Quantifying Language Variation Acoustically with Few Resources

Martijn Bartelds
University of Groningen
The Netherlands
m.bartelds@rug.nl

Martijn Wieling
University of Groningen
The Netherlands
m.b.wieling@rug.nl

Abstract

Deep acoustic models represent linguistic information based on massive amounts of data. Unfortunately, for regional languages and dialects such resources are mostly not available. However, deep acoustic models might have learned linguistic information that transfers to low-resource languages. In this study, we evaluate whether this is the case through the task of distinguishing low-resource (Dutch) regional varieties. By extracting embeddings from the hidden layers of various `wav2vec 2.0` models (including new models which are pre-trained and/or fine-tuned on Dutch) and using dynamic time warping, we compute pairwise pronunciation differences averaged over 10 words for over 100 individual dialects from four (regional) languages. We then cluster the resulting difference matrix in four groups and compare these to a gold standard, and a partitioning on the basis of comparing phonetic transcriptions. Our results show that acoustic models outperform the (traditional) transcription-based approach without requiring phonetic transcriptions, with the best performance achieved by the multilingual `XLSR-53` model fine-tuned on Dutch. On the basis of only six seconds of speech, the resulting clustering closely matches the gold standard.

1 Introduction

Deep acoustic models have improved automatic speech recognition (ASR) substantially in recent years (Schneider et al., 2019; Baevski et al., 2020a,b; Conneau et al., 2020). These models represent linguistic information based on massive amounts of data. While these models are generally evaluated on ASR benchmarks, few studies have addressed what kind of linguistic information is represented by them. The work of Pasad et al. (2021) examined information represented by the `wav2vec 2.0` model (Baevski et al., 2020b) across the various Transformer layers. They showed that different layers encode different types of linguistic information. Specifically, the initial layers appeared to be most similar to the input speech features, whereas the middle layers mostly encoded contextual information. The final layers again turned out to be similar to the input speech features. However, the representations of the final layers changed when the model was fine-tuned, likely because task-specific information was learned. In addition, Ma et al. (2021) investigated several deep acoustic models using phonetic probing tasks, and found that representations from these models capture information useful for distinguishing English phones. Importantly, these deep acoustic models were better able to distinguish English phones than using conventional MFCC or filterbank features. Although they evaluated the transferability of deep acoustic representations across several domains, it remains unclear whether these models learned information that transfers to other languages. This is, however, important when working on more inclusive speech technology. Especially when resources for training these models are lacking, such as for regional languages and dialects. In this paper, we therefore investigate if hidden layers of deep acoustic models incorporate fine-grained information, which can be used to represent differences between, and in turn distinguish, regional language varieties.

Past work on investigating language variation has often been based on computing pronunciation distances that rely on phonetically transcribed speech (Nerbonne and Heeringa, 1997; Livescu and Glass, 2000; Heeringa, 2004). These (edit) distances have been found to match perceptual judgements of similarity well (Gooskens and Heeringa, 2004; Wieling et al., 2014). However, transcribing speech phonetically is time-consuming and prone to errors (Bucholtz, 2007; Novotney and Callison-Burch, 2010). While automatic approaches for computing phonetic transcriptions exist (e.g., Li et al. 2020), they produce lower quality phonetic

transcriptions than human transcribers do. Additionally, (discrete) phonetic transcriptions do not capture all (continuous) aspects of human speech (Liberman, 2018).

To mitigate these shortcomings, acoustic approaches have been developed for investigating language variation (Huckvale, 2007; Ferragne and Pellegrino, 2010; Strycharczuk et al., 2020; Bartelds et al., 2020). However, these studies either exclusively focused on the vowels (ignoring differences in the consonants), or were negatively influenced by non-linguistic variation in the speech signal.

Recently, Bartelds et al. (2022) found that representations from the hidden layers of pre-trained and fine-tuned `wav2vec 2.0` (large) models are suitable to represent language variation. They showed that these representations capture linguistic information that is not represented by phonetic transcriptions, while being less sensitive to non-linguistic variation in the speech signal. Furthermore, this approach seems to provide a better match to human perceptual judgements than phonetic transcription-based approaches.

To investigate if `wav2vec 2.0` acoustic models (including newly trained Dutch models) learn fine-grained linguistic information that can transfer to regional languages and dialects, we will assess whether or not regional languages and dialects spoken in the Netherlands can be distinguished using these models. Our code and newly trained models are publicly available.[1]

2 Dataset

We use Dutch dialect pronunciation recordings from the Goeman-Taeldeman-Van Reenen-Project (Goeman and Taeldeman, 1996). Audio recordings of hundreds of words were obtained (and manually phonetically transcribed) in the 1980s and 1990s and are available for 613 dialect varieties in the Netherlands and Belgium. Unfortunately, the hour-long audio recordings were not segmented, and the metadata with the time stamps we use to extract the audio containing individual word pronunciations were only partially available. In total, therefore, we extract the acoustic recordings (judged to be of sufficient quality) for 10 words (*armen*: 'arms', *deeg*: 'dough', *draden*: 'wires', *duiven*: 'pigeons', *naalden*: 'needles', *ogen*: 'eyes', *pijpen*: 'pipes', *tangen*: 'pliers', *volk*: 'people', *vuur*: 'fire') pronounced in 106 locations in the Netherlands. On average, the duration of these 10 words is only 6.3 seconds for each location. Some example pronunciations are shown in Table 1.

	Standard Dutch	**Frisian** (Joure)	**Low Saxon** (Eelde)	**Limburgish** (Echt)
Arms	ɑrəmən	jɛrəmən	ʔaɪms	æʁəm
Dough	deɪx	deiç	dɛix	deix
Wires	dradən	trɪdn	drɔdn	dʁɵi

Table 1: Phonetic transcriptions of the words 'arms', 'dough', and 'wires' obtained from three locations where different regional languages (Frisian, Low Saxon, and Limburgish) are spoken, as well as in Standard Dutch. The names of the locations are provided between parentheses.

3 Methods

We compute embeddings from the hidden Transformer layers of three fine-tuned deep acoustic `wav2vec 2.0` large models, and subsequently determine pronunciation differences using dynamic time warping (DTW) with these embeddings (Müller, 2007). We use fine-tuned acoustic models in this study as their hidden representations were found to show the closest match with human perceptual judgements of pronunciation variation (Bartelds et al., 2022). For the transcription-based approach, we apply a (phonetically sensitive) Levenshtein distance algorithm to the available corresponding phonetic transcriptions of the 10 words in all locations. After averaging the word-based differences, the result of both approaches is a distance matrix representing the aggregate pronunciation difference between every pair of locations. Both distance matrices are then clustered in four groups and quantitatively compared to a gold standard clustering of four groups (see Figure 1a). These groups correspond to the three regional languages spoken in the Netherlands that are recognised by the European Charter for Regional or Minority Languages (Frisian: light blue in Figure 1a, Low Saxon: dark blue, Limburgish: light green) and standard Dutch (dark green).

We use the fine-tuned English `wav2vec 2.0` large model (abbreviated as `w2v2-en`) released by Baevski et al. (2020b). In addition, we use a new pre-trained Dutch `wav2vec 2.0` large model that is fine-tuned on Dutch labelled data (abbreviated as `w2v2-nl`), and we use the multilingual `XLSR-53` model of Conneau et al. (2020)

[1] https://github.com/Bartelds/language-variation

that is fine-tuned on the same Dutch labelled data (XLSR-nl). We explicitly use models for Dutch because this language is closely related to the different regional languages and dialects spoken in the Netherlands (including Frisian, Low Saxon, and Limburgish; Eberhard et al., 2021). The advantage of having a Transformer-based language model that is linguistically closest was shown by de Vries et al. (2021), albeit for a different task (i.e. part-of-speech tagging). It may therefore be the case that a high degree of language similarity is also beneficial for Transformer-based models that learn speech representations.

Acoustic models w2v2-en is pre-trained on 960 hours of English speech from the Librispeech dataset (Panayotov et al., 2015). The model consists of a convolutional encoder, a quantizer, and a 24-layer Transformer network. Subsequently, the learned representations are fine-tuned on 960 hours of labelled data by adding a randomly initialised linear projection layer on top of the Transformer network. This projection layer is used to predict characters from the labelled data using the connectionist temporal classification loss function (CTC; Graves et al., 2006).

w2v2-nl is obtained by further pre-training the English model on 243 hours (cross-talk and silences removed) of Dutch speech from the Spoken Dutch Corpus (Oostdijk, 2000). This approach converged faster in preliminary experiments compared to a randomly initialised network. Subsequently, the model is fine-tuned on the same 243 hours of (now labelled) Dutch speech using CTC. Pre-training is performed for 2 million steps with 100,000 iterations for warm up, and a linearly decreasing learning rate starting at $5e{-}5$. Fine-tuning is performed on labelled data for 1 million steps, with a linearly decreasing learning rate starting at $1e{-}5$. Other configuration details are similar to those reported in Baevski et al. (2020b).

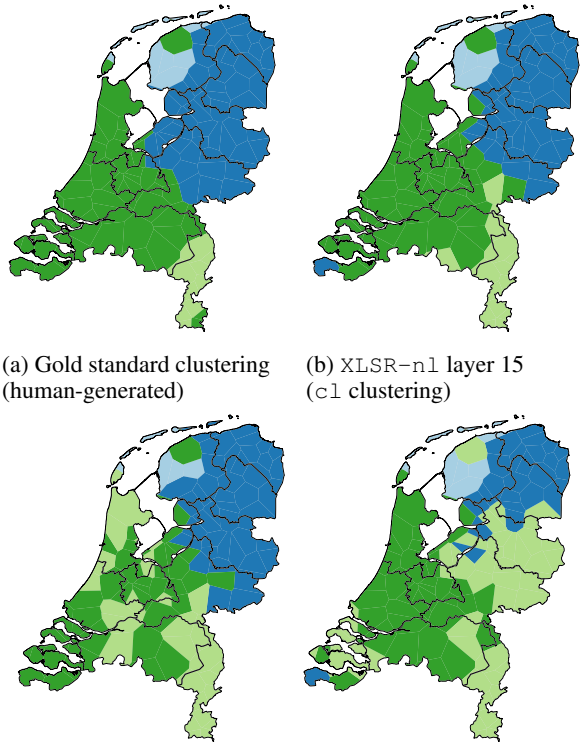
XLSR-53 has the same architecture as the other acoustic models, except that the quantizer has learned a single set of discrete speech representations that is shared across the pre-training languages (which includes Dutch and German, but not Frisian, Low Saxon or Limburgish). This model is pre-trained on 56,000 hours of speech in 53 languages (44,000 hours consists of English speech) obtained from the BABEL, Common Voice and Multilingual Librispeech datasets (Gales et al., 2014; Ardila et al., 2020; Pratap et al., 2020). To obtain XLSR-nl, XLSR-53 is fine-tuned on the same labelled data as w2v2-nl with the same configuration details.

Obtaining pronunciation differences We compute pronunciation differences between all 106 locations in our dataset using both phonetic transcriptions and acoustic embeddings. For determining the phonetic transcription-based distance, we use a variant of the Levenshtein distance (LD) algorithm proposed by Wieling et al. (2012), which includes automatically determined phonetic segment distances. This algorithm matches perception well (Wieling et al., 2014) and is often used for investigating dialect variation.

Given a pair of locations, recordings of the same word are compared using LD (phonetic transcriptions) or DTW (acoustic embeddings), which is a frequently-used algorithm for comparing representations of acoustic sequences (Senin, 2008). The acoustic embeddings are obtained for each model for each of the 24 layers separately (i.e. to determine the optimal layer). The word-based distances between two locations are averaged to determine the single pronunciation distance between a location pair. This process is repeated for all pairs to create a symmetric distance matrix including all locations.

Clustering We classify the phonetic transcription distance matrix and the acoustic distance matrices (three models times 24 layers) from the acoustic embeddings using seven clustering techniques, yielding the four different groups. Of course, the choice of clustering technique may influence the results, but we determine the optimal clustering algorithm by selecting the one best representing the underlying difference matrix. We use clustering techniques that have previously been applied to distance matrices of dialect pronunciations, namely single link (sl), complete link (cl), group average (ga), weighted average (wa), unweighted centroid (uc), weighted centroid (wc) and minimum variance (mv) clustering (Heeringa et al., 2002; Prokić and Nerbonne, 2008).

To select the best clustering algorithm, we calculate the cophenetic correlation coefficient (Sokal and Rohlf, 1962). This coefficient represents the (Pearson) correlation between the original distances and the clustering-based cophenetic distances (i.e. extracted from the dendrogram underlying the clustering). Higher values indicate a better



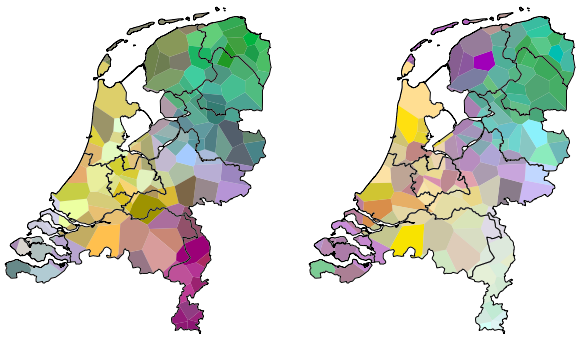
(a) Gold standard clustering (human-generated)

(b) XLSR-nl layer 15 (cl clustering)

(c) w2v2-en layer 13 (cl clustering)

(d) w2v2-nl layer 16 (wa clustering)

Figure 1: Cluster maps visualizing four clusters on the map of the Netherlands. Separate clusters are indicated by the different colours.



(a) XLSR-nl layer 15

(b) LD

Figure 2: MDS maps visualizing pronunciation differences based on Dutch dialect pronunciations. Similar colours correspond to pronunciations that are also similar.

correspondence between the original data and the clustering (with a value of 1 being perfect). We determine the optimal clustering method for each Transformer layer (for the acoustic models) per model by selecting the one with the highest cophenetic correlation coefficient.

Evaluation We compare the layer-based clustering results per model to the gold standard clustering. We do this by computing the CDistance score, which is a clustering comparison measure proposed by Coen et al. (2010). As opposed to other techniques for comparing clustering partitions, this measure incorporates spatial information in the evaluation (i.e. the coordinates of the locations), which is essential for evaluating spatial (i.e. geographical) clustering. The CDistance scores (for the optimal clustering method per layer) are compared across the layers for each model. The layer with the lowest score per model (i.e. most closely matching the gold standard clustering) is selected for the comparison of the three models. In addition, we create multidimensional scaling (MDS) maps (Torgerson, 1952) using the best-performing model and compare it to the frequently used LD algorithm to show the (more fine-grained) relationship between the geographical location of the locations and the pronunciation differences.

4 Results and discussion

Model	Layer	Clustering	CDistance
w2v2-en	13	cl	0.34
w2v2-nl	16	wa	0.34
XLSR-nl	15	cl	**0.20**
LD		ga	0.46

Table 2: CDistance scores for the different models with the optimal clustering algorithm and output layer (if applicable). Lower scores indicate a better match with the gold standard clustering.

In Table 2, we show the CDistance scores associated with the different models. Ideally, the best layer would have been selected using a validation set instead of using all data, but our set of words was unfortunately too small to be adequately split. However, given that the optimal layers reported in Table 2 correspond with the middle hidden layers found to be best representing pronunciation differences in the work of Bartelds et al. (2022), we do not believe this to be problematic.

Our results show that the XLSR-nl model with output layer 15 and complete link clustering shows the best performance among the fine-tuned models. Note that the standard deviation of the performance for the XLSR-nl model across all Transformer layers was equal to 0.09, which highlights the strong performance of this model over the other models. Importantly, all fine-tuned acoustic models improve

over the LD algorithm, which is traditionally used to investigate (dialectal) language variation. Perhaps surprisingly, the w2v2-nl model performs similar to the w2v2-en model. We do not have a clear explanation for this pattern, but it may be caused by the Dutch model being based on the English model, in combination with a smaller amount of Dutch as opposed to English data used for pre-training. In future work we aim to investigate this.

The multilingual XLSR-nl model outperforms both monolingual models. The XLSR-nl model is pre-trained on a variety of languages, including Dutch, English and German. The regional languages and dialects spoken in the Netherlands have clear links to these three languages (i.e. Frisian has some overlap with English, Low Saxon has some overlap with German, and all varieties overlap with Dutch, which is also the fine-tuning language).

To illustrate, Figure 1 visualizes the gold standard together with the fine-tuned acoustic models. The XLSR-nl model clearly classifies pronunciations in the geographical area where Limburgish is spoken (i.e. the light green area) most accurately. While the XLSR-nl model does not perfectly distinguish the Low Saxon pronunciations (i.e. the dark blue area), the other models perform worse in this regard.

To evaluate (albeit subjectively) how well more fine-grained differences are captured by the best-performing model, Figure 2 shows the MDS maps for the XLSR-nl model, as well as the LD algorithm. Both approaches show the relative gradual nature of dialect variation well. However, the XLSR-nl model seems to capture the larger distinctions (e.g., delineating the Limburgish area) better than the LD algorithm. Based on these evaluations, XLSR-nl appears to be the best model when little data is available.

5 Conclusion

We have found that the XLSR-nl model can be effectively used to distinguish between language groups in the Netherlands when only a small amount of data is available. It even outperformed the LD algorithm, which requires time-consuming phonetic transcriptions. Our study further shows that multilingual pre-training and fine-tuning on a similar language (compared to the target languages) is beneficial over using a monolingual model.

Acknowledgements

We gratefully acknowledge the financial support of the Center for Groningen Language and Culture (CGTC). In addition, we thank Wietse de Vries for his help in training the models, and the anonymous reviewers for their insightful feedback. Any mistakes remain our own.

References

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. 2020. Common Voice: A Massively-Multilingual Speech Corpus. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4218–4222, Marseille, France. European Language Resources Association.

Alexei Baevski, Steffen Schneider, and Michael Auli. 2020a. vq-wav2vec: Self-Supervised Learning of Discrete Speech Representations. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020b. wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Martijn Bartelds, Wietse de Vries, Faraz Sanal, Caitlin Richter, Mark Liberman, and Martijn Wieling. 2022. Neural representations for modeling variation in speech. *Journal of Phonetics*, 92:101137.

Martijn Bartelds, Caitlin Richter, Mark Liberman, and Martijn Wieling. 2020. A New Acoustic-Based Pronunciation Distance Measure. *Frontiers in Artificial Intelligence*, 3:39.

Mary Bucholtz. 2007. Variation in transcription. *Discourse Studies*, 9(6):784–808.

Michael H. Coen, M. Hidayath Ansari, and Nathanael Fillmore. 2010. Comparing Clusterings in Space. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10), June 21-24, 2010, Haifa, Israel*, pages 231–238. Omnipress.

Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli. 2020. Unsupervised Cross-lingual Representation Learning for Speech Recognition. *arXiv preprint arXiv:2006.13979*.

Wietse de Vries, Martijn Bartelds, Malvina Nissim, and Martijn Wieling. 2021. Adapting Monolingual Models: Data can be Scarce when Language Similarity

is High. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4901–4907, Online. Association for Computational Linguistics.

David M. Eberhard, Gary F. Simons, and Charles D. Fennig. 2021. Ethnologue: Languages of the World. Twenty-fourth edition. SIL International.

Emmanuel Ferragne and François Pellegrino. 2010. Vowel systems and accent similarity in the British Isles: Exploiting multidimensional acoustic distances in phonetics. *Journal of Phonetics*, 38(4):526–539.

Mark J. F. Gales, Kate M. Knill, Anton Ragni, and Shakti P. Rath. 2014. Speech recognition and keyword spotting for low-resource languages: BABEL project research at CUED. In *Fourth International workshop on spoken language technologies for under-resourced languages (SLTU-2014)*, pages 16–23. International Speech Communication Association (ISCA).

Ton Goeman and Johan Taeldeman. 1996. Fonologie en morfologie van de Nederlandse dialecten: een nieuwe materiaalverzameling en twee nieuwe atlasprojecten. *Taal en Tongval*, 48:38–59.

Charlotte Gooskens and Wilbert Heeringa. 2004. Perceptive evaluation of Levenshtein dialect distance measurements using Norwegian dialect data. *Language Variation and Change*, 16(3):189–207.

Alex Graves, Santiago Fernández, Faustino J. Gomez, and Jürgen Schmidhuber. 2006. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Machine Learning, Proceedings of the Twenty-Third International Conference (ICML 2006), Pittsburgh, Pennsylvania, USA, June 25-29, 2006*, volume 148 of *ACM International Conference Proceeding Series*, pages 369–376. ACM.

Wilbert Heeringa. 2004. *Measuring Dialect Pronunciation Differences using Levenshtein Distance*. Ph.D. thesis, University of Groningen.

Wilbert Heeringa, John Nerbonne, and Peter Kleiweg. 2002. Validating Dialect Comparison Methods. In *Classification, Automation, and New Media*, pages 445–452, Berlin, Heidelberg. Springer Berlin Heidelberg.

Mark Huckvale. 2007. *ACCDIST: An Accent Similarity Metric for Accent Recognition and Diagnosis*, pages 258–275. Springer Berlin Heidelberg, Berlin, Heidelberg.

Xinjian Li, Siddharth Dalmia, Juncheng Li, Matthew Lee, Patrick Littell, Jiali Yao, Antonios Anastasopoulos, David R. Mortensen, Graham Neubig, Alan W. Black, and Florian Metze. 2020. Universal Phone Recognition with a Multilingual Allophone System. In *2020 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2020, Barcelona, Spain, May 4-8, 2020*, pages 8249–8253. IEEE.

Mark Liberman. 2018. Towards progress in theories of language sound structure. In Diane Brentari and Jackson L. Lee, editors, *Shaping phonology*. University of Chicago Press.

K. Livescu and J. Glass. 2000. Lexical modeling of non-native speech for automatic speech recognition. In *2000 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.00CH37100)*, volume 3, pages 1683–1686.

Danni Ma, Neville Ryant, and Mark Liberman. 2021. Probing Acoustic Representations for Phonetic Properties. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 311–315.

Meinard Müller. 2007. Dynamic Time Warping. *Information Retrieval for Music and Motion*, pages 69–84.

John Nerbonne and Wilbert Heeringa. 1997. Measuring Dialect Distance Phonetically. In *Computational Phonology: Third Meeting of the ACL Special Interest Group in Computational Phonology*.

Scott Novotney and Chris Callison-Burch. 2010. Cheap, Fast and Good Enough: Automatic Speech Recognition with Non-Expert Transcription. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 207–215, Los Angeles, California. Association for Computational Linguistics.

Nelleke Oostdijk. 2000. The Spoken Dutch Corpus. Overview and First Evaluation. In *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC'00)*, Athens, Greece. European Language Resources Association (ELRA).

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: An ASR corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2015, South Brisbane, Queensland, Australia, April 19-24, 2015*, pages 5206–5210. IEEE.

Ankita Pasad, Ju-Chieh Chou, and Karen Livescu. 2021. Layer-Wise Analysis of a Self-Supervised Speech Representation Model. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 914–921.

Vineel Pratap, Qiantong Xu, Anuroop Sriram, Gabriel Synnaeve, and Ronan Collobert. 2020. MLS: A Large-Scale Multilingual Dataset for Speech Research. In *Proc. Interspeech 2020*, pages 2757–2761.

Jelena Prokić and John Nerbonne. 2008. Recognising Groups among Dialects. *International Journal of Humanities and Arts Computing*, 2(1-2):153–172.

Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli. 2019. wav2vec: Unsupervised Pre-Training for Speech Recognition. In *Proc. Interspeech 2019*, pages 3465–3469.

Pavel Senin. 2008. Dynamic Time Warping Algorithm Review. *Information and Computer Science Department University of Hawaii at Manoa Honolulu, USA*, 855(1-23):40.

Robert R. Sokal and F. James Rohlf. 1962. The Comparison of Dendrograms by Objective Methods. *Taxon*, 11(2):33–40.

Patrycja Strycharczuk, Manuel López-Ibáñez, Georgina Brown, and Adrian Leemann. 2020. General Northern English. Exploring Regional Variation in the North of England With Machine Learning. *Frontiers in Artificial Intelligence*, 3:48.

Warren S. Torgerson. 1952. Multidimensional scaling: I. Theory and method. *Psychometrika*, 17(4):401–419.

Martijn Wieling, Jelke Bloem, Kaitlin Mignella, Mona Timmermeister, and John Nerbonne. 2014. Measuring Foreign Accent Strength in English: Validating Levenshtein Distance as a Measure. *Language Dynamics and Change*, 4(2):253 – 269.

Martijn Wieling, Eliza Margaretha, and John Nerbonne. 2012. Inducing a measure of phonetic similarity from pronunciation variation. *Journal of Phonetics*, 40(2):307–314.