FT SPEECH: Danish Parliament Speech Corpus

Andreas Kirkedal\(^1,2\*\), Marija Stepanović\(^1\*\), Barbara Plank\(^1\)

\(^1\) IT University of Copenhagen (ITU), Copenhagen, Denmark
\(^2\) Interactions LLC, USA

aski@itu.dk, maste@itu.dk, bapl@itu.dk

Abstract

This paper introduces FT SPEECH, a new speech corpus created from the recorded meetings of the Danish Parliament, otherwise known as the Folketing (FT). The corpus contains over 1,800 hours of transcribed speech by a total of 434 speakers. It is significantly larger in duration, vocabulary, and amount of spontaneous speech than the existing public speech corpora for Danish, which are largely limited to read-aloud and dictation data. We outline design considerations, including the preprocessing methods and the alignment procedure. To evaluate the quality of the corpus, we train automatic speech recognition systems (ASR) on the new resource and compare them to the systems trained on the Danish part of Språkbanken, the largest public ASR corpus for Danish to date. Our baseline results show that we achieve a 14.01 WER on the new corpus. A combination of FT SPEECH with in-domain language data provides comparable results to models trained specifically on Språkbanken, showing that FT SPEECH transfers well to this data set. Interestingly, our results demonstrate that the opposite is not the case. This shows that FT SPEECH provides a valuable resource for promoting research on Danish ASR with more spontaneous speech.

Index Terms: speech corpus, speech recognition, Danish

1. Introduction

The main obstacle to effective automatic speech recognition is the lack of training material, especially for less-resourced languages such as Danish. Fortunately, the recent proliferation of available online content and open-source software has facilitated the collection of such data, lowering the threshold for wider-spread open ASR technology and research. In this regard, recordings of parliamentary sessions and their official reports form a new resource and compare them to the systems trained on the Danish part of Språkbanken, the largest public ASR corpus for Danish to date. Our baseline results show that we achieve a 14.01 WER on the new corpus. A combination of FT SPEECH with in-domain language data provides comparable results to models trained specifically on Språkbanken, showing that FT SPEECH transfers well to this data set. Interestingly, our results demonstrate that the opposite is not the case. This shows that FT SPEECH provides a valuable resource for promoting research on Danish ASR with more spontaneous speech.

2. Related Work

Research into ASR for Danish has been rather limited owing to the scarcity of publicly available speech corpora [1]. At present, there are only two public Danish speech corpora: Danish Språkbanken and DanPASS, which were developed under different research questions and objectives.

The more comprehensive of the two, Danish Språkbanken, is included in Språkbanken (Norwegian Language Bank), a collection of open-access and open-source language resources for Norwegian, Swedish, and Danish compiled by Nordisk Språkteknologi (NST). It contains two subsets designed specifically for the development of ASR systems: NST Danish ASR Database (SBRread) [2] and NST Danish Dictation (SBDictate) [3].

SBRread is the only public data set suitable for training ASR systems. It contains around 390 hours of phonetically balanced read-aloud speech by a total of 616 speakers, as well as general meta-data about the speakers. A standardized version of this data set was released with a recipe to train ASR systems in the Kaldi repository (sprakbanken) [4]. On the other hand, SBDictate is a smaller data set with roughly 54 hours of speech by 151 speakers aimed at acoustic modeling of automatic dictation. However, despite their size and speaker variety, both of these data sets are limited by their highly contrived nature. Namely, the utterances in these data sets constitute read-aloud sentences or phrases such as personal names, place names, acronyms, numerals, spelled out letters, etc.

DanPASS is a phonetically annotated speech corpus primarily intended for acoustic and auditory phonetic analyses [5]. It contains about 9 hours of speech by 27 speakers recorded in a studio using professional recording equipment. Although this corpus may offer a theoretical basis for the development of speech technologies, it is not particularly suitable for ASR due to its small size and artificial setting.

Currently, the Kaldi recipe sprakbanken represents the state-of-the-art on the SBRread test set and DanPASS [6]. However, for reasons outlined above, the performance of these models degrades sharply when they are confronted with large-vocabulary spontaneous speech, as we will show in Section 5.

In order to compile a corpus of utterances more akin to rapid spontaneous speech, we follow the recent trend of converting open parliamentary data into ASR speech corpora. This has been accomplished for languages such as Icelandic [7], Finnish [8], and Bulgarian [9]. In addition, a multilingual speech corpus has been constructed from the debates held in the European Parliament [10]. Meanwhile, the official reports of Folketing meetings have already been used to create a text...
corpus released within CLARIN [11] and proved invaluable for multiple research disciplines [12, 13, 14].

In constructing FT SPEECH, we follow a procedure similar to the one used to create LibriSpeech [15]. Other influential work on automatic alignment methods in the creation or correction of speech corpora includes [16, 17, 18, 19].

3. Corpus Preparation and Alignment

This section presents the corpus preparation and alignment procedures, including the description of the raw source data, audio and text preprocessing, lexicon creation, and alignment.

3.1. Source Data Description

FT SPEECH was created from the recordings of Danish parliamentary sessions and their annotated reports, which are freely available on the Folketing’s official website: ft.dk. The audio recordings are available in two formats: MP3 (audio only) and AAC (as part of the audio and video stream container MP4).

The sessions used to create the corpus include 1,003 meetings of the Parliament recorded in the period from October 5, 2010 (first video broadcast) until December 20, 2019 (last meeting in 2019). This amounted to about 4,960 hours of recorded audio material featuring 447 different speakers.

The reports of the parliamentary meetings are transcribed and published by the Office of the Folketing Hansard. Each report contains a comprehensive account of all parliamentary activities in the course of one meeting, including near-verbatim transcripts of the speeches by Members of Parliament (MPs) accompanied by their corresponding metadata, such as the speaker’s name, role, and political affiliation, as well as the approximate start and end timestamps of the speech.

The reports are released online as XML and PDF documents. Initially, only a preliminary version is released while the report is still subject to revision. From this point forward, it can take up to several months until the final version is published. During this period, the reports, and, in particular, the speech transcripts may undergo a number of modifications to ensure adherence to the formal guidelines established by the Presidium of the Danish Parliament. Therefore, the speeches are not transcribed strictly verbatim, but are instead adapted into standard written form by omitting speech disfluencies, correcting factual errors and slips of the tongue, and adding context to ensure the transcripts reflect the intentions of the speaker clearly and accurately [20].

In addition to prescribing documentation guidelines, the Presidium also enforces observance of parliamentary etiquette, which mandates decorum and the use of formal and respectful language in the Parliament. Some of the official rules state that the MPs must be addressed as either Mr. or Ms. followed by their full name, while the Ministers must be addressed with their minister titles. Furthermore, the MPs may not interject, applaud, express disagreement, or otherwise make noise during speeches and debates [12]. This makes the FT meetings well-suited for aligning the speech from the recordings. This was an arbitrary decision since the two channels were identical. The mono recordings were left unchanged at this stage. Next, we converted the selected single-channel MP3 recordings to WAV using a 16-bit linear PCM sample encoding (PCM_S16LE) sampled at 16 kHz. Finally, to extract speeches by single speakers, we segmented the obtained WAV files according to the timestamps and speaker names provided in the annotated meeting reports in the XML format.

To ensure the speaker names in the annotations refer to unique individuals, we cross-checked them with the biographies of past and present MPs available on the official website. This procedure resulted in 414K speeches whose duration ranged from less than 1 second to 15 minutes (mean \( \approx 40.1 \) s, SD \( \approx 63.68 \) s). However, most speeches did not perfectly align with their corresponding transcripts due to the inaccuracy of the timestamps in the XML annotations, which is one of the issues we try to overcome with the alignment procedure outlined in Section 3.4.

As stated earlier, the textual transcripts of the speeches were extracted from the XML documents containing the reports of the FT meetings. Their preprocessing involved expanding all common abbreviations, numbers, dates, and symbols, as well as removing all punctuation, capitalization, and unspoken parenthetical remarks and references.

3.2. Audio and Text Preprocessing

First, we downloaded the audio recordings of all FT meetings available on the official website up to and including December 2019. All the recordings were in the MP3 format with a bitrate of 128 kbit/s. Their duration ranged from 5 minutes to 16 hours (mean \( \approx 5 \) h, SD \( \approx 3 \) h).

We began the audio processing by extracting the left channel from the stereo recordings. This was an arbitrary decision since the two channels were identical. The mono recordings were left unchanged at this stage. Next, we converted the selected single-channel MP3 recordings to WAV using a 16-bit linear PCM sample encoding (PCM_S16LE) sampled at 16 kHz. Finally, to extract speeches by single speakers, we segmented the obtained WAV files according to the timestamps and speaker names provided in the annotated meeting reports in the XML format. To ensure the speaker names in the annotations refer to unique individuals, we cross-checked them with the biographies of past and present MPs available on the official website.

This procedure resulted in 414K speeches whose duration ranged from less than 1 second to 15 minutes (mean \( \approx 40.1 \) s, SD \( \approx 63.68 \) s). However, most speeches did not perfectly align with their corresponding transcripts due to the inaccuracy of the timestamps in the XML annotations, which is one of the issues we try to overcome with the alignment procedure outlined in Section 3.4.

As stated earlier, the textual transcripts of the speeches were extracted from the XML documents containing the reports of the FT meetings. Their preprocessing involved expanding all common abbreviations, numbers, dates, and symbols, as well as removing all punctuation, capitalization, and unspoken parenthetical remarks and references.

3.3. Alignment Lexicon

The lexicon used for alignment was produced by concatenating the vocabulary created from the preprocessed transcripts of FT speeches with the sprakbanken lexicon containing all words from the SBRread train set. This yielded around 233K unique words (types). Their pronunciations were generated using eSpeak NG, a multilingual rule-based grapheme-to-phoneme converter and speech synthesizer. We stripped these pronunciations of all stress, vowel length, and stød markers. We also made the pronunciations of foreign words consistent.

3 A parliamentary debate is a sequence of monologues on the same topic.

\[2\] A parliamentary debate is a sequence of monologues on the same topic.

\[3\] URL to the video and audio recordings: ft.dk/da/dokumenter/dokumentlister/referater

\[4\] URL to XML transcripts of the proceedings: ftp://oda.ft.dk/ODAXML/Referat/samling

\[5\] URLs to the biographies of FT MPs in Danish and English: https://www.ft.dk/da/medlemmer

\[6\] https://github.com/espeak-ng/espeak-ng
with the Danish phonetic alphabet in eSpeak, and manually added the unstressed forms of common function words.

### 3.4. Alignment Model

Because the FT meeting reports are not transcribed verbatim, we expect a large proportion of words in the aligned utterances to be misaligned. For instance, if a speaker mistakenly stated, *My uh party is against tax-taxation of the air used to to create soft ice*, but the party were, in fact, in favor of such taxation, the transcript would be edited by changing *against to for* and removing filler words (*uh*), restarts (*tax-*), and repetitions (*to*). In this example, if *against* were correctly recognized by an ASR system, the word would be misaligned because it did not match the transcript. For this reason, we need to extract verbatim transcriptions while allowing for word repetitions, restarts, and filler words that occur frequently in spontaneous speech. In our example, we would extract two segments: *My party is and taxation of the air used to create soft ice.*

First, we create a word alignment with a procedure similar to the one used in LibriSpeech. To compare ASR hypotheses to transcripts, we decode the timestamp-segmented FT data with a speaker-independent GMM AM trained with boosted MMI on transcripts, we decode the timestamp-segmented FT data with a such as algorithm classifies a misrecognized word as a word repetition (that start with a correctly recognized word and end with a misrecognized word (*against* in our example). If the segmentation against were correctly recognized by an ASR system, the word would be misaligned because it did not match the transcript. For this reason, we need to extract verbatim transcriptions while allowing for word repetitions, restarts, and filler words that occur frequently in spontaneous speech. In our example, we would extract two segments: *My party is and taxation of the air used to create soft ice.*

Next, we segment the utterances again and keep segments that start with a correctly recognized word and end with a misrecognized word (*against* in our example). If the segmentation algorithm classifies a misrecognized word as a word repetition such as *to*, the word is inserted into the reference and we do not split the segment. If segments contain silences longer than 0.3 seconds, they are split again. All ends of segments are extended in the audio and transcript to include the next word and to reduce edge effects in feature extraction from the core segment. The transcript is paddled with a silence token such as `<UNK>` unless the segment is at the utterance boundaries.

We restrict the utterances we include in the corpus to utterances with a duration ranging from 2 to 60 seconds to ensure they can be used to train AMs. As a result, we discard 487,938 utterances of which 1,320 were longer than 1 minute.

### 4. Corpus Description and Organization

Following the alignment, segmentation, and elimination of overly long or short utterances, the finished corpus, termed FT SPEECH, contains 1,017,244 utterances with a total duration of 1,857 hours produced by 434 speakers.

We partition this corpus into a training, development, and test set with no speaker overlap. To create the development and test set, we select the same number of male and female speakers with at least 150 utterances and 900 seconds per speaker, while trying to minimize training data loss.

Since the total duration per speaker varies significantly, it was impossible to create a speaker-balanced development or test set. Therefore, we decided to further partition both of these sets into two subsets: *balanced* and *other*. The balanced portions (*dev-balanced* and *test-balanced*) contain approximately equal amounts of speech per speaker. They were created by choosing a random sample of utterances from each speaker such that the total duration per speaker was 900 seconds and that the difference in the number of utterances per speaker was kept low. The other portions (*dev-other* and *test-other*) consist of the remaining utterances by the same speakers which had to be removed from the training data to avoid speaker overlap. Detailed statistics of the corpus and each of the partitions are shown in Table 1.

### 5. Speech Recognition Experiments

This section describes the ASR experiments conducted for the purpose of evaluating the new resource, FT SPEECH. We build two acoustic models: one trained on FT SPEECH train (FT AM) and the other on SBRead train data (SB AM), as well as two language models: one trained on FT text data (FT LMs) and the other on SBRead training transcripts (SB LMs), and subsequently evaluate their in-domain and out-of-domain combinations on three different test sets. Since SBRead is an established ASR corpus, we use the acoustic and language models trained on it as a reference point in our performance evaluation.

#### 5.1. Acoustic Models

We follow the sprakbanken recipe⁸ to train monophone and triphone segmentation GMM AMs [21] from scratch on FT SPEECH and generate an alignment to train an iVector model [22] for speaker adaptation of a Time-Delay Neural Network (TDNN) [23] AM. The only modification compared to the sprakbanken recipe is that we do not perform data augmentation with speed-perturbation on FT SPEECH because the size

---

7Implemented in the script `clean_and_segment_data.sh` in the Kaldi repository

8The recipe can be found here: [https://github.com/kaldi-asr/kaldi/blob/master/egs/sprakbanken/s5/run.sh](https://github.com/kaldi-asr/kaldi/blob/master/egs/sprakbanken/s5/run.sh)

---

### Table 1: Corpus partitions and their size in hours, total number of utterances, tokens, types, OOV tokens, and speakers. The speaker counts by gender are marked as F (female) and M (male). OOV tokens comprise all tokens whose types are not part of our ASR lexicon.

<table>
<thead>
<tr>
<th>Subset</th>
<th>Hours</th>
<th>Utterances</th>
<th>Tokens</th>
<th>Types</th>
<th>OOV tokens</th>
<th>Speakers (F+M)</th>
</tr>
</thead>
<tbody>
<tr>
<td>train</td>
<td>1,816.29</td>
<td>995,677</td>
<td>18,865,071</td>
<td>147,326</td>
<td>0</td>
<td>374 (146+228)</td>
</tr>
<tr>
<td>dev-balanced</td>
<td>5.03</td>
<td>2,601</td>
<td>51,497</td>
<td>6,888</td>
<td>2,846</td>
<td>20 (10+10)</td>
</tr>
<tr>
<td>dev-other</td>
<td>14.96</td>
<td>7,595</td>
<td>152,225</td>
<td>12,701</td>
<td>8,248</td>
<td></td>
</tr>
<tr>
<td>test-balanced</td>
<td>10.05</td>
<td>5,534</td>
<td>103,439</td>
<td>10,050</td>
<td>5,871</td>
<td>40 (20+20)</td>
</tr>
<tr>
<td>test-other</td>
<td>10.88</td>
<td>5,837</td>
<td>111,818</td>
<td>10,491</td>
<td>6,145</td>
<td></td>
</tr>
<tr>
<td>total</td>
<td>1,857.21</td>
<td>1,017,244</td>
<td>19,284,050</td>
<td>149,239</td>
<td>23,110</td>
<td>434 (176+258)</td>
</tr>
</tbody>
</table>
of the training data is larger than the size of SBRead augmented with speed-perturbation. For SBRead, we use the training, development, and test split introduced in [6].

We train so-called chain TDNN AMs with the LF-MMI objective on FT SPEECH and SBRead. LF-MMI is a sequence discriminative training criterion that maximizes the log probability of the correct phone sequence [24]. We train the AMs for 4 epochs on minibatches of 128 chunks where each chunk contains 150 frames. The frames are 40-dimensional MFCC features. The feature frames are subsampled so we only train on every third frame, but we create different versions of the training data by shifting the frames by 1 and 2 frames to create 3 versions. The effect is that every training epoch corresponds to 3 epochs. We use HMMs with a single state rather than the classic 3-state topology because of the low frame rate.

The first layer of the TDNN stacks 3 frames and a 100-dimensional iVector and projects the supervector to a 450-dimensional vector with an affine transform. The remaining layers consists of an affine transform, ReLU activation and a renorm component which is a layernorm without the mean term. We use a learning rate that decays from 0.001 to 0.0001 during training and we clip parameters at a Frobenius norm of 2.0. All components which are identical for the two AMs. Note that there are several important differences to the AM in [6] (see section 7.1).

### 5.2. ASR Lexicon and Language Models

To create the lexicon for the ASR experiments, we reuse the alignment lexicon but remove all types that appear only in the preprocessed transcripts of speeches by the speakers placed in the FT SPEECH development and test sets. With SRILM [25], we estimated several 3-gram and 4-gram LMs with Witten-Bell or Kneser-Ney smoothing on both text from the FT meeting reports and on the SBRead transcripts. Ultimately, we choose the trigram models with Witten-Bell smoothing, which we will refer to as FT LM and SB LM, for the final evaluation, as they achieve the best performance on their corresponding development sets. Before training, the transcripts were segmented into sentences using the spaCy sentence segmenter for Danish [26], and then preprocessed as described in Section 3.2.

### 6. Performance Evaluation

We use the standard word error rate metric (WER) to evaluate the performance. Our evaluation spans all four combinations of the two AMs and the two LMs (the lexicon is constant in all cases). We evaluate each of them on three test sets: SBRead test (introduced in [6] as SPTEST), SBDictate test (included in the original SBDictate data [3]), and FT SPEECH test-balanced (Table 1). The WER results for all combinations of AMs, LMs, and test data are shown in Table 2.

From Table 2, we can see that for FT SPEECH, the best WER of 14.01% is obtained with the in-corpus LM and AM. As expected, in-domain AM and LM combinations perform best in all in-domain settings (boldface in Table 2). However, going across domains remains a challenge.

When the LM does not match the domain of the test set, WER rises by 5–14% absolute, presumably due to significant lexical differences between SBRead and FT SPEECH. As stated previously, a large number of SBRead utterances consist solely of either proper nouns, numerals, spelled out names, or imperative sentences, while some also contain articulated punctuation symbols used for modeling automatic dictation. These kinds of utterances, most of which were devised to increase phonetic diversity and type-to-token ratio, do not occur in FT SPEECH nor general spontaneous speech. For these reasons, SBRead is a less challenging resource, reflected in the lower WER (8.81). We see a gap in performance when we fix the test set and LM, but replace the AM. This decrease in performance occurs as a result of the acoustic differences between the FT SPEECH and SBRead utterances, especially, the differences in the speech genre, recording environment and equipment, and audio encodings. Namely, SBRead was recorded in a quiet office and encoded in a lossless format, whereas FT SPEECH was recorded in the FT meeting chamber and encoded into a lossy format.

Most importantly, however, we observe that the combination of FT AM and SB LM evaluated on SBRead test achieves a WER of 13.07%, which is comparable to the results previously published on this test set (13.08–13.38% WER, presented in Table 7 in [6]). On SBDictate, it achieves a WER of 20.71 with FT AM+SB LM, which is 6.25% absolute WER off from in-domain data results. This means that the FT AM generalizes well to SBRead data, and it shows the benefit of the new corpus containing more spontaneous speech in a more realistic environment with disfluencies and background noise. Interestingly, the converse is not the case: the SB AM does not generalize to the FT SPEECH domain, resulting in the worst overall WER. This shows how poorly existing resources generalize, which further underlines the value of the proposed resource. While not strictly comparable, our WER results on FT are in similar ranges to related work on Icelandic, another Northern Germanic language, were a WER of 14.76% was reported on parliamentary speeches [7].

### 7. Conclusion

This paper introduces FT SPEECH, a novel corpus for Danish ASR containing more than 1,800 hours of speech. It enriches the limited landscape of existing resources for Danish with a resource containing more spontaneous speech in challenging realistic conditions. Our baseline results show that a combination of FT SPEECH with in-domain language data provides not only comparable results to prior work, but also a more challenging benchmark for future studies. As the source material expands naturally, we will update the corpus with new data.

### 8. References
