system, if it has to guarantee high lexical coverage on general applications like newspaper text dictation.

REFERENCES

[1] J.L. Gauvain et al., "Speaker-Independent Continuous Speech Dictation," *Speech Communication*, **15**, (1-2), 1994.

[2] D.B. Paul and J.M. Baker, "The Design for the Wall Street Journal-based CSR Corpus," *ICSLP-92*.

[3] S. Young, P. Woodland ,"State clustering in hidden Markov model-based continuous speech recognition," Computer Speech & Language **8**, 1994.

[4] H.J.M. Steeneken, D.A. Van Leeuwen, "Multi-Lingual Assessment of Speaker Independent Large Vocabulary Speech-Recognition Systems: the SQALE Project," *Eurospeech'95*.

[5] L. Lamel, M. Adda-Decker, J.L. Gauvain, "Issues in Large Vocabulary, Multilingual Speech Recognition," *Eurospeech'95*.

[6] Ch. Dugast, X. Aubert, R. Kneser, "The Philips Large-Vocabulary Recognition System for American-English, French and German," *Eurospeech'95*.

[7] D. Pye, P. Woodland , S. Young, "Large Vocabulary Multilingual Speech Recognition using HTK" *Eurospeech'95*.