# Adding Metadata to Existing Parliamentary Speech Corpus

**Phoebe Parsons**[1] **Per Erik Solberg**[2] **Knut Kvale**[3]

**Torbjørn Svendsen**[1] **Giampiero Salvi**[1]

[1]Department of Electronic Systems, NTNU, Trondheim, Norway
[2]National Library of Norway, Oslo, Norway
[3]Telenor Research and Innovation, Oslo, Norway
{phoebe.parsons, torbjorn.svendsen, giampiero.salvi}@ntnu.no,
per.solberg@nb.no, knut.kvale@telenor.com

## Abstract

Parliamentary proceedings are convenient data sources for creating corpora for speech technology. Given its public nature, there is an abundance of extra information about the speakers that can be legally and ethically harvested to enrich this kind of corpora. This paper describes the methods we have used to add speaker metadata to the Stortinget Speech Corpus (SSC) containing over 5,000 hours of Norwegian speech with non-verbatim transcripts but without speaker metadata. The additional metadata for each speech segment includes speaker ID, gender, date of birth, municipality of birth, and counties represented. We also infer speaker dialect from their municipality of birth using a manually designed mapping between municipalities and Norwegian dialects. We provide observations on the SSC data and give suggestions for how it may be used for tasks other than speech recognition. Finally, we demonstrate the utility of this new metadata through a dialect identification task. The described methods can be adapted to add metadata information to parliamentary corpora in other languages.

## 1 Introduction

There has been, historically, a lack of high quality, freely available speech resources for machine learning tasks. Traditionally, these resources have been created to facilitate development of automatic speech recognition (ASR) models, and as such have been expensive to create, requiring human hours for both data collection and then careful, verbatim transcription. Even for "well resourced" languages like English, datasets rarely exceeded 1,000 hours. However, as new ASR technologies loosen the requirements for transcription precision, this allows for even larger datasets that are created from less verbatim sources (Chen et al., 2021; Galvez et al., 2021). These new, more loosely supervised datasets often lack details found in older, more traditional speech resources and are therefore potentially limited in their application.

Many established speech resources are composed of relatively short duration segments with speech from only one speaker at a time. Additionally, this speaker is often known (even if only by an anonymized speaker identifier) and metadata, such as age and gender, is given about them. This richness of metadata allows for speech technology and machine learning tasks beyond ASR — such as language (or dialect) identification, speaker diarization, speaker identification or verification. Crucial to all these tasks is knowledge about who is speaking.

Recently, a number of speech corpora were created from public domain recordings of parliamentary proceedings; for instance, Iceland (Helgadóttir et al., 2017), Denmark (Kirkedal et al., 2020), Finland (Virkkunen et al., 2023), Croatia (Ljubešić et al., 2022) and the European Parliament (Wang et al., 2021). In all of these works it is known, at the very least, who is speaking in each segment (either by name or speaker ID), with most also including gender information. Virkkunen et al. explored their dataset using the rich metadata they were able to pull from an open API providing both distribution information and ASR results along age, gender, and educational background lines. However, it appears that this rich metadata was not released with the final dataset. Ljubešic et al. included name, gender, year of birth, party affiliation and party status for their speakers.

In 2023, the National Library of Norway (NB) developed the Stortinget Speech Corpus (SSC) (Solberg et al., 2023) using data from the Norwe-



*Proceedings of the Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies (NoDaLiDa/Baltic-HLT 2025)*, pages 448–457

gian parliament (called *Stortinget* in Norwegian). In early 2024, NB published the results of their analysis of several ASR systems for Norwegian (Solberg et al., 2024). In this report they showed that Whisper models (Radford et al., 2022), fine-tuned on the SSC and some additional smaller datasets, performed best on an unseen test set created from radio and TV program audio. This fine-tuned model outperformed both the base Whisper model as well as fine-tuned wav2vec (Baevski et al., 2020) models and commercial ASR systems from Google and Microsoft, thus demonstrating the importance of this speech corpus in combination with the Whisper architecture for ASR.

Despite the SSC's obvious utility in training well-performing ASR models, it, as originally created, contains no metadata for each speech segment. We believe the effort to construct the missing metadata has merit as expanding the SSC into other speech technology domains would be a benefit for Norwegian speech research. To that end, we have undertaken the effort of ensuring that each segment in the SSC has been matched to a speaker identifier and that public speaker metadata has been added. As a result of this effort, this new metadata is now included with the SSC and made available by the Norwegian Language Bank at the National Library of Norway. Furthermore, we offer a recommendation for a subset of the SSC that more closely resembles traditional well annotated speech corpora and may be more applicable to other speech tasks. Finally, we believe that the efforts described in this paper can be easily extended and applied to similar corpora in other languages and countries.

## 2 The Pre-Existing SSC Dataset

The SSC contains more than 5,000 hours of natural Norwegian speech paired with non-verbatim transcripts created from the Norwegian parliament. The National Library of Norway created the SSC by following the technique described by (Ljubešić et al., 2022). They first broke the plenary meetings into segments using voice activity detection. Shorter segments were combined resulting in each SSC segment being roughly 30 seconds. In doing this, no concern was given to speaker boundaries. That is, the 30 second segments were created from files containing recordings of a whole day's worth of parliamentary discussion, without awareness of who was speaking or whether there was one or multiple speakers in the segment. Thus for each segment in the SSC, no speaker metadata is available.

After the audio had been segmented, an ASR system was then used to generate transcripts for these segments. The Levenshtein ratio[1] was then used to align the ASR output with the text of the official parliamentary proceedings sourced from the ParlaMint-NO corpus[2]. The proceedings were human-transcribed at the utterance level with some light editing and omissions for standardization and legibility. Because the official proceedings are not a verbatim transcription of the spoken utterances, the ASR transcriptions may deviate considerably from the proceedings. Consequently, only segments where the score produced by the Levenshtein ratio between the proceedings text and the ASR text was above a threshold (0.5) were kept. For the selected 30 second segments, the proceedings text was taken as the transcription. The Levenshtein ratio score was also kept in the SSC. In this manner, the SSC was created.

## 3 Speaker Metadata

### 3.1 Recovering Speaker Information from ParlaMint

The first objective of this work was to recover who is speaking in each segment of the SSC. To do this, we turn our attention to the Norwegian ParlaMint-NO text corpus. As mentioned in Section 2, this corpus contains the proceedings text. Additionally, it is annotated with metadata on speaker identity, gender, date of birth, and which of the two written forms of Norwegian the transcript is in for every utterance.

The task of reconciling the SSC text and the speaker metadata was done using word offsets. When creating the SSC, ASR output was aligned with the proceedings from ParlaMint. Though the metadata available in ParlaMint was discarded during the original creation of the SSC, the word offsets — the index of the starting and ending words in the ParlaMint proceedings — were preserved for each approximately 30 second segment. We can then join the ParlaMint metadata and the SSC segments by reconciling the offsets.

To illustrate how this reconciling of word off-

[1] https://rapidfuzz.github.io/Levenshtein/levenshtein.html#ratio

[2] https://www.nb.no/sprakbanken/ressurskatalog/oai-nb-no-sbr-77/

| Utt ID | Speaker ID | Proceedings Text | Start | End |
|---|---|---|---|---|
| 1 | person.1 | Good morning all | 0 | 2 |
| 2 | person.1 | Today we will be starting with Representative Smith | 3 | 10 |
| 3 | person.2 | As many of you know, moose in Norway are a common sight | 11 | 22 |
| 4 | person.2 | Therefore we propose that | 23 | 26 |

Table 1: A synthetic example of ParlaMint utterances. The starting and ending word indexes have been added for each utterance.

set and metadata occurs, we refer to the the fictional snippet of ParlaMint utterances presented in Table 1. Let us assume that we have an SSC segment where the text offsets with respect to that ParlaMint snippet are between 5 and 24. We first determine which utterance contains the word offset 5, in this case utterance 2 (the start index is smaller than 5 and the ending index is larger). We then determine which utterance contains the index 24, in this case utterance 4. As utterance 3 is between our starting and ending utterances we assume it too aligns with the current SSC segment.

By aligning the text in the SSC with the ParlaMint text, we have recovered the speaker information from each segment in the SSC. In addition to speaker identifiers (represented as person.1 and person.2 in the example), we also now have the date of birth, gender, and Norwegian written form (or forms) used in each segment. The speaker identifiers can be used to add further metadata, as described in the following sections.

### 3.2 Stortinget API

Beyond the metadata available from ParlaMint, we believed it to be useful to add publicly available information to the corpus, including the municipality and county where the speaker was born, as well as the county or counties represented by that speaker. This was accomplished by use of the Stortinget application programming interface (API)[3]. The Stortinget API provides a programmatic way to access data about the Norwegian parliament, including endpoints for bibliographic information on the speakers in the parliament. All metadata from the API is covered by a Norwegian Licence for Open Government Data[4] which permits copying, using and distributing information from the API. The endpoint `kodetbiografi` contains information on the speaker's municipality of birth, county of birth, and counties represented.

[3] https://data.stortinget.no/

[4] https://data.norge.no/nlod/en/2.0

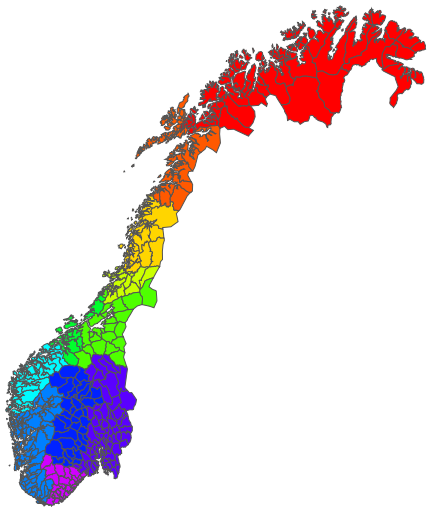

Figure 1: The municipalities of Norway mapped to dialect regions. The eastern dialect regions have been collapsed from (Skjekkeland, 1997).

We called this endpoint for each speaker using the speaker ID from the ParlaMint utterances. Not all speakers have information provided for each of the three fields that we were interested in. If one of these fields lacked information for a speaker, no further efforts were made to find this information, both on practicality and privacy grounds.

### 3.3 Municipality to Dialect

Our aim in gathering this municipality and county information was to enable an automated method of assigning presumed dialect. That is, given that Norwegian dialects are largely decided along geographic lines (Sandøy, 1987, p. 16), we hoped to use the municipality of birth to infer which dialect a person is likely to be speaking in.

The Norwegian language has no official standard speaking style (The Language Council of Norway). Hence, there is a large variety of dialectal realizations manifesting in pronunciation, lexical items, and grammar. Additionally, the culture encourages people to speak with their native

dialect in all situations from the least formal to the most. It is even common for speakers to retain their native dialects and to use dialectal lexical items when speaking in parliament. Thus, we find including dialect information to be both pertinent and, hopefully, useful to machine learning tasks related to speech.

To enable this automatic assignment of dialect, we created a mapping between all municipalities and counties in Norway and their assumed dialects. Using the dialect map created by Skjekkeland (Skjekkeland, 1997) as the ground truth, we manually analyzed maps of each county and their municipalities in order to align them with the boundaries drawn in Skjekkeland's map. Further, as we wished this municipality-to-dialect mapping to be useful with other existing Norwegian resources, historical municipalities and counties were included. As we found this mapping useful and was nontrivial to produce, it has been made available through the Norwegian Language Bank[5].

This inference of dialect from birth municipality does not, of course, account for people who were born in one place then quickly moved to another. Nor does our inference take into account that speakers often tend to adapt their dialect, at least slightly, to the local or national "standard" dialect. Therefore, for speakers who represent the same county they were born in, one could assume that speaker is still, potentially, representative of the dialect label assigned. However, for the working going forward, we will be using dialect labels generated from the speaker's municipality of birth regardless of if they later moved to a new county.

## 4 Data observations

As stated earlier, the SSC was designed for loosely supervised ASR training, and has already been used for this aim[6]. However, other speech tasks require either a greater degree of faithfulness in transcription or audio with only one speaker per segment, or both. In order to understand which, if any, part of the SSC might be useful in these other tasks, an analysis of the data was performed.

[5] https://www.nb.no/sprakbanken/en/resource-catalogue/oai-nb-no-sbr-92/

[6] https://huggingface.co/collections/NbAiLab/nb-whisper-65cb8322877f943912afcd9f

### 4.1 Towards verbatim transcripts

While non-verbatim transcripts work well for weakly-supervised training of ASR models such as Whisper, other ASR frameworks (e.g. wav2vec 2.0 (Baevski et al., 2020)) still require transcripts that align more closely with the audio. Thus, we begin to look at the transcriptions available to us to understand how often they align.

As described in Section 2, each SSC section has a score denoting the similarity between the proceedings text and the verbatim transcripts produced by ASR. It follows that when the proceedings text has a high similarity score to the ASR output the SSC text is presumably verbatim. However, these similarity scores are not infallible as ASR errors could lower the score regardless if the the SSC text was actually verbatim.

Despite the potential for ASR errors, we have observed that low scores are often a result of spoken information being omitted from the proceedings text. During proceedings, the Stortinget president often introduces the next speaker or provides other administrative information. Additionally, other speakers often recognize the president, have false starts in their sentences, or include other unnecessary words. As the proceedings are meant to be read, the transcribers tasked with creating the proceedings omit and lightly editorialize for readability. Thus, as can be seen in the example in Table 2, introductions of the next speaker (which would be obvious from the names associated with each utterance when reading the transcript) are not included in the proceedings.

We have found that, as a general rule of thumb, segments with Levenshtein ratios over 0.8 are highly accurate. While some segments achieve a perfect score of 1.0, they only account for 13.5 hours of the over 5,000 total hours in the SSC. Whereas, if all segments scoring over 0.8 are included, then over 3,300 hours of data is available.

### 4.2 One-speaker segments

As tasks such as speaker identification or dialect recognition generally require audio segments with only one speaker, identifying subsets of the SSC where there is only one speaker is beneficial.

This can be done by either finding segments in the SSC corpus that already contain a single speaker, according to the metadata, or by splitting multiple-speaker segments into a number of single-speaker sub-segments. To assess the impact

| | | | |
|---|---|---|---|
| Speaker ID | person.LHH | person.DTA | person.TRJ |
| Proceedings | dette løftebruddet? | | Nei, jeg tror |
| Transcription | dette løftebruddet | statsråd ris johansen | nei president jeg trur |
| English | this breach of promise | minister ris johansen | No, president I think |

Table 2: An example of different transcription standards.

| Speakers in Segment | Count of Segments |
|---|---|
| 1 | 624337 |
| 2 | 88560 |
| 3 | 11808 |
| 4 | 78 |

Table 3: Counts of segments in the SSC by the number of speakers, according to the new SSC metadata

of those strategies, we aggregated SSC segments by the number of speakers in Table 3. We can see that one-speaker segments account for 86.14% of the SSC segments. Thus, it is feasible to simply discard the segments with multiple speakers and still have over 4,478 hours of audio.

However, as mentioned when discussing the proceeding transcriptions, there are many instances where brief speaker turns are not included in the transcriptions. It follows then, that though the new metadata in the SSC may only recognize one speaker, another speaker could have spoken and simply been omitted from the proceedings on the basis of readability.

### 4.3 Splitting multi-speaker segments

To fully use all data, segments with multiple speakers would ideally be split into one-speaker segments. We explored the simple approach of using forced alignment to align the text and the audio. The speaker utterance boundaries are known from the text and could be used to split the audio. However, in most instances the proceedings text is not verbatim enough for forced alignment. Forced alignment using the more verbatim ASR output could be feasible, however we then need to align the ASR output with the proceedings text (with a high degree of fidelity)—a non-trivial task. Ultimately, for future work, we see speaker diarization as a promising alternative for multi-speaker segments. Additionally, we have yet to explore how much, if any, of the speech in multi-speaker is overlapping, providing yet another avenue for future work.

## 5 Comparison with the NPSC

The Norwegian Parliamentary Speech Corpus (NPSC) (Solberg and Ortiz, 2022) was created using data from 41 days of Norwegian parliament recordings where humans manually segmented and transcribed the data. Thus, the NPSC composes a small subset of the data available in the SSC. After listening to each speaker, the transcribers assigned each speaker in the NPSC a dialect. Five dialect regions were used for this task: Eastern Norway (from here on called East), Western Norway (West), Northern Norway (North), Trøndelag (Mid), and Southern Norway (South). Given this careful human supervision, the NPSC utterances may then serve as a "ground truth" for verbatim text, as well as speaker identities and dialects.

To reconcile the NPSC and the SSC, we could not use word offsets as the words are from fundamentally different sources (verbatim transcription versus official proceedings). However, the millisecond offset from the beginning of the day's recording was preserved in both the NPSC and SSC segments. Therefore, we were able to use these millisecond offsets and the same approach as described with the word offsets to determine which NPSC utterances corresponded to each SSC segment.

### 5.1 One-speaker segments

As discussed in Section 4.2, there are potentially segments that the SSC metadata identifies as single-speaker, but in reality contains speech from multiple speakers. To understand the scope of this potential problem, we compare the speaker counts asserted by the SSC and the NPSC.

By doing this, we can see that when looking at utterances where the SSC metadata claims that only one speaker is present, we find that the NPSC believes there are more speakers 10.6% of the time. On the whole, we find that the NPSC and

SSC disagree on speaker counts 20.8% of the time. This implies that we should remain skeptical about the speaker counts given by the SSC, especially for tasks where it is crutial to have one, and only one, speaker in a segment.

### 5.2 Speaker dialect labels

As the NPSC also contains human prescribed dialect labels for each speaker, we can compare our inferred dialect labels with these ground truth labels.

There are 226 speakers in the NPSC, of which, metadata was available to assign the dialect to 164 of them. The dialect label from the SSC (as generated from municipality of birth) agreed with the NPSC human assigned label approximately 91.5% of the time.

Many of the speakers that we were unable to provide a dialect label for were speakers that spoke only a little or infrequently. Thus, the dialectally labeled speech accounts for 71.3% of the NPSC audio. Further, the duration of labeled audio in the SSC accounts for 78.6% of the audio, (over 4,000 hours), a similar percentage to the NPSC.

## 6 Automatic dialect identification

To demonstrate the utility of these new dialect labels, we have investigated the task of automatic dialect classification.

### 6.1 Model and fine-tuning

For the task of automatic dialect classification, we chose to fine-tune a model instead of creating a model from scratch. As a starting point, we took a model already fine-tuned for the language identification task[7], itself fine-tuned from the Whisper-medium model[8]. The Whisper-medium model contains 769M parameters and was trained for ASR and speech translation on 680,000 hours of speech. The fine-tuning to language identification was done using the FLEURS dataset[9] upon which the model achieved an accuracy of 0.88.

We then further fined-tuned the model from language to Norwegian dialect identification. Two models were trained, one using data from the NPSC, the other data from the SSC. This will allow us to understand the impact of the larger amount of data available in the SSC. For training, the first two convolutional layers in the encoder were fixed and each model was allowed to train for 3 epochs. The resulting model after these 3 epochs was used for the evaluation reported below.

### 6.2 Dataset splits

To prepare the NPSC and SSC for fine-tuning, the datasets were then divided into train, validation, and test sets by speakers. That is, a speaker (and all the utterances they said) would be assigned to one, and only one, of the three splits to ensure that the model was not simply learning the speaker's voice. As the NPSC is smaller, we utilized all of the NPSC where we had a dialect label for the speaker, resulting in a total of approximately 126 hours of speech.

We chose to use a subset of the SSC for the fine-tuning effort so as to have a dialectally balanced dataset. We determined which of the dialect regions contained the smallest amount of data (the South) and sampled data from each of the other regions to a similar size. This resulted in approximately 155 hours of data for each of the five dialect regions, or 774 hours of data in total. The size in both hours and number of speakers for both the NPSC and SSC training sets can be seen in Table 4.

To make a more direct comparison between the NPSC and SSC, we created a test set containing data from both. To do this, we removed speakers from the NPSC test set that appeared in the SSC training and removed speakers from the SSC test set that appeared in the NPSC training set. We then combined the remaining test data into a common NPSC+SSC test set.

### 6.3 Nordavinden og Sola

To evaluate how well these dialect identification models generalize beyond the parliamentary domain, we turned the Nordavinden og Sola (NVOS)[10] (in English, *The North Wind and the Sun*) database. This database consists of speakers reading The North Wind and the Sun fable in Norwegian. Although the task was read speech, participants were allowed to alter the text, both in terms of lexical items and word order, to best fit their native dialects. The municipality for each

[7] https://huggingface.co/sanchit-gandhi/whisper-medium-fleurs-lang-id

[8] https://huggingface.co/openai/whisper-medium

[9] https://huggingface.co/datasets/google/xtreme_s#language-identification---fleurs-langid

[10] https://www.hf.ntnu.no/nos/

| | NPSC | | | | | | | | | SSC | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | train | | | validation | | | test | | | train | | | validation | | | test | | |
| | seg | dur | spk | seg | dur | spk | seg | dur | spk | seg | dur | spk | seg | dur | spk | seg | dur | spk |
| east | 21631 | 40 | 92 | 5826 | 4 | 12 | 2003 | 9 | 12 | 17690 | 127 | 92 | 2781 | 20 | 12 | 1192 | 8 | 11 |
| west | 12589 | 26 | 56 | 3288 | 5 | 7 | 2134 | 7 | 7 | 15481 | 111 | 60 | 3885 | 28 | 8 | 2191 | 15 | 7 |
| mid | 4560 | 9 | 22 | 208 | 2 | 3 | 906 | 0.5 | 3 | 17586 | 127 | 28 | 2298 | 16 | 4 | 1734 | 13 | 4 |
| north | 5230 | 11 | 26 | 549 | 3 | 3 | 1437 | 1 | 4 | 17092 | 123 | 27 | 2030 | 14 | 4 | 2361 | 16 | 3 |
| south | 3254 | 6 | 15 | 103 | 2 | 2 | 803 | 0.25 | 2 | 13245 | 95 | 11 | 3002 | 22 | 2 | 5315 | 36 | 1 |

Table 4: Amount of data available each data split for the NPSC and sub-sampled SSC. Quantities in number of segments, duration of speech in hours, and number of unique speakers.

speaker was recorded and from this we were able to assign one of the five cardinal dialects.

As the NPSC and SSC have different average utterance durations (NPSC utterances being an average of 7 seconds with a standard deviation of 5 seconds versus the SSC's average of 25.8 seconds and standard deviation of 4.3 seconds), we created two test sets with the NVOS data with different utterance durations. In the first, the audio was left unaltered and the whole utterance (on average, about 32 seconds) was given to the model. In the second, we split each audio in half and then asked the model to classify these approximately 15 second audio clips.

### 6.4 Results and discussion

The accuracy, balanced accuracy, and weighted F1 from evaluating each of the two models (trained on NPSC or SSC) can be seen in Table 5. Metrics were calculated using scikit-learn 1.4.2. The model trained on the SSC data performs better than or equally well as the model trained on the NPSC for all test sets. When looking at the NPSC part of the combined test set, we can see that the SSC model performed as well as the in-domain NPSC model. However, the NPSC model performed very poorly when asked to predict using SSC audio.

Metrics for recall (macro and weighted) and F1 (micro and macro) were also calculauted. However, as they follow the same general trend (where the model trained on SSC data performed as well or better than the model train on the NPSC, they are not included in this paper.

We can further observe from Table 5 that the length of the segment in the NVOS data has little impact. The SSC model did perform slightly better when presented with the full audio clips. This could be due to the fact that the segments in the SSC are approximately 30 seconds as well.

Confusion matrices for these tests are presented in Figure 2. We find that both models often perform well on the East and West regions and poorly on the South. In fact, it is only in matrix (f) that a model predicts the South at all (of note as well, the only Southern speaker in the common test set is from the NPSC, meaning that the SSC model is robust enough to predict South for an out-of-domain speaker).

From these results, we can see that having more data even if not necessarily more speakers (211 speakers in the NPSC training set versus 218 in the SSC set) can positively impact model performance both in-domain and out.

While we are encouraged by the results presented here, there are several potentially confounding features. Our methodology for splitting the data along speaker lines does lead to imperfect datasets (for example, the South being represented by only one speaker in the test set, despite having the most hours of data 4). Further, no attention was paid to the content of the utterances. That is, within the parliamentary domain, it is conceivable that there are several set phrases that each speaker is apt to repeat. So, while there is no speaker overlap between the train, validation, and test sets, there is the potential for overlap of spoken content. Further, given the limited number of speakers, it is possible the that the model has learned some speaker-dependent features. Thus, we look forward to further exploring the impact of speaker and content on dialect identification in future works.

## 7 Conclusion

Through the efforts described in the paper, we enrich the SSC with speaker ID, gender, written form, age, dialect, municipality and county of birth and counties represented for each SSC segment.

Although the methods are developed for the Norwegian parliament, we believe they can rela-

| | Trained on | NVOS half | NVOS full | Common test set | | |
|---|---|---|---|---|---|---|
| | | | | Total | NPSC | SSC |
| Accuracy | NPSC | 0.75 | 0.75 | 0.60 | 0.74 | 0.45 |
| | SSC | 0.78 | 0.79 | 0.77 | 0.74 | 0.80 |
| Balanced Accuracy | NPSC | 0.63 | 0.63 | 0.48 | 0.52 | 0.58 |
| | SSC | 0.68 | 0.68 | 0.61 | 0.56 | 0.81 |
| Weighted F1 | NPSC | 0.72 | 0.73 | 0.55 | 0.73 | 0.38 |
| | SSC | 0.77 | 0.78 | 0.76 | 0.73 | 0.81 |

Table 5: Accuracy, balanced accuracy, and weighted F1 of dialect identification models trained on either NPSC or SSC data. Models were evaluated against the NVOS dataset and the common dataset.

tively easily be adapted to parliamentary speech corpora in other languages.

The further aim of our work herein was to provide a subset of the large SSC that could be used for tasks beyond ASR. Thus, we provided observations on the corpus and suggested suitable subsets for different tasks in speech technology.

We demonstrated the utility of this new metadata through a dialect identification task. The model trained using SSC outperformed the model trained with a smaller parliamentary corpus, thus showing an benefit of a corpus of the SSC's size.

Finally, as a continuation of (Ljubešić et al., 2022) and (Solberg et al., 2023), this work provides a general template for how public datasets, such as parliamentary recordings, may be transformed into corpora for machine learning.

## Acknowledgments

This work has been done as part of the SCRIBE project as funded by the Norwegian Research Council, project number: 322964.

## References

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems*, volume 33, pages 12449–12460. Curran Associates, Inc.

Guoguo Chen, Shuzhou Chai, Guan-Bo Wang, Jiayu Du, Wei-Qiang Zhang, Chao Weng, Dan Su, Daniel Povey, Jan Trmal, Junbo Zhang, Mingjie Jin, Sanjeev Khudanpur, Shinji Watanabe, Shuaijiang Zhao, Wei Zou, Xiangang Li, Xuchen Yao, Yongqing Wang, Zhao You, and Zhiyong Yan. 2021. GigaSpeech: An Evolving, Multi-Domain ASR Corpus with 10,000 Hours of Transcribed Audio. In *Proc. Interspeech 2021*, pages 3670–3674.

Daniel Galvez, Greg Diamos, Juan Ciro, Juan Felipe Cerón, Keith Achorn, Anjali Gopi, David Kanter, Maximilian Lam, Mark Mazumder, and Vijay Janapa Reddi. 2021. The people's speech: A large-scale diverse english speech recognition dataset for commercial usage. *CoRR*, abs/2111.09344.

Inga Rún Helgadóttir, Róbert Kjaran, Anna Björk Nikulásdóttir, and Jón Guðnason. 2017. Building an ASR Corpus Using Althingi's Parliamentary Speeches. In *Proc. Interspeech 2017*, pages 2163–2167.

Andreas Kirkedal, Marija Stepanović, and Barbara Plank. 2020. FT Speech: Danish Parliament Speech Corpus. In *Proc. Interspeech 2020*.

Nikola Ljubešić, Danijel Koržinek, Peter Rupnik, and Ivo-Pavao Jazbec. 2022. ParlaSpeech-HR - a freely available ASR dataset for Croatian bootstrapped from the ParlaMint corpus. In *Proceedings of the Workshop ParlaCLARIN III within the 13th Language Resources and Evaluation Conference*, pages 111–116, Marseille, France. European Language Resources Association.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust Speech Recognition via Large-Scale Weak Supervision. *arXiv*.

Helge Sandøy. 1987. *Norsk dialektkunnskap*. Novus Forlag.

Martin Skjekkeland. 1997. *Dei norske dialektane : tradisjonelle særdrag i jamføring med skriftmåla*. Høyskoleforl.

Per Erik Solberg, Pierre Beauguitte, Per Egil Kummervold, and Freddy Wetjen. 2023. A large Norwegian dataset for weak supervision ASR. In *Proceedings of the Second Workshop on Resources and Representations for Under-Resourced Languages and Domains (RESOURCEFUL-2023)*, pages 48–52, Tórshavn, the Faroe Islands. Association for Computational Linguistics.

Per Erik Solberg and Pablo Ortiz. 2022. The Norwegian parliamentary speech corpus. In *Proceedings of*

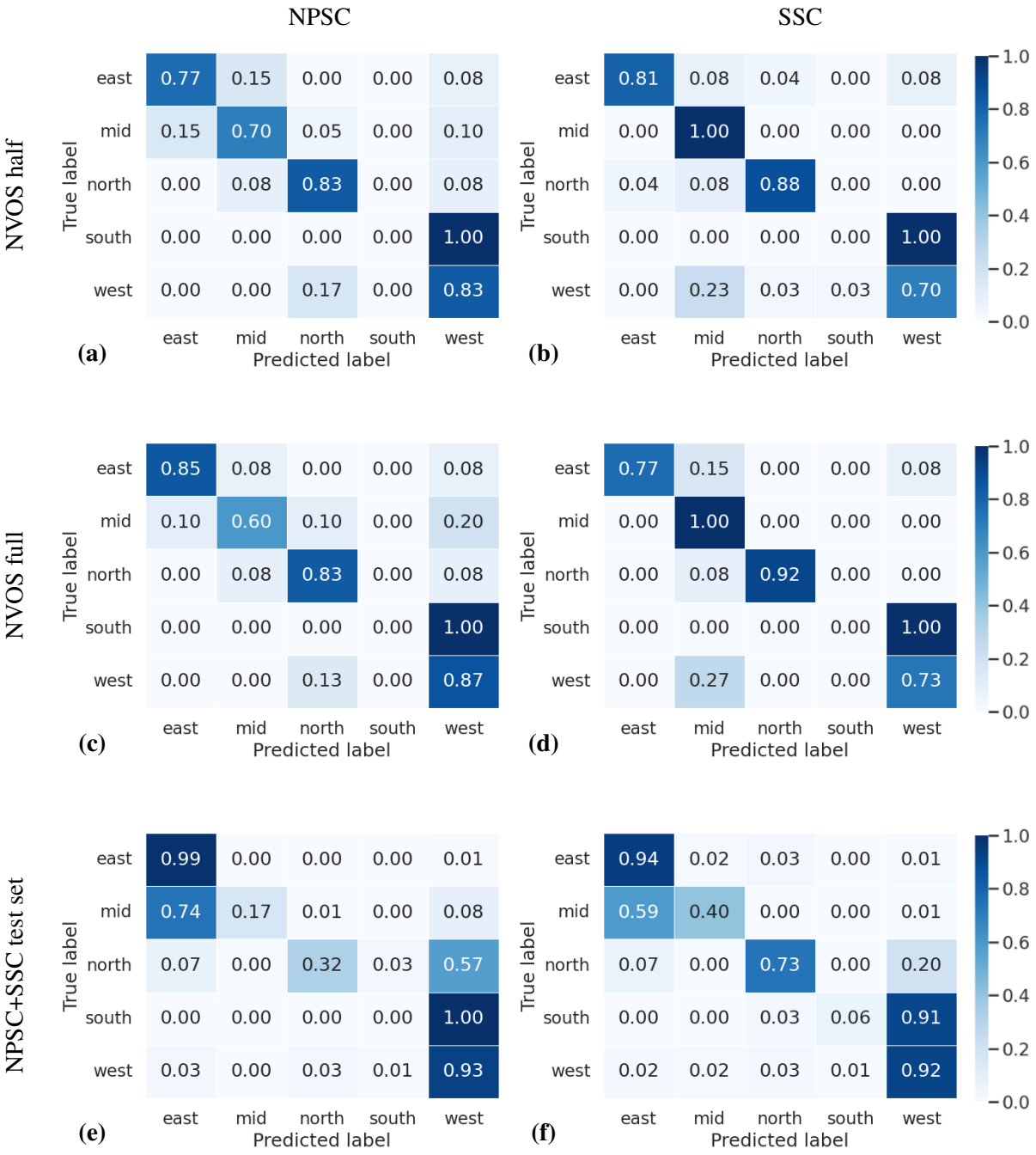


Figure 2: Normalized confusion matrices showing classifier performance on the three shared test sets. The first column (a, c, e) are from the model fine-tuned using NPSC data. The second column (b, d, f) are from the model fine-tuned using the SSC. The first row are the results when evaluated using the NVOS halves set, the second row the NVOS full set, and the third row the test set comprised of both NPSC and SSC data.

*the Thirteenth Language Resources and Evaluation Conference*, pages 1003–1008, Marseille, France. European Language Resources Association.

Per Erik Solberg, Marie Røsok, Ingerid Løyning Dale, and Arne Martinus Lindstad. 2024. Status for norsk talegjenkjenning. Technical report, National Library of Norway.

The Language Council of Norway. Uttaleråd. https://sprakradet.no/godt-og-korrekt-sprak/praktisk-sprakbruk/uttalerad/. Accessed: 2025-01-07.

Anja Virkkunen, Aku Rouhe, Nhan Phan, and Mikko Kurimo. 2023. Finnish parliament asr corpus: Analysis, benchmarks and statistics. *Lang. Resour. Eval.*, 57(4):1645–1670.

Changhan Wang, Morgane Riviere, Ann Lee, Anne Wu, Chaitanya Talnikar, Daniel Haziza, Mary Williamson, Juan Pino, and Emmanuel Dupoux. 2021. VoxPopuli: A Large-Scale Multilingual Speech Corpus for Representation Learning, Semi-Supervised Learning and Interpretation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 993–1003, Online. Association for Computational Linguistics.