Creation of an Annotated German Broadcast Speech Database for Spoken Document Retrieval

Stefan Eickeler, Martha Larson, Wolff Rüter, Joachim Köhler

Fraunhofer Institute for Media Communication IMK
Schloss Birlinghoven
53754 Sankt Augustin, Germany
{stefan.eickeler,martha.larson,wolff.rueter,joachim.koehler}@imk.fraunhofer.de

Abstract

In this paper we present a semi-automatic method for creating annotated data sets from German-language broadcast resources for which audio files as well as transcripts are available on the Internet. The transcripts are required to be reasonably accurate, but not perfect. Our approach is implemented by a integrated bundle of data processing tools, which support the human annotator in the creation of an annotated data set specialized for research in the area of spoken document classification and retrieval. Annotation decisions that would require prohibitively large amounts training data or system development time to make automatically are taken over by the human annotator. Annotation decisions which are easily automated and tedious for humans are shouldered by the computer. Using our method we can process and annotate the data approximately ten times faster that it was possible by hand. The data is downloaded and the transcripts are normalized by a series of filters as well as a semi-automatic digit to text conversion. Then, the system makes use of the Bayesian Information Criterion (BIC) to segment the audio data and Automatic Speech Recognition (ASR) to forced-alignment of the speech signal with written transcripts. We demonstrate the method with the concrete example of our *Deutsche Welle* database of programs from the *Kalenderblatt* radio series.

1. Introduction

The highest hurdle in the development of pattern recognition applications is often the creation of annotated data sets of a size adequate to test and train the system. Progress in the area of German-language broadcast news has been impeded by the difficulty of compiling an audio database with corresponding reference transcriptions of the quality and scope required for spoken document retrieval.

The Internet is a valuable source of audio data, and many available audio resources have accompanying transcripts. These transcripts tend to be radio broadcast or parliamentary transcripts, and are intended to represent the semantic content, rather than a literal transcription of the audio. In this paper we present a semi-automatic method that allows such data sets to be upgraded with a modest commitment of human hours, into audio databases with aligned literal transcriptions. Our approach consists of an integrated bundle of data processing tools, which support the human annotator in the creation of an annotated data set specialized for research in the area of spoken document classification and retrieval. We demonstrate these tools on the concrete example of our *Deutsche Welle Kalenderblatt* database. This database was created for use within a multimedia information portal project called Pi-AVIda (Personalized Interactive Audio, Voice and Video Information for Media Analysis and Multimedia Portals) funded by the German Federal Ministry of Education and Research. *Deutsche Welle* has expressed its interested support of our use of the resource.

We developed our semi-automatic approach to data annotation with an eye to optimizing our investment of human effort. At the onset we carefully considered the potentials and limitations of available automatic technologies. Annotation decisions that would require prohibitively large amounts training data or system development time to make automatically are taken over by the human annotator. Annotation decisions which are easily automated and tedious for humans are shouldered by the computer.

We also paid close attention to the balance between invested human hours and the final research value of the resource. Those parts of the database necessary for development and testing were segmented by hand and subjected to rigorous checking, whereas those parts of the database that were reserved for training, were processed only to the resolution necessary for model training.

The section 2 gives a brief overview of *Deutsche Welle Kalenderblatt* website and describes the download process by which we acquired the raw contents of our database. Section 3 concerns the processing of the transcripts including the text normalization filter and the semi-automatic digit-to-text conversion that we have developed. Section 4 describes the processing of the audio documents including their segmentation with the Bayesian Information Criterion (BIC). Section 5 explains the alignment of the transcripts to the audio, supported by Automatic Speech Recognition (ASR) and a graphic user interface, which allows the human annotator to easily check and correct the recognizer output. Section 6 introduces the manner in which the data set was annotated with topic categories. Conclusions and outlook are presented in Section 7.

2. An Internet resource: *Deutsche Welle Kalenderblatt*

The *Deutsche Welle Kalenderblatt* website (http://www.kalenderblatt.de) contains radio programs of the *Kalenderblatt* radio series starting from 1999. This Internet

source is appealing, since the amount audio data available is generous and the texts are nearly word-for-word transcriptions of the spoken audio. Each program is about 5 minutes long and contains approximately 650 running words.

The *Kalenderblatt* series broadcasts programs related to a broad range of topics of current, historical and cultural interest, and this variation makes it an interesting resource for spoken document classification and retrieval experiments. This material represents a truly non-trivial task for broadcast speech recognition, since the radio reports contain music and other sound effects, which are often layered over the speakers. The speech of the major reporters is interspersed with recorded interviews and comments from other speakers. The audio data is available in streaming (Real) format and is compressed with lossy compression techniques. Each radio program is accompanied by a transcript, which is intended to render the semantic, rather than the literal content of the radio broadcast. These transcripts are close enough to word-for-word transcripts, however, to be useful for training a spoken document classification system. With our semi-automatic methods, they can be upgraded to the perfect literal transcripts needed for training and testing.

The audio and accompanying text files for the *Deutsche Welle* database were downloaded from the website of *Kalenderblatt* using an automatic scripts. The text of each report was extracted by parsing the HTML pages. The appropriate audio stream in real format is digitally grabbed and stored in a PCM-based speech file. The audio stream is mono and has a rate of 31.1 Kbit/s. The sampling rate of the original stream is 22.05 kHz, which has to be resampled to 16 kHz.

In the first step of database preparation, the collected corpus is divided into a development set (10 %), a test set (10 %) and a training set (80 %). The text of all three sets is normalized and digits are converted to text as described in section 3. The audio of all three sets is segmented by the BIC algorithm as described in section 4. The test and the development set are further segmented by hand down to the sentence level. The audio and the transcriptions of all three sets are aligned using the ASR system in forced alignment mode as described in section 5 and corrected semi-automatically by human annotators. The test and the development sets are checked again by hand in order to insure the highest standard of accuracy. All three data sets were annotated with topic categories as described in section 6.

3. Processing the transcripts

After the transcripts are extracted from the original HTML, they are subjected to normalization and to digit-to-text conversion.

3.1. Normalization of the text transcripts

The first step in preparing a database for research is to map all elements (characters, words) occurring in the text onto the set of elements which have been defined for use in the system. Our needs on this front, however, are quite complex, since the very choice of the basic elements a spoken document classification or retrieval system should use must be optimized during system development and more than one set of elements might be needed to develop the same system.

Experiments might show, for instance, that the speech recognition module of the spoken document classification system functions better if all distinctions between capitals and small letters are leveled. This outcome is plausible because capitalization does not affect the pronunciation of the word. The classification or retrieval component, on the other hand, might function better if distinctions between capital and small letters are retained. Capitalization carries semantic weight in German since all nouns are written with capital letters.

Instead of normalizing our text in all the necessary different forms, we have created an intermediary XML format in which it is stored. A series of filter scripts take the text and normalize it into the particular form (all lower case, all syllables) that is needed at any given time.

At this time we are experimenting with even more radical normalizations of our data bank needed to develop and test spoken retrieval systems based on subword units. In particular we are experimenting with syllables and strings of phonemes, which we generate using the transcription module of the Boss II speech synthesis system of the Institute of Communication Research and Phonetics at the University of Bonn (http://www.ikp.uni-bonn.de/~kst/boss_ii.htm) (Stöber et al., 2000; Klabbers et al., 2001). The transcription module can be comfortably called from a script which will output a flexible text normalization according to the need of the moment.

3.2. Digit to text conversion

In order keep the vocabulary finite and the coverage as high as possible it is necessary to convert digits occurring in the texts into words units. This problem is especially acute in German because of the gender and case inflections, which means that a single written number might correspond to several different inflected forms, and thus to multiple pronunciations. The separate forms must be generated individually in order that they each receive a separate phonemization in the lexicon of the speech recognition system.

For some written expressions containing digits in German, it is possible to deterministically deduce from the immediate context, how they are inflected in the spoken language. For other forms, the information needed to deduce the correct form of the word is located an indeterminate number of words a way. This fact can be illustrated on an example:

im 5. fünften und 6. sechsten Jahrhundert

"in the 5+6 century"

das 5. fünfte und 6. sechste Jahrhundert

"the 5+6 century"

The decision of the pronunciation of "6." as "sechsten" or as "sechste" is based on the first word in the line: "im" and "das".

Technically speaking it would be possible to train a fi-

nite state transducer to do the digit-to-text task. The size of a database needed for training of the transducer would have to be large to insure the 100 % accuracy necessary for the ground truth. For this reason we have developed a tool which allows semi-automatic mapping of digits to words. In those cases in which the form is uniquely determined by the immediate context, the word is substituted in exactly. In other all other cases a human user is consulted to provide the correct identification of the form. The human user has to do about 548/8267=7 % of the cases manually. This decrease represents a significant reduction of time resources necessary for the creation of the database compared to a fully manual conversion.

The mapping of digits to words is further complicated due to the convention of German orthography that specifies that a single number, no matter how long, must be written as a single word. In orthographically correct German it takes 2003 different words to represent the years from 0 until 2002. We supplement the digit-to-word mapping by a heuristic decomposition of German word compounds. This decomposition allows us to represent all numbers using only a finite set of about 50 elements.

4. Segmentation of audio data using BIC

The Bayesian Information Criterion (BIC) (Tritschler and Gopinath, 1999) is used to divide the audio stream into segments. A BIC-based algorithm creates segment boundaries for each change of speaker and for each transition of speaker to music and vice versa.

The algorithm determines the BIC of a global Gaussian model for the audio stream. Then it moves segmentation hypotheses iteratively through each position in the audio. At each potential segment boundary it determines a composite BIC of the Gaussian model of the audio preceding and of the audio following the boundary. If the composite BIC is lower than the global BIC plus a penalty, a boundary is set.

$$\begin{aligned} \Delta BIC_i &= \frac{N}{2}\log|\Sigma_{\mathrm{w}}| - \frac{i}{2}\log|\Sigma_1| - \frac{N-i}{2}\log|\Sigma_1| \\ &\quad -\frac{1}{2}\lambda\left(d + \frac{d(d+1)}{2}\right)\log N \end{aligned} \tag{1}$$

Σ is the covariance matrix of the whole audio stream. Σ_1 the covariance matrix of the first segment and Σ_2 of the second. λ is a penalty factor which we choose to be 5. d is the dimensionality of the feature vector. We use a 12 dimensional Mel-warped cepstral vectors of the audio signal.

5. Alignment of transcription to audio

The result of the segmentation process are segments of the original audio which contain homogeneously either music or the speech of a single person. These segments are algined with the transcript text and are depicted in a visual environment for machine-supported correction or transcript errors.

5.1. Alignment with Automatic Speech Recognition (ASR)

Speech recognition technologies are used to assign the map segments of the transcripts to the corresponding segments of the audio. Our speech recognition functionality derives from the HMM-based ISIP public domain speech recognition toolkit (Ganapathiraju et al., 1999). In order to accomplish the automatic assignment we extended the functionality of the ISIP toolkit with some improvements in the handling of lattices. Because we are interested in preserving the generality of our semi-automatic approach to database annotation, we did not adapt the acoustic models we used in the speech recognition module to the domain of the *Kalenderblatt* database. Instead we used very basic monophone acoustic models trained on 33K sentences from the Phondat/Siemens 100 speech database. These models proved to be quite satisfactory in allowing us to align the transcripts with the audio segments.

In the transcription of the programs, tags sometimes indicate the presence of music inserted into the audio. These tags do not occur consistently and are not coordinated with the results of the BIC based segmentation. Furthermore the reporter narrating the report often drops a sentence, to respect the time limit of 5 min. For these reasons is not possible to align audio segments with transcript segments using only the order of their occurrence.

A Markov Model (lattice) is build from the transcription of the program and is used to restrict the search space in the recognition. Each state of the model represents a word and the following punctuation marks of the transcription. Transitions from one word to the following word in the transcription are allowed. Additionally transitions from the start of the lattice to each word and from each word to the end of the lattice ensure that the recognition process can assign audio segments belonging to text anywhere in the middle of the transcription. For the sentences dropped by the speaker, a skip from the end of each sentence to the sentence after the next is introduced. Figure 5.1. shows an example lattice for a short text: “Das ist der Durchbruch: Die atmosphrische Gaskraftmaschine findet reienden Absatz. Man kann damit bohren, sgen, hmmern usw.” (That is the breakthrough: The atmospherical gas-powered machine sells like hot cakes. It allows drilling, sawing, hammering, etc.)

To preserve any punctuation marks and special characters like umlauts of the transcription, which cannot be handled by the recognition software, a unique label is assigned to each word and the punctuation marks in the transcription. After the alignment the labels can be replaced by the original words with the original punctuation marks. The dictionary of the speech recognition system is the link between the phonemization of the original words and the labels. Phonemization of the words for the speech recognizer is effected by the transcription module of the BOSS II (Stöber et al., 2000; Klabbers et al., 2001) software, which also accomplished the phonemization and syllabification necessary for some of the normalizations described in subsection 3.1. The BOSS II system uses three different approaches to determine the phonemes for each word. An exception dictionary is first consulted to check if the word has been frequently phonemized or has proven particular tricky to phonemize correctly. Rules formulated on the basis of German phonotactics are applied if the word is not found in the exception dictionary. If the rules are not able to decom-

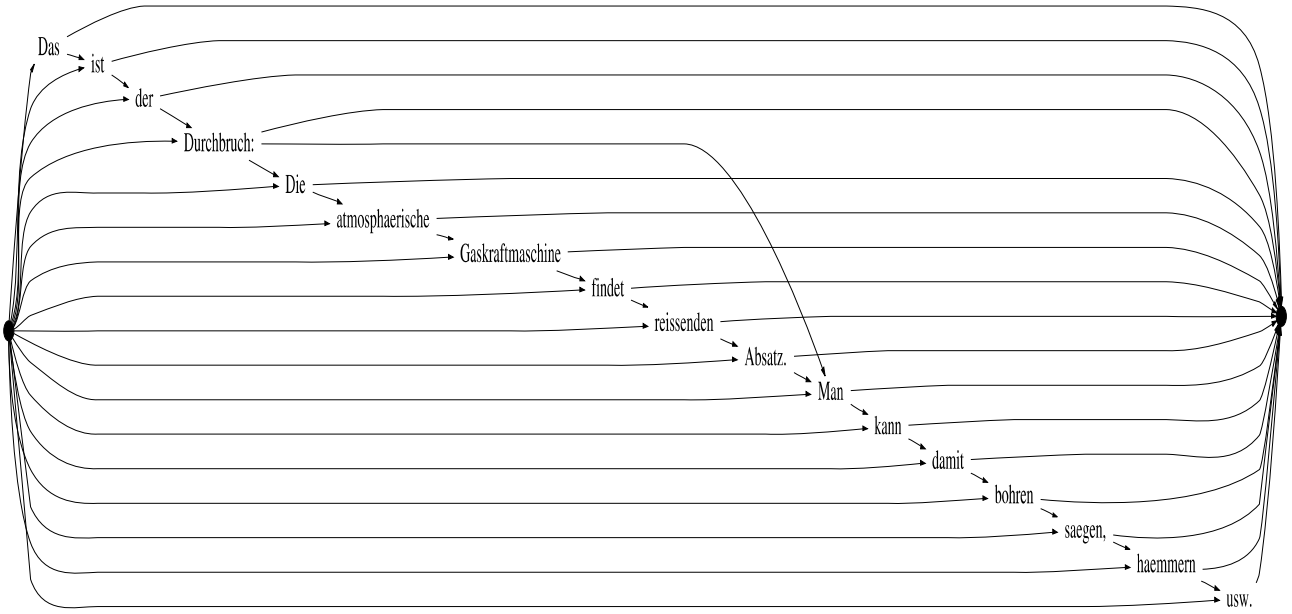

Figure 1: Example lattice for a short text.

pose the word, a statistical decomposition is applied as last resort.

5.2. Calculation of confidence measures

In order to focus in on unreliable regions, we have developed a word and utterance verification algorithm which calculates a word confidence. The posterior probability for each aligned word is calculated using the Bayes formula and the two-best approximation (Dolfing and Wendemuth, 1998):

$$P(W|\mathbf{O}) = \frac{P'(\mathbf{O}|W)P(W)}{P'(\mathbf{O}|W)P(W) + P'(\mathbf{O}|W_{\mathrm{alt}})P(W_{\mathrm{alt}})} \quad (2)$$

W is the aligned word, W_{alt} is a alternative word used for the normalization of the probability. $\mathbf{O}$ is the sequence of 39 dimensional feature vectors containing the MFCC coefficients, the energy, delta features, and delta-delta features. The a priori probabilities $P(W)$ and $P(W_{\mathrm{alt}})$ are equal. The probabilities $P'(\mathbf{O}|W)$ and $P'(\mathbf{O}|W_{\mathrm{alt}})$ are scaled by the number of vectors in the $\mathbf{O}$ times 0.8. This is essential because the independence assumption, which does not hold, is used in speech recognition and results in probabilities near zero.

$$\log P'(\mathbf{O}|W) = \frac{\log P(\mathbf{O}|W)}{0.8 \cdot T} \quad (3)$$

with T the number of feature vectors in $\mathbf{O}$.

The probability $P(\mathbf{O}|W_{\mathrm{alt}})$ of the second best word is determined by extracting the feature vectors belonging to the aligned word from the vector sequence and performing a phone recognition on these features.

5.3. Visualization of results

The words scores are then graphically depicted in an annotation visualization environment directly underneath the speech signal which they represent. For this purpose we use the Transcriber annotation toolkit (Barras et al., 2000) with a modified Document Type Definition (DTD) and the TCL/TK visualization program. The confidence of the aligned word is displayed with a graduation of colors. Words that were recognized with high confidence are depicted in black and the human annotator can skip them. Words with low confidence are depicted in red, and the annotator is alerted to the fact that there is a definite mismatch between the audio and the transcripts at these places that needs to be corrected. Blue and green words are words that were recognized with medium context, and the annotator can click into the audio and listen only to these sections to determine if there indeed is a word that needs to be inserted, deleted or otherwise corrected. The corrections are made by the annotator in the text window of the visual environment, but the changes are automatically added to the XML file format in which the transcripts are stored.

The visual environment can also be used by the annotator to insert or correct segment boundaries. As was mentioned above, the test set and the development set were carefully checked through again by hand, and any boundary mistakes that might have been committed by the BIC-based segmentation algorithm were corrected. Additionally, in the test and the development set, sentence boundaries were all set manually, in order to allow the spoken document retrieval system to be thoroughly baselined in the test phase. The annotation visualization environment played an essential role in this work, allowing the annotator to see the exact time correspondences between the word and the signal. When the annotator moves the segment boundary by dragging with the mouse, the appropriate change in the time stamp is made in the XML file in which the correspondences between text and signal are stored.

In this semi-automatic manner a fast and cost-effective annotation is possible. Figure 5.3. shows the Transcriber tool with the different colors for the confidence of the

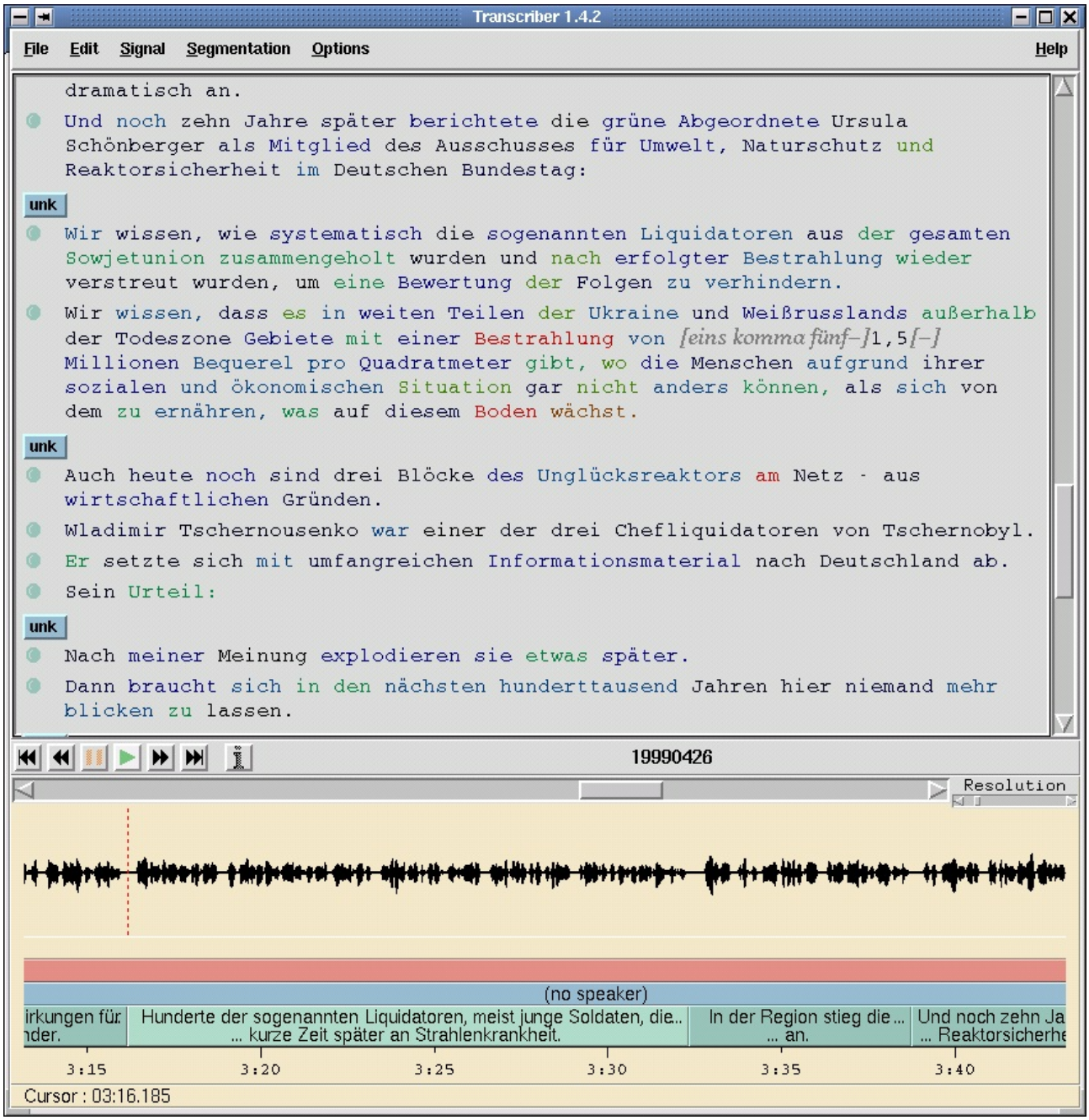

Figure 2: Result of automatic transcription displayed in modified transcriber tool.

aligned words.

6. Classification of topics

Since the databank is to be used to develop classification and information retrieval methods, it is necessary to classify all the texts into reference categories in order to train and test such systems. In final step of database preparation, we use the first level categories of the subject reference system of the International Press Telecommunications Council (IPTC) (http://www.iptc.org) to tag the documents with reference categories. A voting procedure between three human taggers decides which category is chosen as the reference class for each document in the database in order to insure that reference tags reflect average human intuitions about category membership.

7. Conclusions

In this paper we have presented a semi-automatic method of producing a fully annotated audio database from audio data and imperfect transcripts from the Internet. We implement our method with a series of tools that successively download data, normalize text, segment audio, align audio and text. These steps are carried out in a semi-automatic manner, with the annotations decisions which can be easily made automatically left to the computer, and the annotation decisions for which an automatic system cannot be easily trained being taken over by a human annotator. In a final step, topic classes are assigned to all doc-

uments by human annotators.

The method we present here reduces the amount to time necessary to annotate a database for research and development activities in spoken document classification and retrieval by a factor of ten. This decrease in annotation time allows us to annotate more data for the given research budget. An additional benefit of our semi-automatic method, is that the work of annotation is less tedious, making the often dreary task of annotation an all together more pleasant experience.

8. References

Claude Barras, Edouard Geoffrois, Zhibiao Wu, and Mark Liberman. 2000. Transcriber: Development and Use of a Tool for Assisting Speech Corpora Production. *Speech Communication special issue on Speech Annotation and Corpus Tools*, 33(1-2), jan.

J. G. A. Dolfing and A. Wendemuth. 1998. Combination of Confidence Measures in Isolated Word Recognition. In *5th International Conference on Spoken Language Processsing*, pages 3237–3240, Sydney, December.

A. Ganapathiraju, N. Deshmukh, J. Zhao, X. Zhang, Y. Wu, J. Hamaker, and J. Picone. 1999. The ISIP Public Domain Decoder for Large Vocabulary Conversational Speech Recognition. Technical report, Institute for Signal and Information Processing, Mississippi State University, May.

Esther Klabbers, Karlheinz Stöber, Raymond Veldhuis, Petra Wagner, and Stefan Breuer. 2001. Speech Synthesis Development Made Easy: The Bonn Open Synthesis System. In *Proc. EUROSPEECH*, Aalborg, Denmark.

Karlheinz Stöber, Petra Wagner, Jörg Helbig, Stefanie Köster, David Stall, Matthias Thomae, Jens Blauert, Wolfgang Hess, Rüdiger Hoffmann, and Helmut Mangold. 2000. Speech Synthesis by Multilevel Selection and Concatenation of Units from Large Speech Corpora. In Wolfgang Wahlster, editor, *Verbmobil: Foundations of Speech-to-Speech Translation*, Artificial Intelligence, pages 519–537. Springer, Berlin.

Alain Tritschler and Ramesh Gopinath. 1999. Improved Speaker Segmentation and Segments Clustering Using the Bayesian Information Criterion. In *Proc. EUROSPEECH*, volume 2, pages 261–264.