Effects of Speaker Bias in Dialect Identification and Automatic Transcription with Self-Supervised Speech Models

Olli Kuparinen
Faculty of Information Technology and Communication Sciences
Tampere University
olli.kuparinen@tuni.fi

Abstract

A major issue in audio modeling is speaker bias, in which the models learn language external traits, such as a speaker's timbre or pitch, and use this information as a shortcut to a language task. This is especially problematic for dialectology, as it is typical in dialect corpora that only a few speakers represent a complete dialect area. In this paper, we explore the effects of speaker bias in two dialectal tasks: dialect identification and automatic dialectal transcription. We build two different data partitions of dialect interviews in Finnish and Norwegian: 1) a speaker dependent partition in which all of the speakers appear in training, development, and test sets, and 2) a speaker independent partition where each speaker only appears in exactly one set. We further experiment with modifications of the training data by augmenting the original audio with pitch shifts and noise, as well as changing the original speakers' voices with voice conversion models. We show that the dialect identification models are highly affected by speaker bias, whereas automatic dialectal transcription models are not. The audio modifications do not offer major performance gains for either of the languages or tasks.

1 Introduction

Natural language processing (NLP) has long been focused on texts, mostly collected from the internet. The recent development of self-supervised speech models such as wav2vec2.0 (Baevski et al., 2020) and Whisper (Radford et al., 2022) has however shifted the focus more towards audio data, offering possibilities for automatic speech recognition (ASR), speech synthesis, and spoken language identification, for instance. A similar shift can be seen in dialectologically inclined NLP, for which data has typically been text in the form of (phonetically) transcribed speech or user-generated content from social media.

One major difference between text and audio in dialectal tasks is the nature of the medium: speech consists of many speaker-related effects (timbre, pitch, duration, etc.) in addition to the linguistic content, whereas text is formally more consistent (written mostly in standardized alphabets). This raises potential issues, as the speech models learn speaker-specific traits instead of (or at least in addition to) dialectal traits. Since dialectal datasets often include only a few speakers per dialect, this can lead trained models to neglect dialectal information and only focus on the speaker effects as a shortcut to dialect identification, for instance. This effect can be called **speaker bias** (or speaker leakage) in audio models (Abdullah et al., 2025).

In this work, we analyze the effects of speaker bias in two dialect-focused NLP tasks: **dialect identification** and **automatic dialectal transcription**. We use interview data from two unrelated languages in Finnish and Norwegian, and create two different data partitions to showcase the potential issues in data processing. Based on these partitions, we analyze how speaker bias can alter the perceived performance of models in dialectal audio tasks. We further explore typical methods used to mitigate speaker bias in speech modeling, such as audio augmentation and voice conversion, as possible solutions to the raised issues. Working on pre-trained models, the focus of the paper is not in best possible performance, but the performance differences introduced in data preprocessing. The main contributions of the paper are thus:

- show and analyze the effects of speaker bias in dialect identification and automatic dialectal transcription,
- explore possible solutions in audio augmentation and voice conversion, and
- analyze performance in the two tasks in two unrelated languages.

Proceedings of the Thirteenth Workshop on NLP for Similar Languages, Varieties and Dialects, pages 32–44
March 29, 2026.

Style	SKN (fi)	LIA (no)
Transcript	jos vuav ver lähttöö	denn fysste kjirrkjå så va byggde # sto på enn annja plass
Standard	jos vain veri lähtee	den første kyrkja som var bygd stod på ein annan plass
English	if only the blood bleeds	the first church that was built was in another place

Table 1: Examples from the two datasets with the dialectal transcription on top and a standard language alternative below. Our Norwegian dataset is standardized to Nynorsk. The # in the Norwegian example denotes a pause in speech. An English gloss is presented at the bottom.

2 Related Work

2.1 Dialect Identification

Language and dialect identification from text is a standard task in natural language processing. The automatic distinction between distant languages has been declared solved (McNamee, 2005), but for similar languages and dialects the task is still relevant.

Dialect identification has been extensively studied in the the VarDial workshops, that have often included a shared task in discriminating between similar languages and dialects (e.g., Gaman et al., 2020; Chakravarthi et al., 2021; Aepli et al., 2023). For a long time, traditional linear classifiers such as support vector machines, naïve Bayes, and logistic regression, offered the best performance in the dialect identification tasks (e.g., Wu et al., 2019; Jauhiainen et al., 2019; Camposampiero et al., 2022). Another popular option has been to fine-tune BERT for classification tasks (e.g., Zaharia et al., 2020; Bengoetxea et al., 2025). Related to the languages concerned in this paper, Hämäläinen et al. (2021) train a text only and text+audio dialect classifiers on the same Finnish dataset that we use. However, they split their data based on utterances only, corresponding to the speaker dependent set of our work (see Section 3.1), and use a more fine-grained dialect division.

Dialect identification from audio files has gained more interest after the release of the large pre-trained audio models. Systems utilizing Whisper (e.g., Elleuch et al., 2025) and wav2vec2.0 (e.g., Gutscher and Pucher, 2025) for dialect identification have become increasingly popular for instance in the Interspeech conferences. Many works comment on the problems with identifying dialects directly from the original audio and propose different workarounds, such as low-pass filtering and F0 monotonization (Parsons et al., 2025) or voice conversion (Abdullah et al., 2025; Fischbach et al., 2025), as well as model modifications (Luo and Zhou, 2023). Kakouros and Hiovain-Asikainen (2023) present results on North Sami dialect identification, and also experiment with splitting their data into speaker dependent and speaker independent partitions, which is also a part of this study (see Section 3.1.)

2.2 Automatic Speech Recognition on Dialects

Most works on automatic speech recognition (ASR) focus on producing standard language text, even if the spoken language would be non-standard. This is a natural goal, given that many downstream applications need commands in the standard language. There is a broad field of studies on making dialectal speech automatically recognized to the standard language (e.g., Plüss et al., 2022; Miwa and Kai, 2023; Lin et al., 2024).

Another popular direction in automatic speech recognition has been automatic phoneme recognition, aiming to train (universal) systems that recognize the phonemes in speech and output corresponding IPA symbols (e.g., Li et al., 2020, 2022; Glocker et al., 2023). Automatic dialectal transcription can be characterized by being somewhere in between the standard language ASR and phoneme recognition: transcriptions aim to be phonetically precise, but are often language-specific and might not make some distinctions that are not relevant for the language (e.g., sibilants in Finnish are transcribed as /s/ irrespective of their true phonetic nature). The difference between dialectal transcriptions and standard languages in our datasets is presented in Table 1.

Works that aim to automatically transcribe speech to this domain instead of the standard language or the phoneme level are scarce. Suwanbandit et al. (2023) release a dataset of Thai dialects with transcriptions and translations to standard Thai, and further report on ASR experiments with the dataset. Kuparinen (2025) trains automatic dialectal transcription models on the same Finnish and Norwegian datasets that are described

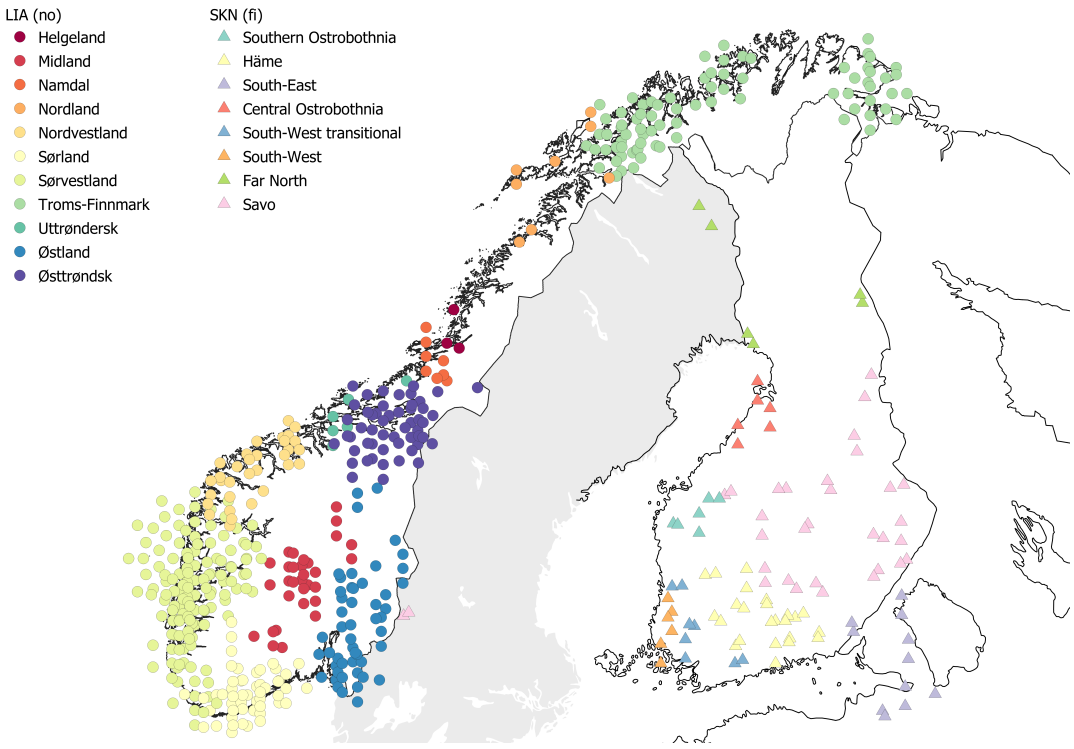

Figure 1: The speakers of both datasets on a map with dialect information as color. Norwegian speakers are presented with circles and Finnish speakers with triangles. If there are multiple speakers per location, the points are dislocated (e.g., Bergen on the West coast of Norway). In the Norwegian classification, Namdal and Uttrøndersk are concatenated as well as Helgeland and Nordland. Map is made with QGIS.

in this work, but only utilizes the speaker dependent data partition. Some works also aim to produce both standard language text and dialectal text and compare the systems (Nigmatulina et al., 2020; Blaschke et al., 2025). Since dialectal transcriptions tend to have varying styles and quality, the task can be compared to low resource ASR, where data volume is likewise low and available texts might not be highly standardized.

2.3 Audio Modification

A typical way to mitigate speaker bias in speech modeling is augmenting the audio with modifications, such as changing the pitch of the voice or incorporating time or frequency masking. Speed perturbation (Ko et al., 2015) and SpecAugment (Park et al., 2019) are well known systems designed to augment the feature inputs with modifications of speech, pitch, and frequency.

Augmentations have been found to be beneficial, for example, in automatic speech recognition of dysarthric speech (Bhat et al., 2022), emotion recognition (Wu and Lee, 2023) and language assessment (Lun et al., 2024). Ullah et al. (2024) experiment with different augmentation techniques for pre-training low resource speech models, and find that the combination of noise and pitch shifting offers best performance. We use similar augmentations in this work.

Besides augmentation of the speech features, another option for speaker bias mitigation has been the synthetic creation of new voices with systems such as HiFi-GAN (Kong et al., 2020) or using voice conversion. Voice conversion aims to transform a spoken sequence to the voice characteristics of another speaker. In essence, it makes an utterance sound as if it was spoken by someone else.

Voice conversion has been utilized, for instance, in automatic speech recognition (Casanova et al., 2023) and keyword recognition (Wubet and Lian, 2022). It has been found to be helpful in low-resource settings, where the number of natural speakers tends to be small (Baas and Kamper, 2022). This is similar to many dialectal datasets, where each dialect is often presented by only a handful of speakers. Abdullah et al. (2025) present results on Arabic dialect identification, comparing the effects of different modifications on the original audio. They show that both data augmentation

and voice conversion enhance performance both in domain and out of domain. Given their positive results, our setup follows their work.

3 Data

We use dialectal datasets of two unrelated languages, namely Finnish and Norwegian.[1] For Finnish, we use the Samples of Spoken Finnish (Institute for the Languages of Finland, 2014), which includes 99 interviews from 50 Finnish-speaking locations that present all of the Finnish dialects. The interviews were mostly recorded in the 1960s, and have since been digitized, annotated and transcribed phonetically using the Uralic phonetic alphabet. We filter the dataset to only include audio segments with a duration between 1 and 20 seconds. The filtered data includes approximately 65 hours of audio from the dialect speakers. We use the dialect division put forth by Itkonen (1989) with 8 dialect areas. The interview locations and their traditional dialect area are presented in Figure 1.

For Norwegian, we use the LIA Norwegian speech corpus (Norwegian University of Science and Technology et al., 2019), which is a joint effort of four Norwegian universities, aiming to collect the old dialectal interviews into one corpus (Hagen et al., 2021). We only use the interviews that were labeled as free talk (*fritale*) and exclude the ones that had other tasks or a special focus (such as place names). After the filtering, the dataset includes 465 speakers from 159 locations, and around 173 hours of audio from the dialect speakers. The interviews are transcribed phonetically using the Dano-Norwegian alphabet with # as a pause marker. We use the same duration limit for Norwegian as for Finnish. Furthermore, for the computationally heavy task of automatic transcription, we only use half of the available Norwegian data to make training possible with our resources.

We use the dialect division presented by Skjekkeland (1997)[2]. The dataset is unbalanced on dialects however, with 137 speakers of the South-Western (Sørvestland) dialects and only three speakers from Helgeland. We thus combine some of the smaller dialect areas based on their top-level dialect: Namdalsk and Uttrøndersk (both part of the Trøndersk dialects), and Nordlandsk and Helgelandsk (both part of the Northern dialects). The interview locations and their dialect areas are presented in Figure 1.

3.1 Data Splitting

In natural language processing tasks, (dialectal) data are often split into training, development, and test sets by splitting interviews in proportions (e.g., 80% to training, and 10% to both development and test sets). Since all speakers appear in all data splits, the tasks might be easier than actual use cases in the wild. This is especially true for audio tasks, where models learn speaker-specific traits, and can use these traits as a shortcut to the actual task (e.g., dialect identification). We will construct a data split with the described basic setup, with each speakers' utterances divided 80/10/10, which will be called a **speaker dependent** set (following Kakouros and Hiovain-Asikainen, 2023).

As a comparison, we also construct a **speaker independent** setup. In this setup, we split the data into training, development, and test sets on full interviews, assigning 80% of interviews to training, 10% to development, and 10% to testing (see also Kuparinen et al., 2023). Furthermore, as one of our tasks is dialect identification, we make sure that all dialects are always presented in all of the data splits, assigning interviews to the sets based on the amount of speakers per dialect as presented in Figure 1. We will also do further augmentation and voice conversion on this data split, described in Sections 3.2 and 3.3.

3.2 Audio Augmentation

For audio augmentation, we use both pitch shift and additional noise. Using torchaudio (Hwang et al., 2023), we take the original audio sample and randomly shift the pitch of the voice with either -4, -2, +2, or +4 steps. Furthermore, we add low level noise to the pitch shifted waveform matching the duration of the signal. Thus, the original audio will have a slightly different pitch than before, and the whole segment will have added noise on the background. To keep the setup relatively simple, we do not use any further augmentations (such as frequency or time masking). Both pitch shift and noise were found to be beneficial for Arabic dialect identification by Abdullah et al. (2025).

[1]The segmented data are available at https://huggingface.co/collections/okuparinen/dialectal-transcription-fi-no.

[2]We utilize the mapping of municipality names and dialect areas by Phoebe Parsons, available at https://www.nb.no/sprakbanken/en/resource-catalogue/oai-nb-no-sbr-92/.

Setup	Training	Train size	Orig.	Aug.	VC	Test
Speaker dependent						
Original	all speakers	N	✔	✘	✘	all speakers
Voice per dialect	all speakers	N	✘	✘	✔	all speakers
Speaker independent						
Original	80% of speakers	N	✔	✘	✘	10% of sp.
Orig. + Pitch shift and noise	80% of speakers	2 x N	✔	✔	✘	10% of sp.
Orig. + VC with one voice	80% of speakers	2 x N	✔	✘	✔	10% of sp.
Orig. + VC with four voices	80% of speakers	5 x N	✔	✘	✔	10% of sp.

Table 2: Summary of the different data splits and operations on the training set. The first two setups serve as benchmarks for speaker bias in speech models. N=the size of original training utterances. Orig.=original audio, Aug.=audio augmentation with pitch shifts and noise, VC=voice conversion of the original audio. Sp. = speakers. Development set has the same setup as the test set. Exact sizes of the different sets are presented in Appendix A.

3.3 Voice Conversion

We use nearest neighbor voice conversion (kNN-VC) presented in Baas et al. (2023).[3] The approach utilizes reference segments from another speaker to convert the original voice, while still maintaining the dialectal content. As reference speakers, we use external dialectal datasets: for Finnish, we utilize the Finnish Dialect Corpus of the Syntax Archive (University of Turku and Institute for the Languages of Finland, 2021), and for Norwegian we use readings of the North Wind and the Sun in different dialects.[4]

For our voice conversion setups, all of the training samples are conversed to the voice of another speaker. These voice conversed training sets are then concatenated with the original training samples, effectively multiplying the size of the training data (as is done with the augmented data as well). We experiment with one voice (training data twice the size of the original) and with four voices (training five times the size of the original).

As a final experiment, we also construct a voice conversed version of the speaker dependent dataset, where each dialect is represented by exactly one voice. This setup is called **voice per dialect** and is designed to highlight what happens when speaker traits and dialect traits are combined. If our assumption is correct, and the models use speaker identification as a shortcut to dialect identification, classifiers trained on the voice per dialect set should perform poorly when tested on new voices. The different data setups are summarized in Table 2. The development and test sets are composed of the original speech samples throughout the experiments without any modifications.[5]

4 Task Setup

To evaluate the effect of speaker bias in dialectological speech modeling, we set up two different dialect-focused tasks: dialect identification and automatic dialectal transcription. As the point of the work is not in achieving best possible results for each task, we do not experiment with different base architectures and hyperparameters. Instead, we use wav2vec2.0 (Baevski et al., 2020) based models with basic settings for both tasks. For Finnish, we use the base model trained with 150,000 hours of Finnish speech[6], including 2740 hours of colloquial Finnish (Getman et al., 2025). Such large base models were not available for Norwegian, which is why we use the base version of the multilingual MMS model[7] (Pratap et al., 2023). We further report MMS results of the Finnish speaker independent set with original audio to facilitate base model comparison.

4.1 Dialect Identification

For the dialect identification task, we extract embeddings of our audio samples encoded by the base models. As it is known that the different layers of the models encode different aspects of the speech signal, we experimented with the first layer, middle layers (6 for Finnish, 6 and 12 for Norwegian), and

[3]The implementation is available at https://github.com/bshall/knn-vc.

[4]Available at https://www.hf.ntnu.no/nos/.

[5]The code for the paper is available at: https://github.com/okuparinen/DialectSpeakerBias.

[6]https://huggingface.co/GetmanY1/wav2vec2-base-fi-150k

[7]https://huggingface.co/facebook/mms-300m

the final layer (12 for Finnish, 24 for Norwegian) for the embedding extraction. We found that layer number 6 offers best performance in our classification task for both languages. We thus train our final classifiers on the embeddings extracted from this layer. A comparison of the different layers is presented in Appendix B.

The embeddings have dimensions of N x 768 for the Finnish model, and N x 1024 for the MMS model, where N is the number of 25 ms frames in the utterance. Since the utterances have differing length, we aggregate the embeddings over the utterances. We take the mean and standard deviation over the full utterance (resulting in 1536 dimensions per utterance for Finnish and 2048 dimensions for Norwegian), and use these utterance-based embeddings to train a linear support vector machine classifier.

For our classifier, we first scale the data for zero mean and unit variance before optionally applying principal component analysis on the scaled data. The data is then fed to a linear, class-balanced SVM using one-vs-rest classification strategy. We experiment with the input embedding dimensions (using mean and standard deviation, or only mean or only standard deviation), as well as with using PCA or not using it. For PCA, we also experiment with the number of components (128 or 256 for Finnish and 128, 256 or 512 for Norwegian), and select the best design for each dataset. As a text comparison, we train a similar SVM model with a tf-idf vectorizer on character n-grams of 2 to 4 characters, without applying PCA, on the manual transcriptions.

We use the classifier to predict dialect labels for the development and test sets based on their embedding representations. We evaluate the results on macro F1 and accuracy. We also report the 95% confidence intervals for the results.[8]

4.2 Automatic Dialectal Transcription

We finetune the base models with our transcribed dialectal data for a maximum of 15 epochs (early stopping of 10 epochs) with connectionist temporal classification (CTC) loss. We freeze the feature extractor before finetuning and use a learning rate of 5-e4 for the smaller Finnish dataset and 1e-4 for the larger Norwegian dataset. The finetuning is done with the Huggingface Transformers toolkit (Wolf et al., 2020). We evaluate the systems on character error rate (CER), which in transcription text is more informative than word error rate.[9]

5 Results

5.1 Dialect Identification

The results for the dialect identification experiments are presented for both languages in Table 3. The first rows of the tables have the same data structure (each interview split into training, development, and test sets based on utterances), as do the bottom rows (each interview appears in exactly one set).

Starting with the speaker dependent setup, very clear evidence of speaker bias can be observed. Using only the original audio, the scores are very high, indicating that the classifier learns to connect the speaker traits to a dialect in the training phase. This effect becomes evident when comparing it to the voice per dialect set, which uses the exact same data split but with converted voices for each dialect. When evaluated against the same test set, the classification performance collapses as the classification model has learned to connect the speakers to the dialects, but is then faced with unheard voices. The text baseline is solid for Finnish, but worse than the audio based model. For Norwegian, the text model is far behind the audio model.

It is also noticeable that for the speaker dependent set, both languages have very similar performance in the audio models: around 90% accuracy scores for the original audio and around 17% for the voice per dialect setup. This is not true for the speaker independent setup, where Norwegian has much higher scores throughout. This also indicates that, for the speaker dependent setup, the linguistic factors are not as important as the speaker traits for the classification. In essence, the model is more of a speaker recognition system than a dialect identification system.

For the speaker independent data split, we see a performance drop of around 60 points on original audio for Finnish, and around 40 points for Norwegian. This further indicates there is a major speaker related effect in the classification. For the text based models, there is also a considerable effect of the data split change for Finnish, but not so much for Norwegian. For Finnish, the text based model clearly outperforms audio, whereas for Norwegian the audio models are slightly better.

[8] Calculated with https://github.com/luferrer/ConfidenceIntervals with 1000 bootstrap sets.

[9] We use the implementation provided in https://github.com/nsmartinez/WERpp.

SKN (fi)		
Setup	**Macro F1 ↑**	**Accuracy ↑**
Speaker dependent		
Transcription text	$74.85^{73.10–76.53}$	$77.83^{76.48–79.12}$
Original audio	$88.77^{87.44–90.04}$	$89.79^{88.84–90.73}$
Voice per dialect	$16.07^{14.72–17.41}$	$17.01^{15.75–18.28}$
Speaker independent		
Transcription text	$61.88^{60.32–63.51}$	$65.23^{63.74–66.70}$
Original audio	$24.29^{23.06–25.52}$	$33.38^{31.83–34.90}$
Orig. + Pitch shift and noise	$27.55^{26.10–28.89}$	$32.85^{31.25–34.40}$
Orig. + VC with one voice	$21.05^{19.81–22.26}$	$26.85^{25.49–28.37}$
Orig. + VC with four voices	$\mathbf{29.63}^{28.21–30.98}$	$\mathbf{38.09}^{36.65–39.64}$

LIA (no)		
Setup	**Macro F1 ↑**	**Accuracy ↑**
Speaker dependent		
Transcription text	$53.01^{51.89–54.05}$	$58.91^{57.98–59.76}$
Original audio	$91.69^{91.04–92.30}$	$92.56^{92.09–93.01}$
Voice per dialect	$12.58^{11.95–13.19}$	$17.45^{16.75–18.18}$
Speaker independent		
Transcription text	$47.17^{46.03–48.25}$	$54.30^{53.34–55.24}$
Original audio	$50.47^{49.59–51.37}$	$64.52^{63.55–65.47}$
Orig. + Pitch shift and noise	$\mathbf{51.65}^{50.89–52.44}$	$64.97^{64.12–65.90}$
Orig. + VC with one voice	$51.32^{50.58–52.15}$	$\mathbf{65.32}^{64.39–66.21}$
Orig. + VC with four voices	$49.31^{48.55–50.07}$	$62.91^{61.93–63.88}$

Table 3: Macro F1 and accuracy scores for the classification task with 95% confidence intervals in superscript. The top table shows the SKN (Finnish) results, and the bottom table shows the LIA (Norwegian) results. Per-chance accuracy is 12.5 for Finnish and 11.11 for Norwegian. For comparison, using the MMS model for the Finnish speaker independent set with original audio achieves a macro F1 score of $27.05^{25.81–28.19}$ (+2.79 vs. monolingual). Results on the development set and model details (embedding dimensions, PCA parameters) are provided in Appendix B.

Setup	**SKN (fi)**	**LIA (no)**
Speaker dependent		
Original audio	$9.65^{9.43–9.87}$	$19.30^{19.02–19.55}$
Voice per dialect	$16.81^{16.47–17.19}$	$29.38^{29.04–29.72}$
Speaker independent		
Original audio	$\mathbf{15.89}^{15.46–16.31}$	$20.21^{19.91–20.50}$
Orig. + Pitch shift and noise	$16.43^{16.00–16.83}$	$\mathbf{19.53}^{19.31–19.87}$
Orig. + VC with one voice	$16.35^{15.91–16.76}$	$20.17^{19.88–20.46}$
Orig. + VC with four voices	$17.55^{17.16–17.95}$	$20.99^{20.71–21.26}$

Table 4: Character error rate % (↓) results of the automatic transcription on the different data splits with 95% confidence intervals in superscript. The MMS model finetuned for the Finnish speaker independent set with original audio achieves a character error rate of $19.45^{19.08–19.85}$ (+3.56 vs. monolingual).

The performance of the audio models is poor for Finnish across the board, whereas for Norwegian the models fare better. Regarding the audio modifications, there is a lack of a consistent and meaningful effect. For both languages, the augmentation with pitch shifts and noise offers a tiny performance gain in macro F1 over the original audio, whereas the voice conversions seem unstable. For Finnish, the model with four additional voices is clearly the best of the audio models, but for Norwegian it is the worst. Likewise the system with one additional voice is the best model for Norwegian in terms of accuracy but clearly the worst for Finnish.

The classification results are very different from the ones reported by e.g., Abdullah et al. (2025), who present a consistent performance gain from augmentation and voice conversion. Possible causes for this are that we are working with more dialect classes (8 and 9 vs. 5) and older data with possibly inconsistent quality, but the lack of positive results is still surprising.

5.2 Automatic Dialectal Transcription

The results for the automatic dialectal transcription task are presented in Table 4. For the speaker dependent models, the original audio still offers better performance than the voice per dialect system. For Finnish, the difference is not as large as for the classification task, however. In fact, the voice per dialect system is only slightly worse than the original audio based system in the speaker independent split for Finnish (although the speaker dependent split is easier in that the same topics appear in both training and testing, even if the voices are different).

For Norwegian, the difference between the original audio and voice per dialect systems in the speaker dependent set is very large. Interestingly, the difference between the dependent and independent sets is quite small (0.91 points in CER % for the original audio). For Finnish the difference is larger, but still not as big as for the classification task. This indicates that automatic transcription is not as dependent on the speakers as the classification.

In terms of character error rate, the modifications of the original audio are consistently harmful for the transcription quality in the Finnish experiments. Adding more voices and thus more training data with the same transcriptions seems to make the models worse, possibly overfitting to the training data. For Norwegian, however, the scores are similar for all versions, but the augmented data offers best results. This was also the case in the classification task.

Comparing the two languages, results for Finnish are consistently better than for Norwegian. This is most likely a result of at least two causes: 1) for Finnish, we could use a language-specific base model but for Norwegian we had to use a multilingual one[10], and 2) for the Norwegian data, we noticed inconsistent quality in both the audio and the transcriptions, most likely resulting from the fact that the corpus is collected from multiple sources. This is also reflected in the transcription-based classification, where the Finnish models are performing better than the Norwegian ones, indicating possible variation in the Norwegian transcription quality. Finally, in the audio classification the results were converse: Finnish models are much worse than Norwegian. This could be a result of clearer dialectal differences in the Norwegian data, which makes identification easier, but building a unified transcription model harder.

5.3 Classifying on Automatic Transcriptions

As a final experiment, we analyze if the classification scores for Finnish can be elevated by classifying the utterances based on the automatically created transcriptions presented in Section 5.2. We use the text models trained on manual transcriptions for the original classification experiments in Section 5.1, but infer on the automatically created transcriptions of the test set. We only analyze the results on the Finnish data, as the Norwegian audio models outperformed the text based systems. The results are presented in Table 5.

The classification on the automatic transcriptions enhances performance dramatically with macro F1 scores around 25 points higher compared to the audio based classification. Even the highly speaker dependent system of voice per dialect achieves a macro F1 score of 45.82. In the speaker independent setup, the ASR model based on original audio was the best in terms of character error rate, but the worst in terms of classification F1 (albeit the differences are small for both cases). In conclusion, the text based models massively outperform audio in speaker independent setups for Finnish, even if the text is automatically created.

[10] Using MMS for Finnish enhanced performance in the classification task, but worsened it in the transcription task, following the Norwegian results.

Setup	Audio	ASR
Speaker dependent		
Original audio	$88.77^{87.44–90.04}$	$62.86^{60.96–64.59}$
Voice per dialect	$16.07^{14.72–17.41}$	$45.82^{44.06–47.49}$
Speaker independent		
Original audio	$24.29^{23.06–25.52}$	$48.10^{46.53–49.72}$
Orig. + Pitch shift and noise	$27.55^{26.10–28.89}$	$\mathbf{50.53}^{48.99–52.12}$
Orig. + VC with one voice	$21.05^{19.81–22.26}$	$49.94^{48.30–51.59}$
Orig. + VC with four voices	$29.63^{28.21–30.98}$	$49.66^{48.12–51.22}$
Manual transcription		
Speaker dependent	$74.85^{73.10–76.53}$	
Speaker independent	$61.88^{60.32–63.51}$	

Table 5: Macro F1 scores ↑ for Finnish SVM classifiers with 95% confidence intervals in superscript. On the left, we present the results for classifiers trained and evaluated on audio embeddings, and on the right, the models trained on manual transcriptions and evaluated on automatic transcriptions. The corresponding classification results for the manual transcriptions are presented on the bottom. The Norwegian results are omitted, as the audio models already outperformed the manual transcription based systems.

6 Conclusion

In this paper, we have shown how speaker bias severely affects audio modeling in dialect identification and to a lesser extent in automatic dialectal transcription, with data from two unrelated languages. The speaker dependent setup highlighted how audio models shortcut to speaker recognition instead of dialect identification: using original audio in training achieved excellent scores for both languages, whereas using the same data split with a single voice per dialect ended in collapse.

We further aimed to mitigate the effects of bias with traditionally used techniques in audio augmentation and voice conversion. While there were some positive effects, the overall usefulness of these methods on the tasks remained negligible. This is in contrast to previous findings on, for instance, Arabic dialect identification (Abdullah et al., 2025). Especially for Finnish, dialect identification proved to be difficult for all models, and classification on transcriptions provided far better results. Conversely for Norwegian, identification from audio outperformed text, but automatic dialectal transcription performed worse than for Finnish.

This work has focused on speaker bias mitigation solutions that are applied on the input (i.e., the waveform itself). More elaborate systems could also be applied post-hoc, by targeting the embedding dimensions that hold the most information on speaker traits, and filtering or down-weighting such dimensions. Thebaud et al. (2024) use Integrated Gradients to trace which phonemes affect speaker recognition the most, but a similar system could be applied also to trace speaker effects from model embeddings. Zhu et al. (2025) train ECAPA-TDNN (Desplanques et al., 2020) based speaker embeddings and use SHAP values to trace the speaker-affected dimensions from the content embeddings of different self-supervised models. They further experiment how filtering the speaker information affects ASR accuracy. Systems based on explainability methods could thus provide interesting possibilities for dialectal audio modeling as well, but they are beyond the scope of this paper and thus left for future work.

Limitations

The building of the speaker independent set is a possible source of variation. As the data splits are constructed from full interviews, the scores are highly affected by the interviews chosen to the development and test sets. A possible way to undermine this effect would be to build several folds of the data split. We restricted our experiments to one fold due to resource limitations, but included the 95% confidence intervals to show possible fluctuation.

This work focuses on two languages spoken in the Nordic countries. Even though the languages represent different families, the datasets themselves are largely collected following similar dialectological and cultural ideologies.

Acknowledgments

This work is supported by the Research Council of Finland through project No. 360356 "Speech as Speech – Acoustic Modeling in Variational Linguistics". The author also wishes to acknowledge CSC – IT Center for Science, Finland, for computational resources.

References

Badr M. Abdullah, Matthew Baas, Bernd Möbius, and Dietrich Klakow. 2025. Voice Conversion Improves Cross-Domain Robustness for Spoken Arabic Dialect Identification. In *Interspeech 2025*, pages 2790–2794.

Noëmi Aepli, Çağrı Çöltekin, Rob Van Der Goot, Tommi Jauhiainen, Mourhaf Kazzaz, Nikola Ljubešić, Kai North, Barbara Plank, Yves Scherrer, and Marcos Zampieri. 2023. Findings of the VarDial Evaluation Campaign 2023. In *Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023)*, pages 251–261, Dubrovnik, Croatia. Association for Computational Linguistics.

Matthew Baas and Herman Kamper. 2022. Voice Conversion Can Improve ASR in Very Low-Resource Settings. In *Interspeech 2022*, pages 3513–3517.

Matthew Baas, Benjamin van Niekerk, and Herman Kamper. 2023. Voice Conversion With Just Nearest Neighbors. In *Interspeech 2023*, pages 2053–2057.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. *Advances in Neural Information Processing systems*, 33:12449–12460.

Jaione Bengoetxea, Mikel Zubillaga, Ekhi Azurmendi, Maite Heredia, Julen Etxaniz, Markel Ferro, and Jeremy Barnes. 2025. HiTZ at VarDial 2025 NorSID: Overcoming Data Scarcity with Language Transfer and Automatic Data Annotation. In *Proceedings of the 12th Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 209–219, Abu Dhabi, UAE. Association for Computational Linguistics.

Chitralekha Bhat, Ashish Panda, and Helmer Strik. 2022. Improved ASR Performance for Dysarthric Speech Using Two-stage DataAugmentation. In *Interspeech 2022*, pages 46–50.

Verena Blaschke, Miriam Winkler, Constantin Förster, Gabriele Wenger-Glemser, and Barbara Plank. 2025. A Multi-Dialectal Dataset for German Dialect ASR and Dialect-to-Standard Speech Translation. In *Interspeech 2025*, pages 913–917.

Giacomo Camposampiero, Quynh Anh Nguyen, and Francesco Di Stefano. 2022. The Curious Case of Logistic Regression for Italian Languages and Dialects Identification. In *Proceedings of the Ninth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 86–98, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Edresson Casanova, Christopher Shulby, Alexander Korolev, Arnaldo Candido Junior, Anderson da Silva Soares, Sandra Aluísio, and Moacir Antonelli Ponti. 2023. ASR Data Augmentation in Low-Resource Settings Using Cross-Lingual Multi-Speaker TTS and Cross-Lingual Voice Conversion. In *Interspeech 2023*, pages 1244–1248.

Bharathi Raja Chakravarthi, Gaman Mihaela, Radu Tudor Ionescu, Heidi Jauhiainen, Tommi Jauhiainen, Krister Lindén, Nikola Ljubešić, Niko Partanen, Ruba Priyadharshini, Christoph Purschke, Eswari Rajagopal, Yves Scherrer, and Marcos Zampieri. 2021. Findings of the VarDial Evaluation Campaign 2021. In *Proceedings of the Eighth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 1–11, Kiyv, Ukraine. Association for Computational Linguistics.

Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck. 2020. ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification. In *Interspeech 2020*, pages 3830–3834.

Haroun Elleuch, Salima Mdhaffar, Yannick Estève, and Fethi Bougares. 2025. ADI-20: Arabic Dialect Identification Dataset and Models. In *Interspeech 2025*, pages 2775–2779.

Lea Fischbach, Akbar Karimi, Caroline Kleen, Alfred Lameli, and Lucie Flek. 2025. Improving Low-Resource Dialect Classification Using Retrieval-based Voice Conversion. In *Interspeech 2025*, pages 2780–2784.

Mihaela Gaman, Dirk Hovy, Radu Tudor Ionescu, Heidi Jauhiainen, Tommi Jauhiainen, Krister Lindén, Nikola Ljubešić, Niko Partanen, Christoph Purschke, Yves Scherrer, and Marcos Zampieri. 2020. A Report on the VarDial Evaluation Campaign 2020. In *Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 1–14, Barcelona, Spain (Online). International Committee on Computational Linguistics (ICCL).

Yaroslav Getman, Tamás Grósz, Tommi Lehtonen, and Mikko Kurimo. 2025. Is Your Model Big Enough? Training and Interpreting Large-Scale Monolingual Speech Foundation Models. In *Interspeech 2025*, pages 231–235.

Kevin Glocker, Aaricia Herygers, and Munir Georges. 2023. Allophant: Cross-Lingual Phoneme Recognition with Articulatory Attributes. In *Proc. Interspeech 2023*.

Lorenz Gutscher and Michael Pucher. 2025. Audio-Based Classification and Geographic Regression of Austrian Dialects. In *Interspeech 2025*, pages 2765–2769.

Kristin Hagen, Gjert Kristoffersen, Øystein A. Vangsnes, and Tor A. Åfarli, editors. 2021. *Språk i arkiva: Ny forsking om eldre talemål frå LIA-prosjektet*. Novus forlag.

Mika Hämäläinen, Khalid Alnajjar, Niko Partanen, and Jack Rueter. 2021. Finnish Dialect Identification: The Effect of Audio and Text. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8777–8783, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jeff Hwang, Moto Hira, Caroline Chen, Xiaohui Zhang, Zhaoheng Ni, Guangzhi Sun, Pingchuan Ma, Ruizhe Huang, Vineel Pratap, Yuekai Zhang, Anurag Kumar, Chin-Yun Yu, Chuang Zhu, Chunxi Liu, Jacob Kahn, Mirco Ravanelli, Peng Sun, Shinji Watanabe, Yangyang Shi, and 5 others. 2023. TorchAudio 2.1: Advancing Speech Recognition, Self-Supervised Learning, and Audio Processing Components for PyTorch. *Preprint*, arXiv:2310.17864.

Institute for the Languages of Finland. 2014. Samples of Spoken Finnish, Downloadable Version.

Terho Itkonen. 1989. *Nurmijärven murrekirja*. Suomalaisen Kirjallisuuden Seuran toimituksia ; 498. Suomalaisen kirjallisuuden seura, Helsinki.

Tommi Jauhiainen, Krister Lindén, and Heidi Jauhiainen. 2019. Discriminating Between Mandarin Chinese and Swiss-German Varieties Using Adaptive Language Models. In *Proceedings of the Sixth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 178–187, Ann Arbor, Michigan. Association for Computational Linguistics.

Sofoklis Kakouros and Katri Hiovain-Asikainen. 2023. North Sámi Dialect Identification with Self-supervised Speech Models. In *Interspeech 2023*, pages 5306–5310.

Tom Ko, Vijayaditya Peddinti, Daniel Povey, and Sanjeev Khudanpur. 2015. Audio Augmentation for Speech Recognition. In *Interspeech 2015*, pages 3586–3589.

Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. 2020. HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis. *Preprint*, arXiv:2010.05646.

Olli Kuparinen. 2025. Automatic Dialectal Transcription: An Evaluation on Finnish and Norwegian. In *Interspeech 2025*, pages 2390–2394.

Olli Kuparinen, Aleksandra Miletić, and Yves Scherrer. 2023. Dialect-to-Standard Normalization: A Large-Scale Multilingual Evaluation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13814–13828, Singapore. Association for Computational Linguistics.

Xinjian Li, Siddharth Dalmia, Juncheng Li, Matthew Lee, Patrick Littell, Jiali Yao, Antonios Anastasopoulos, David R Mortensen, Graham Neubig, Alan W Black, and Metze Florian. 2020. Universal Phone Recognition with a Multilingual Allophone System. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8249–8253. IEEE.

Xinjian Li, Florian Metze, David R. Mortensen, Alan W Black, and Shinji Watanabe. 2022. Phone Inventories and Recognition for Every Language. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1061–1067, Marseille, France. European Language Resources Association.

Jiayan Lin, Shenghui Lu, Hukai Huang, Wenhao Guan, Binbin Xu, Hui Bu, Qingyang Hong, and Lin Li. 2024. MinSpeech: A Corpus of Southern Min Dialect for Automatic Speech Recognition. In *Interspeech 2024*, pages 2330–2334.

Tin Mei Lun, Ekaterina Voskoboinik, Ragheb Al-Ghezi, Tamas Grosz, and Mikko Kurimo. 2024. Oversampling, Augmentation and Curriculum Learning for Speaking Assessment with Limited Training Data. In *Interspeech 2024*, pages 4019–4023.

Qibao Luo and Ruohua Zhou. 2023. Exploring the Impact of Back-End Network on Wav2vec 2.0 for Dialect Identification. In *Interspeech 2023*, pages 5356–5360.

Paul McNamee. 2005. Language Identification: A Solved Problem Suitable for Undergraduate Instruction. *Journal of Computing Sciences in Colleges*, 20(3):94–101.

Shogo Miwa and Atsuhiko Kai. 2023. Dialect Speech Recognition Modeling using Corpus of Japanese Dialects and Self-Supervised Learning-based Model XLSR. In *Interspeech 2023*, pages 4928–4932.

Iuliia Nigmatulina, Tannon Kew, and Tanja Samardzic. 2020. ASR for Non-standardised Languages with Dialectal Variation: the case of Swiss German. In *Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 15–24, Barcelona, Spain (Online). International Committee on Computational Linguistics (ICCL).

Norwegian University of Science and Technology, University of Bergen, University of Oslo, and The Arctic University of Norway. 2019. Lia norsk - korpus av eldre dialektopptak.

Daniel S. Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D. Cubuk, and Quoc V. Le. 2019. SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition. In *Interspeech 2019*, pages 2613–2617.

Phoebe Parsons, Heming Strømholt Bremnes, Knut Kvale, Torbjørn Svendsen, and Giampiero Salvi. 2025. Effects of Prosodic Information on Dialect Classification Using Whisper Features. In *Interspeech 2025*, pages 2785–2789.

Michel Plüss, Manuela Hürlimann, Marc Cuny, Alla Stöckli, Nikolaos Kapotis, Julia Hartmann, Malgorzata Anna Ulasik, Christian Scheller, Yanick Schraner, Amit Jain, Jan Deriu, Mark Cieliebak, and Manfred Vogel. 2022. SDS-200: A Swiss German Speech to Standard German Text Corpus. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3250–3256, Marseille, France. European Language Resources Association.

Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, Alexei Baevski, Yossi Adi, Xiaohui Zhang, Wei-Ning Hsu, Alexis Conneau, and Michael Auli. 2023. Scaling Speech Technology to 1,000+ Languages. *arXiv*.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust Speech Recognition via Large-Scale Weak Supervision. *Preprint*, arXiv:2212.04356.

Martin Skjekkeland. 1997. *Dei norske dialektane : tradisjonelle særdrag i jamføring med skriftmåla*. Høyskoleforlaget, Kristiansand.

Artit Suwanbandit, Burin Naowarat, Orathai Sangpetch, and Ekapol Chuangsuwanich. 2023. Thai Dialect Corpus and Transfer-based Curriculum Learning Investigation for Dialect Automatic Speech Recognition. In *Interspeech 2023*, pages 4069–4073.

Thomas Thebaud, Gabriel Hernández, Sarah Flora Samson Juan, and Marie Tahon. 2024. A Phonetic Analysis of Speaker Verification Systems through Phoneme selection and Integrated Gradients. In *The Speaker and Language Recognition Workshop (Odyssey 2024)*, pages 59–66.

Asad Ullah, Alessandro Ragano, and Andrew Hines. 2024. Reduce, Reuse, Recycle: Is Perturbed Data Better than Other Language Augmentation for Low Resource Self-Supervised Speech Models. In *Interspeech 2024*, pages 77–81.

University of Turku and Institute for the Languages of Finland. 2021. The Finnish Dialect Corpus of the Syntax Archive, Downloadable Version.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Nianheng Wu, Eric DeMattos, Kwok Him So, Pin-zhen Chen, and Çağrı Çöltekin. 2019. Language Discrimination and Transfer Learning for Similar Languages: Experiments with Feature Combinations and Adaptation. In *Proceedings of the Sixth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 54–63, Ann Arbor, Michigan. Association for Computational Linguistics.

Ya-Tse Wu and Chi-Chun Lee. 2023. MetricAug: A Distortion Metric-Lead Augmentation Strategy for Training Noise-Robust Speech Emotion Recognizer. In *Interspeech 2023*, pages 3587–3591.

Yeshanew Ale Wubet and Kuang-Yow Lian. 2022. Voice Conversion Based Augmentation and a Hybrid CNN-LSTM Model for Improving Speaker-Independent Keyword Recognition on Limited Datasets. *IEEE Access*, 10:89170–89180.

George-Eduard Zaharia, Andrei-Marius Avram, Dumitru-Clementin Cercel, and Traian Rebedea. 2020. Exploring the Power of Romanian BERT for Dialect Identification. In *Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 232–241, Barcelona, Spain (Online). International Committee on Computational Linguistics (ICCL).

Xiaoxu Zhu, Junhua Li, Aaron J. Li, Yiming Ren, and Baoxiang Li. 2025. Speaker Disentanglement of Speech Pre-trained Model Based on Interpretability. *Preprint*, arXiv:2507.17851.

A Data Size

Table 6 presents the data sizes in whitespace separated tokens and audio duration in hours and minutes. Because the speaker independent sets consist of full interviews, they show more variation in size than the speaker dependent sets that are split on utterances.

B Development Set Results

Table 7 presents the results of the classification task on the different layers of the development set. The results clearly show that layer 6 offers best performance for all setups, and it was thus used for the final classification.

Table 8 presents the development set results for the classification task, as well as the inputs for the SVM model (using the mean and standard deviation (std), or just mean or just standard deviation of the utterance embeddings). If PCA was used, the number of components is also presented. The development set results are broadly 5–10 points higher than the test set results for Finnish, and around 5 points higher for Norwegian. For the development set, the setup with original audio and 4 voices achieves the best score of the speaker independent set for both languages.

Table 9 presents the development set results for the automatic dialectal transcription task. For

		Train		Dev.		Test	
Dataset	**Split**	**Tokens**	**Audio**	**Tokens**	**Audio**	**Tokens**	**Audio**
SKN (fi)	Dependent	498,587	52:02	62,907	06:35	62,356	06:33
	Independent	489,168	51:12	64,031	06:36	70,651	07:23
LIA (no)	Dependent	1,562,453	138:16	194,773	17:19	200,969	17:53
	Independent	1,575,923	140:04	206,627	17:59	175,645	15:24

Table 6: Statistics of the used data. Audio duration in hh:mm. Dev. = development set.

	SKN (fi)			LIA (no)		
Setup	**L1**	**L6**	**L12**	**L1**	**L6**	**L12**
Speaker independent						
Original audio	20.34	**32.40**	23.16	42.33	**53.01**	47.44
Orig. + Pitch shift and noise	21.06	**32.43**	21.97	45.85	**54.08**	47.72
Orig. + VC with one voice	24.31	**33.82**	24.89	45.56	**54.68**	47.83
Orig. + VC with four voices	25.87	**34.88**	25.70	43.07	**54.87**	44.97

Table 7: Layer-wise macro F1 scores in the classification task for the development set. L = layer.

	SKN (fi)		LIA (no)	
Setup	**SVM**	**Macro F1 ↑**	**SVM**	**Macro F1 ↑**
Speaker dependent				
Transcription	–	74.15	–	52.77
Orig. audio	mean+std	90.04	mean+std	92.26
Voice per dialect	std, PCA 256	15.55	std, PCA 528	13.04
Speaker independent				
Transcription	–	66.85	–	49.39
Orig. audio	mean, PCA 256	32.40	mean+std	53.01
Orig. + Aug.	std	32.43	mean+std	54.08
Orig. + VC1	mean, PCA 128	33.82	mean+std	54.68
Orig. + VC4	mean+std	**34.88**	mean+std	**54.87**

Table 8: Macro F1 scores on the development set in the classification task, as well as the SVM model inputs for the best model. The number of PCA components are presented if PCA was used.

Finnish, the results are very slightly better than for the test set, whereas for Norwegian the test set results are better by around 2 points across the speaker independent set.

Setup	**SKN (fi)**	**LIA (no)**
Speaker dependent		
Original audio	9.74	19.37
Voice per dialect	16.75	29.75
Speaker independent		
Original audio	**15.00**	22.00
Orig. + Aug.	15.26	**21.51**
Orig. + VC1	15.42	22.12
Orig. + VC4	16.99	23.01

Table 9: Character error rate % (↓) results of the automatic transcription on the development set.