published as: HJ. Yang, C. Oehlke, Ch. Meinel, German Speech Recognition: A Solution for the Analysis and Processing of Lecture Recording. In Proc. of 10th IEEE/ACIS International Conference on Computer and Information Science (ICIS 2011) , Sanya, Heinan Island, China, May 2011

German Speech Recognition: A Solution for the Analysis and Processing of Lecture Recordings

Haojin Yang, Christoph Oehlke, Christoph Meinel
Hasso Plattner Institut (HPI), University of Potsdam
P.O. Box 900460, D-14440 Potsdam
e-mail: {Haojin.Yang, Meinel}@hpi.uni-potsdam.de; Christoph.Oehlke@student.hpi.uni-potsdam.de

***Abstract*—Since recording technology has become more robust and easier to use, more and more universities are taking the opportunity to record their lectures and put them on the Web in order to make them accessable by students. The automatic speech recognition (ASR) techniques provide a valueable source for indexing and retrieval of lecture video materials. In this paper, we evaluate the state-of-the-art speech recognition software to find a solution for the automatic transcription of German lecture videos. Our experimental results show that the word error rates (WERs) was reduced by 12.8% when the speech training corpus of a lecturer is increased by 1.6 hours.**

***Keywords*-automatic speech recognition; e-learning; multimedia retrieval; recorded lecture videos.**

I. Introduction

In the past decade, more and more universities recorded their lectures and presentations by using state-of-the-art recording systems such as ”tele-TASK” [3] and made them available over the internet. Recording this kind of content quickly leads to very large amounts of multi-media data. Thus, the challenge of finding lecture videos on the internet or within video portals has become a very important and challenging task. Content-based retrieval within lecture video data requires textual metadata that has to be provided manually by the users or has to be extracted by automated analysis. Many research projects have experimented with these data and determined two main questions: How can we access the content of multi-media lecture videos easily [1][4][6][7][8] and how can we find the appropriate semantical information within them [4][5]?

Speech is the most natural way of communication and also the main carrier of information in nearly all lectures. Therefore it is of distinct advantage that the speech information can be used for automatic indexing of lecture videos. The studies described in [1][4] are based on ”out-of-the-box” commercial speech recognition software. Their proposed indexing and searching functionalities are closely related to recognition rates. Concerning such commercial software, it is not easy to adapt it for a special working domain and custom extensions are rarely possible. [2] and [9] focus on English speech recognition for TED (technology entertainment and design) lecture videos and webcasts. Their approaches do not use the vocabulary extension method. Therefore their training dictionary can not be extended or optimized periodically. [8] proposed a solution for creating an English speech corpus basing on lecture audio data. But, they have not dealt with a complete speech recognition system. [7] introduced a spoken document retrieval system for Korean lecture search. Their automatically generated search index table is based on lecture transcriptions. However, they do not consider the recognition of topic-related technical foreign words which are important for keyword-search. Overall, most of these lecture recognition systems have a low recognition accuracy, the WERs of audio lectures are approximately 40%-80% [1][2][7][8][9]. The poor recognition results limit the usability of their approaches.

If we regard German speech, we see that it is much harder to be recognized than English speech. This is because of the different language characteristics. Compared to the English language, German has a much higher lexical variety. The out-of-vocabulary (OOV) rate of a 20k German lexicon is about 7.5% higher than it is for an appropriate English lexicon [10]. German is a compound language and has a highly inflective nature. Because of these peculiarities, a German recognition vocabulary is several times larger than a corresponding English one and it is hard to resolve word forms that sound similar. In addition, German lecture videos in specific domain e.g. computer science are more difficult to recognize than common contents like TV news. This is because there are many topic-related technical terms which are out of the standard vocabulary. A lot of them are foreign words, sometimes even pronounced with English pronunciation rules, e.g. the words ”server” and ”World Wide Web” are often mentioned in German lectures about computer science.

We have compared the current state-of-the-art speech recognition software and developed a solution for the automatic transcription of German lecture videos. Our solution enables a continued improvement of recognition rate by creating and refining new training data. The topic-related technical terms have been added to the training vocabulary. In addition, we developed an automatic vocabulary extension procedure for adding new speech training resources. In the experiments, we have determined that the training time of our speech corpus influences the recognition accuracy

Table I
COMPARISON BETWEEN STATE-OF-THE-ART SPEECH RECOGNITION SOFTWARE [12]

Criteria	IBM ViaVoice	Dragon Naturally Speaking	Sphinx 4	Julius	HTK
Recognition rate (60% of the total score)	5	8	6.5	8	6
Plattform independence (15%)	6	2	8	9	8
Cost (5%)	7	4	10	10	7
Modularity (15%)	5	1	10	10	10
Actuality (5%)	0	8	7	7	6
Total score	5	5.85	7.45	8.5	6.35

significantly.

The paper is organized as follows. In Section II we discuss several state-of-the-art speech recognition software in more detail and evaluate them for our task. In Section III the development of our German lecture speech recognition system is presented. Experimental results are provided in Section IV. Section V concludes the paper with an outlook on future work.

II. SPEECH RECOGNITION SOFTWARE

After a detailed study of state-of-the-art speech recognition software, we selected IBM ViaVoice, Dragon Naturally Speaking, CMU (Carnegie Mellon University) Sphinx, Julius and HTK (Hidden Markov Model Toolkit) for our evaluation. In [12] a comparison between several pieces of speech recognition software is given. We have referenced their evaluation in our research results and illustrated them in Table I, where higher score means better performance.

The evaluation of recognition rates was taken from [12][13]. IBM ViaVoice is basing on Windows and Mac OS. The active vocabulary includes 30000 words plus an add-on capacity of further 34000 words. For control and command applications the average WER of ViaVoice is lower than 7%. The software costs about 45$, but its development has been terminated for several years. Dragon Naturally Speaking (DNS) is another competing commercial software, where costs go from 120$ to over 1000$. Its active vocabulary includes 30000 words. The average WER is about 5% for command, control and small dictation applications. DNS is based on Windows only and programming interfaces are not supported at all. HTK (Hidden Markov Model Toolkit) is a portable toolkit for building and manipulating hidden Markov models. It is used mostly for speech recognition and copyrighted by Microsoft. Altering the software for the licensee's internal use is allowed. CMU Sphinx is an open-source toolkit for speech recognition and is an ongoing project by Carnegie Mellon University. As the same as HTK the Sphinx toolkit can be used for training both, acoustic models and language models, which will be discussed in detail in section III. Sphinx 4 has a generalized, pluggable language model architecture and its average WER for a 5000 words vocabulary is lower than 7%. Julius is an open-source large vocabulary continuous speech recognition engine that is developed primarily for Japanese language. To run Julius recognizer, a language model and an acoustic model have to be provided. Julius accepts acoustic models in HTK ASCII format. Moreover, both Julius and Sphinx accept n-gram language models in ARPA format.

In addition to the results in Table I, there is another important reason that affected our evaluation: Because of the particular characteristics of lecture videos (mentioned in section I), we need to update the recognition vocabulary and retrain the system from time to time for including newly added words. This process is much more limited by commerical software than by open-source software. Moreover, lecture recognition is a long-term task: There might be more and more special requirements that need to be met, so it is important that we have the absolute control over the refining process. Therefore, we eventually decided not to consider the commercial software for our research.

After the comparison between HTK and Sphinx according to [14][15][16] (on Chinese Mandarin, English and Hindish), the following conclusions have been mentioned:

- Sphinx 3 and HDecode provide comparable levels of WER and xRT (the real-time factor) [16].
- Sphinx 3 and HTK are comparable in recognizing isolated words, however Sphinx 3 performs much better than HTK in recognizing continuous sentences [14].
- The acoustic models for Sphinx were much better trained than the ones for HTK [15].

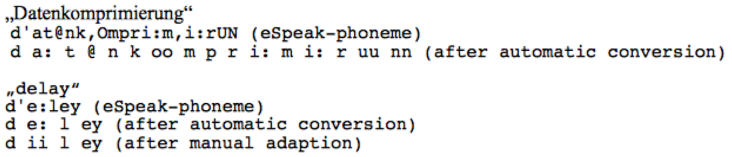
Concerning Julius, we found out that it has good recognition rates and works very fast. However, it lacks support for German language, so we would have to train the acoustic model using HTK in a way that it suits Julius. Therefore we finally chose Sphinx for our recognition task and started to create a speech corpus by using our lecture videos in order to increase the recognition rate.

III. SYSTEM DEVELOPMENT

Both, the acoustic model and the language model, are the most important parts of a modern statistic-based speech recognition system. An acoustic model is created by taking speech recordings and their appropriate text transcriptions,

Table II
CONVERSION OF ESPEAK-PHONEME-FORMAT INTO OUR ASCII-PHONEME-FORMAT

eSpeak phoneme	Our ASCII-phoneme	Description
a	a:	long ’a’
@	@	modification of the vowel ’e’
O	oo	short ’o’
i:	i:	long ’i’
U	uu	short ’u’
N	nn	modification of the consonant ’n’



```
„Datenkomprimierung“
 d'at@nk,Ompri:m,i:rUN (eSpeak-phoneme)
 d a: t @ n k oo m p r i: m i: r uu nn (after automatic conversion)

„delay“
d'e:ley (eSpeak-phoneme)
d e: l ey (after automatic conversion)
d ii l ey (after manual adaption)
```

Figure 1. Phoneme-Representation of German and English Words

using software to create statistical representations of the sounds that make up each word. A language model tries to capture the properties of a language and to predict the next occurrence in a speech sequence. Our speech recognition system makes use of phone-based continuous HMMs (Hidden Markov Models) for acoustic modeling and n-gram statistics based on German plaintext that has been extracted from different databases for language modeling. The language vocabulary is created by refining and extending an open-source German vocabulary. This is discussed in more detail in the next subsection.

A. *Vocabulary*

Vocabulary is a very important part of an ASR system. We have investigated three open-source German dictionaries: HADI-BOMP lexicon[1] that has been developed by speech and communication workgroup of University Bonn, Voxforge lexicon[2] and Ralf’s German dictionary[3]. Unfortunately none of them is immediately useable for our task. This is the case because in our preliminary training we only have a small speech corpus (about 25 hours) and an appropriate vocabulary size should be about 5000-15000 words. Too many foreign entries will lead to higher WER, because of confusion with similar sounding words. Finally we decided to create a new vocabulary basing on Ralf’s German dictionary for our working domain. The creation is based on the following steps:

- 45 phonemes have been defined in ASCII-format.
- Trimming the Ralf’s German dictionary by dropping all words neither included in our transcription texts nor in other topic-related documents. Hereby the dictionary size was reduced from over 380000 words down to 13300 words.
- Adding 413 technical terms related to ’computer science’ published on German Wiktionary[4] into the dictionary.
- Generating a phoneme representation for each word by using eSpeak[5] and automatically adapting these eSpeak phonemes into our ASCII-phoneme-format.

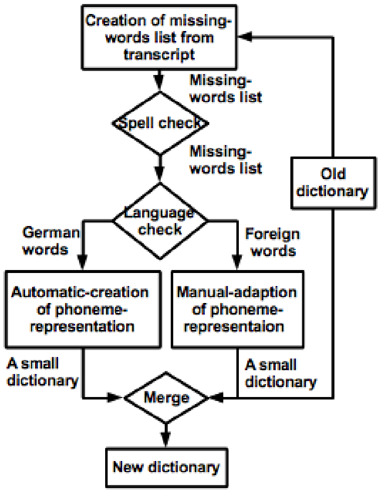


Figure 2. Dictionary Extending Procedure

The phoneme creation steps are illustrated by two sample words from our dictionary in Figure 1. The phonetic output of eSpeak is in a special format, containing additional information, like how to stress each syllable. Table II shows the

[1] http://www.sk.uni-bonn.de/forschung/phonetik/sprachsynthese/bomp
[2] www.voxforge.org
[3] http://spirit.blau.in/simon/2010/05/13/ralfs-german-dictionary-0-1-9-3/
[4] http://de.wiktionary.org/wiki/Verzeichnis:Informatik
[5] http://espeak.sourceforge.net/

Table III
Recognition results are shown as %WER using bigram and trigram LMs.

LM	Corpora	Corpora Size	WER Bigram	WER Trigram
LM1	Leipzig-Wortschatz + transcripts	3.05M words	81.2%	81%
LM2	DBPedia + transcripts	2.6GB texts	81.9%	82.3%
LM3	DBPedia + Leipzig-Wordschatz + German daily news + transcripts	3.2GB texts	81.2%	82%
LM4	German daily news + transcripts	1.05M words	76%	75.5%
LM5	transcripts	50K words	81.9%	81.9%

* AM: 3000 senones, 16 tied-state.

conversion of consonants and vowels of the German word ”Datenkomprimierung” from eSpeak format to our ASCII-phoneme-format. The phoneme-representations of English technical terms like ”delay” in Figure 1 were manually adapted. In addition, we have trained technical terms more carefully, because a high recognition rate of technical terms can improve the accuracy of speech-transcript-based keyword search for lecture videos [1]. Before adding our newly created speech corpus to our training set an automatic vocabulary extension algorithm is used. Figure 2 shows its workflow. First, the new transcript file is automatically parsed for determining all words which are not included in our current dictionary. Subsequently the spell check, word correction and the language check processes are performed. This is essential, since we want to avoid spelling mistakes and need to separate German words and foreign words. All foreign words have to be processed manually, whereas the conversion of German words to the correct phonemes is handled automatically. Finally, all collected words plus their phonemes are merged into a new dictionary.

B. Acoustic Model

The Acoustic Model (AM) is trained using speech utterances and their transcriptions. A good speech corpus is the basis for developing speech recognition systems and it is usually characterized by the following features: recording hours, speech type, number of speakers, background environment, etc.

The AM training is usually performed by dictating (e.g. reading and recording a predefined text), but the actual input that has to be recognized is an audio track containing free speech, which differs significantly from the dictation text. Experimental results from [1] show that the quality and the format of the speech recording strongly affect the recognition accuracy. [1] also concluded that ”the training is useful but only if done under the same conditions.” This implies that better training results can be achived by using audio input which has the same signal frequency and the same background noise as the recognition audio. [8] pointed out that the alignments of the 16KHz broadband speech provide optimal quality for the training data. According to these conclusions we decided to create our speech corpus by using our lecture videos in an appropriate format (16KHz-16Bit). In order to ensure that the trainer can detect most of our lecture audio data, we have segmented it to many pieces of speech utterances that are mostly between 1 and 3 seconds long, then selected utterances that have good audio quality, manually transcripted them and finally added them to the training corpus. This work is very time-consuming. It is therefore very difficult to generate large amounts of speech data in a short time. In order to carry out the preliminary experiment, we decided to add about 23 hours of the Voxforge open-source German speech corpus[6] to our training set. The AM is trained using CMU Sphinx Acoustic Model trainer.

C. Language Model

A Language Model (LM) is trained using a large and structured set of texts. After a detailed investigation we have collected text corpora from the following sources:

- Extracted plain texts from German DBPedia[7]. The extracting process has been performed according to [17]. About 2.6 GB texts were extracted.
- Text corpus from Leipzig-Wortschatz[8], 3M words.
- German daily news 1996-2000 from radio, about 1M words.
- Audio transcripts, 50K words.

Because there are still many foreign sentences and special characters in extracted DBPedia corpus and the Leipzig-Wortschatz corpus, we have performed the following refinements:

- text preprocessing:
  - Normalized numbers and characters.
  - Embedded all single sentences in a context cue ”<s>” and ”</s>”.
  - Replaced all commas and semicolons with context cue ”<sil>”.
  - Replaced all the other punctuations with whitespaces.

[6] http://www.voxforge.org

[7] http://wiki.dbpedia.org/Downloads351

[8] http://corpora.informatik.uni-leipzig.de/download.html

- Foreign words and filtering of special characters: All the words in the DBPedia text corpus are filtered with the help of a predefined large German vocabulary (about 420000 words, including all technical terms).

We have trained several LMs using the CMU statistical language modeling toolkit (SLM toolkit)[9]. Their experimental evaluation is discussed in the next section.

IV. EXPERIMENTAL RESULTS

The preliminary experiments were carried out on 50 undetected speech sentences of our lecturer that were randomly selected from his computer science lecture. Our experiments follow two main goals:

- evaluation of different language models
- find out how training time of our speech corpus influences recognition accuracy

We use the common metric WER% for the evaluations. It can be calculated as follows:

$$WER = \frac{S + D + I}{N}$$

where S is the number of subsitutions, D is the number of the deletions, I is the number of insertions, N is the number of words in the groundtruth.

In experiment I, the AM is trained using a 24.5 hours speech corpus (23 hours voxforge speech corpus plus 1.5 hours of our corpus) and a 59K words vocabulary. LMs are trained using different combinations of text corpora. Table III shows the results of this experiment. The LM that has been trained using the German daily news corpus plus transcripts has the best recognition accuracy. Using the trigram language model tends to improve the recognition performance. The results show that the recognition accuracy is not directly related to the size of text corpus.

In the second experiment, we have used the best LM from experiment I. A pruned 1.3K words vocabulary is used. AMs are trained with different sizes of speech corpora. The test results are illustrated in Table IV. We have found out that the recognition rate is very closely related to the amount of training time performed with the lecturer. The WER is reduced by 12.8% when the speech training corpus of a lecturer is increased by 1.6 hours.

As already mentioned in section 3.2, the creation of our lecture speech corpus is very time-consuming work. We have only built a small training corpus so far. The AM that is trained with this corpus, therefore still does not meet the requirements of a real application. However, our experiments show that through a continuous extension of the speech corpus in our working domain, a considerable increase of the recognition acurracy is predictable.

V. CONCLUSION

Since the increasing amounts of the lecture videos and recorded presentations were made accessable, the indexing of these multi-media data has become a great challenge. The speech information is one of the best resources for semantic indexing. However, the speech recognition is still an active research area and almost none of the existing lecture speech recognition systems have achieved a good recognition rate. In this paper, we have evaluated the state-of-the-art speech recognition software and proposed a solution for a recognition system for German lecture videos.

As further work, we plan to continue generating and collecting speech data from our lectures and retrain the system periodically. We also consider to make the speech resource available to the research community. In order to achieve a better recognition accuracy, a word decomposition algorithm can be utilized in combination with our training, so the German vocabulary size and the OOV rates can also be reduced. The actual use of the automatically transcripted audio for a lecture video portal, such as transcript-based semantic search, automatically created video subtitles, a recommendation system based on speech information, etc. will be developed in the future.

REFERENCES

[1] W. Hürst, T. Kreuzer, and M. Wiesenhütter, ”A qualitative study towards using large vocabulary automatic speech recognition to index recorded presentations for search and access over the web”. In *Proc. of IADIS WWW / Internet (ICWI)*, pp.135-143, 2002.

[2] E.Leeuwis, M.Federico, and M. Cettolo, ”Language modeling and transcription of the TED Corpus lectures”. In *Proc. of the IEEE ICASSP*, 2003.

[3] V. Schillings, C. Meinel, ”tele-TASK - Teleteaching Anywhere Solution Kit”. In *Proc. of ACM SIGUCCS* 2002, Providence (Rhode Island), USA, pp.130-133, 2002.

[4] S. Repp, C. Meinel, ”Semantic indexing for recorded educational lecture videos”. In *Proc. of International Conference on Pervasive Computing and Communications Workshops (PERCOMW)*, pp.240, 2006.

[5] H. Sack, J. Waitelonis, ”Automated annotations of synchronized multimedia presentations”. In *Proc. of Workshop on Mastering the Gap: From Information Extraction to Semantic Representation, CEUR Workshop*, 2006.

[6] R. Anderson, C. Hoyer, C. Prince, J. Su, F. Videon, S. Wolfman, ”Speech, ink, and slides: the interaction of content channels”. In *Proc. of the Twelfth ACM International Conference on Multimedia*, 2004.

[7] D. Lee, G.G.Lee, ”A Korean Spoken Document Retrieval System for Lecture Search”. In *Proc. of the SSCS speech search workshop at SIGIR*, 2008.

[9] http://www.speech.cs.cmu.edu/SLM/toolkit.html

Table IV
RECOGNITION RESULTS ARE SHOWN AS %WER USING TRAINED AMS.

AM	Corpora	Corpora Size	WER
AM1	voxforge speech corpus	23.3 hours	82.9%
AM2	voxforge + 30 minutes hpi corpus	23.8 houts	78.5%
AM3	voxforge + 1 hour hpi corpus	24.3 hours	77.1%
AM4	voxforge + 1.6 hours hpi corpus	24.9 hours	70.1%

* LM: trained with German daily news corpus and transcripts (1.05M words)

[8] J. Glass, T. J. Hazen, L. Hetherington, C. Wang, ”Analysis and Processing of Lecture Audio Data: Preliminary Investigations”. In *HLT-NAACL Workshop on Interdisciplinary Approaches to Speech Indexing and Retrieval*, 2004.

[9] C. Munteanu, G. Penn, R. Baecker, Y. C. Zhang, ”Automatic Speech Recognition for Webcasts: How Good is Good Enough and What to Do When it Isnt”. In *Proc. of the 8th international conference on Multimodal interfaces*, 2006.

[10] M.Adda-Decker, G Adda, L. Lamel, J.L. Gauvain, ”Developments In Large Vocabulary, Continuous Speech Recognition of Germa”. In *Proc. IEEE ICASSP-96*, 1996.

[11] R. Hecht, J Riedler, G. Backfried, ”Fitting German into N-Gram Language Models”. In *Proc. of the 5th International Conference on Text, Speech and Dialogue*, pp.341–346, 2002.

[12] S. Franz, ”Technical Concept of Simon: A Speech Recognition Software for Physically Disabled People”. http://www.simon-listens.org/ (last access: 16/12/2010).

[13] C. Horisberger ”Language Models in Speech Recognition”. http://informatik.unibas.ch/lehre/ws05/cs505 /abschlusspraesentationen.html (last access: 16/10/2010).

[14] G. Ma, We. Zhou, J. Zheng, X. You, W Ye, ”A Comparison between HTK and SPHINX on Chinese Mandarin”. *jcai, 2009 International Joint Conference on Artificial Intelligence*, pp.394–397, 2009.

[15] Samudravijaya, K. / Barot, Maria, ”A Comparison of Public-Domain Software Tools for Speech Recognition”. *In WSLP-2003*, pp.125–131, 2003.

[16] Keith Vertanen, ”BASELINE WSJ ACOUSTIC MODELS FOR HTK AND SPHINX: TRAINING RECIPES AND RECOGNITION EXPERIMENTS”. *Technical Report, Cavendish Laboratory*, 2006.

[17] Internet: http://blog.afterthedeadline.com/2009/12/04/ (last access: 16/12/2010).