

The 300k LIMSI German Broadcast News Transcription System

Kevin McTait and Martine Adda-Decker

LIMSI-CNRS, BP 133
91403 Orsay cedex, FRANCE
{mctait,madda}@limsi.fr

Abstract

This paper describes improvements to the existing LIMSI German broadcast news transcription system, especially its extension from a 65k vocabulary to 300k words. Automatic speech recognition for German is more problematic than for a language such as English in that the inflectional morphology of German and its highly generative process of compounding lead to many more out of vocabulary words for a given vocabulary size. Experiments undertaken to tackle this problem and reduce the transcription error rate include bringing the language models up to date, improved pronunciation models, semi-automatically constructed pronunciation lexicons and increasing the size of the system's vocabulary.

1. Introduction

This paper reports on ongoing developments to the LIMSI German broadcast news (BN) transcription system in the context of the LE-4 OLIVE project [4]. The LIMSI BN transcription system [5] consists of an audio partitioner and a speech recognizer. From the audio flux, non-speech segments such as music are rejected and homogenous speech segments are identified using an iterative Viterbi segmentation/clustering procedure based on Gaussian mixture models (GMMs). The speech segments are then labelled for gender and bandwidth. The speech recognizer makes use of continuous density Hidden Markov Models (CD-HMMs) with Gaussian mixture for acoustic modelling and 4-gram backoff language models (LM). Word recognition is performed in multiple passes where current hypotheses are used for cluster based acoustic model adaptation prior to the next decoding pass. A more thorough description can be found in [6].

The word error rate (WER) for German broadcast news transcription is generally less impressive than for English. The reason is that German is a highly inflected language. German also has a highly generative compounding process where words (defined as sequences of characters delimited by whitespace) are concatenated to form longer word compounds. The compounding process provides a theoretical infinite upper bound to the size of the vocabulary of German. For example, a relatively new English compound nominal *the speech recognition problem* becomes *das Spracherkennungsproblem* in German. Both these phenomena lead to a much higher out of vocabulary (OOV) rate than English for a given vocabulary size.

Previous attempts to maximize lexical coverage of a highly inflected and generative language such as German include dynamic adaptive vocabularies [8] and morphological decomposition [7] [9]. The LIMSI German system does make use of morphological decomposition, but only to break down the morpheme boundaries of compounds for more successful application of the grapheme to phoneme rules in pronunciation lexicon construction [1]. This paper aims to tackle the OOV problem

Berliner Tageszeitung	1986-99	147M words
Die Welt	1996-98	20M words
Donau Courier	1992-93	7.4M words
Frankfurter Rundschau	1992-93	34.3M words
VDI Nachrichten	1990-91	0.2 M words
Agence France Press	1994-96	36.2M words
Associated Press Worldstream	1993-96	40.5 M words
Deutsche Presse Agentur	1993-96	28.8M words
ARTE transcripts	1994-98	210K words
Total		315M words

Table 1: Training Material

n Words	20k	60k	100k	200k	500k	1000k
OOV (%)	11.5	6.1	4.4	2.7	1.3	0.7

Table 2: OOV rates for n most frequent words over training data

and improve word recognition by updating the LMs with more recent audio transcriptions (section 2), improving the pronunciation lexica and pronunciation models (section 3) and increasing the size of the vocabulary of the LMs (section 4).

2. Updating the Language Models

The German newspaper, newswire and audio broadcast news transcriptions used to train the LM in the OLIVE project are presented in Table 1. They total approximately 315M words. The audio transcripts from ARTE comprise approximately 11 hours of news broadcasts and 22 hours of documentaries. Table 2 shows the OOV rate of the n most frequent words over the training data from 20k to 1M words. The OOV rate for German is higher than that of English, which tends to reach a lexical coverage of over 99% when using a vocabulary of 65k words. The OOV rate for German is usually about 5% when using an LM vocabulary of 65k words.

Since there are no newspaper/newswire training material posterior to 1999 and no audio transcriptions available beyond 1998, new and more recent transcriptions were sought to bring the OLIVE LM up to date. In the absence of new manually transcribed audio data, summaries of transcriptions of news broadcasts were downloaded from the *Tagesschau* archived news website (<http://www.tagesschau.de>). After cleaning and normalisation, approximately 1.5M words, representing approximately 140 hours, were obtained, representing news broadcasts from the period May 2001 to January 2002. However, as summaries they more closely represent written rather than spoken language and are not 'true' transcriptions.

LM	WER (%)	OOV (%)	PPX
Olive 65k	22.1 (20.0)	3.9	612
Olive+Tagesschau 65k	20.0 (18.2)	3.3	500
Olive+Tagesschau 300k	18.9 (17.6)	1.7	466

Table 3: WER, OOV and perplexity for 3 LMs

A new interpolated LM with a vocabulary of 65k words was constructed using the training data in Table 1 and the new data from Tagesschau (Olive+Tagesschau). An additional LM was constructed using the same training material but with a vocabulary of 300k words in order to determine whether increasing the vocabulary size improves performance. Also, a new method of establishing the LM vocabulary was developed. Previously, the n most frequent words from the newspaper/newswire training data were combined with the m most frequent words in the audio transcriptions above a certain frequency threshold. The threshold is varied in order to minimise the OOV rate over an independent development text of the genre of the expected test conditions. A new approach was developed by merging the most *probable* words from both the newspaper/newswire genre of training data and the audio transcriptions. The OOV rate over the development text is minimised by applying a weight to both of these lists. In the same way, it is also possible to weight and merge word lists from various time periods.

The recognizer was tested on 6 news broadcast shows from Tagesschau, independent of and posterior to the training data in Table 1 and the Tagesschau training data, comprising the period April to May 2002. The six shows total approximately 100 minutes. The non-weighted average WER, OOV and perplexity of all six shows, tested with each of the three LMs, are shown in Table 3. Perplexity was measured with respect to a theoretically infinite vocabulary of 100M words, using the method in [3]. The WER, OOV and perplexity of the Olive+Tagesschau 65k model are lower than that of the Olive 65k model (absolute gain of 2.1% in WER). The improved result may be on account of the proximity, in terms of time period and genre, between the Tagesschau training and test data. The lowest WER (and OOV and perplexity) was produced when using the 300k LM (absolute gain of 3.2% over the Olive 65k model), indicating the importance of increasing the size of a LM vocabulary.

The WER given in parentheses indicates the use of a rewrite file in scoring that normalises errors caused by the incorrect transcription of compound nouns. For example, if *Tourismusminister* were (erroneously) transcribed as *Tourismus Minister*, it is scored as correct as opposed to an insertion plus substitution. This is carried out for two reasons. First, errors of this nature are less ‘serious’ but also it indicates how important it would be to tackle the problem of compounds in German.

3. Acoustic Modelling

The German BN transcription system makes use of CD-HMMs with Gaussian mixture for acoustic modelling. They were trained on 24 hours of audio data from ARTE. The number of triphone contexts remained the same as for the OLIVE system, where models of approximately 5k phone contexts, 16 Gaussians in the first decoding pass and 11k phone contexts for the second and third passes with 16 and 32 Gaussians respectively were used. No gender-specific and no telephone bandwidth models were created. Varying the sizes of the acoustic models in terms of phone contexts brought no appreciable gain.

φ	example	φ	example	φ	example
b	bald	i	wieder	Q	eins
d	die	I	wider	q	aus
g	gar	y	föhlen	c	neun
p	paar	Y	föhllen	R	über
t	tun	e	zehn	l	speziell
k	kalt	E	Zelt	4	vor
f	vor	9	zählt	l	los
s	dass	X	kalte	r	rasch
S	schon	a	aber	m	mehr
J	gleich	A	ab	n	noch
v	wann	x	öffnen	G	eng
z	so	@	Öfen	N	kalten
Z	Genre	o	ober	M	kaltem
j	jung	O	ob	L	übel
K	doch	u	tun	&	filler
h	hier	U	nun	H	breath
				.	silence

Table 4: The 49 German phones (φ) including syllabic /NML/. Glottal stop has been discarded.

3.1. Pronunciation Modelling

A set of 46 and 49 phones were used for German. The set of 49 phones is given in Table 4, where for each phone, an example of its occurrence in a German word is given. The phone set includes three special phones for breath, silence and filler words. The difference between the set of 46 and 49 phones is the addition of the three syllabic phones /NML/. They are used as an alternative to /schwa-vowel+[nml]/ in the case of morpheme-final unstressed syllables. For example, the pronunciation of *haben* which is normally /habXn/ would become /habN/. HMMs are not particularly good at modelling duration, and using one model as opposed to two halves the minimum duration required. Therefore, the 3 pronunciation variants were designed in order to produce shorter and better adapted pronunciation models. Building on the study undertaken in [2], experiments were undertaken to observe the effect on the WER when including /NML/ as pronunciation variants in the lexicon.

3.2. Pronunciation Lexicons

The pronunciation lexicons are derived from a grapheme-to-phoneme converter developed at LIMSI to provide full or canonical pronunciations. It is a PERL script including approximately 350 rules for standard German words, most common exceptions, foreign words and acronyms. Automatic creation of pronunciations is important given the potential size of the vocabulary (300k words here). However, it is possible to intervene manually in the process of lexicon creation. Manual verification of the n most frequent words in the training texts was carried out in order to either resolve errors in the grapheme-to-phoneme converter or to add pronunciation variants.

One source of errors arise from grapheme-to-phoneme ambiguities such as *st* \rightarrow /St/ or /st/. If no morphological decomposition is applied, a word such as *Morgenstunde* is transcribed as /mORgNstUndX/ as opposed to its correct pronunciation /mORgNStUndX/. Further ambiguities arise from letter sequences such as *ge* \rightarrow /ge/ or /gX/. Another problem concerns reduction where very frequently occurring words and word sequences, in particular frequently occurring long words,

LM	46 auto	46 semi	49 auto	49 semi
Olive 65k	24.0	22.3	24.7	22.1
Olive+Tagesschau 65k	21.1	20.3	21.7	20.0
Olive+Tagesschau 300k	19.9	18.9	19.9	18.9

Table 5: auto and semi-auto lexicons, 46 & 49 phones

are shortened due to poor or loose articulation (words in German are longer on average than in English). In severe cases, a speaker may omit an unstressed syllable completely. In this case, alternative reduced pronunciations are required. For example, a frequent word such as *gehen*, whose pronunciation becomes /geXn/ as a result of the grapheme to phoneme converter, is more often found as /gen/ in real speech. Long words, particularly numbers and dates, are problematic where, for example, *neunundneunzig* (/ncnUntncntsIJ/) may be reduced to /ncncn-sIJ/. Frequent word sequences such as *haben wir* (/habXn vi4/) are often reduced to /ham vX/ or even /ham X/.

Therefore, two types of pronunciation lexicon are available: a lexicon created automatically using only the grapheme to phoneme rules and a semi-automatic lexicon created using the same grapheme to phoneme converter plus a degree of manual intervention as outlined above. This enables us to study the impact of manual intervention on the WER. Also available are two different phone sets for German: one containing 46 phones and the other containing 49, the difference between them being the addition of the pronunciation variants /NML/. This enables us to make a 4-way comparison using an automatic lexicon using 46 phones, an automatic lexicon using 49 phones, a semi-automatic lexicon using 46 phones and a semi-automatic lexicon using 49 phones. Four new sets of acoustic models were trained corresponding to the the phone set and lexicon production method (automatic or semi-automatic). The size of the acoustic models was kept at approximately the same size as the previous models. The same test data from Tagesschau was also used. The WER for each configuration is given in Table 5.

Table 5 shows that an absolute gain of between 0.8% and 2.6% can be achieved by using a semi-automatic as opposed to an automatic lexicon when using the 65k LMs. However, when comparing the WER between equivalent lexicons created using 46 or 49 phones, the situation is less clear. When using automatic lexicons, a lower WER is observed using 46 as opposed to 49 phones. However, when using semi-automatic lexicons, an absolute gain in WER of between 0.2% and 0.3% is observed when using 49 phones as opposed to 46 phones. In any case, when using the 65k models, the lowest WER was achieved using a semi-automatic lexicon with a set of 49 phones (Olive+Tagesschau LM). When a 300k LM is used, there is no difference in WER when either 46 or 49 phones are used. Vocabulary size appears to be the more important factor. However, there is still an absolute gain of 1.0% when using a semi-automatic as opposed to automatic lexicon. Therefore, in conclusion, a semi-automatic lexicon would appear to perform better than an automatic lexicon, but the the set of 49 phones improves performance only for “smaller” vocabularies.

4. Increasing LM Vocabulary Size

Given that the lowest WER in the experiments in section 2 was observed with a LM of 300k words, it was logical to con-

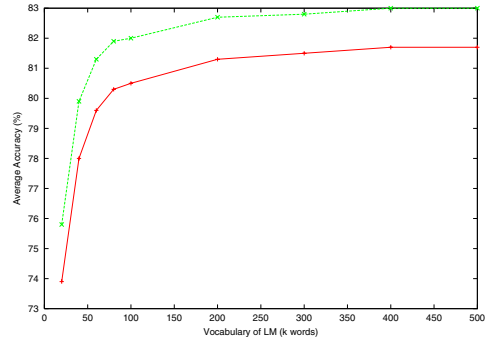


Figure 1: LM vocabulary size vs. performance

	WER	OOV	PPX	Speed
20k	26.1 (24.2)	6.9	701	7.25xRT
40k	22.0 (20.1)	4.5	558	7.52xRT
60k	20.4 (18.7)	3.5	514	7.88xRT
80k	19.7 (18.1)	2.8	489	7.86xRT
100k	19.5 (18.0)	2.6	478	8.00xRT
200k	18.7 (17.3)	1.8	456	8.50xRT
300k	18.5 (17.2)	1.4	444	8.93xRT
400k	18.3 (17.0)	1.2	441	9.41xRT
500k	18.3 (17.0)	1.1	440	9.87xRT

Table 6: Performance of LMs (20-500k words)

sider increasing vocabulary further in order to establish the relationship between LM vocabulary size and performance. LMs with vocabularies ranging from 20k to 500k words were built using the training data outlined in Table 1 and the Tagesschau transcription summaries. Semi-automatic pronunciation lexica based on a 49 phone set were used. The same 6 Tagesschau shows formed the test data. The size of the acoustic models remained constant at 5k phone contexts, 16 gaussians, in the first decoding pass and 11k phone contexts in the second and third passes with 16 and 32 gaussians respectively. Table 6 shows the average non-weighted WER, OOV and perplexity of all 6 Tagesschau shows for each of the LMs. The WER in brackets indicates the use of a compound normalisation rewrite file in scoring. The recognizer speed is also included. The table clearly shows that WER, OOV and perplexity drop as the LM vocabulary size increases, decoding time increases, but remains with the 10xRT limit.

Figure 1 shows that the average performance (100 - WER) of a German LM improves as the size of its vocabulary increases, but levels off after a certain point and does not degrade. Moving from a 400k to 500k word vocabulary has no effect on the WER. The curve at the top of the graph indicates the performance if the problem of compounds were solved using the rewrite file discussed above. This curve corresponds to the WER scores in parentheses in Table 6.

The graph in Figure 2 shows how the perplexity of a German LM drops as the vocabulary size increases. Furthermore, the perplexity begins to level off after a certain size. Addition of further training material has little or no effect.

The OOV rate also drops as the vocabulary size increases, levelling off after a certain point (see Figure 3). However, even

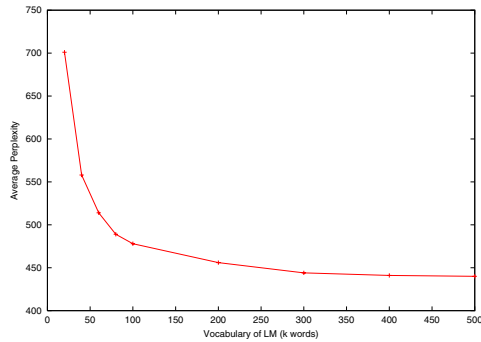


Figure 2: LM vocabulary size vs. perplexity

at 500k words, the OOV rate (0.9%) still remains above that reported for U.S. English using a 65k vocabulary.

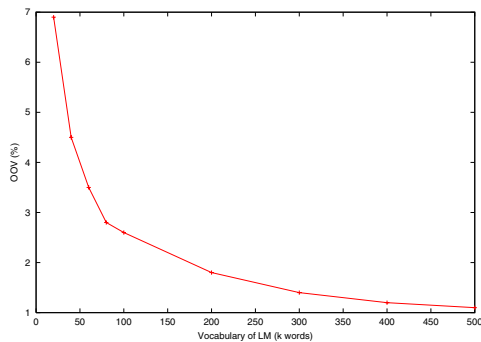


Figure 3: LM vocabulary size vs. OOV rate

As outlined in Table 6, the speed of the decoder increases as the size of the vocabulary of the LMs increases (see Figure 4). The relationship appears to be linear.

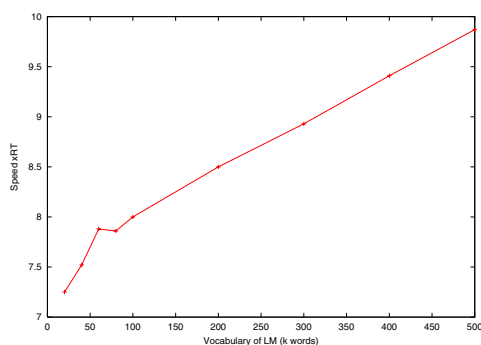


Figure 4: LM vocabulary size vs. decoding speed

The graph in Figure 5 shows the (more or less) linear relationship between the OOV rate and the WER, indicating the need to address the high OOV rate of German.

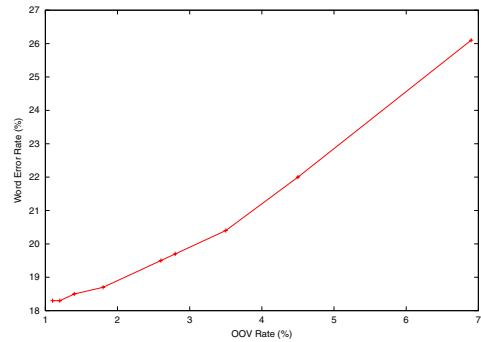


Figure 5: OOV rate vs. WER

5. Conclusion

This paper presents improvements to our German BN system, obtained by adding more recent training data for language modelling, by vocabulary increase and manual pronunciation variants for reductions on frequent words. Adding the more recent 1.5 M words from the *Tagesschau* web archives to the older 315 M BN training corpus highlights the importance of time and topic related training material resulting in a 2-3% absolute reduction in word error rate. The gain achieved by extending the vocabulary size from 65k to 300k items represent another 1-1.5% absolute reduction. Manually added pronunciation variants account for another 1% absolute word error reduction as compared to completely automatic pronunciation lexica.

6. References

- [1] Adda-Decker, M., Adda, G. and Lamel, L., "Investigating Text Normalization and Pronunciation Variants for German Broadcast Transcription" Proc. ICSLP, pp. 266-269, Beijing, Oct. 2000.
- [2] Adda-Decker, M. and Lamel, L., "Modeling Reduced Pronunciations in German" Proc. Workshop on Phonetics and Phonology in ASR, Saarbrücken, March 2000.
- [3] Federico, M. and Bertoldi, N., "Broadcast News LM Adaptation using Contemporary Texts" Proc. Eurospeech, Aalborg, Denmark, 2001.
- [4] Garnier-Rizet, M., Prouts, B., et al., "Progress Report on Speech Recognition" Project Olive, (LE4-8364), deliverable report D3.41, 1999.
- [5] Gauvain, J-L., "The LIMSI 1999 Hub-4E Transcription System" Proc. DARPA Speech Transcription Workshop 2000.
- [6] Gauvain, J-L., Lamel, L. and Adda, G., "The LIMSI Broadcast News Transcription System" Speech Communication 37(1-2):89-108, 2002.
- [7] Geutner, P., "Using Morphology Towards Better Large-Vocabulary Speech Recognition Systems" Proc. ICASSP 1995.
- [8] Geutner, P., Finke, M. and Scheytt, P., "Adaptive Vocabularies for Transcribing Multilingual Broadcast News", Proc. IEEE-ICASSP, Seattle, May 1998.
- [9] Lungen, H., Pampel, M., Drexel, G., Gibbon, D., Althoff, F. and Schillo C., "Morphology and Speech Technology" Proc. 2nd ACL-SIGPHON Workshop, pp. 25-30., University of California, Santa Cruz, 1996.