



Effects of prosodic information on dialect classification using Whisper features

Phoebe Parsons¹, Heming Strømholtholt Bremnes¹, Knut Kvale², Torbjørn Svendsen¹, Giampiero Salvi¹

¹Department of Electronic Systems, NTNU, Norway

²Telenor Research and Innovation, Telenor, Norway

{phoebe.parsons, heming.s.bremnes, torbjorn.svendsen, giampiero.salvi}@ntnu.no,
knut.kvale@telenor.com

Abstract

In dialect identification (DID), a model needs to attend to subtle cues to distinguish between highly similar linguistic variants. However, the knowledge of which cues are important and why is limited. Inspired by the literature on human DID, we fine-tuned a Whisper model with modified audio to see how deprivation of various signal components would impact performance. Specifically, the audio manipulation sought to either isolate or remove (tonal) prosodic information, by either low-pass filtering or monotonizing F_0 , respectively.

Results indicate that fine-tuning on low-pass filtered data produces a significant improvement over unmodified data. Utilizing sensitivity maps in the frequency domain, we argue that the low-pass model is able to devote more attention to lower frequency bands, thus exploiting task-relevant pitch dynamics. Though only evaluated with Norwegian, we suggest that our methodology should generalize, encouraging improvement in DID and its downstream applications.

Index Terms: dialect classification, sensitivity maps, Whisper, Norwegian

1. Introduction

Dialect identification (DID) exists between language identification (LID) and speaker identification (SID). Whereas LID may utilize the varying phonemic inventories between languages to make decisions, the differences between dialects are often much more subtle. Conversely, SID often aims to discriminate between speakers with similar, or identical, phonemic inventories and must instead look for features unique to the individual. Dialect classification must remain above the level of individual variability, while still finding the often subtle patterns that define a dialect. Systems able to perform DID have been desirable for many years, often with the goal of running a more specialized downstream model [1]. Approaches span from hidden Markov models and support vector machines [2, 3, 4] to more recent work utilizing representations from large multi-lingual foundation models. In [5], in addition to more traditional acoustic features such as mel-frequency cepstral coefficients (MFCC), the authors also use representations from HuBERT, WavLM, and wav2vec 2.0's XLS-R. They then use a simple feed-forward network to perform classification on the acoustic features. Of the acoustic features tested, representations from the multi-lingual models result in the best classification performance. The authors of [6] use a similar approach (the addition of feed-forward classification layers to a foundation model), but instead uses the encoder from the Whisper [7] model. They, too, find good performance with such an architecture. In both [8, 9], the authors propose methods to improve training (parameter efficient training and a more robust back-

end, respectively) to further exploit the benefits of the representations in the foundation models.

Prosody is a feature that is known to vary between dialects in many languages [10], and numerous studies have found reliance on prosodic cues in dialect classification by humans [11, 12, 13]. Previous studies looking at prosody in relation to dialect classification, such as [1, 3, 14], have often utilized specific prosody-related features (e.g., fundamental frequency contours, syllable duration, etc.) as input to their models. Nevertheless, little is known about how, if at all, transformer classifiers use prosodic information to classify dialects. The authors in [15] attempt to probe what phonemic and prosodic knowledge is contained in different layers of a wav2vec 2.0 model. They investigate prosody by attempting to predict word-level prominence and boundaries using representations from various network layers. In [16], the authors utilize the representations from wav2vec 2.0 and Whisper to classify the vocal intensity in audio into one of four levels. However, to the authors' knowledge, this paper is the first time Whisper embeddings have been utilized to examine the impact of pitch specifically.

In this paper, we evaluate the performance impacts of fine-tuning a model for Norwegian dialect classification using audio manipulated to either enhance or reduce reliance on prosodic information. More explicitly, we do the following:

- Utilize the Whisper feature encoder to train a 4-way dialect classification system.
- Train and evaluate models using data modified in two different ways (low-passed and monotonized) to focus the model on specific pitch features.
- Use sensitivity maps in the frequency domain to help to explain model performance.

2. Methods

To understand both how well transformers can perform DID, as well as how modified audio can impact them, we fine-tuned a version of Whisper for dialect classification.

To analyze the effects of prosody, we (1) low-pass filtered the audio to preserve prosodic contours and remove phonemic information, and (2) monotonized the audio to do the opposite. Models were then trained and tested on all three audio types.

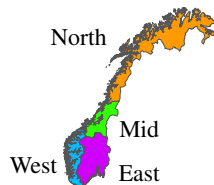
To help us understand which features the model is using to make classification decisions, we generated sensitivity maps on the inputs. We found the two dimensional map (that is, spectrogram-like) to be sparse and hard to interpret. Therefore, as our work is concerned with the effects of prosody on classification, we summed the gradients along the time axis to display the cumulative attention paid to different frequencies by different models for different audio types.

Table 1: Amount of data available for each data split of the SSC. Quantities in number of segments (*seg*), duration (*dur*) of speech in hours, and number of unique speakers (*spk*).

	train			validation			test		
	seg	dur	spk	seg	dur	spk	seg	dur	spk
East	28657	204	98	2470	18	13	2582	19	12
West	23554	169	64	3649	26	9	3870	28	8
Mid	17586	127	28	2298	16	4	1734	13	4
North	17092	123	27	2030	14	4	2361	16	3

Table 2: Amount of data available for the NVOS full. Quantities in number of segments (*seg*), duration of speech in minutes (*dur*), number of unique speakers (*spk*).

	seg	dur	spk
East	14	8	14
West	17	9	17
Mid	10	5	10
North	12	6	12



3. Experiments

We use Norwegian as a test-case for prosodically-focused DID. While Norwegian dialects can be grouped by several morpho-phonological features [17], Norwegian speakers self-report predominantly relying on prosodic features for dialect classification [18]. There is also some evidence to support the hypothesis that prosody carries relevant information for human dialect classification in Norwegian [19].

Norwegian is a pitch-accent language, meaning that stress is indicated by the onset of a specific tonal pattern [20]. There are two lexically determined pitch-accents in Norwegian, and all pitch contours associated with stressed syllables are modulations of these. Importantly, the realization of these pitch contours is dialect dependent. Thus, the prosodic realizations can be used to group dialects into four major classes: East, Mid, North, and West [21].

3.1. Datasets

We used the Stortinget Speech Corpus (SSC) [22] for fine-tuning and evaluation of our DID models. The SSC is publicly available¹ through the National Library of Norway and consists of approximately 5,000 hours of recordings from the Norwegian parliament. Non-verbatim transcriptions, created from the official parliamentary proceedings, are provided.

For our work, we utilized the approximately 770 hour subset of the SSC created by [23] which includes meta-information about the likely dialect for each speaker. This subset was created to have a balanced number of ~ 30 second speech segments for each of the five dialect regions. However, as the South dialect region does not have its own distinctive prosody (speakers have either Eastern or Western prosody) [21], the Southern speakers were reassigned as either East and West based on their prosody, resulting in the four regions. Amounts of data for each dialect and dataset can be seen in Table 1.

Beyond the SSC, we also used the Nordavinden og Sola (NVOS) (English: "the North Wind and the Sun") dataset² as a small, out-of-domain test set. The NVOS consists of speakers reading the fable of the North Wind and the Sun. The average

¹<https://www.nb.no/sprakbanken/en/resource-catalogue/oai-nb-no-sbr-91/>

²<https://www.hf.ntnu.no/nos/>

duration of each reading is 32 seconds (SD: 5 seconds). Only native Norwegian speakers (53 of the 55 available recordings) were used. Details about data quantities can be seen in Table 2. Participants were given the text ahead of time and encouraged to make changes as needed so it sounded natural in their dialect. Each speaker's dialect is annotated with one of 23 regions from [24]. Using the dialect region mapping³, these regions were converted into the four dialect regions.

3.2. Audio manipulation

For manipulation of all audio files, the fundamental frequency (F_0) was estimated using the REAPER package⁴ and audio manipulation was done in Praat [25].

In human experiments, prosody is normally isolated by low-pass filtering the data at somewhere between 225 and 400Hz [26, 27]. Though we wanted to follow the literature as much as possible, when manually inspecting the results of low-pass filtering at 400Hz, we found that speakers with low average F_0 (\bar{F}_0) were perceived as more intelligible than speakers with high \bar{F}_0 . Therefore, to ensure equal obfuscation across speakers, we wished to dynamically change the cutoff frequency based on the speaker's \bar{F}_0 . Through experimentation and manual inspection, we settled on $420.2 * (1 - e^{-0.0124 * \bar{F}_0})$ to determine the cutoff frequency. The smoothing (transition bandwidth) of the filter was set at one quarter of the \bar{F}_0 .

To flatten the pitch, the audio was monotonized to the \bar{F}_0 of the segment. Monotonization was accomplished in Praat by creating a new, flat pitch tier, and then re-synthesizing the audio with the new tier.

3.3. Model fine-tuning

The `nb-whisper-medium`⁵ was used as our foundation model. This was originally fine tuned from Open AI's Whisper model by the National Library of Norway's AI Lab to perform better on speech recognition tasks for Norwegian [28].

Using the HuggingFace Whisper API, a model was fine-tuned for each type of audio manipulation (e.g., the low-pass model was trained on audio that had been low-pass filtered)⁶. All model trainings used the train and validation sets from the SSC. For our classification task, only the encoder part of the Whisper model was used. Outputs from the encoding layer were mean pooled and then passed through a projection layer to go from 1024 features to 256. Finally, a classifier layer processed the 256 features into our four output classes. The whole model was trained for 3 epochs with a learning rate of $3e^{-5}$ and a device batch size of 16. Training was performed on 6 Nvidia RTX A5000 GPUs and each training took between one hour and fifteen minutes and one hour and thirty minutes.

After training, each model was then tested against the SSC test set for each type of audio manipulation. Models were further tested against the NVOS for each audio manipulation type.

3.4. Sensitivity maps

To generate sensitivity maps, we used the SmoothGrad [29] technique of adding noise to the model's input image (in this case, mel-spectograms) and then averaging the resulting gradients. The gradients from 25 iterations of adding noise were averaged. For each iteration, we took the gradients from the

³<https://www.nb.no/sprakbanken/en/resource-catalogue/oai-nb-no-sbr-92/>

⁴<https://github.com/google/REAPER>

⁵<https://huggingface.co/NbAiLab/nb-whisper-medium>

⁶<https://github.com/scribe-project/did-prosody-whisper>

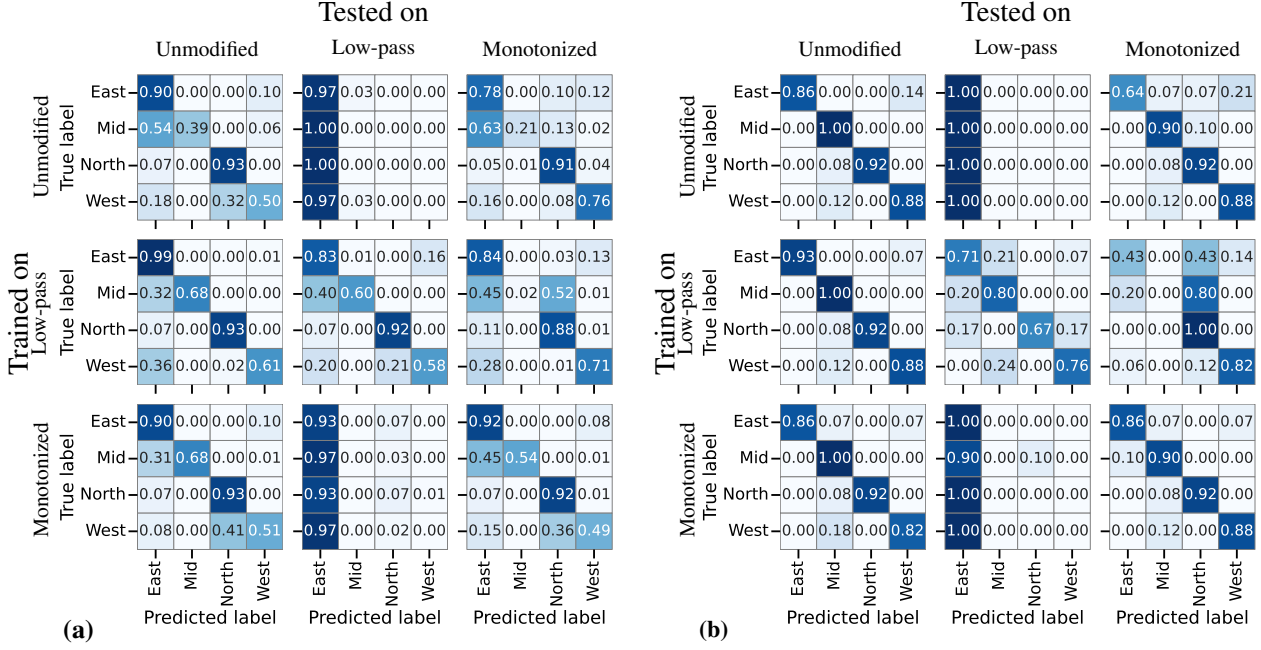


Figure 1: Normalized confusion matrices showing classifier performance on the SSC test set (a; left) and NVOS (b; right). The columns of matrices denote the type of data the model was tested on. The rows of matrices denote the type of audio the model was trained on.

Table 3: The weighted F1 values for the SSC test set and NVOS test set. Rows represent the type of audio modification for the training data for each model; unmod(ified), LP (low-pass), and mono(tonized). Columns represent the test set as well as the modification to the test audio.

Trained on	SSC test set			NVOS test set		
	Unmod	LP	Mono	Unmod	LP	Mono
Unmod	0.66	0.10	0.69	0.91	0.11	0.83
LP	0.79	0.72	0.62	0.93	0.74	0.55
Mono	0.72	0.12	0.69	0.89	0.11	0.89

correct, rather than predicted, class. The sensitivity map values were summed across the time domain to result in a plot of sensitivity over frequency. For normalization, these values were divided by the sum of their values.

4. Results

4.1. Classification

Results from testing on the SSC test set are presented in Table 3 (left) and the confusion matrices are in Figure 1a. Results from testing on the NVOS can be seen in Table 3 (right) and Figure 1b.

The results in Table 3 and Figure 1 show that low-pass filtering has the greatest impact on model performance. Models trained on unmodified or monotonized audio are unable to predict from low-pass filtered data, evidenced by the low F1 values and nearly universal prediction of East. Monotonized audio, in contrast, has a much less dramatic effect, with all models able to achieve above chance levels on monotonized audio. From observing the performance of the low-pass model on low-passed data, we can see that the model is able to predict above chance. Although the low-pass model performs worse than either of the

other two models on the monotonized data, we can observe that the low-pass model surprisingly outperforms the other models on unmodified audio.

Beyond individual model performance, we can observe that nearly all models perform better under nearly all conditions on the NVOS dataset compared to the SSC (the low-pass model predicting from monotonized data being the exception).

In order to assess for statistical significance, we used R to run a logistic regression on accuracy. We evaluated dialect and test data set as well as training data manipulation and test data manipulation and their interaction as independent variables. For all test, $p < 1.15E-6$, thus confirming the significance of the previously reported results after Bonferroni correction ($\alpha = \frac{0.05}{13} = 0.004$). Most pertinently, the regression model demonstrates that the classifier trained on low-pass filtered data is significantly better than the classifier trained on unmodified data ($\beta = 0.697$; $SE = 0.035$; $z = 19.950$; $p < 2E-16$).

4.2. Sensitivity maps

Sensitivity maps are plotted in Figure 2. In all the plots, we can see that the low-pass model, unsurprisingly, pays more attention to the low frequencies (below 500 Hz) than either the model trained on unmodified or monotonized data. In addition to the low-pass model's attention on the lower frequencies, it appears that the low-pass model has the greatest generalizability to different audio inputs. By looking across the top row of plots in Figure 2, we can see that while the low-pass model generally attends more to lower frequencies, for the unmodified audio and the monotonized audio, it largely overlaps with the other two models. However, for low-pass audio, the low-pass model devotes markedly more attention to the lower frequencies and is visibly less attentive in the remaining high frequencies.

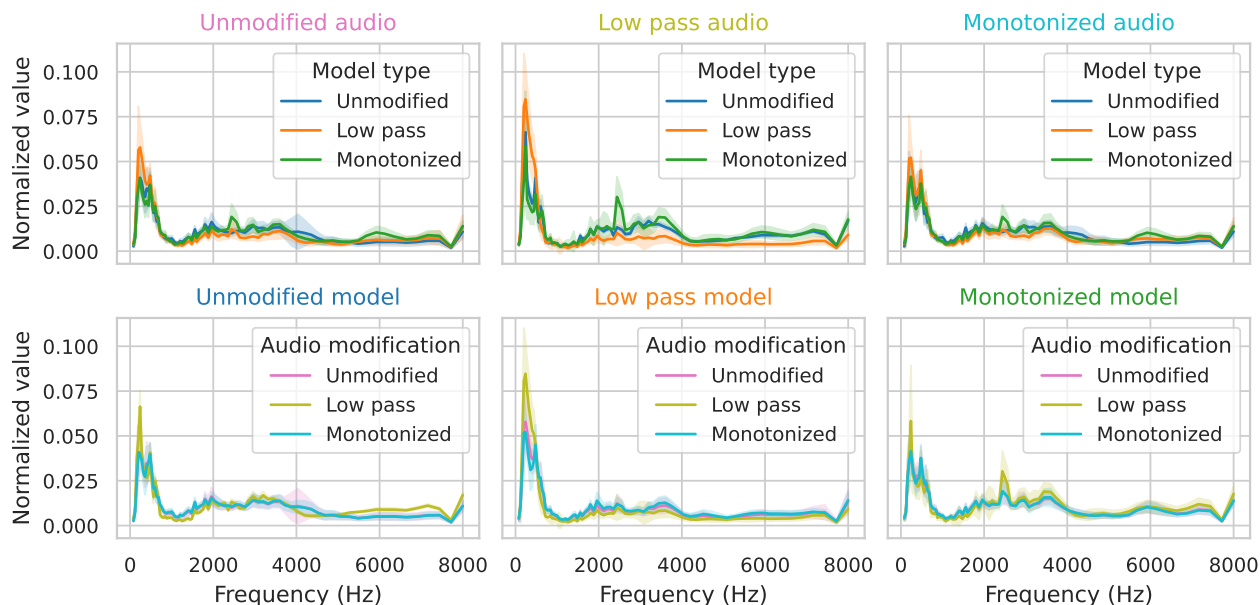


Figure 2: Two visualizations of the normalized SmoothGrad sensitivity maps. Values were summed along the time axis for each model and audio type. The test data for all plots was the NVOS dataset. In the first row, each plot denotes the type of audio, each line the model. In the second row, each plot is for the type of model, each line the type of audio. The halo around each line denotes 1 SD.

5. Discussion

The performance of the low-pass model on low-pass audio supports the idea that prosodic contours contain enough information to infer dialect—at least to some extent. Additionally, the (relatively) low performance of the low-pass model on the monotonized audio demonstrates that when prosodic contours are flattened (even though the frequency band containing F_0 is retained) important information for the low-pass model is removed. This indicates that the model is using information from the pitch dynamics, not merely the presence of the pitch itself, to perform inference.

The better performance on NVOS than the SSC is rather contrary to received wisdom regarding out of domain performance—that models will perform best on data most similar to that which they were trained on. This could be due to the differences between the SSC and NVOS. Firstly, in the NVOS, all speakers are reading the same text which potentially reduces the confusion that may arise from linguistic context. Additionally, the different numbers of speakers and segments per speaker between the two datasets may also contribute to the performance difference. From looking at Tables 1 and 2, we can see that NVOS has more speakers per region than SSC. Further, each NVOS speaker contributes a roughly similar amount of data, whereas individual speaker contributions in the SSC may range from less than an hour to over a dozen hours. Thus, for the SSC test set, if the model finds one speaker particularly challenging to classify, the effect will be much more dramatic than it would be in the NVOS set.

Focusing then on what we might assume to be a more robust result (NVOS), we can see in Figure 1b, that the low-pass model confuses the Mid region with the North when asked to predict on monotonized audio. Given the extant linguistic knowledge about the North and Mid regions—that they differ in prosodic traits but share many other phonemic, morphological, and lex-

ical features [21]—this lends credence to an interpretation that prosodic contour played a substantial role to our classifier in distinguishing these dialects.

Furthermore, the high performance of the low-pass model implies that knowledge from the training of the foundational model is retained despite fine-tuning on low-pass audio. While the model may focus on the lower frequencies, it retains the ability to utilize information in the higher frequencies, when such information is present. Finally, the ability of the low-pass model to outperform the unmodified model on unmodified data indicates, firstly, that prosodic information is, in fact, helpful to the dialect classification task for Norwegian, and, secondly, that encouraging the model to devote more attention to pitch contours does aid classification.

6. Conclusion

We have shown that forcing a DID model to focus on prosodic information by fine-tuning a foundation model on low-pass filtered data improves model performance. Importantly, the low-pass model outperforms the model trained on unmodified data even when the test data was unmodified. By contrast, the removal of pitch information by monotonizing F_0 during fine-tuning was not associated with increased performance. Our results demonstrate that the specific deprivation of higher frequencies during training enables the model to make better use of the task relevant pitch dynamics, by devoting more attention to the lower frequency bands. At the same time, the low-pass model retains knowledge from its foundation model, making it more adaptable than the other two models.

Our method was exemplified in a Norwegian context, where prosody is isoglossic, but our expectation is that the results will generalize to DID in the many other languages where prosody is an important dialect marker.

7. Acknowledgments

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

8. References

- [1] F. Biadsy and J. Hirschberg, "Using prosody and phonotactics in Arabic dialect identification," in *Interspeech 2009*, 2009, pp. 208–211.
- [2] G. Salvi, "Advances in regional accent clustering in Swedish," in *Interspeech 2005*, 2005, pp. 2841–2844.
- [3] A. Etman and A. A. Louis, "American dialect identification using phonotactic and prosodic features," in *2015 SAI Intelligent Systems Conference (IntelliSys)*, 2015, pp. 963–970.
- [4] N. B. Chittaragi, A. Prakash, and S. G. Koolagudi, "Dialect identification using spectral and prosodic features on single and ensemble classifiers," *Arabian Journal for Science and Engineering*, vol. 43, no. 8, pp. 4289–4302, Aug 2018. [Online]. Available: <https://doi.org/10.1007/s13369-017-2941-0>
- [5] S. Kakouros and K. Hiovain-Asikainen, "North Sámi dialect identification with self-supervised speech models," in *Interspeech 2023*, 2023, pp. 5306–5310.
- [6] S. Saranya, B. Bharathi, S. Gomathy Dhanya, and A. Krishnakumar, "Real-time continuous Tamil dialect speech recognition and summarization," *Circuits, Systems, and Signal Processing*, Dec 2024. [Online]. Available: <https://doi.org/10.1007/s00034-024-02950-5>
- [7] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," 2022. [Online]. Available: <https://arxiv.org/abs/2212.04356>
- [8] S. Radhakrishnan, C.-H. H. Yang, S. A. Khan, N. A. Kiani, D. Gomez-Cabrero, and J. N. Tegner, "A parameter-efficient learning approach to Arabic dialect identification with pre-trained general-purpose speech model," in *Interspeech 2023*, 2023, pp. 1958–1962.
- [9] Q. Luo and R. Zhou, "Exploring the impact of back-end network on wav2vec 2.0 for dialect identification," in *Interspeech 2023*, 2023, pp. 5356–5360.
- [10] M. Armstrong, M. Breen, S. Gooden, E. Levon, and K. M. Yu, "Sociolectal and Dialectal Variation in Prosody," *Language and Speech*, vol. 65, pp. 783–790, 2022.
- [11] M. Barkat, J. Ohala, and F. Pellegrino, "Prosody as a distinctive feature for the discrimination of Arabic dialects," in *Proceedings of the 6th European Conference on Speech Communication and Technology*, 1999, pp. 395–398.
- [12] S. Frota, M. Vigário, and F. Martins, "Language Discrimination and Rhythm Classes: Evidence from Portuguese," in *Proceedings of Speech Prosody 2002*, 2002, pp. 319–322.
- [13] R. Fuchs, "The Perception of Speech Rhythm in Indian English and British English," in *Speech Rhythm in Varieties of English*, R. Fuchs, Ed. Berlin: Springer, 2016, pp. 113–162.
- [14] S. Kakouros, K. Hiovain, M. Vainio, and J. Šimko, "Dialect identification of spoken North Sámi language varieties using prosodic features," in *Speech Prosody 2020*, 2020, pp. 625–629.
- [15] M. Yang, R. C. M. C. Shekar, O. Kang, and J. H. L. Hansen, "What can an accent identifier learn? probing phonetic and prosodic information in a wav2vec2-based accent identification model," in *Interspeech 2023*, 2023, pp. 1923–1927.
- [16] M. Kodali, S. R. Kadiri, and P. Alku, "Classification of vocal intensity category from speech using the wav2vec2 and whisper embeddings," in *Interspeech 2023*, 2023, pp. 4134–4138.
- [17] A. Nesse and B. Høyland, "Norwegian dialect classifications," *Dialectologia*, vol. Special Issue X, pp. 255–298, 2023.
- [18] R. van Ommeren and P. M. Kveen, "Det folkelingvistiske konseptet "tonefall": Ei sosiolingvistisk utforskning av prosodiens indeksikalitet," *Maal Og Minne*, vol. 111, 2019.
- [19] C. Gooskens, "How well can Norwegians identify their dialects?" *Nordic Journal of Linguistics*, vol. 28, pp. 37–60, 2005.
- [20] G. Kristoffersen, *The Phonology of Norwegian*. Oxford: Oxford University Press, 2000.
- [21] B. Mælum and U. Røyndland, *Det norske dialektlandskapet*, 2nd ed. Oslo: Cappelen Damm akademisk, 2023.
- [22] P. E. Solberg, P. Beauguitte, P. E. Kummervold, and F. Wetjen, "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)*. Tórshavn, the Faroe Islands: Association for Computational Linguistics, May 2023, pp. 48–52. [Online]. Available: <https://aclanthology.org/2023.resourceful-1.7/>
- [23] P. Parsons, P. E. Solberg, K. Kvale, T. Svendsen, and G. Salvi, "Adding metadata to existing parliamentary speech corpus," in *Proceedings of the Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies (NoDaLiDa/Baltic-HLT 2025)*. Tallinn, Estonia: University of Tartu Library, Mar. 2025, pp. 448–457. [Online]. Available: <https://aclanthology.org/2025.nodalida-1.49/>
- [24] I. Amdal and H. Ljøen, "TABU.0 - en norsk telefonedatabase," Telenor, Tech. Rep., 06 1995.
- [25] Boersma, Paul and Weenink, David, "Praat: doing phonetics by computer." [Online]. Available: <http://www.praat.org/>
- [26] S. Alcorn, K. Meemann, C. G. Clopper, and R. Smiljanic, "Acoustic cues and linguistic experience as factors in regional dialect classification," *The Journal of the Acoustical Society of America*, vol. 147, no. 1, pp. 657–670, 2020.
- [27] A. Leemann, M.-J. Kolly, F. Nolan, and Y. Li, "The role of segments and prosody in the identification of a speaker's dialect," *Journal of Phonetics*, vol. 68, pp. 69–84, 2018.
- [28] P. E. Kummervold, J. de la Rosa, F. Wetjen, R.-A. Braaten, and P. E. Solberg, "Whispering in Norwegian: Navigating orthographic and dialectic challenges," in *Interspeech 2024*, 2024, pp. 3984–3988.
- [29] D. Smilkov, N. Thorat, B. Kim, F. Viégas, and M. Wattenberg, "SmoothGrad: removing noise by adding noise," in *ICLM Workshop on Visualization for Deep Learning*, Sydney, Australia, 2017.