Methods and Tools for Speech Data Acquisition exploiting a Database of German Parliamentary Speeches and Transcripts from the Internet

Konstantin Biatov, Joachim Köhler

Fraunhofer Institute for Media Communication
Schloß Birlinghoven
53754 Sankt Augustin
Germany
biatov@imk.fraunhofer.de, Joachim.Koehler@imk.fraunhofer.de

Abstract

This paper describes methods that exploit stenographic transcripts of the German parliament to improve the acoustic models of a speech recognition system for this domain. The stenographic transcripts and the speech data are available on the Internet. Using data from the Internet makes it possible to avoid the costly process of the collection and annotation of a huge amount of data. The automatic data acquisition technique works using the stenographic transcripts and acoustic data from the German parliamentary speeches plus general acoustic models, trained on different data. The idea of this technique is to generate special finite state automata from the stenographic transcripts. These finite state automata simulate potential possible correspondences between the stenographic transcript and the spoken audio content, i.e. accurate transcript. The first step is the recognition of the speech data using finite state automaton as a language model. The next step is to find, to extract and to verify the match between sections of recognized words and actually spoken audio content. After this, the automatically extracted and verified data can be used for acoustic model training. Experiments show that for a given recognition task from the German Parliament domain the absolute decrease of the word error rate is 20%.

1. Introduction

Currently large vocabulary speech recognition technology is still domain dependent and it is necessary to collect data for the training of the acoustic and language models. Data acquisition is normally very time consuming and costly. The recording and the generation of perfect transcriptions takes a lot of human effort. Fortunately, now a lot of audio sources and speech database are available for different applications. For example, there are unlimited broadcasting radio and television, medical transcriptions generated in the normal course of a medical operation and also parliamentary debates available in the Internet (Lamel et al., 2002; Pakhomov et al., 2001; Biatov et al., 2001). Some of the audio resources have available transcriptions, which exactly represent the semantic content of the audio documents, but are not perfect as accurate audio source transcriptions. Therefore these transcriptions can not be used directly for the acoustic model training.

There are some techniques for accurate transcription acquisition for acoustic model training. All of them in different ways use a speech recognizer to transcribe audio data and then filter the results of recognition and generate the transcriptions for an audio segments. One main difference between these approaches is that some of them don't use transcribed data or use only small amount of transcribed data for initial training of the acoustic model.

There are some experiments of using untranscribed data for the acoustic model training (Kemp et al., 1998; Wessel et al., 2001). The performance of such approaches could be measured by the amount of untranscribed data that is needed to get recognition performance (word error rate) comparable to that was obtained with transcribed data. Kemp reports that starting from initial training using only 30 minutes of transcribed data and then using 45 hours of untranscribed speech he got performance comparable to the results which were obtained after training using 15.5 hours transcribed data (Kemp, 1999).

There are also published series of the experiments where transcribed data are available (Gauvian et al., 2001; Lamel et al., 2002). The idea of this approach is to use a speech recognizer at the first step for automatic generation of audio data transcription. Then automatically transcribed data are compared with the available stenographic transcription using dynamic programming alignment. The results of alignment are filtered to find possible accurate matchs between audio data and automatically transcribed data. For filtering some phone duration criteria could be used. For example a phone duration longer than 500 ms is likely to indicate to the error (Lamel et al., 2002). After filtering, the stenographic transcriptions are used to train the acoustic models.

The method described in this paper is similar to the methodology used when transcribed data are available (Lamel et al., 2002). One important difference is the use of special finite state automata that model the match between the stenographic transcription and the accurate transcript. These finite state automata are used as a language model for the speech recognizer. Another difference is the method for filtering potentially incorrect words from stenographic transcription. Before presenting the method of automatic extraction of relevant training sequences, the speech database exploited in these experiments will be described.

2. The Audio data corpus of the German Parliamentary Speeches

The speeches of the German Parliament are captured and broadcasted to other TV and radio formats by the German Parliament TV. The TV station Phoenix broadcasts the speeches over a TV satellite or cable channel including the video and audio stream. In parallel the speeches are also distributed over the Internet using the RealNetwork platform. It is possible to access the live stream and also to retrieve older debates in the archive of the Bundestag. The existing RealNetwork files in the archive can be played and stored automatically as PCM files with a special recording software programmed for the Linux operating system. The amount of available speech data is very large. There are an average of 60 debates each year in the German Parliament. Each debate contains about 6 hours of speech from several representatives. The average duration of a speech of one representative is 15 minutes. The entire duration for each year is 360 hours. Because the RealNetworks recordings are not complete for the period between 1999 and 2001 we have collected about 250 hours of speech. The database of the German Parliament contains the speeches of more than 600 different speakers. The size is large enough to train the acoustic models. The focus of our project is to use this database for spoken document retrieval. The goal is to search for special speeches of representatives about predefined topics. Although the raw speech data is already there, we need the transliteration of the speech material for our investigations. Because the post annotation of this huge database is not feasible, the automatic transcription techniques are desirable. Fortunately, for all parliamentary speeches stenographic transcripts are available electronically over the Internet.

Normally available transcriptions exactly represent the content of parliamentary debates, but are not perfect for audio data transcription. We have compared the stenographic transcriptions of 6 speeches with the corresponding accurate transcripts as a reference using NIST tools. As shown in Table 1 stenographic transcriptions have sufficient difference in comparison with accurate transcription for one speaker. The largest component in the error structure is deletion of words. This means that the stenographer skips words more often than substitutes or inserts them.

Speech number	Deletions	Insertions	Substitution	Word accuracy
1	16.7%	4.6%	5.6%	73.1%
2	6.6%	2.4%	2.2%	88.8%
3	15.6%	7.7%	10.8%	65.9%
4	7.8%	2.2%	3.6%	86.4%
5	5.1%	2.8%	3.5%	88.6%
6	6.3%	1.8%	2.7%	89.2%

Table 1: Comparison of the stenographic transcription with the accurate transcription.

On the other side the evaluation of out of vocabulary words shown in Table 2 demonstrates that in most cases near 5% word types from accurate transcriptions are not present in the stenographic transcripts.

Speech chunks	1	2	3	4	5	6
Number of word types	1195	453	924	1048	1250	1061
OOV	6%	4%	19%	5%	4%	4%

Table 2: Information about Out Of Vocabulary words.

In contrast to the broadcast news which are produced by professional speakers, the parliamentary speeches in most cases are produced by non-professional speakers and have repetitions and corrections of words.

Sometimes the speakers are emotional and the speeches include spontaneous speech. Information about repetitions, corrections, presence of the emotional speech fragments is not labeled in the stenographic transcriptions.

The parliamentary speeches very often include non-speech segments such as applause and sometimes include laughter and breath sounds. Usually accurate information about such sound events is also not present in the stenographic transcripts.

3. Automatic extraction of the relevant training sequences using the stenographic transcripts

In this paper we describe the method of speech data acquisition for acoustic model training. As was mentioned in the introduction, speech data could include non-speech sounds. At the beginning of the acquisition process we detect non-speech fragments using linear discriminate classification and extract homogeneous speech fragments between non-speech sounds. Normally the duration of one speech is 15-30 minutes. Sometime applause or laughter as natural delimiters are absent in speech. In this case we use long size silence detection to chunk speech into smaller fragments with average duration 1.5 - 2 minutes.

Next we recognize each speech fragment using speaker independent speech recognition. For recognition we use the open source ISIP toolkit (Picone et al., 1998) and speaker independent acoustic models trained on another German speech database. These monophone acoustic models are trained on 33K sentences from the Phondat and Siemens100 database distributed by ELRA. We don't know the correspondence between small homogeneous audio fragments and stenographic transcripts. We exploit both the results of the speech recognition, which is a text, and the stenographic transcripts to find the approximate match between audio fragments and the stenographic transcripts. Based on dynamic programming, an alignment procedure approximates the correspondence between audio and transcripts fragments.

Availability of the correspondence between speech fragments and the transcripts gives the possibility of generating a finite state automaton for each audio fragment. This finite state automaton models potential possible matches between the stenographic transcripts and accurate transcription generated by a professional transcription agency for test purposes. Table 1 shows the comparison between 5 stenographic transcriptions and accurate transcriptions. Some operations could be applied to make stenographic transcriptions closer to the accurate transcriptions i.e. actually spoken audio content.

These operations are skipping of some words, which are inserted or substituted in the stenographic transcripts, but are absent in the accurate transcription. Since we don't know which words were inserted or substituted, we permit each word to be skipped in the stenographic transcription. One exception is that we skip articles together with nouns or adjectives as one unit.

Another operation is filling space in the audio fragments that correspond to spoken words deleted in the stenographic transcripts. For this goal we use an acoustic filler model. An acoustic filler is also useful for filling speaker words repetitions and self-corrections.

It will be also useful to evaluate the probabilities of all mentioned events to accommodate stochastic finite state automata for the current task.

The finite state automata are based on the stenographic transcriptions and include the elements and transformations described above. In this way the automata generate a set of sentences which we compare with speech fragments using speech recognition decoder. A finite state automaton is used as a language model for the recognition of each speech fragment.

Our choice to use finite state automata for the language models is based on experiments on comparison of different language models for the current speech recognition task. Table 3 shows the result of recognition with this finite state automaton in comparison with the results obtained with a 2-gram language model. In the experiments the language model was generated from the stenographic transcripts and also from accurate transcription obtained from transcription office. For language models generation was used the CMU-Cambridge toolkit (Clarkson et al, 1997).

Language Model	2-gram model from stenographic transcription	2-gram model from accurate trascription	Finite state automata based on stenographic transcription
Word accuracy	39.7%	52.6%	55.5%

Table 3. Comparison of speech recognition word accuracy for different language models.

We recognize all acoustic fragments using a finite state automata as a language models. The results of recognition are sequences of words which potentially are the transcriptions of the speech fragments.

The next step is filtering of the results of the recognition. Only words sequences which have at least two words and which do not include a filler element are considered as potentially valid transcriptions. Then we extract these speech fragments that correspond to potentially valid transcriptions.

After the extraction we make a verification of each extracted speech fragment. Gorin showed experiments of comparison of the length of the phonetic transcription of audio data with the length of the phone sequence that is the result of the phoneme recognition of the same audio data (Gorin et al., 1999). For known utterances the phonetic transcription was generated from the orthographic transcription by replacing each word with its most likely dictionary pronunciation and deleting word-delimiters. The same utterances were recognized. The lengths of recognized phone sequences were compared with the appropriate lengths of the generated phonetic transcriptions. As shown, in most cases the transcribed and recognized utterance have approximately the same length. We use this technique for verification.

For the verification we do a task-independent phoneme recognition of each extracted speech fragment using the general 2-gram phoneme language model. The result of the recognition is a phoneme sequence. From the other side for each speech fragment we obtained the sequence of the words as a result of recognition using finite state automaton as a language model. Using these results for each speech fragment we generate a phonetic transcription from the appropriate word sequence deleting word-delimiters. Then we compare the length of recognized phone sequence obtained as a result of recognition using general 2-gram phoneme language model and the length of the phonetic transcription of the corresponding words sequence, obtained as a result of recognition using finite state automaton. If the difference between these two lengths is less than a threshold we consider the speech fragment to be valid and use the result of recognition as a phonetic transcription of the speech fragment.

Finally after extraction, filtering and verification we create a speech database containing the speech segments and the corresponding transcriptions. This automatically generated speech database is used to retrain the domain dependent acoustic models. All processing steps used to extract speech fragments and the corresponding word transcriptions are fully unsupervised.

The described methods are implemented as tools. These tools are written in Perl and C and use CMU-Cambridge toolkit, Missisipi State University toolkit and NIST tools. They use audio data and stenographic transcripts from the Internet as an input. The input looks like the list of raw file and the list of stenographic transcripts from the Internet. In the process of acquisition the tools execute the following groups of operations.

1. Stenographic transcripts normalization. There operations encode symbols with the umlaut, remove non-alphabetic symbols from the transcripts, convert all numbers to the text, generate normalized vocabulary, generate pronunciation of all words from the vocabulary.

2. Audio data normalization . There are operations are used for recognition and deletion of non-speech fragments from an audio data, segmentation an audio data into small 1.5-2 minutes speech chunks using natural delimiters (applause) or long silence periods.

3. Alignment of the small audio chunks with the normalized stenographic transcripts. The results of these operations are the list of small speech chunks and the list of the text fragments, which approximately correspond to the speech chunks.

4. Audio data transcription, text and speech fragments extraction. These operations exploit speech recognizer with finite state automaton as language model, filter the results of recognition and then extract perspective speech fragments and corresponding text fragments for further verification.

5. Verification speech and text fragments. The final result is the list of speech fragments and the list of text fragments which are similar to the accurate transcriptions and which are ready to be used in the training procedure.

4. Experiments

The first test for the evaluation of new acoustic models was to compare these models with the speaker independent monophone mixture acoustic models trained on 33K sentences from the Phondat/Siemens100 speech database. 5 speeches (2.5 hours) of one speaker were used for acquisition. The speaker dependent and domain dependent acoustic models were trained on 1K short sentences extracted from 5 speeches. In both cases, for the word recognition task, the same 2-gram language model was used. For testing 30 minutes of parliamentary speech was used. The accurate transcription of this speech was generated for the test purposes. The result of the evaluation shows 20% absolute improvement in the word accuracy.

The second test was to compare retrained acoustic models with the acoustic models prepared by using accurate transcription from the same amount of the data. The accurate transcriptions were obtained from the transcription office. Our experience shows that even after manual transcription by the professional transcription agency, the transcripts are still not absolutely free from errors. The forced alignment procedure was used for mapping accurate transcription to the speech data for further segmentation into smaller speech chunks. The resulting speech chunks and its accurate spoken content were used for retraining acoustic models.

In the training data extracted from the stenographic transcripts were 2053 word types. In the training data extracted from the accurate transcription were 2641 word types.

The acoustic models were tested using domain independent phoneme recognition. The result of recognition using the adapted models was 54.7% phoneme accuracy. The result when accurate transcription was used was 59.4% phoneme accuracy.

5. Conclusions and future work

The experiments show that the proposed method is effective for unsupervised acquisition of speech data from the Internet when stenographic transcripts are available. We have tested it on relatively small amount of data. Now the netto size of the downloaded corpus is 120 hours and testing could be done for all available data. Besides the improvement and optimization of the speech recognition technology using more sophisticated adaptation methods, the general goal is to make the preprocessed transcripts available for other research groups. Further, the database of the German Parliament also contains the video signal. The collection and preparation of a huge multimodal database is our research focus for the next years.

6. Acknowledgements

This work was funded by the German Federal Ministry for Research and Education.

7. References

Biatov K., Larson M. & Köhler J., 2001. Aufbau and Optimierung eines deutschsprachigen Spoken Document Retrieval Systems für Bundestagsreden. Elektronische Sprachsignalverarbeitung, 61-68.

Clarkson P. & Rosenfeld R., 1997. Statistical language modelling using CMU-Cambridge toolkit. EUROSPEECH 97 Proceedings.

Gauvian J.L. & Lamel L., 2001. Large-Vocabulary Continuous Speech Recognition: Advances and Applications. Proceedings of the IEEE, 88:1181-1200.

Gorin A., Petrovska-Delacretaz D., Riccardi G. & Wright J., 1999. Learning spoken language without transcription. ASRU 1999 Proceedings.

Kemp T. & Waibel A., 1998. Unsupervised training of a speech recognizer using TV broadcasts. ICSLP 98, 1998 Proceedings.

Kemp T., 1999. Unsupervised training of a speech recognizer: recent experiments. EUROSPEECH 99 Proceedings.

Lamel L., Gauvian J.L. & Adda G., 2002. Lightly supervised and unsupervised acoustic model training. Computer Speech and Language, 16:115-129.

NIST Spoken Language Technology Evaluation and Utility, http://www.nist.gov/speech/tools/

Pakhomov S., Schonwetter M. and Bachenko J., 2001. Generating training data for medical dictation. NAACL 2001 Proceedings.

Picone et al., 1998. A public domain decoder for large vocabulary conversational speech recognition, Missisippi State University, 1998.

Wessel F. and Ney H., 2001. Unsupervised training of acoustic models for large vocabulary continuous speech recognition. *ASRU 2001 Workshop Proceedings*.