DiSCo — A Speaker and Speech Recognition Evaluation Corpus for Challenging Problems in the Broadcast Domain

Doris Baum, Barbara Samlowski, Thomas Winkler, Rolf Bardeli, and Daniel Schneider
Fraunhofer IAIS
Schloss Birlinghoven
Sankt Augustin
Germany

***Abstract*—Systems for speech and speaker recognition already achieve low error rates when applied to high-quality audiovisual broadcast data, such as news shows recorded in a studio environment. Several evaluation corpora exist for this domain in various languages. However, in actual applications for broadcast data analysis, the data requirements are more complex. There are many data types beyond the planned speech of the news anchorperson. For example, interesting live recordings from prominent politicians are often recorded in an environment with challenging acoustic properties. Discussions typically expose highly spontaneous speech, with different speakers talking at the same time. The performance of standard approaches to speech and speaker recognition typically deteriorates under such data characteristics, and dedicated techniques have to be developed to handle these problems. Corresponding evaluation corpora are needed which reflect the challenging conditions of the actual applications.**

Currently, no German evaluation corpus is available which covers the required acoustic conditions and diverse language properties. This contribution describes the design of a new speaker and speech recognition evaluation corpus for the broadcast domain, reflecting the typical problems encountered in actual applications.

I. Introduction

With the computing power, annotated training material, and refined recognition systems available today, speech and speaker recognition produce sufficiently good results for setting up a useful spoken document retrieval system for restricted domains. Current systems for German data such as [1] achieve satisfying error rates for speech recognition and spoken term detection on a test set of broadcast news data recorded in a studio environment. However, the test set used in the evaluation of this system only contains recordings with no background music or noise, no cross-talk and no telephone data. About half of it is planned speech from professional speakers, i.e., anchorpersons reading news. This leaves out a large part of material contained in broadcasts which is of particular interest to audio search engine users in media archives. Examples for such relevant material include:

- Spontaneous speech from emotionally charged situations, often containing hesitations and stammering
- Debates with speakers interrupting each other
- People with foreign or regional accents
- Voice-overs on foreign language interviews
- Live recorded interviews made in noisy environments
- Telephone interviews
- Public speeches, often containing reverberation
- Background music

Performance evaluations including such challenging problems are required to develop and compare new robust algorithms for speech and speaker recognition. Moreover, such evaluations are often asked for by professional users of spoken document retrieval systems, who need these figures in order to assess the business value of a system.

Although evaluation corpora with some of the required characteristics exist for other languages [2], [3], no sufficiently annotated corpus exists for the German language which covers the required range of material. This paper describes efforts to design a new speech and speaker evaluation corpus, DiSCo (Difficult Speech Corpus), with the goal of measuring and improving a system's performance by testing on representative material from the broadcast domain. Section II contains a summary of related work on speech and speaker recognition on broadcast data and in difficult conditions. Section III describes the most important adverse conditions in broadcast data we identified and the problems they pose. Section IV details the considerations and the decisions made during corpus design and the transcription process, and Section V gives results from experiments carried out on a preliminary version of the new corpus.

II. Speech and speaker recognition in broadcast data

Automatic speech recognition has a wide area of potential applications. Accordingly, the number and diversity of difficult environments for speech and speaker recognition is equally high. For example, speech recognition in car [4], [5] or motorcycle environments [6], in meetings [7], or in broadcast data are areas of active and busy research.

In this context, the broadcast domain is especially interesting for two reasons. First, there is ample demand for automatic analysis of speech in broadcast data. Applications

reach from content-based search and browsing in television, movie, and radio archives [1], [8] to content-enrichment tasks like automatic subtitling [9]. Second, although there are a number of challenging problems for speech technology in this domain, there are also large portions of material which are feasible for automatic methods and thus allow realistic applications to be built.

During the 1990s, broadcast news data has been seen as appropriate material for fostering research in speech recognition, see for example the 1996–1999 NIST Broadcast News Recognition Evaluation [10]. Such high quality broadcast data with large amounts of planned speech is no longer considered sufficiently difficult for the evaluation and promotion of speech recognition tasks. It is therefore often enriched by more difficult conversational speech. This can be seen, for example, in the NIST Rich Transcription Evaluation Project [11]. Also, additional languages move into the focus of attention, e.g., Arabic [12] and Chinese [13].

One problem in the broadcast domain that does not stem from a specific acoustic situation is vocabulary size. In addition to the fact that the vocabulary for this domain is usually quite large, it is also subject to perpetual change. No matter how large the dictionary of a speech recognizer is, new words will always move into the focus of interest and often become the most important to be recognized. There are various approaches for coping with such *out of vocabulary* (OOV) words. One very flexible approach, here, is to not only create word transcriptions but also to retain syllable or other subword transcriptions. In this way, retrieval applications can search for out of vocabulary words by using their subword transcriptions [1], [8].

Speaker recognition is a valuable additional tool for the analysis of broadcast data. Often, users are interested in searching for information provided by specific interesting speakers like politicians or celebrities. In addition to this gain in metadata, speaker recognition allows the application of high-performance acoustic models for individual speakers.

The current standard speaker recognition techniques, such as [14], work very well for clean, studio-recorded, wideband speech, even for large sets of speakers [15]. However, the performance declines dramatically for bad recording or transmission channel conditions (e.g., for telephone data [15]) or when there is mismatch between training and test data capturing conditions. This is due to the fact that they use spectral features to capture the shape of a speaker's vocal tract in order to identify him or her, an approach vulnerable to channel variation and spectral noise. In broadcast data, these kinds of problems are often found, thus making reliable speaker recognition a challenging task. To overcome the limitations imposed by spectral features, a number of speaker recognition approaches using high-level features which try to capture the speakers' intonation, pronunciation, and style, have been proposed [16], [17], [18], [19], [20]. High-level features often require more training and test data but are less susceptible to channel variation and varying acoustic conditions. In order to test which of these techniques might be applicable to a system for German broadcast data, development and test data for speaker recognition from the domain is needed.

III. Adverse conditions in broadcast data

For our purposes, broadcast data falls into three categories:

First, data produced in a studio environment with professional equipment and trained speakers, for which the quality of the speech and audio data is rather high. Even in this controlled environment, the speech information can suffer from certain influences, which makes an automatic analysis of speech more difficult.

Second, data from non-studio productions, like live broadcasts from sports events or documentary features, for which environmental conditions can be even more manifold and adverse. As news and documentaries cover real-life situations, practically all environmental noise conditions might also occur in broadcast and have to be taken into account.

Finally, in both situations an overlap of speakers, i.e., either various speakers speaking at the same time or the situation of voice-overs, poses a considerable challenge to existing speech technology.

However, it can be assumed that some conditions are more likely for the broadcast domain than others. For the development of DiSCo the following conditions are considered to be most dominant and representative for broadcast data, and, hence, should be covered by the corpus:

- **Additive Noise.** Additive noise is the main source of degradation for many speech recognition systems and the most manifold as well. Thus, many scientific publications broach the issue of development and evaluation of algorithms for the reduction of additive noise (e.g, [21], [22]). Every sound which is recorded but which is not part of the analyzed speech can be considered as noise as it generally leads to degradation of the speech or speaker recognition performance.
  Typical additive noise in broadcast can be traffic noise, camera clicking, noise from machines, stadium noise during sports events, etc. Music and speech in the background of a speaker are also additive noise in terms of the previous definition. Due to their specific characteristics, both, music and speech, are classified separately in this corpus, as they might introduce additional challenges for speech analysis. Additive noise can be present in every program, but it is more likely to occur in programs like infotainment shows, talk shows, sports event coverage, news event coverage, and light programs.
- **Music in the Background.** Music in the background of a speaker is a common type of additive noise in broadcast programs. But due to its specific harmonic characteristics, the influence of background music on speech analysis is often severe and, therefore, of particular interest in speech recognition for the broadcast domain [23], [24]. Hence, music is classified separately for this corpus. Music is often mixed artificially into the background of a speaker to create a certain atmosphere. But music can also be part of the real acoustic environment of a recording. Music

in the background is used or can be present in several programs, e.g., infotainment shows and documentaries.

- **Speech in the Background.** Background talk is very critical for speech analysis, as it is rather difficult to separate two or even more speakers [25]. Another speaker in the background – or, even worse: cross-talk situations, i.e., two speakers speaking at the same time with about the same volume – dramatically decreases the performance of speech and speaker recognition systems. Background speech is often present in interviews or in voice-over situations like translations of original speech. Typical programs for background speech, voice-over, and cross-talk are mainly political debates, talk shows, and news.
- **Reverberation.** Reverberation and its effect and compensation in robust speech recognition is a seperate field of research [26]. Reverberation is caused by acoustic characteristics of the room. Studio data is generally low in reverberation, but speech in political debates of the parliament, for example, often suffers from reverberation effects. Similar challenges are echos caused by acoustic feedback. A prominent example is telephone speech in the broadcast environment. Echos mainly occur when the speaker on the telephone uses handsfree devices or listens to the delayed channel of his broadcast device while calling a live show. Parliament debates and call-in shows are qualified for providing data with distortions caused by reverberation and echos.
- **Telephone Speech.** Telephone speech has specific channel characteristics and provides much worse speech quality than high quality studio recordings. Additionally, the channel characteristics also vary for different phone channels (GSM, ISDN, analog connections, etc.). Additive noise and echos can also be present in telephone speech. Thus, telephone speech suffers from many different sources of degradation [27]. In the broadcast domain, telephone speech can mainly be found for telephone interviews and for some live coverages from foreign correspondents (often with additive noise in the background). A sufficient quantity of telephone speech in the broadcast domain is covered by adding a call-in show to the corpus.
- **Speech Diversity.** A more generalized challenge in automatic speech recognition and speech analysis is the diversity of speech. Though most speech in the broadcast domain is quite clear and planned, fast speakers, speakers with different accents and dialects as well as spontaneous speech can also be present in specific programs. All these variations and individual characteristics in speech complicate a reliable automatic speech recognition [28]. A broad selection of different speech and speaker characteristics is achieved by capturing a variety of different programs including news, talk shows, sports shows, etc.

IV. Corpus Design

A. Intended Use and Applications of the Corpus

There are many different types of corpora, each having its own set of demands on the data and the annotations. In Llisterri's guidelines for building spoken corpora [29], two large groups are identified according to their applications and user communities:

The first group consists of corpora developed by the so-called "corpus linguistics community" in order to provide data for linguistic research. Topics of interest include conversation and discourse analysis, children's or child-directed speech, and the development of lexica. Corpora of this type require the data to be as natural as possible. In many cases, spontaneous conversation is preferred. Annotations may include grammatical tagging as well as prosodic information, while exact information about word pronunciation can often be disregarded.

The second group comprises corpora compiled by the "speech community" which focuses on theories of phonetics and phonology as well as on technical and technological applications thereof. Traditionally, corpora developed by this user group are produced in a very controlled environment. Often, prompt sentences are read aloud and recorded under laboratory conditions. The speech community tends to place more emphasis on the pronunciation of words than on prosody or grammatical issues.

As an evaluation corpus for automatic speech and speaker recognition, our database belongs to the second category. However, in order to simulate the real-life situations our recognition system is intended for, we use natural and spontaneous speech gathered from reports and interviews transmitted on television and the internet, rather than controlled recordings of phonetically balanced sentences. Our database is designed to cover a wide range of acoustic situations so as to reflect the many challenges confronting automatic speech and speaker recognition and term detection outlined in the previous sections. It includes speech samples from a number of well-known public figures of interest to train and test speaker recognition in adverse conditions.

B. Types of Data Included

One difficulty in putting together a broadcast corpus suitable for our purposes is the uncertainty in predicting which television programs will contain what type of data. Therefore, a good coverage of programs containing the different adverse situations targeted by the corpus is vital. The following list gives an overview of the recorded material and the special acoustic situations they cover:

- News Broadcasts
  Daily news programs contain different types of data, but the speaking style, in general, is formal and planned. Often, texts are read by professional newscasters. There are longer passages of clean speech, which can be used in comparisons against more complicated data. During reports and commentaries from experts and politicians, however, background noise is often present, and in many cases news summaries are read against a background of music.
- In-depth News Commentaries

Programs of this type provide detailed analysis and discussion of current events. The topics are similar to those dealt with in news broadcasts, but there are also longer interviews with prominent public figures and celebrities. Overlapping tends to occur in discourse between interview partners as well as in passages of foreign speech which are superimposed with simultaneous translations.

- Sports Commentaries
  These shows, which are similar in structure to the programs in the foregoing category, feature news from the world of sports, with German shows often focusing on soccer events. These shows contain informal interviews, on occasion with voice-over translations, as well as a considerable amount of audience and stadium noise.
- Infotainment Shows
  Popular science shows are conducted in a planned but informal speech style. They also contain short passages of spontaneous speech from street interviews. Background music is especially prevalent here, so these recordings serve as test material for dealing with voice over music.
- Political Talk Shows
  In these discussion rounds, politicians, public figures, and other guests debate specific topics. They contain passages of heated argumentation with spontaneous speech and considerable speaker overlap. Moreover, they are a challenging test instance for speaker recognition.
- Parliamentary Debates
  The speeches in these debates are often planned, but the recordings include much background noise from the audience as well as a high level of reverberation. Furthermore, as the speakers are important politicians, the data is a challenging test case for speaker recognition.
- Call-in Shows
  One important application for robust speech recognition is telephone speech. Short telephone interviews can occasionally be found in news broadcasts and commentaries. To increase this type of data in our corpus, we decided to include recordings from a call-in show. The informal style of this type of show increases the spontaneous speech part of the corpus.
- Crime Fiction Series
  As a final especially challenging test case, we included a few installments of crime series. In these programs, several kinds of complex speech material are combined - speaker overlap, excessive background noise, and background music.

C. The Annotation Process

The manual annotation process is designed to be iterative: A preliminary set of annotations is produced and then reviewed by the human annotators in terms of content and formal aspects so that mistakes can be corrected. Where expedient, the annotation guidelines are modified in order to obtain better results in the next cycle.

The recordings are annotated in three phases. In the first phase the data is segmented into utterances and transcribed orthographically. For this process, we use the program Transcriber[1]. During the following two phases, the utterances are classified into groups using an especially developed annotation program, called DIVE. In the second phase, each utterance is labelled according to speaker. The utterances are also classified according to a specific set of features from a given list (see Table I).

TABLE I
ANNOTATION FEATURES AND ATTRIBUTES FOR THE DISCO CORPUS

Feature	Attributes
background noise	yes / no
channel quality	studio / telephone / other
type of speech	spontaneous / planned
speech rate	low / medium (default setting) / high

During the third phase, the data groups are analysed to refine the classes for re-annotation. Different types of background noise are specified.

This procedure is devided into the following steps:

- **Step 1** - Recording the data:
  In the first step, the television programs are recorded by a digital video recorder. The resulting files are saved into separate directories according to program name and into subdirectories indicating the time and date of recording. Each recording comprises three different types of files: an index file, the video files themselves and a text file with additional information such as program subheadings or summaries. At this stage, a first quality check secures that the programs have been recorded properly.
- **Step 2** - Producing the scripts:
  In order to further process the data, several scripts have to be created to convert the files into the required formats. These will be described together with the step in which they are used. Unlike the other tasks described here, this step does not have to be repeated for every recording.
- **Step 3** - Producing the audio files:
  The audio files used for transcribing the recordings have to be extracted from the video file with the help of a script. For automatic speech recognition as well as for human text transcriptions a wave file (16 kHz, 16 bit, stereo) is needed. The following annotation according to classes will be done on the basis of an mp4 audio/video file.
- **Step 4** - Gathering the metadata:
  Another script is necessary to gather information about the recording and the program into an XML-file.
- **Step 5** - Creating the text transcription framework
  Optionally, the transcribers can use an automatically computed transcription and segmentation as a basis for their work.
- **Step 6** - Creating the orthographic transcriptions
  In the first annotation phase, the data are transcribed orthographically. Silence or pure background noise, unintelligible or foreign speech, and speaker overlap are

[1]http://trans.sourceforge.net/

not transcribed, but marked separately. For this step, annotation guidelines detail the transcription conventions for, among other things, numbers, compound words, words with different possible spellings, contractions and hesitations.

- **Step 7** - Combining transcription and metadata
  The orthographic transcription and the collected information about the respective recording are combined into a single XML file.
- **Step 8** - Creating the classification framework
  As in step 5, a skeleton classification file is automatically created to aid the annotators in their task of dividing the data according to speaker and the selected features.
- **Step 9** - Classifying the data
  In the second annotation phase, the utterances are tagged according to speaker as well as to a set of predetermined features.
- **Step 10** - Analysing the annotated data
  The resulting groups of data are analysed and a new set of classes are determined according to which the utterances are to be annotated a second time.
- **Step 11** - Reiteration of classification
  During the third annotation phase, the refined class features are used to re-annotate the utterances.
  Following these steps, a well annoted corpus with rich information for the evaluation and development of speech technology is derived.

V. Experimental results

A. Linguistic analysis

Preliminary linguistic analysis has been performed for a subset of the recorded programs in order to gain insight into the distribution of important parameters for speech recognition. This preliminary corpus contains approximately four hours of speech from five different German television programs covering a political discussion show, a foreign affairs report, an interview show, a regional infotainment show, and a sports show. Figure 1 shows the fraction of the corpus covered by each of these programs.

Transcribable speech accounts for 77.6 percent of the total time, i.e., three hours and ten minutes. The remaining part comprises 16.5 percent of silence or pure background noise, 4.7 percent of unintelligible speech, and 1.8 percent of speaker overlap, with two or more people speaking at the same time. The amount of time taken up by periods of silence, unintelligible speech, and speaker overlap vary from program to program. As can be seen in Table II, discussion shows contain more speaker overlap than news commentaries and a program featuring international news includes larger amounts of foreign speech, which has been tagged as unintelligible.

One important aspect of linguistic corpus analysis is the assessment of word type distributions. Frequency lists can be produced which record the different word forms or types that the corpus consists of together with the number of tokens belonging to each of these word types, i.e., the number of times that one particular word form appears in the corpus. Depending on the aim of the analysis, the definition for distinguishing word types can vary. For some studies, e.g., determining the vocabulary of a language, it may be advisable to count different grammatical forms of a word as one word type or to differentiate between words that are spelled in the same way but have different meanings [30]. Generally the word types of a corpus are not distributed equally. On the contrary, studies often show that while a few word types appear very often, a large number occur seldom or only once, i.e., they follow a Zipfian distribution [31].

The fact that corpora regularly contain a few strongly represented words and a large percentage of ”hapax legomena”, i.e., word types which appear only once, poses a challenge for corpus-based research and applications of speech technology, where representative data are required [32]. This is also true for evaluation corpora such as the DiSCo database. Besides word transcripts, some speech recognizers can also produce transcripts on the subword level, allowing for vocabulary independent speech search. As our speech recognition system produces both word and syllable transcripts, the type-token relations for this corpus will be analyzed on the level of both, syllables as well as words.

The database collected so far contains 34,387 word tokens that can be divided into 6,305 orthographic or 6,067 phonological word types. The latter are distinguished by their standard pronunciation. Accordingly, homophones are counted as one type and stammered words are not treated as separate types.

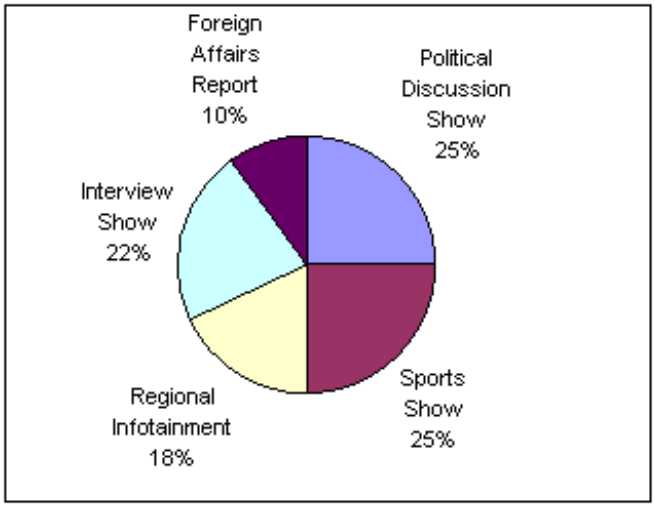

Fig. 1. Contribution of different broadcast formats to the total length of the preliminary corpus.

TABLE II
Distribution of silence, unintelligible speech and overlapping speeech according to program type

Program	Silence	Unintelligible	Overlap
Interview show	14.54%	3.21%	0.88%
Political discussion show	11.83%	1.14%	4.89%
Foreign affairs report	17.89%	15.19%	0.34%
Regional infotainment with dialect	18.49%	2.60%	0.93%
Sports show	20.88%	6.40%	0.82%

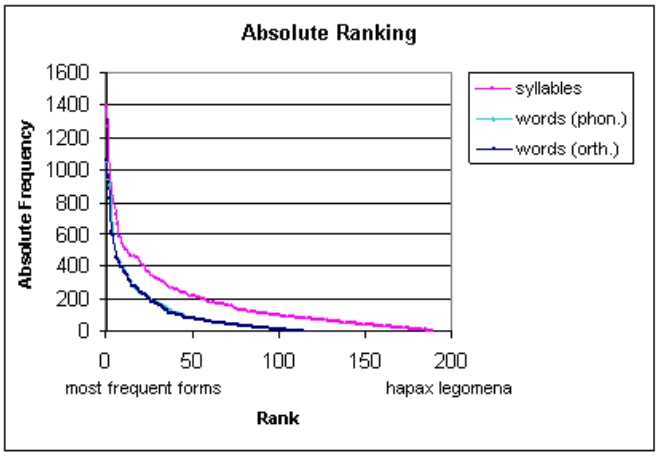

Fig. 2. Absolute word and syllable frequencies according to their frequency rank.

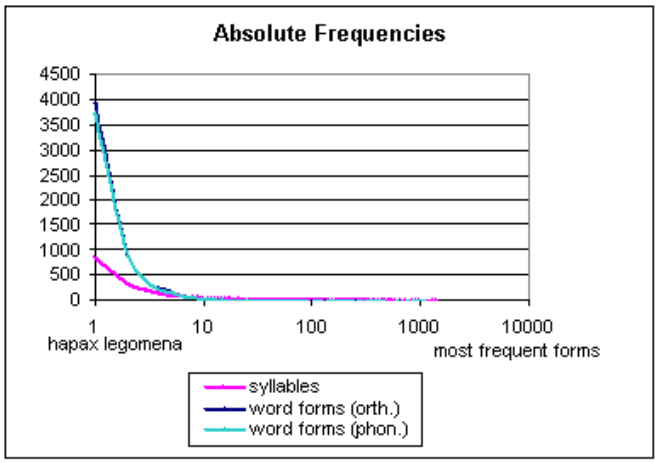

Fig. 3. Word type frequency vs. occurrence per frequency.

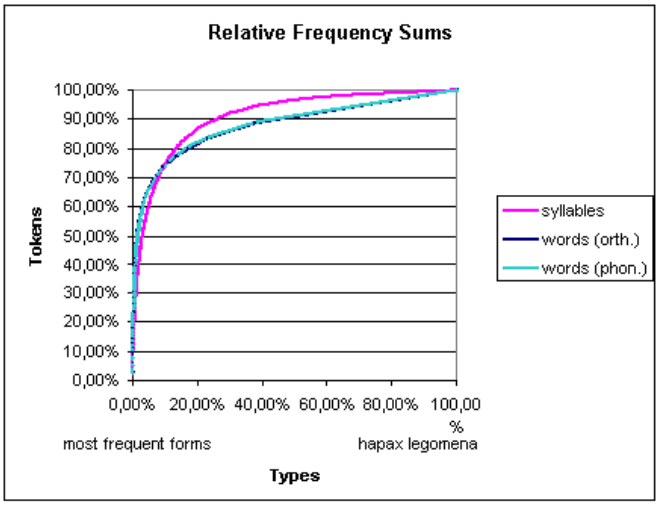

Fig. 4. Running total of relative type and token frequencies.

The corpus can be divided into 59,788 syllable tokens, which belong to 2,653 types. As for the phonological word forms, syllable types are defined by pronunciation.

A chart representing the absolute word and syllable frequencies according to their rank shows that although there are more different syllables than words, the syllable frequencies decline less rapidly than word frequencies and require more ranking steps to reach the lowest syllable frequency (Figure 2). One reason for this is that they start out on a higher level, as the most frequent word form, the German article *die*, appearing 1059 times, is subsumed by the corresponding syllable, which occurs 1396 times. It becomes apparent here as well as in the following charts that the curves for orthographic and phonological words follow very similar paths.

Another way of visualizing the results is by plotting the frequency of the word types against the number of occurrences for this frequency, e.g., how many word types appear only once in the corpus (Figure 3). This puts more emphasis on uncommon word types than the frequency ranking approach [31]. Here, it becomes obvious that the number of syllables that occur only once is significantly lower than the number of singly occurring word forms.

The last fact is also confirmed by an analysis of relative frequencies. While about 60 percent of the corpus's word types occur only once, comprising 10 percent of the total corpus, the percentage of hapax legomena syllables is slightly above 30 percent. Only 1.4 percent of the corpus is made up of unique syllables. Figure 4, which represents the running total of relative type and token frequencies, starting with the most frequent types, shows that both on word and on syllable level, 75% of the corpus can be represented by the top ten percent of word or syllable types. Furthermore it can be seen that uncommon syllable types make up less of the corpus in comparison to rare word types.

On the whole, both the phonological and the orthographic word forms of our corpus follow the expected Zipfian distribution. In absolute numbers, there are less different syllable types that have to be recognized, while a relative analysis shows that frequent syllables comprise more and infrequent syllables less of the corpus than the corresponding word types.

B. Automatic Speech Recognition

Preliminary experiments using automatic speech recognition (ASR) have been carried out on a subset of the recorded programs. The goal of this first pre-evaluation was to gain insight into the challenges of the individual recording types. We used an ASR setup based on the configuration described in [1] with an increased recognition vocabulary of 200,000 words and a trigram language model.

In order to eliminate the effect of automatic segmentation errors on the speech recognition result, a manual segmentation into speech segments was carried out before the actual transcription. Table III shows the word error rate on the manually segmented speech portion of the selected evaluation files.

The resulting error rates reflect the varying level of acoustic and linguistic complexity of the recordings. The lowest overall

TABLE III
OVERALL WORD ERROR RATE (WER) ON SELECTED PROGRAMS

Program	WER
News show, only planned speech	16.1 %
Interview show	29.4%
Political discussion show	39.9%
Foreign affairs report	41.8%
Regional infotainment with dialect	52.0%
Sports show	64.5%

word error rate can be observed on the planned speech portion of a broadcast news show, read by a professional speaker in a silent studio environment. The results on the interview program indicate that spontaneous speech presents an additional challenge to speech recognition, even if the interviewed people in the analyzed show are *media professionals* such as politicians. The recognition rate degrades further if the prevalent speech type changes from interview to discussion. Here, the speech of the participants is not only highly spontaneous, it can also be emotional, and vary greatly in speed. Speakers interrupt each other frequently, and this cross-talk makes speech recognition even harder.

As stated above, an additional challenge for speech recognition algorithms is background noise. A large part of the evaluated foreign affairs report contains voice-overs with the translation of a non-German recording, with two active voices confusing the speech recognizer. Moreover, the number of OOV words in the report is higher than the average OOV rate observed in the data annotated so far.

The performance of an ASR system drops if the mismatch between training and evaluation data increases. This is particularly the case when dialectal speech is to be recognized and the dialect was not present in the training set. Without any additional acoustic adaptation [33] the word error rate increases significantly.

The sports show has some challenging acoustic and linguistic properties, posing additional problems for the speech recognition system. There is a large portion of highly spontaneous speech in interviews, as well as a high number of OOVs due to the frequent occurrence of proper names of athletes or sports clubs. Moreover, recordings from sport events usually take place in a rather loud acoustic environment, including intense crowd noise or noise from the sport itself (such as motor car noise).

Although the observed word error rates are rather high, it is still possible to use the resulting transcripts for spoken document retrieval [34]. Corresponding information retrieval experiments with Spoken Term Detection comparable to [1] will be carried out in future evaluations.

To gain further insight into the challenges of the various programs, they were manually labeled according to the type of speech used. One of the labels *spontaneous*, *planned*, and *unsure* was assigned to each segment of speech manually. Only those segments labeled as *spontaneous* or as *planned* were used for the further analysis, to investigate the difficulty they pose for automatic speech recognition. Table IV shows the amount of planned and spontaneous speech in each of the programs. It becomes apparent that the political discussion show contains mostly spontaneous speech and that the foreign affairs report has mostly planned speech – from reporters and interpreters. For the rest of the programs, there is a rather balanced proportion of both speech types.

TABLE IV
TIME OF SPEECH TYPES IN THE SELECTED PROGRAMS

Program	Minutes planned	Minutes spontaneous
Interview show	8:06	8:17
Political discussion show	3:39	31:32
Foreign affairs report	17:05	0:49
Regional infotainment with dialect	20:50	15:19
Sports show	19:40	27:31

TABLE V
WORD ERROR RATES (WER) ON PLANNED AND SPONTANEOUS SPEECH

Program	WER planned	WER spontaneous
Interview show	18.30%	39.00%
Political discussion show	27.90%	40.50%
Foreign affairs report	39.70%	76.10%
Regional infotainment with dialect	48.50%	54.40%
Sports show	58.70%	68.30%
Weighted sum	44.60%	52.00%

The word error rates were recalculated for both classes, the results are shown in Table V. Speech type alone can not explain the differences in error rates between the programs – other factors have to be taken into account. While for the interview show the word error rate of planned speech is almost as low as for the news broadcast in Table III (18.3% vs. 16.1%), the other programs have a much higher word error rate on their planned speech part. We suspect that the reasons for this are background music, noise, and overdubbing of translations in the case of the foreign affairs report, dialect and talking speed in the case of the regional infotainment show, and stadium and audience noise in the case of the sports show. For spontaneous speech, the word error rate is about 40% in clean acoustic environments with practised speakers talking standard German, such as in the interview and political discussion shows. This rate deteriorates further if dialect is used or in noisy environments.

Altogether, spontaneous speech poses severe problems for automatic speech recognition, increasing the word error rate, often by at least 20% – but it is, of course, not the only challenge to be tackled. So a corpus for evaluation of difficult speech must facilitate more annotation categories, like dialect, talking speed, and background noise.

VI. CONCLUSION

Taking speech and speaker recognition in real world scenarios to the next level is only possible with a corpus documenting exactly those challenges which are just out of reach of the current state of the art. For the German language, no such corpus has been available. Our experiments show that the types

of broadcast material selected for our corpus covers very well the kind of material that is difficult to handle by state-of-the-art algorithms. Once completed, the DiSCo corpus will serve as a solid foundation for the evaluation of progress in the domains it covers and will thus help developing more robust speech and speaker recognition algorithms.

VII. Acknowledgments

The work for this contribution was supported by the projects CONTENTUS[2], MoveOn[3] and VITALAS[4].

References

[1] D. Schneider, J. Schon, and S. Eickeler, "Towards large scale vocabulary independent spoken term detection: Advances in the Fraunhofer IAIS Audiomining System," in *Proceedings of the ACM SIGIR Workshop "Searching Spontaneous Conversational Speech" held at SIGIR '08*, J. Köhler, M. Larson, F. Jong de, W. Kraaij, and R. Ordelman, Eds., Singapore, 24 July 2008 2008. [Online]. Available: http://ilps.science.uva.nl/SSCS2008/Proceedings/sscs08_proceedings.pdf

[2] J. Garofolo, E. Voorhees, C. Auzanne, and B. Stanford, V.and Lund, "Design and preparation of the 1996 hub-4 broadcast news benchmark test corpora," in *Proceedings of the DARPA Speech Recognition Workshop.*, 1997, pp. 15–21.

[3] S. Galliano, E. Geoffrois, G. Gravier, J.-F. Bonastre, D. Mostefa, and K. Choukria, "Corpus description of the ester evaluation campaign for the rich transcription of french broadcast news," in *Proceedings of the 5th international Conference on Language Resources and Evaluation (LREC 2006).*, 2006, pp. 315–320.

[4] A. Moreno, B. Lindberg, C. Draxler, G. Richard, K. Choukri, S. Euler, and J. Allen, "SPEECHDAT-CAR. a large speech database for automotive environments," in *Proceedings of the 2nd International Conference on Language Resources and Evaluation, (LREC 2000)*, Athens, Greece, 2000.

[5] Y. Gong, "Speech recognition in noisy environments: a survey," *Speech Communication*, vol. 16, no. 3, pp. 261–291, 1995.

[6] T. Winkler, T. Kostoulas, R. Adderley, C. Bonkowski, T. Ganchev, J. Köhler, and N. Fakotakis, "The moveon motorcycle speech corpus," in *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, ELRA, Ed., Marrakech, Morocco, May 2008.

[7] E. Shriberg, R. Dhillon, S. Bhagat, J. Ang, and H. Carvey, "The ICSI Meeting Recorder Dialog Act (MRDA) Corpus," in *Proceedings of 5th SIGdial Workshop on Discourse and Dialogue*, M. Strube and C. Sidner, Eds., 2004, pp. 97–100.

[8] M. Akbacak, D. Vergyri, and A. Stolcke, "Open-vocabulary spoken term detection using graphone-based hybrid recognition systems," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2008).*, 2008, pp. 5240–5243.

[9] P. Wambacq, P. Vanroose, X. Yang, J. Duchateau, and D. H. V. Uytsel, "Speech recognition for subtitling purposes," in *Proceedings 5th International Conference Languages & The Media*, November 2004, p. 46.

[10] "The 1999 NIST Evaluation Plan for Recognition of Broadcast News, in English," 1999. [Online]. Available: http://www.nist.gov/speech/tests/bnr/1999/bnews_99_spec.html

[11] "Rich transcription evaluation project." [Online]. Available: http://www.nist.gov/speech/tests/rt/

[12] D. Vergyri, A. Mandal, W. Wang, A. Stolcke, J. Zheng, M. Graciarena, D. Rybach, C. Gollan, R. Schlater, K. Kirchoff, A. Faria, and N. Morgan, "Development of the sri/nightingale arabic asr system," *to appear in Proceedings of Interspeech 2008, Brisbane, Australia*, 2008.

[13] S. Chu, H. kwang Kuo, Y. Y. Liu, Y. Qin, Q. Shi, and G. Zweig, "The IBM Mandarin Broadcast Speech Transcription System," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2007)*, vol. 2, April 2007, pp. II–345 – II–348.

[2]http://theseus-programm.de/scenarios/de/contentus.html

[3]http://www.m0ve0n.net/

[4]http://vitalas.ercim.org/

[14] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted gaussian mixture models," in *Digital Signal Processing*, vol. 10, 2000, pp. 19–41.

[15] D. A. Reynolds, "Large population speaker identification using clean and telephone speech," *IEEE Signal Processing Letters*, vol. 2, no. 3, pp. 46–48, March 1995.

[16] D. Reynolds, W. Andrews, J. Campbell, J. Navratil, B. Peskin, A. Adami, Q. Jin, D. Klusacek, J. Abramson, R. Mihaescu, J. Godfrey, D. Jones, and B. Xiang, "The SuperSID project: Exploiting high-level information for high-accuracy speaker recognition," in *Proceedings of the International Conference on Audio, Speech, and Signal Processing*, vol. 4, Hong Kong, 2003, pp. 784–787. [Online]. Available: http://www.clsp.jhu.edu/ws2002/groups/supersid/icassp03_overview.pdf

[17] D. A. Reynolds, J. P. Campbell, W. M. Campbell, R. B. Dunn, T. P. Gleason, D. A. Jones, T. F. Quatieri, C. B. Quillen, D. E. Sturim, and P. A. Torres-Carrasquillo, "Beyond cepstra: Exploiting high-level information in speaker recognition," in *Workshop on Multimodal User Authentication*, Santa Barbara, California, December 2003, pp. 223–229.

[18] M. K. Sönmez, E. Shriberg, L. Heck, and M. Weintraub, "Modeling dynamic prosodic variation for speaker verification," in *International Conference on Spoken Language Processing (ICSLP98)*, vol. 7, Sydney, Australia, 1998, pp. 3189–3192.

[19] F. Weber, L. Manganaro, B. Peskin, and E. Shriberg, "Using prosodic and lexical information for speaker identification," in *IEEE International Conference on Acoustics, Speech, and Signal Processing 2002 (ICASSP '02)*, vol. 1, 2002, pp. 141–144. [Online]. Available: http://www.icsi.berkeley.edu/ftp/pub/speech/papers/icassp02-spid.pdf

[20] B. Peskin, J. Navratil, J. Abramson, D. Jones, D. Klusacek, D. A. Reynolds, and B. Xiang, "Using prosodic and conversational features for high-performance speaker recognition: report from JHU WS'02," in *IEEE International Conference on Acoustics, Speech, and Signal Processing 2003 (ICASSP '03)*, vol. 4, April 2003, pp. 792–795. [Online]. Available: http://www.icsi.berkeley.edu/ftp/global/pub/speech/papers/icassp03-peskin2.pdf

[21] H. Hirsch and D. Pearce, "The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *6th International Conference on Spoken Language Processing (ICSLP'00)*, Bejing, China, October 2000.

[22] J. Ming, "Noise compensation for speech recognition with arbitrary additive noise," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 14, no. 3, pp. 833–844, May 2006.

[23] B. Raj, V. Parikh, and R. Stern, "The effects of background music on speech recognition accuracy," in *ICASSP '97: Proceedings of the 1997 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '97)-Volume 2*. Washington, DC, USA: IEEE Computer Society, 1997, p. 851.

[24] P. Vanroose, "Blind source separation of speech and background music for improved speech recognition," in *Proceedings of the 24th Symposium on Information Theory in the Benelux*, 2003, pp. 103–108.

[25] J. R. Hershey, S. J. Rennie, P. A. Olsen, and T. T. Kristjansson, "Super-Human Multi-Talker speech recognition: A graphical modeling approach," *Computer Speech & Language*, vol. In Press, Accepted Manuscript, 2009. [Online]. Available: http://www.sciencedirect.com/science/article/B6WCW-4V8VS6S-1/2/e7f3b944847579 52bb02a525b3b44772

[26] S. Thomas, S. Ganapathy, and H. Hermansky, "Recognition of reverberant speech using frequency domain linear prediction," *Signal Processing Letters, IEEE*, vol. 15, pp. 681–684, 2008. [Online]. Available: http://dx.doi.org/10.1109/LSP.2008.2002708

[27] P. Moreno and R. Stern, "Sources of degradation of speech recognition in the telephone network," *Acoustics, Speech, and Signal Processing, 1994. ICASSP-94., 1994 IEEE International Conference on*, vol. i, pp. I/109–I/112 vol.1, Apr 1994.

[28] M. Nakamura, K. Iwano, and S. Furui, "Differences between acoustic characteristics of spontaneous and read speech and their effects on speech recognition performance," *Computer Speech & Language*, vol. 22, no. 2, pp. 171–184, Apr. 2008. [Online]. Available: http://www.sciencedirect.com/science/article/B6WCW-4PCPFHV-1/2/17f5b4d652317795d7e57a65744c1c97

[29] J. Llisterri, "Preliminary recommendations on spoken texts," Expert Advisory Group on Language Engineering Standards, Tech. Rep., 1996. [Online]. Available: http://www.ilc.cnr.it/EAGLES96/spokentx/spokentx.html

[30] G. Kennedy, *An introduction to corpus linguistics*, ser. Studies in language and linguistics. London: Longman, 2003.
[31] E. Leopold, "Das Zipfsche Gesetz," *Künstliche Intelligenz*, no. 2/02, p. 34, 2002.
[32] J. Sinclair, "Corpus and text: Basic principles," in *Corpus and Text: Basic Principles*. Oxford: Oxbow Books, 2005, pp. 1–16.
[33] P. C. Woodland, "Speaker adaptation for continuous density HMMs: a review," in *ITRW on Adaptation Methods for Speech Recognition*, 2001, pp. 11–19. [Online]. Available: http://publications.eng.cam.ac.uk/2000/
[34] J. Garofolo, G. Auzanne, and E. Voorhees, "The TREC spoken document retrieval track: A success story," in *Proceedings of the Content Based Multimedia Information Access Conference*, 2000. [Online]. Available: http://citeseer.ist.psu.edu/garofolo00trec.html