# DiSCo — A German Evaluation Corpus for Challenging Problems in the Broadcast Domain

**Doris Baum*, Daniel Schneider*, Rolf Bardeli*, Jochen Schwenninger*, Barbara Samlowski†, Thomas Winkler*, Joachim Köhler***

*Fraunhofer IAIS
Sankt Augustin, Germany
{firstname.lastname}@iais.fraunhofer.de

†Institut für Kommunikationswissenschaften
University of Bonn
Bonn, Germany
bsamlows@uni-bonn.de

**Abstract**

Typical broadcast material contains not only studio-recorded texts read by trained speakers, but also spontaneous and dialect speech, debates with cross-talk, voice-overs, and on-site reports with difficult acoustic environments. Standard approaches to speech and speaker recognition usually deteriorate under such conditions. This paper reports on the design, construction, and experimental analysis of DiSCo, a German corpus for the evaluation of speech and speaker recognition on challenging material from the broadcast domain. One of the key requirements for the design of this corpus was a good coverage of different types of serious programmes beyond clean speech and planned speech broadcast news. Corpus annotation encompasses manual segmentation, an orthographic transcription, and labelling with speech mode, dialect, and noise type. We indicate typical use cases for the corpus by reporting results from ASR, speech search, and speaker recognition on the new corpus, thereby obtaining insights into the difficulty of audio recognition on the various classes.

## 1. Introduction

In order to develop robust methods for automatic multimedia indexing and retrieval of broadcast data, a representative evaluation corpus is needed to assess new approaches. Although there is ample demand for robust methods for indexing and retrieval of broadcast data, hardly any speech corpus exists covering both clean speech and more challenging speech segments. Such a multi-purpose broadcast corpus should contain material reflecting common difficult acoustic conditions present in general broadcast data, such as spontaneous speech, background noise, and dialect, as well as clean studio material for comparison. It should also have a sufficient number of speakers to cover variations in speaking style and dialect and to be able to meaningfully test speaker recognition systems. Although there are German speech corpora with broadcast data for a number of tasks (Grimm et al., 2008; Hecht et al., 2002), no corpus fulfilling all aforementioned requirements is currently available.

Also in other languages, evaluation corpora for rich transcription often focus on the broadcast domain (NIST, 2003; Galliano et al., 2006). Some of these corpora (NIST, 1999) deliberately exclude difficult material like sports broadcasts that has been judged as too difficult at the time of corpus creation. Incorporating material that is likely to become the next level of difficulty that can be managed is a main task for new speech recognition corpora. Annotating in depth to cover such effects is highly time-consuming, and thus such corpora usually comprise only a few hours of material (e.g. 3 hours for the NIST-RT03-BN-English corpus and 10 hours for the ESTER corpus).

This paper describes the design and characteristics of a new speech and speaker evaluation corpus, DiSCo (Difficult Speech Corpus), with the goal of measuring and improving audio analysis performance on broadcast material beyond planned and clean speech data.

## 2. Corpus Design

One of the key requirements for the corpus design was a good coverage of different types of serious programmes beyond clean speech and planned speech broadcast news. As it was not feasible to record hundreds of different shows, common categories of programmes and typical speech, speaker and background noise characteristics of information shows were analysed. The major background noise and speech characteristics predefined several of the annotation classes described in detail in Section 3. Furthermore, we used these characteristics and the relevant categories of programmes to select representative broadcasts in order to cover all important conditions. We recorded 18 hours of video material, comprising 29 broadcasts from 8 types of programmes. They fall into the following categories:

- **News** mostly have a formal and planned speaking style because texts are read by professional newscasters. They contain long passages of clean speech, which can be used for 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.
- **Political interview shows** provide detailed analysis and discussion of current events. They contain interviews with politicians, where cross-talk tends to occur in discourse between interview partners.

Table 1: Amount of material in DiSCo for the 8 programme types.

| Programme Type | Material (hh:mm) | Percentage |
|---|---|---|
| News broadcasts | 01:12 | 6% |
| Political interview shows | 02:23 | 13% |
| Sports commentaries | 02:01 | 11% |
| Science shows | 02:00 | 11% |
| Political talk shows | 04:45 | 25% |
| Regional reports | 01:30 | 8% |
| Foreign affairs reports | 03:05 | 16% |
| Television news magazines | 01:45 | 9% |
| All | 18:40 | 100% |

- **Sports commentaries** feature news from the world of sports, with German shows often focusing on football events. They contain interviews, sometimes with voice-over translations, as well as a considerable amount of audience and stadium noise. A difficulty they pose for spoken document retrieval is the higher out-of-vocabulary (OOV) rate due to the large number of sports terms and foreign names of athletes not present in the speech recogniser's dictionary.

- **Science shows** are often conducted in a planned speech style during the documentary part and in spontaneous style during announcements and discussions. Background music is particularly prevalent here. Similar to sports commentaries, the technical terms and topics of science shows differ from those present in normal broadcast news training material, thus potentially increasing the OOV rate.

- **Political talk shows** contain debates of politicians, public figures and others, displaying passages of spontaneous speech and considerable speaker overlap.

- **Regional reports** cover news from specific regions and thus have a lot of accented or dialectal material, often including on-scene interviews with noise or added music. Our corpus contains broadcasts from Bavaria, where the local dialect contrasts strongly with standard German.

- **Foreign affairs reports** are similar in structure to regional reports. However, instead of dialect they contain a lot of foreign speech with voice-over translations, adding more material with background speech to the corpus.

- **Television news magazines** contain a broad mix of material: prepared reports, often with background music and on-site interviews, planned speech announcements, spontaneous speech and speaker overlap, and background noise from the audience.

Table 1 shows the amount of material for each programme type in the corpus. All recorded material was manually segmented and orthographically annotated with Transcriber (Barras et al., 2001), according to a set of transcription rules defined in advance. Non-speech, unintelligible speech, and cross-talk segments were labelled with special markers. The resulting annotations were then labelled with speech type, noise type, dialect usage, and speaker name. Figure 1 shows the steps of the annotation process and the possible labels for each category of annotation.

## 3. Corpus Annotation

In the first phase of the annotation process, the corpus was segmented and transcribed by a professional typist. The program Transcriber used for this purpose provides a graphical interface that allows users to view and listen to the audio signal in question, create and move boundaries, and transcribe the resulting segments. The output file consists of an XML-structure containing points in time for the segment boundaries as well as the transcribed utterances. During transcription boundaries were placed at prompted speech pauses, speaker change and changes from and to different types of non-speech. Separate segments were created for speech pauses with a minimum length of 0.5 seconds as well as for periods of untranscribable speech. Maximum length of speech segments was set at 20 seconds.
Segments not orthographically transcribed were marked as

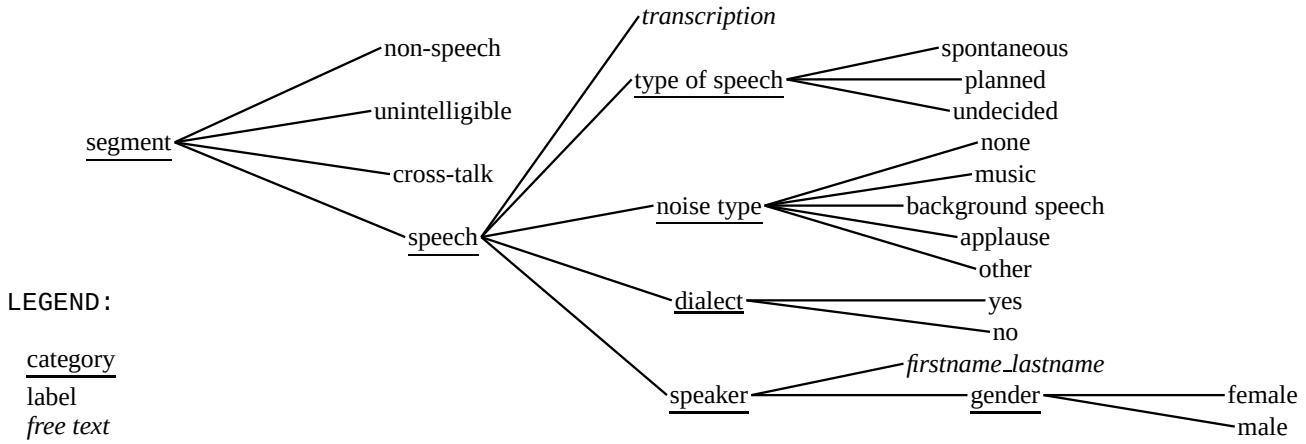
- **non-speech** if they consisted of silence, mere background noise or music only.

- **unintelligible** if they contained foreign speech or speech that was otherwise indiscernible or unintelligible to the transcriber.

- **cross-talk** if two or more speakers were talking simultaneously in conversation so that no main speaker could be identified.

In order to avoid inconsistencies, we created guidelines to determine how the transcriber should deal with phenomena such as numbers, compound words, hesitations, contractions, and mispronounced words. As the transcriptions had to model an ideal speech recognition output and conform to the conventions used in the pronunciation dictionary, no distinction was made between word pronunciation variations due to dialect or speech style. Colloquial expressions or contractions were only included in the transcriptions if they were listed in the pronunciation dictionary. In general, the transcriptions contained little information below the level of standardised words. Speech disfluencies, however, were transcribed as follows: A single tag was used to indicate hesitations such as "um", "err" or "hmm". For stuttered, mispronounced, or cut-off words, the intended word was transcribed and marked with an asterisk in order to allow further analysis of these speech items. Repeated whole words were also repeated in the transcription.
In the second step of the annotation process, the transcribed segments were manually labelled according to speaker name and various aspects that tend to influence speech and speaker recognition performance. A specifically developed annotation tool, DIVE, was used for this task, which allows users to watch and listen to MP4-recordings and annotate them on different levels or tiers. As output, an XML-file following the MPEG7 standard is produced, which enumerates for each tier separately the starting times, durations,

Figure 1: Steps of transcription and annotation. Multiple labels are possible as long as attributes are not contradicting.




and labels of all created segments. In this case, four tiers were used, each having a particular set of possible attributes with which the data could be classified. The various annotation levels were distributed among different annotators, allowing them to concentrate on one aspect only.

- **Type of Speech:** Here the possible labels were "planned" and "spontaneous". Only read or thoroughly prepared speech segments were classified as planned; others were tagged as "spontaneous". For unclear cases no label was given.
- **Noise Type:** In this tier, background noise was labelled and subdivided into the categories "music", "applause", "background speech" and "other". Segments with no background noise were left unlabelled. Because of the heterogeneous nature of background noise, it was possible and often necessary to apply more than one label to a speech segment.
- **Dialect:** The decision here was binary. If segments were characterized by regional or foreign accents they were marked as "dialect"; otherwise they were not labelled. Dialect in this context was defined as an accent sufficiently distinct to give an indication of the speaker's place of origin.
- **Speaker Names:** Prominent speakers and other speakers known by name were identified in the annotations by forename and surname. Unknown speakers were labelled with a combination of the name and date of the program in which they appeared and numbered consecutively. Speaker gender was also encoded in this tier.

The terms used for the labels were chosen to represent the desired categories as closely as possible while at the same time remaining nontechnical and easily understandable by the annotators. Nonetheless, it proved difficult to avoid inconsistencies. Annotators were biased in their decisions by varying contexts as well as their own experience. For this reason, the annotation process was iterative. In order do define the labelling categories more clearly, we analysed a subset of the annotations and refined the annotation rules detailing the type of data the categories should include. Where expedient, the annotations were then revised according to the new guidelines.

## 4. Evaluation Plan

We prepared an evaluation plan for several interesting analysis tasks in order to standardise future evaluations on the new corpus. This will enable the comparison of different approaches to the same problem using the exact same data and task setup. As a starting point, we identified three areas of audio processing where DiSCo should serve as an evaluation corpus, which are defined as follows:

- Automatic speech recognition (ASR) is the task of transcribing each segment with the label speech, either producing a standard word transcription or sub-word units (such as syllables or phonemes).
- Spoken term detection (STD) is the task of searching for a written query in a set of spoken utterances. Spoken term detection is closely connected to automatic speech recognition.
- Speaker recognition (SR) predicts the name of the speaker of a given speech segment using a finite set of pre-defined speaker models. The system should detect whether the speaker is known or not, i.e., the set of possible speakers is not closed.

### 4.1. ASR and STD

We extracted data subsets for evaluating automatic speech recognition and spoken term detection as shown in Table 2. We started with a set containing only planned speech, without any additional background noise and no dialect speech. Then, we isolated the individual challenges by selecting subsets with planned speech and only one of the additional attributes. As individual challenges we chose background music, background speech, background applause, and dialect speech. In addition, we defined a subset containing only spontaneous speech without any additional attributes. Finally, we included the set of all speech utterances in the corpus.

Table 2: ASR and STD evaluation data.

| Attributes | Speech Material (hh:mm) | Utterances | Query Occurr. |
|---|---|---|---|
| planned, none | 00:55 | 1364 | 268 |
| planned, music | 01:11 | 1789 | 317 |
| planned, bg. speech | 00:29 | 727 | 89 |
| planned, applause | 00:06 | 115 | 41 |
| planned, dialect | 00:13 | 318 | 54 |
| spontan., none | 01:55 | 2861 | 425 |
| all speech segments | 11:44 | 17152 | 2736 |

For evaluating the retrieval approaches, a query set was designed semi-automatically from the manual reference transcriptions. First, we applied the Term Selection Tool provided by NIST[1], which automatically extracted a set of queries from the training corpus. In order to augment the training set with more real-world queries, five individuals were asked to manually select ten queries from the transcription text. These 50 queries were added to the automatically generated list, yielding a total set of 501 queries. Table 2 shows the distribution of the query occurrences among the data sets.

### 4.2. Speaker Recognition

For the speaker recognition evaluation, we compiled different evaluation sets for assessing the effect of typical broadcast background noise on speaker recognition performance, one for each noise category.

There are 1073 speakers in the corpus (772 male, 301 female), with an average of approximately 1 minute of material (clean *and* noisy, also including crosstalk) per speaker. The apparent gender imbalance is typical for broadcast material, thus we did not try to alter the male/female-ratio in the test set. The amount of material available is unevenly distributed, with only 183 speakers actually having a minute or more. The others are, for example, unnamed speakers from short street interviews, occuring only for a few seconds in one broadcast.

For speaker recognition evaluation, we selected a set of 70 (53 male, 17 female) test speakers which had sufficient clean (at least 30 seconds used for training) and noisy material. Many of these speakers are well-known politicians, anchorpeople, journalists, or celebrities. 897 of the other speakers were used as impostors. From each of the test speakers, we reserved at least 30 seconds of clean data for training the speaker models (see section 5.2.). From the remaining data we selected the test material including clean and noise segments belonging to only one noise category using at most 5 segments per category and speaker. Crosstalk segments were not used for the evaluation. The respective sets shown in Table 3 include both planned and spontaneous speech. The average segment length is 2.3 seconds.

[1] http://www.itl.nist.gov/iad/mig/tests/std/tools/

Table 3: Speaker recognition evaluation data.

| Noise categories | # of Test Segments |
|---|---|
| none | 769 |
| music | 426 |
| background speech | 252 |
| applause | 120 |
| other | 886 |

Table 4: Word error rate (WER) and spoken term detection (STD) performance on the individual subsets.

| Data Set | WER | Precision | Recall |
|---|---|---|---|
| planned, none | 26.4 % | 0.96 | 0.85 |
| planned, music | 32.7 % | 0.93 | 0.74 |
| planned, bg. speech | 32.6 % | 0.89 | 0.75 |
| planned, applause | 63.5 % | 0.88 | 0.34 |
| planned, dialect | 51.2 % | 0.82 | 0.59 |
| spontan., none | 33.5 % | 0.92 | 0.79 |
| All speech segments | 38.5 % | 0.90 | 0.72 |

## 5. Experimental Results

Based on the evaluation plan presented in Section 4. we evaluated the three tasks of ASR, STD, and SR as described below. The results give information about the challenge each task has to meet for the defined evaluation categories.

### 5.1. ASR and STD

We applied our large vocabulary German ASR decoder as described in (Schneider et al., 2008), using a vocabulary of 200k words. With the given large vocabulary, the OOV rate was 1.4%. For STD, we performed various word- and subword-based approaches as described in (Mertens and Schneider, 2009) and (Mertens et al., 2009).

Table 4 shows the ASR results on the individual subsets, as well as the results for word-based STD on the 1-best transcript. The best performance for ASR and STD is obtained for planned speech without any dialect and background noise. The considered sources of degradation of WER and STD in this evaluation fall into two categories: additive noise (background speech, music, applause) and speaker induced mismatch (dialect, spontaneity). As the acoustic models and language models are mainly trained on planned speech of standard German, both speaker induced sources of mismatch yield to a decreasing performance for ASR and STD.

### 5.2. Speaker Recognition

Experiments with a spectral speaker recognition system, similar to the one in (Reynolds et al., 2000), were carried out on the speaker recognition evaluation set of the corpus. The tests were done with a 512-mixture Gaussian Mixture Model (GMM) system with a Universal Background Model (UBM). We used MFCCs with energy, deltas, and deltadeltas as features, which were normalised with cepstral mean substraction. The background model was a combination of two 256-mixture gender-dependent background models, trained with 1 hour of male and 1/2 hour of female speech, respectively, taken from held-out material. The 70

speaker models were derived from the UBM with maximum a posteriori (MAP) adaptation with at least 30 seconds of clean speech training material per speaker. The evaluation was done on the test segments described in Section 4.2..

Figure 2: DET-curves for the noise categories.

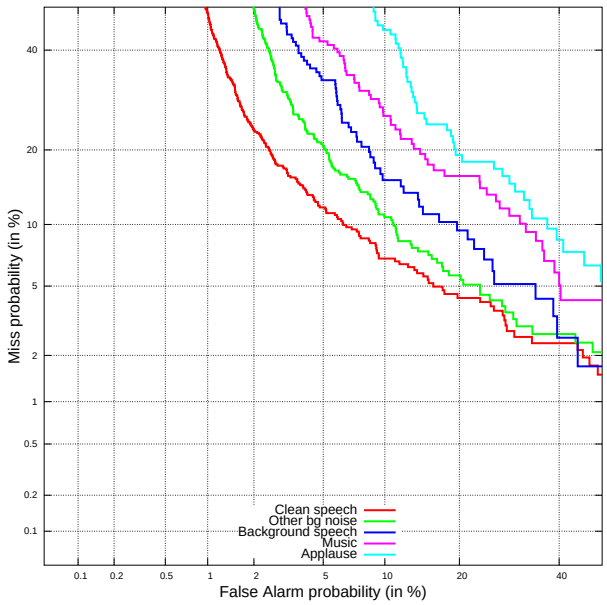



Table 5: Equal error rates for the noise categories (speaker models trained on clean data).

| Test Material Noise Categories | EER |
|---|---|
| none | 8.4 % |
| music | 16.8 % |
| background speech | 13.7 % |
| applause | 19.3 % |
| other | 10.7 % |

The equal error rates (EERs) and detection error trade-off curves (DET curves) given in Table 5 and Figure 2, respectively, show how typical broadcast noise decreases speaker recognition performance. Music and applause appear to be the greatest challenges, as they have the highest EERs (16.8% and 19.3%). Of those two, music is the more prevalent problem: Although the degradation is not as strong, there is a lot more broadcast material with music in the background than with applause (cf. Table 2).

## 6. Conclusion

A new speech corpus with a heterogeneous set of broadcast data is presented. The corpus design is focused on covering a variety of different serious programmes in German including various typical and challenging conditions for speech analysis. Typical background noises, cross-talk situations, spontaneous speech, and dialect speech are covered. We evaluated the effect of the major conditions on the task of automatic speech recognition, spoken term detection, and speaker recognition. The evaluation demonstrates that the presented corpus is very valuable for research and development in applications of speech and speaker recognition for broadcast programmes beyond the rather controlled conditions of broadcast news with planned speech.

## 7. Acknowledgment

The work presented here was funded by the German Federal Ministry of Economy and Technology (BMWi) under the THESEUS project[2].

## 8. References

C. Barras, E. Geoffrois, Z. Wu, and M. Liberman. 2001. 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):5–22.

S. Galliano, E. Geoffrois, G. Gravier, J.-F. Bonastre, D. Mostefa, and K. Choukria. 2006. 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).*, pages 315–320.

M. Grimm, K. Kroschel, and S. Narayanan. 2008. The Vera am Mittag German audio-visual emotional speech database. In *IEEE International Conference on Multimedia and Expo 2008*, pages 865–868.

Robert Hecht, Jürgen Riedler, and Gerhard Backfried. 2002. German broadcast news transcription. In *7th International Conference on Spoken Language Processing (INTERSPEECH 2002)*, pages 1753–1756.

Timo Mertens and Daniel Schneider. 2009. Efficient subword lattice retrieval for german spoken term detection. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2009*, pages 4885–4888. IEEE, April.

Timo Mertens, Daniel Schneider, and Joachim Köhler. 2009. Merging search spaces for subword spoken term detection. In *Interspeech 2009*.

NIST. 1999. The 1999 nist evaluation plan for recognition of broadcast news.

NIST. 2003. The rich transcription spring 2003 (rt-03s) evaluation plan.

Douglas A. Reynolds, Thomas F. Quatieri, and Robert B. Dunn. 2000. Speaker verification using adapted gaussian mixture models. In *Digital Signal Processing*, volume 10, pages 19–41.

Daniel Schneider, Jochen Schon, and Stefan Eickeler. 2008. Towards large scale vocabulary independent spoken term detection: Advances in the Fraunhofer IAIS Audiomining System. In J. Köhler, M. Larson, F.M.G. Jong de, W. Kraaij, and R.J.F. Ordelman, editors, *Proceedings of the ACM SIGIR Workshop "Searching Spontaneous Conversational Speech" held at SIGIR '08*, Singapore, July.

[2]http://www.theseus-programm.de/en-US/home/default.aspx