A Semi-Automatic Workflow for Transcribing and Annotating Broadcast News

Christoph Draxler, Sven Grawunder, Jürgen Trouvain, Felicitas Kleber
Institute of Phonetics and Speech Processing, LMU Munich; Halle University; Max Planck Institute for Evolutionary Anthropology, Leipzig; Saarland University, Saarbrücken
draxler@phonetik.uni-muenchen.de, grawunder@sprechwiss.uni-halle.de, {trouvain, kleber}@lst.uni-saarland.de

Abstract

Audio data archived in radio broadcast stations represent a rich source for various research purposes from phonetic questions up to training and test data for speech modelling. We present an efficient semi-automatic workflow for pre-processing, transcribing and analysing large linguistic-phonetic audio corpora. As a pilot study, we process radio broadcast news from a German public radio station containing recordings from 1956 until 2017. The workflow consists of basic preprocessing, automatic speech recognition, manual word correction, automatic generation of pairs of audio chunks and transcripts, plus an automatic word-, syllable- and phoneme-level segmentation of these chunks. The workflow is organised using the Octra Backend management tool, manual validation and correction of transcripts and chunking are performed using the Octra editor, and the BAS web services perform the segmentation. In an example analysis we show with our specific radio corpus how to use it for comparative longitudinal structure analyses of broadcast news, and for text- and signal-based studies on changes of speech and articulation rate.

Keywords: broadcast news, workflow, automatic speech processing, prosody, speech rate

1. Introduction

A large fraction of audio data in archives worldwide probably consists of speech signals which are neither transcribed nor annotated. Thus, despite their rich potential in terms of content, they cannot be easily accessed for research purposes. However, the creation of a corpus of broadcast news from public radio stations is promising: public broadcast data can be collected, transcribed, and stored under fair use conditions for research purposes.

Also, because the material was intended for public broadcast, the analysis of the audio does not require the consent of the speakers. Furthermore, the recordings are performed in acoustically clean environments with relatively constant technical configurations, and therefore detailed synchronic and diachronic analyses can be performed on a large variety of phonetic and linguistic levels (Kupietz et al., 2010). Thus, audio recordings archived at radio broadcast stations can be regarded as a rich source for various research purposes.

Creating corpora for phonetic or linguistic research in general is time-consuming and thus expensive. Automatic processing promises and often produces quick results, but may miss out on rare linguistic phenomena or produce erroneous transcripts. Thus, manual validation and correction is needed – it has to be optimised to cope with the large amounts of raw transcripts generated by Automated Speech Recognition (ASR).

In sections 2 and 3 we present an optimised semi-automatic workflow for pre-processing, orthographic transcription, manual validation and chunking plus automatic segmentation for broadcast news as well as results from this workflow.

In section 4 we describe an example with public data as a test case, where we process 49 broadcast news from the years 1956 until 2017. The broadcasts are from the same German radio station (Saarländischer Rundfunk): 32 recordings (years 1956 to 2002 with missing years, random day in each year) stem from the archive of the station itself, 17 recordings (2003 to 2017) are from the 'news ark' collection (Schwiesau et al., 2011). We use this corpus for diachronic structural analysis of radio news, and for text- and signal-based longitudinal analyses of speech and articulation rate.

The compilation of this corpus and the following analysis could be performed, because not only is the data publicly broadcasted, but also publicly funded. German Copyright law act ("Urheberrechtsgesetz" UrhG), in particular section 60d UrhG, permits automatic and systematic reproduction of source material to create a corpus for scientific research purposes (European Commission. Directorate General for Research and Innovation., 2022; Margoni and Kretschmer, 2022).

2. Workflow

For data processing, transcription and later analysis we developed and evaluated a workflow in which as many tasks as possible were automated. This tutorial-like section describes the tasks, tools and scripts used.

The workflow consists of the following tasks:

1. scripted pre-processing of the audio files

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 6408–6417
11–16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

2. automatic generation of an orthographic raw transcript using ASR
3. manual correction of the content segments and the transcript
4. automatic chunking of audio files and transcripts
5. automatic segmentation on word, syllable and phoneme level via BAS web services

The result is a systematically structured corpus, with element-wise chunked pairs of audio and transcripts with word, syllable and phoneme level segmentations in several file formats.

For the administration of the workflow, the Octra Backend (Draxler and Pömp, 2024) was used. This web service, provided by the Bavarian Archive for Speech Signals at LMU Munich, is available for free to academic users; a standalone version is currently under development (Pömp and Draxler).

2.1. Pre-processing the audio files

The original names of the audio files were unsystematic, and they contained characters that prohibit automatic processing via web services, e. g. blanks or non-ASCII characters. A simple python script was used to replace these characters systematically. For comparability with the original data, the overall name was not changed.

Some audio files were already in wav format. Files in mp3 format were converted to wav via ffmpeg. As a result, we obtained 49 audio files with sample rates of 48 and 44.1 kHz, with 16 bit linear quantisation and mono. The total corpus size is approx. 1.2 GB and the total duration is 03:36:54, with approx. 04:25 per broadcast.

2.2. Generation of an orthographic raw transcript

The following command was used to automatically transcribe the recordings with whisperX (Bain et al., 2023) (in Version 3.1.1) with the language-independent large-v3 model:

```
whisperx --model large-v3 --lang de
--compute_type float32 --output_dir
results *.wav
```

The result of the ASR was an orthographic transcript for every audio file, in three formats:

`.txt` plain text transcript without timestamps

`.srt` multi-line orthographic transcript with time-based segments

`.json` transcript with hierarchical segmentation and word-wise recognition scores

See Figure 1 for an example of the latter two.

2.3. Manual validation and correction

ASR generates a segmented text transcript with punctuation. For further analysis, the transcript was then resegmented according to content criteria using the the web-based transcription editor Octra (Pömp and Draxler, 2017). Specifically, segment boundaries were adjusted, markers added and text corrected. Additionally, segments were classified into news elements in the same editor.

Segment classification is based on so-called *broadcast (or news) elements* (Table 2), which structure the broadcasts and are often characterised by specific speaking styles (Iivonen et al., 1995; Kern and Trouvain, 2018; Trouvain, 2015; Mok et al., 2014; Savino et al., 2024; de Mareüil et al., 2012).

In view of the planned analyses, transcribers were instructed to place the right segment boundary as close as possible to the end of the visible and audible speech in the signal. This ensures that segment boundaries do not split pauses and thus makes pause duration measurements reliable (cf. Figure 2 (bottom) for examples).

2.4. Transcription guidelines

The transcription guidelines are optimised for rapid and consistent transcription.

- Verbatim transcription in standard orthography, with capitalization according to word type, not its position in the text
- Punctuation as provided by ASR
- Acronyms as words in upper case if spoken as words, e. g. 'EON', in isolated capital letters otherwise, e. g. 'R W E'
- Expansion of digits and numbers to text, e. g. spoken '3' written as 'drei' (*'three'*); numbers higher than hundred are split into thousands, hundreds, etc. For example, the year 1972 is written as 'neunzehn hundert zweiundsiebzig' (*'nineteen hundred seventy two'*)
- Codes for broadcast elements are placed in angled brackets $<$ (e.g. $< COR >$), there is only one code per element, and it is always in initial position

Octra supports project-specific verification of transcripts. This ensures that only formally correct transcripts are stored.

To manage the workflow, a project was set up in the Octra Backend (Draxler and Pömp, 2024) and so-called configurations were created. A *configuration* defines the type of a task. It specifies, for example, the tool used, available keyboard shortcuts, and what formally correct transcripts look like.

```
# .srt format
6
00:04:35,549 --> 00:04:39,271
Saarländische Rundfunk, Nachrichtenredaktion Monika Seel.

# json format
{
 "start": 275.549,
 "end": 279.271,
 "text": " Saarländische Rundfunk, Nachrichtenredaktion Monika Seel.",
 "words": [
  {"word": "Saarländische", "start": 275.549, "end": 275.989, "score": 0.748},
  {"word": "Rundfunk,", "start": 276.009, "end": 276.509, "score": 0.731},
  {"word": "Nachrichtenredaktion", "start": 276.549, "end": 277.77, "score": 0.878},
  {"word": "Monika", "start": 277.81, "end": 278.311, "score": 0.923},
  {"word": "Seel.", "start": 278.351, "end": 279.271, "score": 0.737}
 ]
}
```

Figure 1: Final segments of the broadcast `NWV21_02.10.1990` in ASR-generated transcripts. Note the different content: The `.srt` format contains segment number, segment boundaries and transcript text (top), the `.json` format also contains a recognition score for each word (bottom)

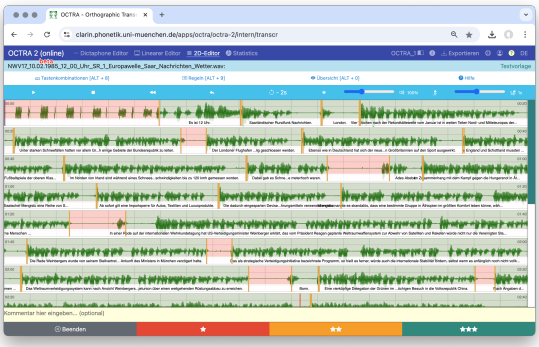

Figure 2: Automatically generated raw transcript in the 2D editor in Octra. Green transcription units contain transcript text, red ones are empty. Boundaries were set by whisperX. Some of these boundaries are so close to each other that they are visible only after zooming in (cf. Figure 5).

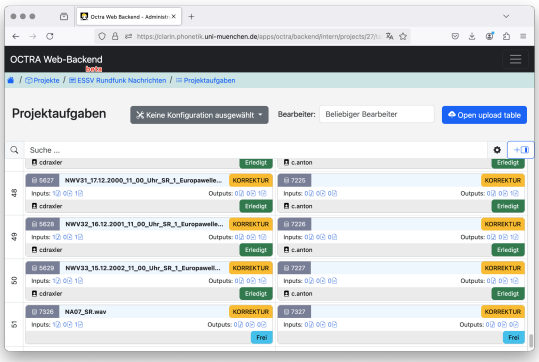

Figure 3: Overview of the available tasks with status information. A row represents a `task`, which may consist of consecutive subtasks.

A specific task consists of a pair of audio and segment file. The task window in the Octra Backend shows the current processing status of a project. Tasks are arranged in a list, and their respective status (e. g. *free*, *in progress*, *completed*) is immediately visible (Figure 3).

In the current project, a second correction process was defined because the transcription guidelines had changed since the first pass. In the Octra Backend, this second correction was defined as a task linked to the first correction, with the output of the first correction being the input for the second. In the list of jobs, such a link between tasks is indicated by several tasks in one line (Figure 3).

After the manual correction was completed, the project was exported. The export process creates a systematically structured project directory named with the project name and the current date. This contains the output, result, metadata, and documentation files in separate subdirectories.

2.5. Chunking the audio files

Octra features a tool to cut audio files according to the segment boundaries. This tool generates an edit decision list containing the signal address in timestamps and samplepoints, so that not only the audio file currently held in the editor, but also original source files can be cut (Table 1).

For cutting the audio files, a separate project was defined in the Octra Backend, containing pairs of cut audio files and the manually corrected transcripts.

2.6. Automatic segmentation via BAS web services

The next workflow task is performed outside the Octra Backend. It uses pairs of audio and transcript text files.

For the analysis, the signal files are segmented on the word, syllable and phoneme level using a pipeline web service provided by BAS (Kisler et al., 2012), (Schiel et al.): the pipeline 'G2P → MAUS →

Fragment	Source	Start (s)	Duration (s)	Transcript
NA03_SR_1_0001.wav	NA03_SR_1.wav	0.000	6.000	PAC SR eins ...
NA03_SR_1_0002.wav	NA03_SR_1.wav	6.000	8.898	TOP Anschlussfinanzierung ...
...	...	...	...	...
NA03_SR_1_0021.wav	NA03_SR_1.wav	235.865	5.108	OUT soweit die ...

Table 1: Edit decision list of an audio file (edited for readability).

PHO2SYL' generates a segmentation in the specified output format.

BAS web services may be called via a graphical user interface or via an API-call from a script. The latter allows automating the process. Because of the communication overhead, it is slower than the batch processing available in the GUI, but the advantage of automatic processing by far outweighs this disadvantage.

In the current project, a python script first generates for every audio file a matching transcript file according to the edit decision list. Then it calls the web service with this file pair and saves the result on the local computer. Figure 4 shows the python code for the API call.

The pipeline returns files in the BAS Partitur Format. These files were then — also automatically via a script – converted to Praat TextGrid (Boersma, 2001) and `annot.json` for the EmuSDMS (Winkelmann et al., 2017) via the web service AnnotConv.[1]

3. Results

After transcription and export the audio and transcript files are held in a single folder on the local computer. Additionally, a distinct folder for every broadcast contains the cut audio signals plus the corresponding edit decision list. The project folder 1) serves as a backup, and 2) contains the transcripts in `.txt` format which are used to compute the word error rate of the ASR.

3.1. Word error rate

To compute the word error rate (WER), the automatically generated and the manually corrected transcripts need to be normalized: remove the broadcast element codes, delete punctuation and special characters, and convert to lower case. This was performed by a python script.

For the computation of the WER, the `jiwer` library (Vaessen, 2018) was used.

[1]To generate a full Emu database from the `annot.json` files, a database configuration file has to be created first. This can be achieved by processing one file pair via the graphical user interface, saving the result as a zip archive file and copying the resulting configuration file in the folder that contains all other database folders.

Element	Description	WER	#files
COR	core news	5.43%	291
REP	report	5.67%	49
TOP	topic	5.90%	32
SPO	sports	9.17%	6
WET	weather	10.99%	39
TRA	traffic	13.51%	25
LOC	location	21.51%	168
INT	intro	31.94%	114
TIM	time	35.65%	68
OUT	outro	42.32%	56
PAC	jingle	53.57%	105

Table 2: Broadcast elements as defined by Grawunder (2011) sorted by their respective WER

Given the high acoustic quality and the standard pronunciation, some of the WER values seem surprisingly high. However, they can be explained by the special properties of broadcast news: many numerical expressions (which are expressed using digits in ASR output, and as text in manual transcripts), acronyms, or proper names. Location and person names were recognized quite well (including names of historical or politically important places, e. g. 'Douaumont', 'San Son'), but names of broadcast stations or with specific writing conventions which occur frequently in this corpus rather badly (e. g. 'Studiowelle Saar' as 'Studio Velesa' or 'Studio Velesar', 'Saar-Lor-Lux' as 'Saarlouis-Lux'). Furthermore, and this is especially true for TIM, OUT and PAC elements, some broadcast elements contain a lot of non-speech content such as chimes, jingles, etc.

3.2. Segment boundaries

The segment boundaries proposed by whisperX had to be manually corrected. Very often, boundaries were very close to each other, and not in speech pauses, but somewhere in a word (cf. Figure 5 top).

One of the goals of the subsequent analysis was to measure pauses, both between and within broadcast elements. Transcribers were thus instructed to move the left boundary of a broadcast element to the beginning of a pause between elements, and to remove all internal boundaries until the left boundary of the next element (cf. Figure 5 bottom).

```
import os
import requests
import xml.etree.ElementTree as ET

def callPipeline (audiopath, textpath, language, outformat, pipeline):
  (_, audiofile) = os.path.split(audiopath)
  (_, textfile) = os.path.split(textpath)

  url = "https://clarin.phonetik.uni-muenchen.de/BASWebServices/services/runPipeline"
  formdata = {
    "SIGNAL": (audiofile, open(audiopath, "rb"), "audio/x-wav"),
    "TEXT": (textfile, open(textpath, "r"), "text/txt"),
    "LANGUAGE": (None, language),
    "OUTFORMAT": (None, outformat),
    "PRESEG" : (None, "true"),
    "PIPE" : (None, pipeline)
  }

  result = requests.post(url, files=formdata)
  if result.ok:
    tree = ET.fromstring(result.text)
    downloadLink = tree.find('downloadLink').text
    if downloadLink:
      downloadResult = requests.get(downloadLink)
      if downloadResult.ok:
        return downloadResult.text

  return None
```

Figure 4: Python code to call the pipeline web service G2P→MAUS→PHO2SYL. The service returns the segmentation text in the specified output format, e. g. TextGrid, csv or emuDB

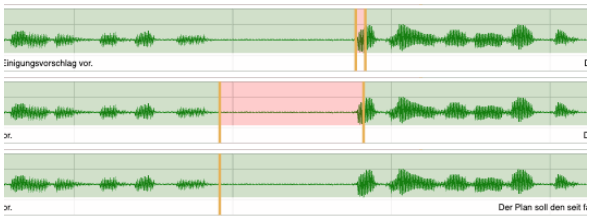

Figure 5: Problematic placement of segment boundaries by ASR: very close to each other (top). Manual correction moved the left boundary to the beginning of the pause (middle) and removed the right boundary (bottom).

3.3. Workflow summary

At the end of the workflow, the 49 original broadcast news were split into 953 broadcast element sized audio snippets with corresponding transcript and automatically generated segmentations.

Exporting the project from Octra Backend generated a systematic file structure on the local computer. A single project directory contains all 49 original recordings and their ASR-generated transcripts. For each broadcast, a separate folder is created to hold the chunked audio and corresponding transcript files of the elements of this broadcast. Furthermore, the edit decision list for each broadcast is stored in the broadcast's folder.

The script that calls the automatic segmentation pipeline web service traverses this project directory, selects matching pairs of audio and transcript files, and stores the resulting segmentation files in the same directory as the audio files.

4. Example of a corpus analysis

4.1. Introduction

Below we present a condensed and slightly modified version of a sample analysis of diachronic prosodic change in speech tempo. For a detailed version in German, see Kleber et al. (2026) where change was examined directly over a 60-year period. Here, we analyse the decades of that period separately (see 4.3.2).

4.1.1. News as spoken genre

News as a spoken genre appears rather recently compared to other genres such as poetry recital, sermons, lectures or parliamentary speeches. Newsreading is in use just for about 100 years when radio broadcasting started to become a mass medium.

Typical features of news speech are that it is monological and read (pre-planned) speech with a quite complex syntax and a correct grammar whereas sentence modes like questions are missing (Iivonen et al., 1995). Normally, news announcers articulate very clearly while they have a 'neutral' tone of voice in terms of their attitudinal expression. As text *and* in its spoken form, news can be considered 'well-formed'. Thus, newsreadings show stylistic features that are in stark contrast to spontaneous speech in conversations.

News as it occurs in radio or TV broadcasts is not only a concatenation of news items but usually has a composed structure consisting of elements such as news bulletins, reports, sports news, business items, weather reports, and traffic announcements.

The mix of various textual elements requires a discourse structure that is also reflected in 'paragraph intonation'.

4.1.2. Prosodic characteristics of professional speakers

The speaking style of news announcers is a prime example of *clear speech*, which is generally more oriented to hyper- than to hypospeech (Lindblom, 1990). Usually, news announcers received a speech training specialized for the required genres in radio broadcast. Typically, they focus on fluency, ease of comprehension, clarity of vocal performance, and standard pronunciation. News announcers can also reach a status for a good standard pronunciation, as it was for example 'BBC pronunciation' for British English (Ladefoged, 2001). That means that news announcers often act as role models reflecting the *current* standard pronunciation.

Regarding pausing, Trouvain et al. (2020) found for German radio news announcers that they produced on average 16 pauses per minute speaking time (low value compared to other genres) of which more than 80% were pauses with an audible inhalation noise (more than in other genres). As for the difference in speaking rate between elements, Grawunder and Kettel (2015) found basically no difference between German newsreading of core news and reports (both 4.3-4.4 syllables per second, henceforth s/s) (see Table 2) but a drop (3.3 s/s) for the transitional intro element in between.

4.1.3. Changes over time

Speech in general is subject to change over time. Sound change is usually studied as deviating pronunciations at the segmental level over longer periods spanning few (sound change in progress) to many generations (completed sound change). However, sound change can also affect prosodic characteristics. Particularly, changes in progress can be expected within a spoken genre situated in a new communicative setting such as radio broadcast, with the new task of transporting best spoken information from one (speaker) to many (listeners). Those changes over time in broadcast are documented for other radio genres such as sports live commentaries (Trouvain, 2015). We thus can also expect changes regarding prosodic features within the genre of radio news over a period of 60 years.

De Mareüil et al. (2012) found in a 50-years period of audio-visual news in France (from 1945 to 1997) that a number of prosodic features have decreased since the 1940s: average pitch (actually fundamental frequency), pitch rise associated with initial stress, vowel duration characterizing an emphatic initial stress, duration of inter-pause intervals while speech rate has not changed.

In a diachronic study, Savino et al. (Savino et al., 2024) compared the prosodic style of Italian newsreading originally recorded in the late 1960s with the productions of two newscasters reading the same texts in 2005. Results show that the intonation styles in terms of time-varying pitch, in terms of slope changes (wiggliness) and pitch excursions at largest rises and falls (spaciousness) were similar across the two eras. However, in the modern era average pitch was higher and speech rate was faster, mainly caused by reducing the number and duration of pauses.

The reported studies give evidence for rather heterogeneous observations. Thus, it can neither be expected that the changes of this spoken genre behave in a certain direction, e.g. shorter pauses in recent times, nor that the developments take place in a gradual way, e.g. observing continuously shorter pauses.

4.2. Methods

For the distribution of broadcast elements we looked at their order of appearance during selected years and their proportional duration.

Measuring speech tempo can entail two types: speech rate and articulation rate. Both, speech rate (including pauses) and articulation rate (excluding pauses) are usually measured as syllables per second (s/s). We tested two different (semi-)automatic approaches of measuring both speech and articulation rate: a signal-based method and a transcription-based method. It is important to note that in neither of both methods the syllables necessarily correspond to canonical syllables, although in manual approaches the usual way is to assume canonical syllables based on the written form of the spoken words.

The signal-based type of tempo measurement is based on the pure acoustic signal by means of the very popular Praat script by de Jong et al.[2] (de Jong and Wempe, 2009; de Jong et al., 2021). The script detects assumed voicing by periodicity analysis and detects intensity peaks 'preceded by dips in intensity', which are considered as potential syllable nuclei. The script subsequently deletes peaks that are not voiced. Thus, low intensity peaks are sometimes missed, whereas high intensity passages sometimes receive assignment of double peaks.

To cross-validate the signal-based assessment of speaking rates and the above workflow, we additionally measured speech rate and pauses based on the MAUS segmentations into phonemes and syllables. To this end, we queried the EMU speech

[2]https://github.com/FieldDB/Praat-Scripts/blob/main/praat-script-syllable-nuclei-v2file.praat

database for various labels (e.g. text elements, pauses, syllables). While this approach can be considered annotation-based, it is to be noted that the MAUS algorithms used for segmentation and labeling are also based on acoustic features extracted from the signal.

4.3. Results

4.3.1. Distribution of news elements

We first assessed the distribution of all 953 broadcast elements detected within our corpus.

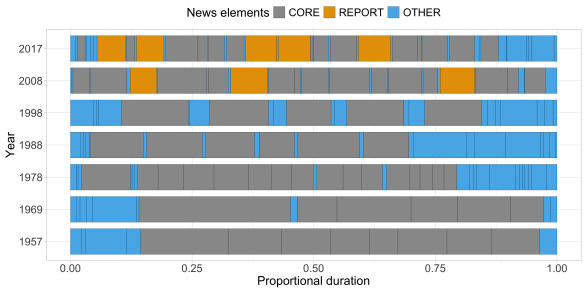

Figure 6: Proportional duration of news items arranged by occurrence within selected broadcasts.

Figure 6 shows the proportional duration of news elements within a sample of news broadcasting years that is representative of the entire corpus and all decades. Each year represents exactly one news broadcast. The single items are grouped according to their appearance within the news broadcast (separated by vertical lines). The category OTHER includes all elements listed in Table 2 except core news and reports. While the proportional duration of the individual core news has decreased over time, their number has increased. Like many changes over time, this development does not appear to have been linear. The duration of broadcasts varied non-linearily, too, from 2 min 17 s (1991) to 6 min 45 s (1979). Moreover, reports represent a recent development in this sample, while core news used to be presented en bloc at the end of a broadcast in earlier years.

4.3.2. Tempo-based analyses of core news

The prosodic analyses focus on the subset of all 291 core news, where the speaker remains the same within one broadcast. Results are presented in periods of decades where the first 'decade' actually spans twenty years (1956-1975) for reasons to do with sparse data within this period and comparability between decades. Figure 7 shows speech rate derived from the signal-based and the MAUS-based analysis as a function of decade. Both analyses indicate a gradual, but non-linear increase in speech rate from the oldest to the newest recordings as indicated by weak positive correlations in some but not all decades (Pearson's r in signal-based analyses: r = .25 (1957-1975), r = .05 (1976-1985), r = .02 (1986-1995), r = .04 (1996-2005), r = .00 (2006-2017); Pearson's r in MAUS-based analyses: r = .11 (1957-1975), r = .16 (1976-1985), r = .15 (1986-1995), r = .15 (1996-2005), r = .00 (2006-2017). Specifically, periods of change in speech rate are followed by periods of stability. The trajectories of change as a function of decade are overall similar between the two analyses despite speech rate being generally higher in the MAUS-based (ranging from 4 to 6.1) compared to the signal-based (ranging from 3.43 to 5.55) approach.

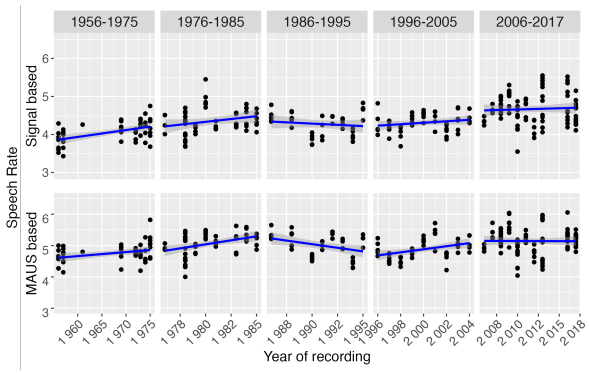

Figure 7: Speech rate (s/s) as a function of year of recording split by decade.

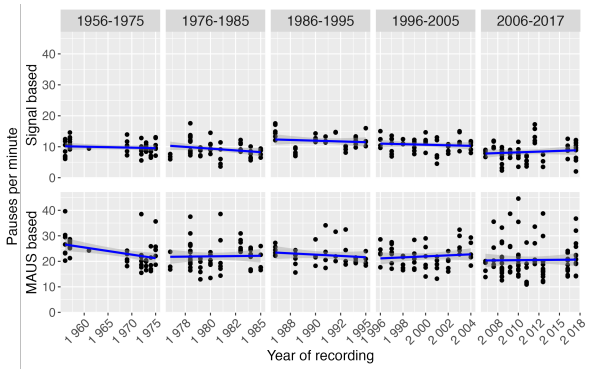

Figure 8: Number of pauses normalized for duration of COR element as a function of recording year split by decade.

Similarly, and commensurate with Figure 8, the number of pauses per minute is overall lower in the signal-based compared to the MAUS-based approach. As opposed to speech rate, however, no meaningful correlation between pause per minute and year of recording was found, regardless of decade or underlying analysis.

Given the overall similarity between the results of the two analyses, we will focus on either one of

them in the remainder of this section.

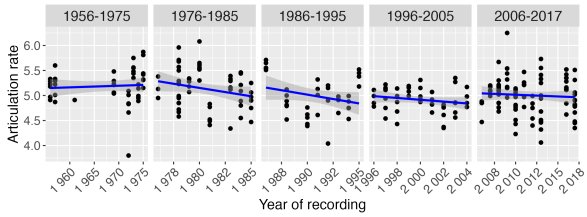

Figure 9: Articulation rate (s/s) as a function of year of recording split by decade.

Signal-based articulation rates, derived from the signal-based analysis, range between a minimum of 3.7 and around a mean of 4.99 s/s (median 5.02 s/s and sd 0.537 s/s). These rates remain more or less stable over the decades with a slightly higher start of 5.15 s/s in the earliest period (1956-1975), followed by 5.10 s/s (1976-1985), then a small decrease of 4.99 s/s (1986-1995) and 4.77 s/s (1996-2005) and finally 4.99 s/s (1996-2017), cf. Figure 9. None of the correlations were, however, meaningful.

4.4. Discussion

Although the current sample of 49 broadcast news with 291 core news elements across nearly 60 years certainly allows to hypothesize about the reported trends of speaking and pause rate, the individual year sample may be driven by individual speakers. The measured speaking rates and articulation rates, however, correspond overall to those previously found for German news (Grawunder, 2011; Pellegrino et al., 2011; Tivadar, 2017).

In line with previous reports on Italian (Savino et al., 2024), the signal- and MAUS-based analyses of the core news showed that speech rate (including pauses) but not articulation rate (excluding pauses) increased over time, possibly due to fewer and shorter pauses. Despite the overall clear trend, analyses by decade suggest that the parameters do not show a constant uniform change, in line with other generational changes (Zellou and Tamminga, 2014). Moreover, diachronic trends based on automatic segmentations can be considered robust indicators of such trends (Kisler and Kleber, 2019).

Pauses play a central role in our prosodic analysis. However, assessing the quality of pause detection remains a task for the future (compare the absolute numbers in Figure 8). This concerns detected pauses that are part of articulation (in stop consonants) but also missed pauses. A manual adjustment of the pauses could also imply the labeling of pauses as containing inhalation noises or not.

5. Summary and outlook

The current project is the first larger and well-documented use case for the Octra Backend transcription management tool. During the course of the project, a new version of the backend was released, which integrates ASR as a background process into the workflow and offers useful new features such as cutting audio files and batch changes, without leaving the backend.

Manual correction cannot by definition be automated, but can be supported by the machine. Classifying chunks with broadcast element codes should be feasible automatically. However, sufficient training data is not yet available. Similarly, automatic recognition of audible inhalation noise would be another step in data preprocessing, as well as reduction and subtraction of music sound beds for better results in pause and syllable detection.

Overall, radio broadcasts represent a good fundament to study diachronic changes over periods longer than half a century and recorded at a high audio quality (Kupietz et al., 2010). Preprocessing of those data for prosodic analysis requires great effort. A massive reduction of manual work, as demonstrated here, is more than welcome.

Acknowledgement

Special thanks go to Camilla Anton for manually correcting the ASR-generated transcripts at very short notice. Work on Octra Backend is partly funded by the German national data infrastructure initiative Text+ (DFG funding no. 460033370). We are also grateful to Saarländischer Rundfunk for providing us the material as mp3 data.

Ethics statement

The data used to compile our corpus stem from professional speakers who worked as news broadcasters for a publicly funded radio station in Germany. German copyright law act permits automatic and systematic reproduction of source material to create a corpus for scientific research purposes (European Commission. Directorate General for Research and Innovation., 2022; Margoni and Kretschmer, 2022).

Limitations of data

The authors are aware of the fact that the data size for some decades and some years are strongly limited and that data used here also represent individual speaker traits that should not be confounded with representative data where multiple speakers are taken into account.

6. Bibliographical References

Max Bain, Jaesung Huh, Tengda Han, and Andrew Zisserman. 2023. WhisperX: Time-accurate speech transcription of long-form audio. In *Proc. Interspeech 2023*, Incheon, Korea.

Paul Boersma. 2001. Praat, a System for doing Phonetics by Computer. *Glot International*, 5(9/10):341–345.

Nivja H. de Jong, Jos Pacilly, and Willemijn Heeren. 2021. PRAAT scripts to measure speed fluency and breakdown fluency in speech automatically. *Assessment in Education: Principles, Policy & Practice*, 28(4):456–476.

Nivja H. de Jong and Ton Wempe. 2009. Praat script to detect syllable nuclei and measure speech rate automatically. *Behavior Research Methods*, 41(2):385–390.

Philippe Boula de Mareüil, Albert Rilliard, and Alexandre Allauzen. 2012. A diachronic study of initial stress and other prosodic features in the French news announcer style: Corpus-based measurements and perceptual experiments. *Language and Speech*, 55(2):263–293.

Christoph Draxler and Julian Pömp. 2024. Octra Backend - eine skalierbare Infrastruktur für Transkriptionsprojekte. In *Proc. ESSV 2024*, pages 102–107, Regensburg.

European Commission. Directorate General for Research and Innovation. 2022. *Study on EU Copyright and Related Rights and Access to and Reuse of Data.* Publications Office, LU.

Sven Grawunder. 2011. Die Erforschung des Sprechens mittels Nachrichtenkorpora: Die Nachrichtenarche der ARD. In Ines Bose and Dietz Schwiesau, editors, *Nachrichten Schreiben, Sprechen, Hören: Forschungen zur Hörverständlichkeit und Radionachrichten*, pages 147–155. Frank & Timme, Berlin.

Sven Grawunder and Sonja Kettel. 2015. Anmoderieren/Überleiten/Antexten. *SPIEL*, 2015(1):151–170.

Antti Iivonen, Tuija Niemi, and Minna Paananen. 1995. Comparison of prosodic characteristics in English, Finnish and German radio and TV newscasts. In *Proc. 13th International Congress of Phonetic Sciences, Stockholm*, pages 382–385.

Friederike Kern and Jürgen Trouvain. 2018. Zur Historie der Inszenierung von Spannung in Fußball-Livereportagen im Radio. *Aptum. Zeitschrift für Sprachkritik und Sprachkultur*, 14(2):101–118.

Thomas Kisler and Felicitas Kleber. 2019. Zur Validität automatisch segmentierter Daten: Eine akustische Analyse der mittelbairischen Lenisierung im Deutsch Heute-Korpus. In Sebastian Kürschner, Mechthild Habermann, and Peter O. Müller, editors, *Methodik moderner Dialektforschung: Erhebung, Aufbereitung und Auswertung von Daten am Beispiel des Oberdeutschen*, number 241-243 in Germanistische Linguistik, pages 279–311. Olms, Hildesheim/Zürich/New York.

Thomas Kisler, Florian Schiel, and Han Sloetjes. 2012. Signal Processing Via Web Services: The Use Case WebMAUS. In *Proc. Digital Humanities*, pages 30–34, Hamburg.

Felicitas Kleber, Christoph Draxler, Jürgen Trouvain, and Sven Grawunder. 2026. Rundfunknachrichten als linguistisch-phonetische ressource für die sprachwandelforschung: Ein halbautomatisiertes verfahren und eine demosprachdatenbank. *IDSopen*, 16:3–18.

Marc Kupietz, Cyril Belica, Holger Keibel, and Andreas Witt. 2010. The German reference corpus DeReKo: A primordial sample for linguistic research. In *Proc. 7th International Conference on Language Resources and Evaluation (LREC)*, Valletta, Malta. European Language Resources Association (ELRA).

Peter Ladefoged. 2001. *Vowels and Consonants – An Introduction to the Sounds of Languages*. Blackwell, Maldon, MA, Oxford.

Björn Lindblom. 1990. Explaining phonetic variation: A sketch of the H&H Theory. In William J. Hardcastle and Alain Marchal, editors, *Speech Production and Speech Modelling*, pages 403–439. Springer Netherlands, Dordrecht.

Thomas Margoni and Martin Kretschmer. 2022. A Deeper Look into the EU Text and Data Mining Exceptions: Harmonisation, Data Ownership, and the Future of Technology. *GRUR International*, 71(8):685–701.

Peggy P.K. Mok, Holly S.H. Fung, and Jingwen Li. 2014. A preliminary study on the prosody of broadcast news in Hong Kong Cantonese. In *Proc. Speech Prosody, Dublin*, pages 1072–1075.

François Pellegrino, Christophe Coupé, and Egidio Marsico. 2011. Across-language perspective on

speech information rate. *Language*, 87(3):539–558.

Julian Pömp and Christoph Draxler. 2017. Octra – A configurable browser-based editor for orthographic transcription. In *Proc. Phonetik und Phonologie*, pages 145–148, Berlin.

Michelina Savino, Simon Wehrle, and Martine Grice. 2024. The prosody of Italian newsreading: a diachronic analysis. In *Proc. Speech Prosody, Leiden*, pages 836–840.

Dietz Schwiesau, Sven Grawunder, and Ines Bose. 2011. Die Nachrichtenarche der ARD. In Ines Bose and Dietz Schwiesau, editors, *Nachrichten Schreiben, Sprechen, Hören: Forschungen Zur Hörverständlichkeit von Radionachrichten*, pages 147–155. Frank & Timme, Berlin.

Hotimir Tivadar. 2017. Speech rate in phonetic-phonological analysis of public speech using the example of political and media speech. *Journal of Linguistics/Jazykovedný casopis*, 68(1):37–56.

Jürgen Trouvain, Raphael Werner, and Bernd Möbius. 2020. An acoustic analysis of inbreath noises in read and spontaneous speech. In *Proceedings of Speech Prosody*, pages 789–793.

Jürgen Trouvain. 2015. Notes on the development of speaking styles over decades — the case of live football commentaries. In *Proc. ISCA-Workshop on the History of Speech Communication Sciences, Dresden*, page 160–166.

Nik Vaessen. 2018. JiWER: Similarity measures for automatic speech recognition evaluation. *Python Package Index*.

Raphael Winkelmann, Jonathan Harrington, and Klaus Jänsch. 2017. Emu-SDMS: Advanced Speech Database Management and Analysis in R. *Computer Speech and Language*.

Georgia Zellou and Meredith Tamminga. 2014. Nasal coarticulation changes over time in Philadelphia english. *Journal of Phonetics*, 47:18–35.

7. Language Resource References

The following language resources require authentication as a member of an academic institution, or a CLARIN account. For commercial use, contact `bas@phonetik.uni-muenchen.de`.

Pömp, Julian and Draxler, Christoph. *Octra Backend Management Tool*. Bavarian Archive for Speech Signals, Institute of Phonetics and Speech Processing, LMU Munich. PID https://doi.org/11022/1009-0000-0008-0081-5.

Schiel, Florian and Jochim, Markus and Draxler, Christoph. *Bavarian Archive for Speech Signals Web Services*. Bavarian Archive for Speech Signals, Institute of Phonetics and Speech Processing, LMU Munich. PID http://hdl.handle.net/11858/00-1779-0000-0028-421B-4.