

Representing Low-Resource Language Varieties: Improved Methods for Spoken Language Processing

Martijn Bartelds

bartelds@stanford.edu

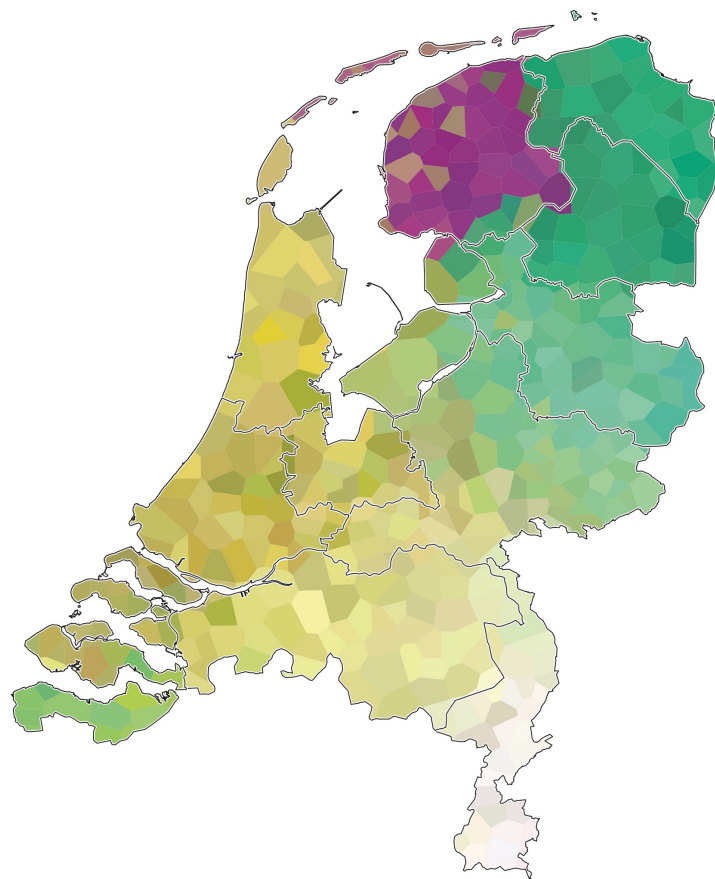
Stanford



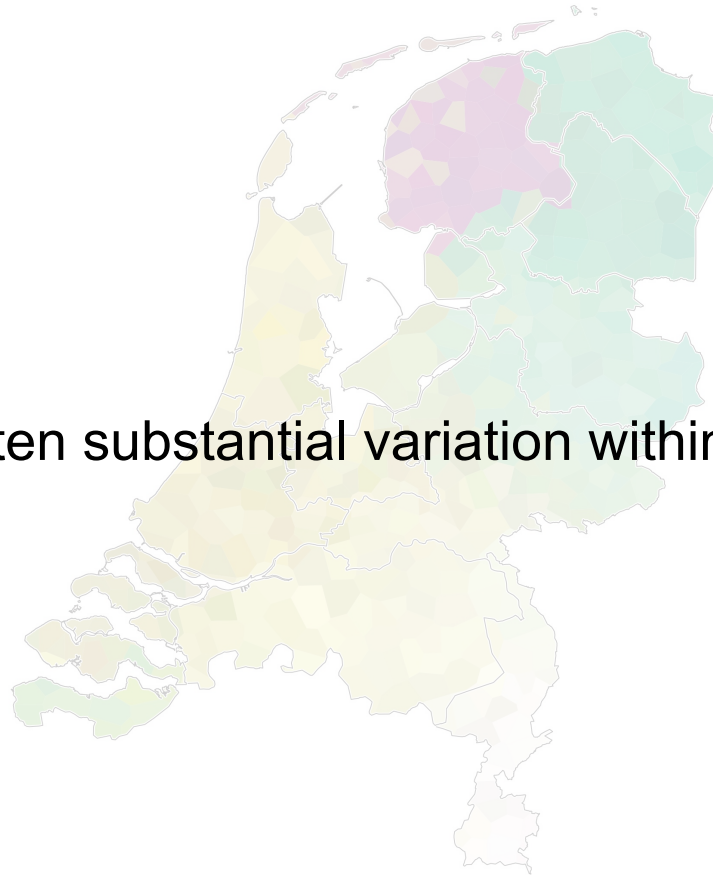
university of
 groningen

faculty of arts

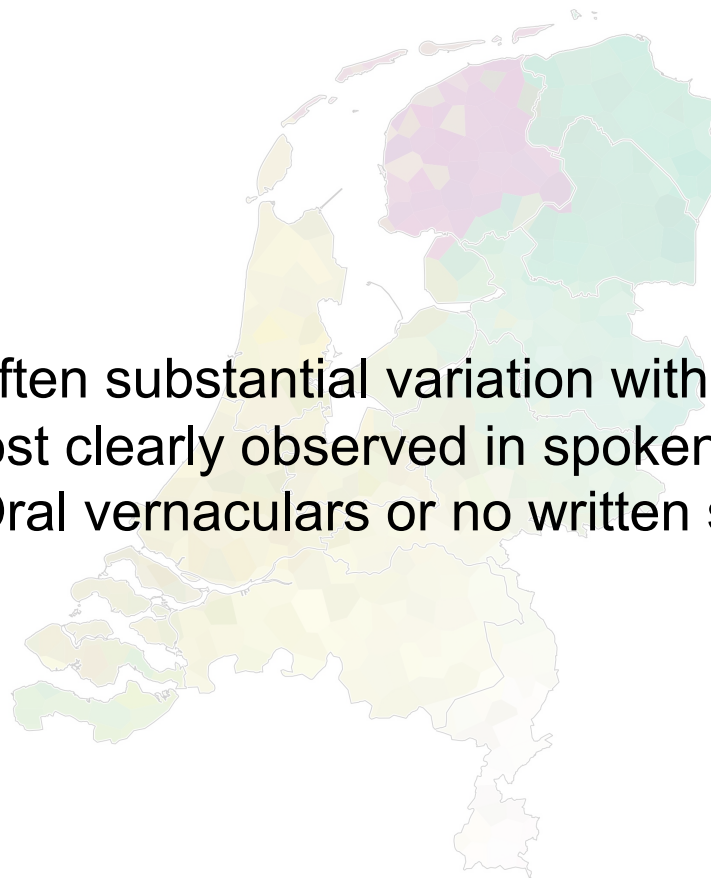


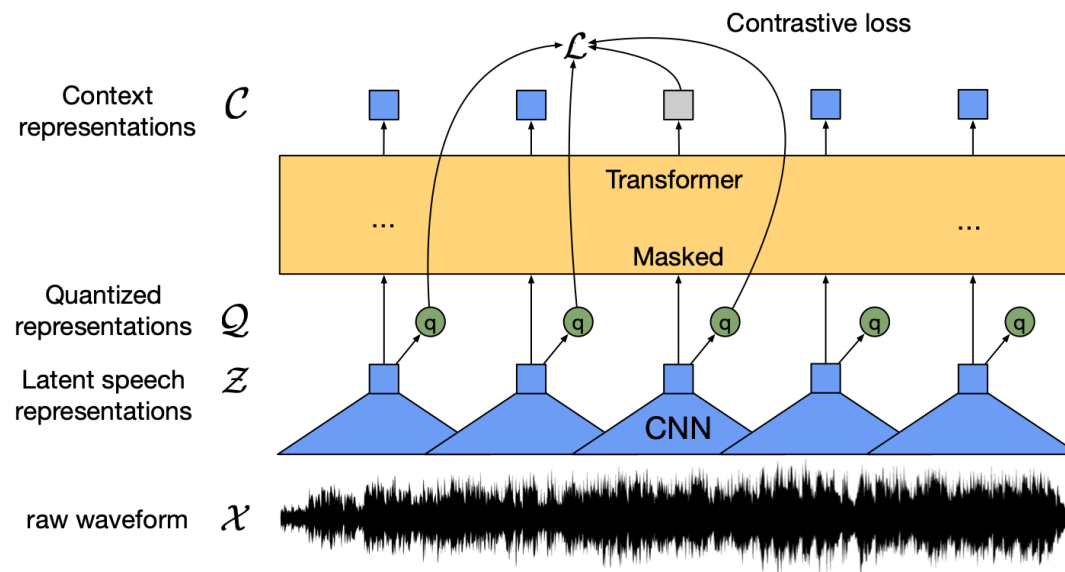


There is often substantial variation within languages



- There is often substantial variation within languages
 - This most clearly observed in spoken language
 - Oral vernaculars or no written system





They do not perform well when handling languages and dialects with limited resources



Can we use these models to describe and model language variation?

Can we use these models to describe and model
language variation?

**Quantifying differences between language
variants automatically without requiring
transcriptions**

Studying language variation may give us
important insights into how varieties relate to their
linguistic communities

k i n d a

k r n d ə

1 **1**

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

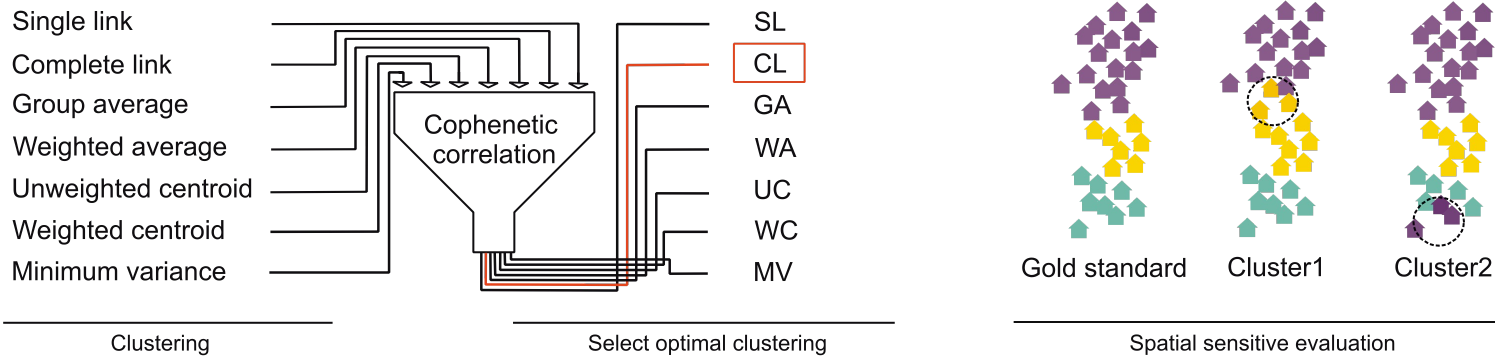
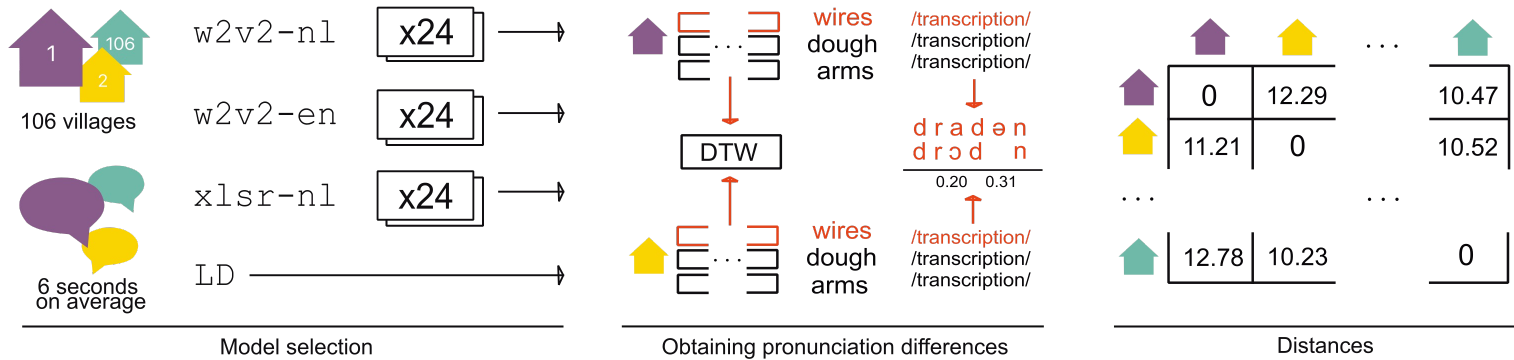


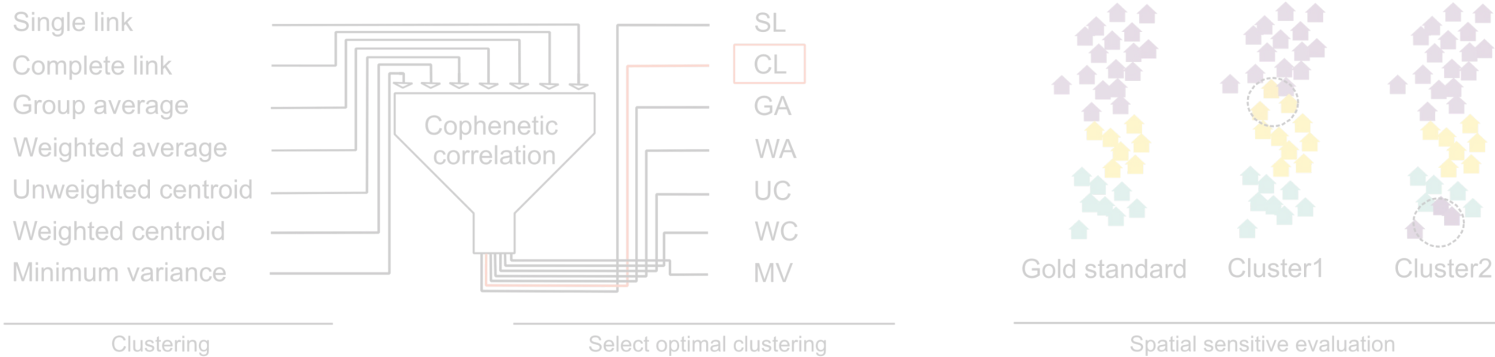
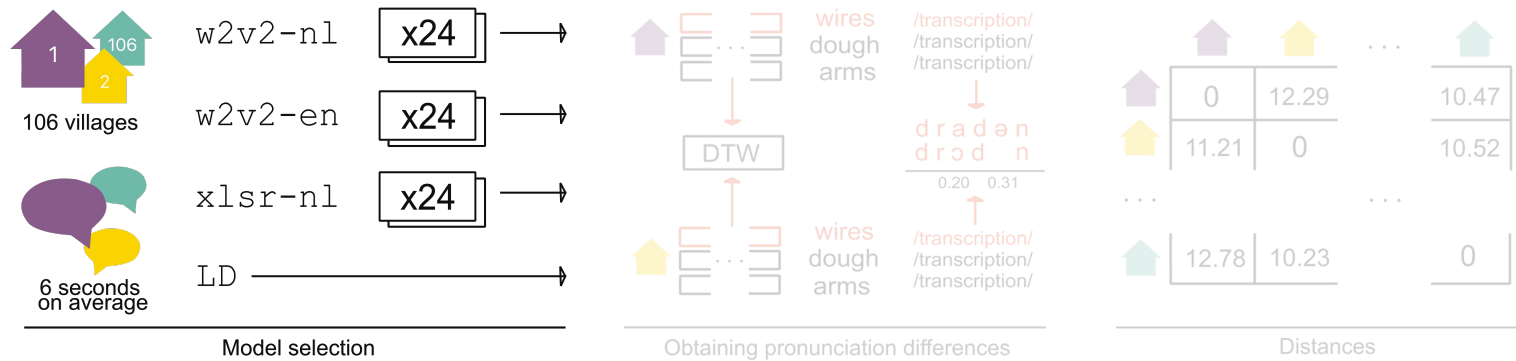
<i>Armen</i>	-	arms
<i>Deeg</i>	-	dough
<i>Draden</i>	-	wires
<i>Duiven</i>	-	pigeons
<i>Naalden</i>	-	needles
<i>Ogen</i>	-	eyes
<i>Pijpen</i>	-	pipes
<i>Tangen</i>	-	pliers
<i>Volk</i>	-	people
<i>Vuur</i>	-	fire

	Standard Dutch	Frisian (Joure)	Low Saxon (Eelde)	Limburgish (Echt)
Arms	ɑrəmən	jɛrəmən	ʔɑɪms	æɐəm
Dough	deɪx	deɪç	dɛix	deix
Wires	dradən	trɪdn	drɔdn	drøi



Only 6 seconds of speech for each location

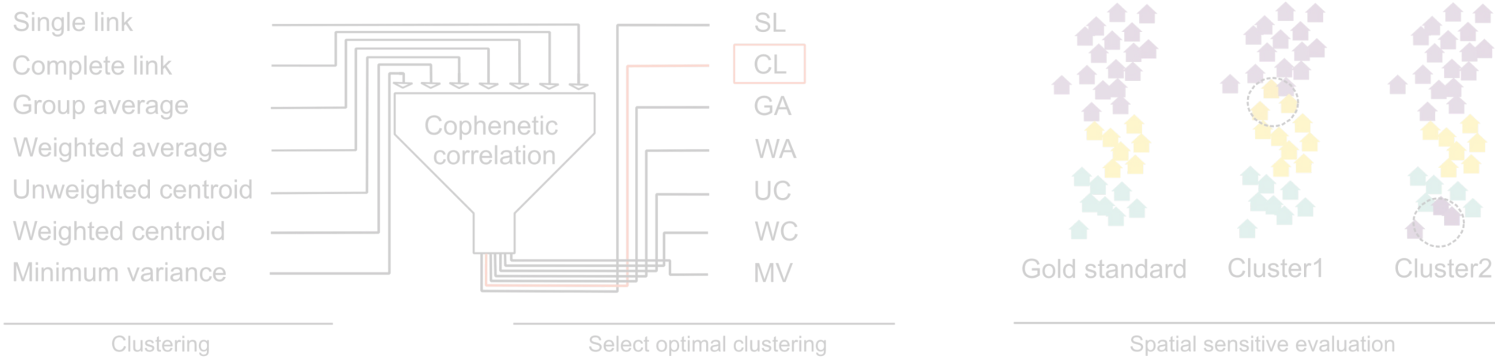
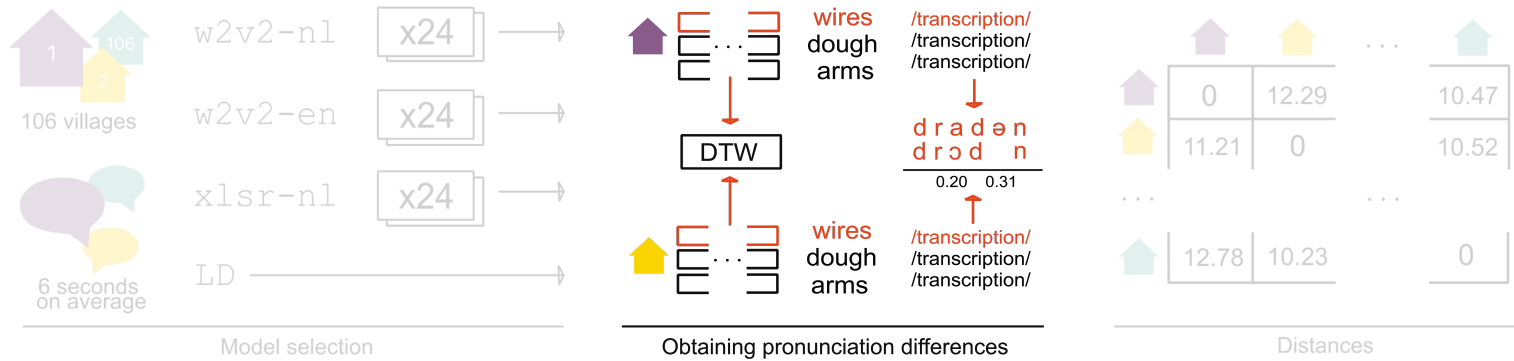


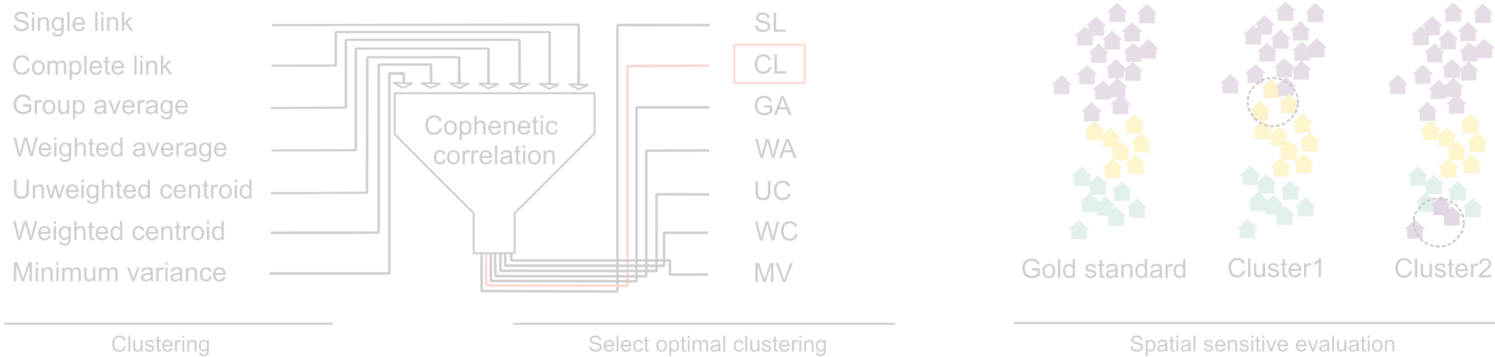
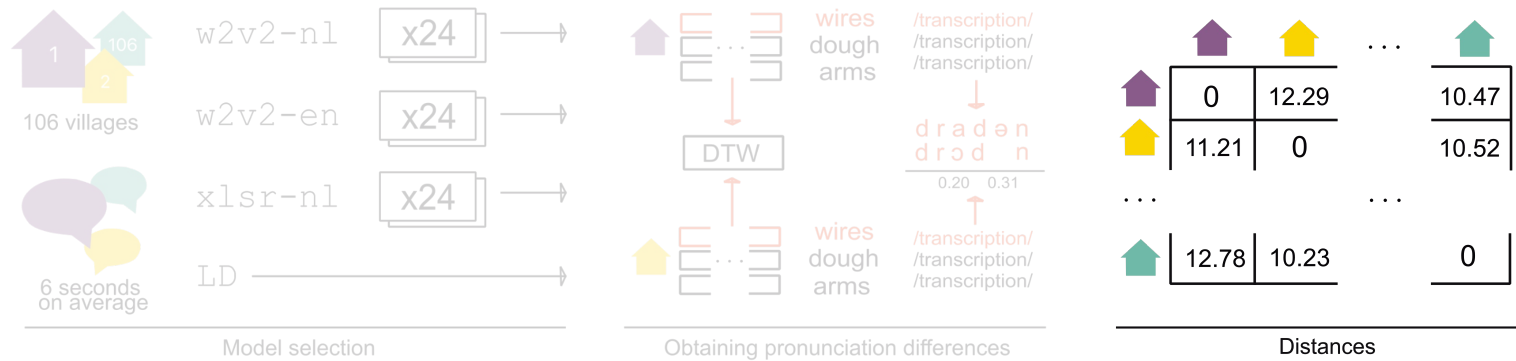


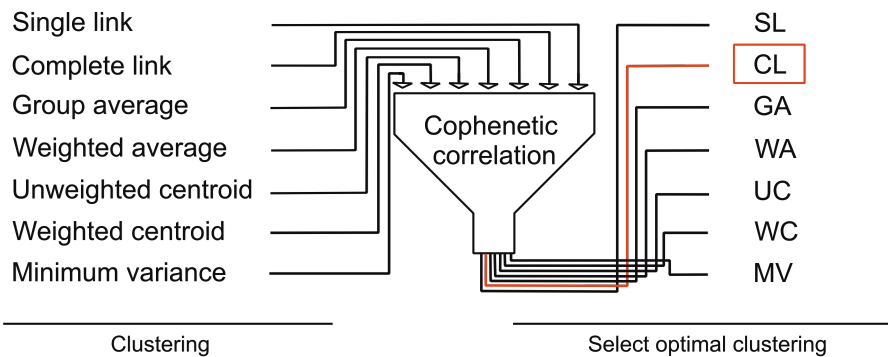
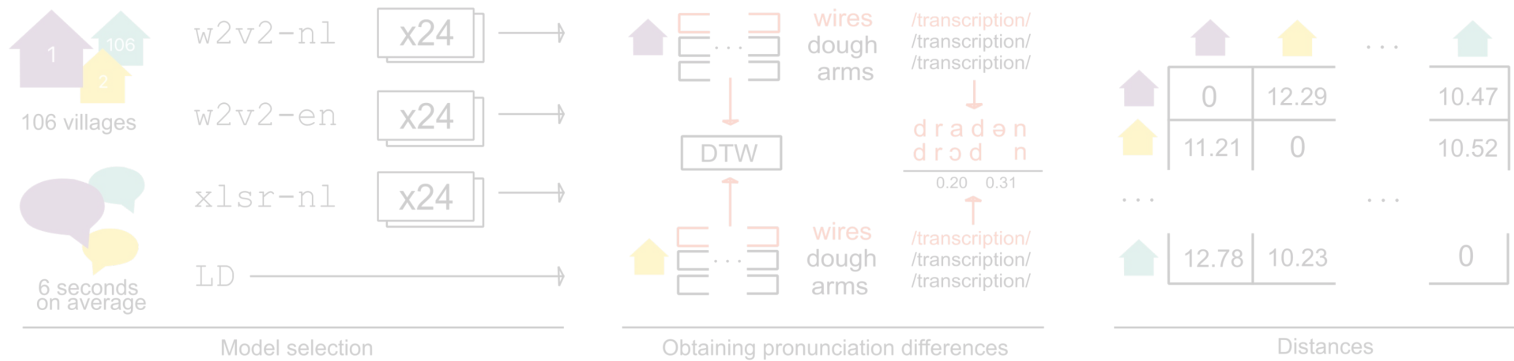


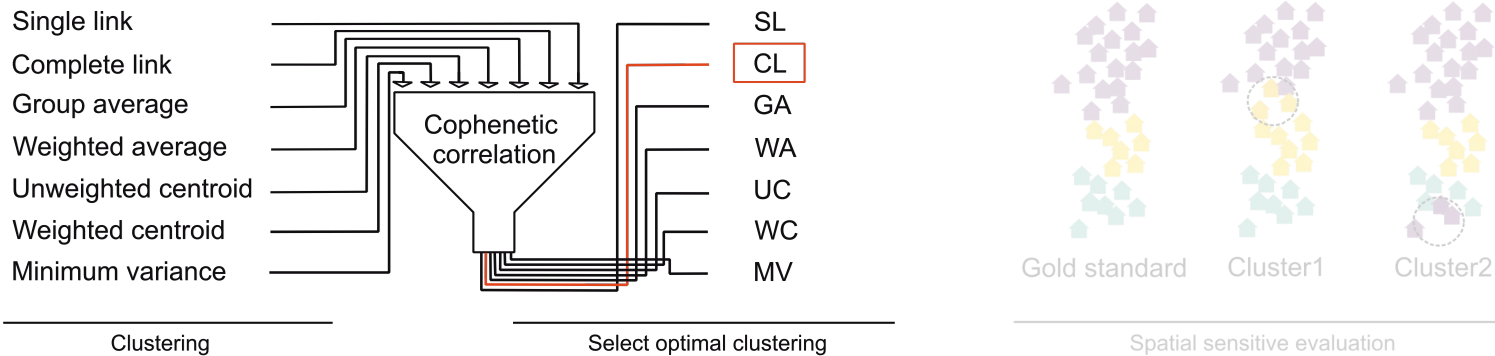
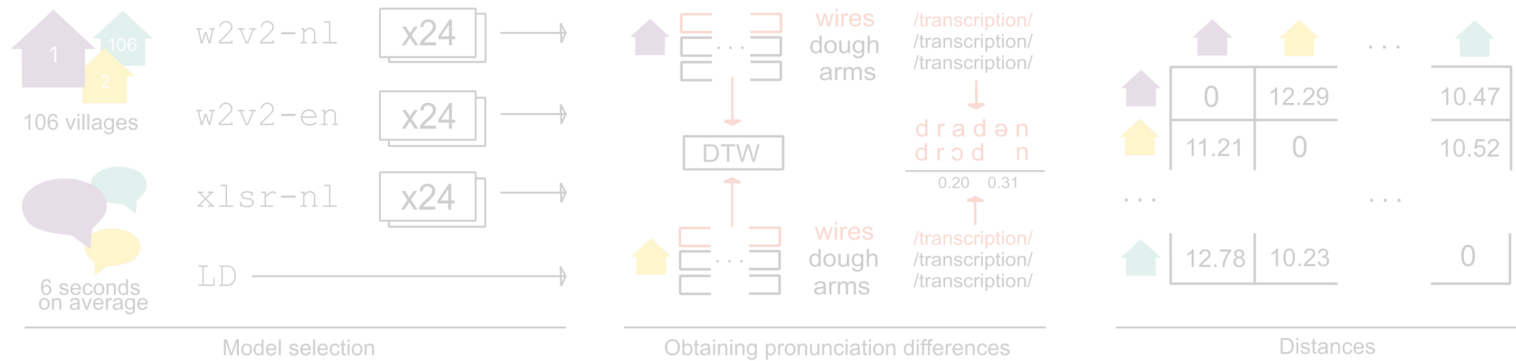
Acoustic models

- `w2v2-en` pre-trained and fine-tuned on 960 hours of English speech from Librispeech (Baevski et al., 2020)
-  `w2v2-nl` further pre-trained and subsequently fine-tuned on 243 hours of Dutch speech from the Spoken Dutch Corpus
-  `XLSR-nl` `XLSR-53` fine-tuned on 243 hours of Dutch speech from the Spoken Dutch Corpus

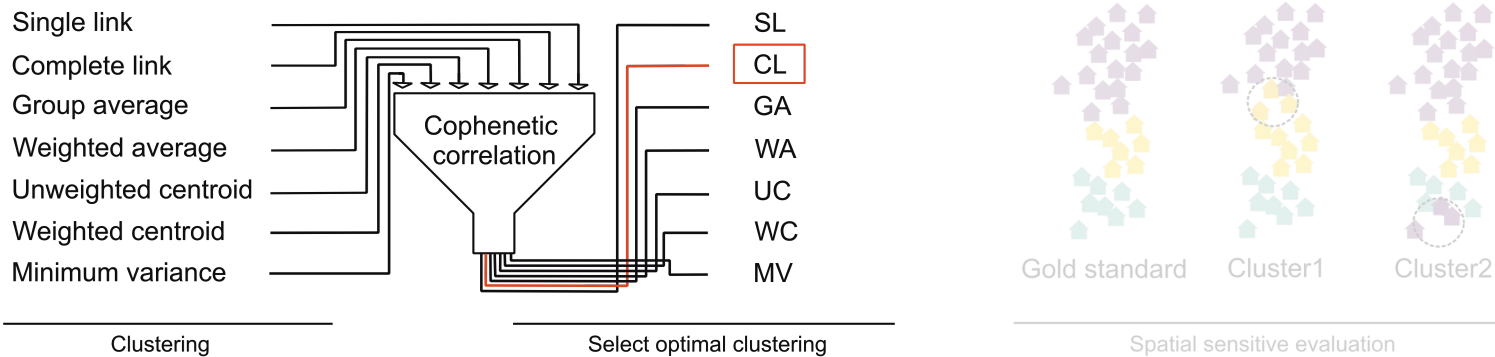
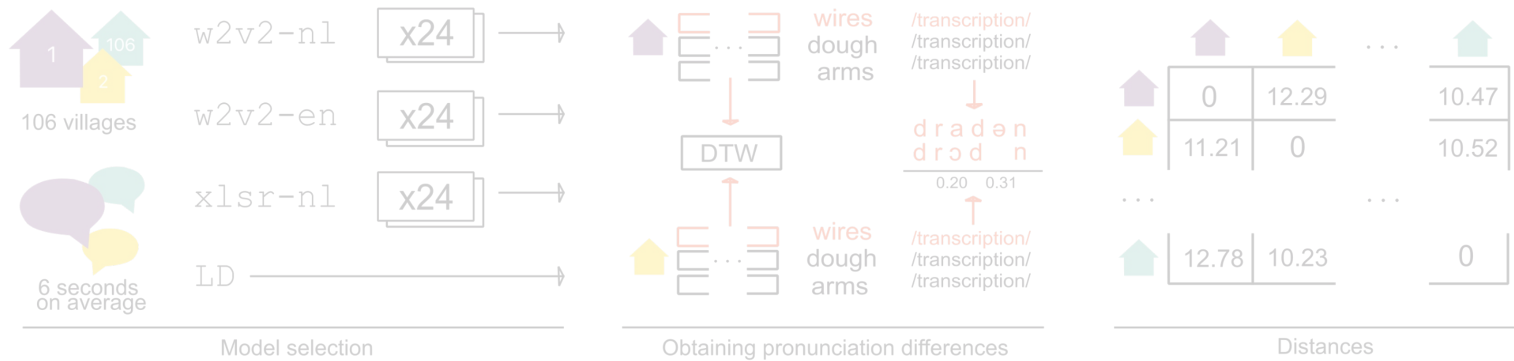




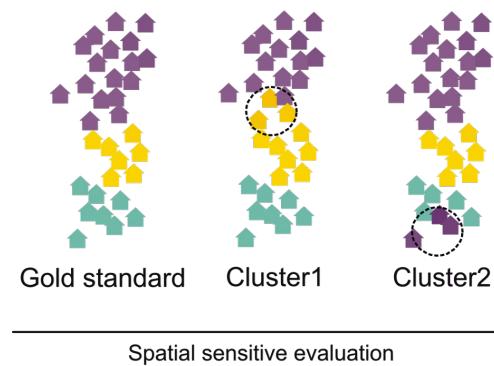
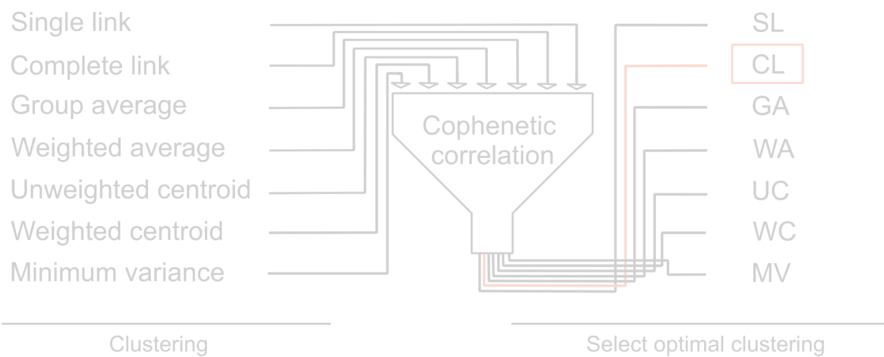
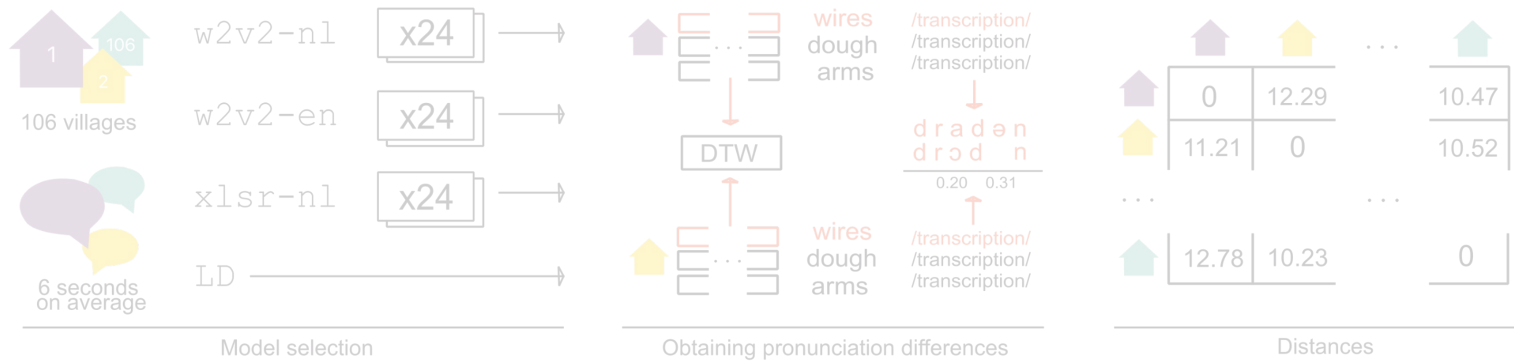




Previously applied to distance matrices of dialect pronunciations (e.g., Heeringa et al. (2002); Prokić and Nerbonne (2008))

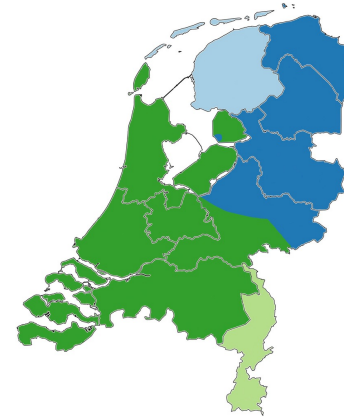


Correlation between the original distances and the clustering-based cophenetic distances (extracted from the dendrogram underlying the clustering)



Perceptual data

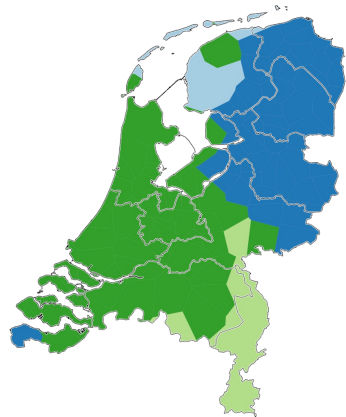
- Online survey
- Listen to the 10 words per village and select the (regional) language
- Approximately 600 ratings per location



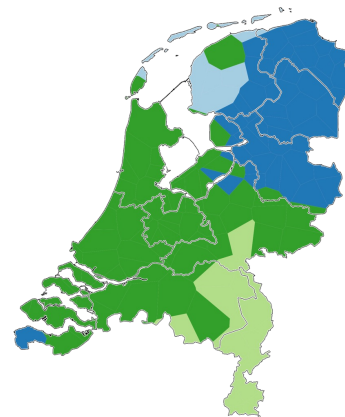
Model	Layer	Method	CDistance	
			Gold	Perception
w2v2-en	13	cl	0.34	0.47
w2v2-nl	16	wa	0.34	0.44
xlsr-nl	15	cl	0.20	0.39
LD		ga	0.46	0.58



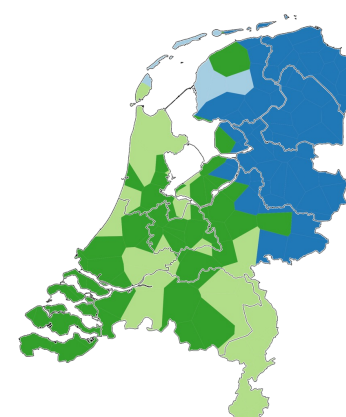
Gold standard
clustering



xlsr-nl layer 15
(cl clustering)



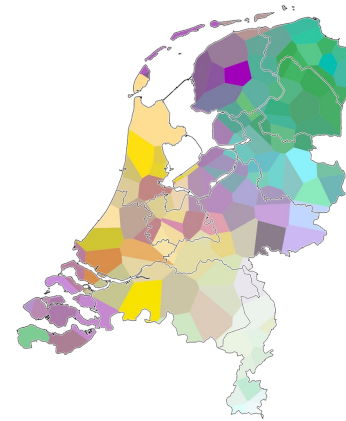
w2v2-nl layer 16
(wa clustering)



w2v2-en layer 13
(cl clustering)



xlsr-nl layer 15
(cl clustering)



LD (ga clustering)

Conclusions

- XLSR-nl can be effectively used to distinguish between language groups in the Netherlands
- Outperformed the LD algorithm that needs time-consuming phonetic transcriptions
- Multilingual pre-training and fine-tuning on a similar language (compared to the target languages) is beneficial over using a monolingual model

Can these models help empower low-resource languages and their varieties?

Can we use these models to improve low-resource speech recognition performance?

Making More of Little Data: Improving Low-Resource Automatic Speech Recognition Using Data Augmentation

Martijn Bartelds¹ Nay San² Bradley McDonnell³
Dan Jurafsky² Martijn Wieling¹

¹University of Groningen ²Stanford University ³University of Hawai'i at Mānoa
m.bartelds@rug.nl

Abstract

The performance of automatic speech recognition (ASR) systems has advanced substantially in recent years, particularly for languages for which a large amount of transcribed speech is available. Unfortunately, for low-resource languages, such as minority languages, regional languages or dialects, ASR performance generally remains much lower. In this study, we investigate whether data augmentation techniques could help improve low-resource ASR performance, focusing on four typologically diverse minority languages or language variants (West Germanic: Gronings, West-Frisian; Malayo-Polynesian: Besemah, Nasal). For all four languages, we examine the use of self-training, where an ASR system trained with the available human-transcribed data is used to generate transcriptions, which are then combined with the original data to train a new ASR system. For Gronings, for which there was a pre-existing text-to-speech (TTS) system available, we also examined the use of TTS to generate ASR training data from text-only sources. We find that using a self-training approach consistently yields improved performance (a relative WER reduction up to 20.5% compared to using an ASR system trained on 24 minutes of manually transcribed speech). The performance gain from TTS augmentation for Gronings was even stronger (up to 25.5% relative reduction in WER compared to a system based on 24 minutes of manually transcribed speech). In sum, our results show the benefit of using self-training or (if possible) TTS-generated data as an efficient solution to overcome the limitations of data availability for resource-scarce languages in order to improve ASR performance.

1 Introduction

Self-supervised learning (SSL) enables speech representation learning without the need for (manually) labeled data. Although this approach is very effective, pre-training an SSL model is costly. This cost (e.g., training time, resources, and memory)

increases with the number of languages added to the model. Furthermore, transferring information across languages, or extending a pre-trained model to new data or to a different domain is computationally expensive, and catastrophic forgetting may occur (Goodfellow et al., 2013). To alleviate this, SSL models are therefore often fine-tuned on the target task with target domain data. For the task of automatic speech recognition (ASR), fine-tuning approaches generally require less data, but training ASR systems that perform well for languages with very little data remains challenging. This leads to (digitally) underrepresented communities and domains such as minority languages, regional languages and dialects not profiting sufficiently from most recent technological advancements.

Recent studies explored fine-tuning of pre-trained self-supervised models for ASR using speech from low-resource languages (e.g., Coto-Solano et al. 2022; Guillaume et al. 2022), and difficulties of modeling resource-scarce languages and dialects were acknowledged in previous work (Aksénova et al., 2022). It remains an open question to what extent model performance is dependent on the amount of fine-tuning data and the type of language, when the total amount of available data for a language is limited. Having a better understanding of how limited training data affects model performance paves the way for creating meaningful speech technology for a wider range of languages.

In this paper, we fine-tune pre-trained SSL models for ASR using varying amounts of data from four typologically diverse minority languages or language variants: Gronings, West-Frisian, Besemah and Nasal, which have a limited amount of data available. We specifically investigate whether data augmentation approaches can be used to generate additional training data to improve the performance of these models, particularly when very little resources are available. By using data from (ongoing) language documentation projects, we evaluate

wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations

Alexei Baevski Henry Zhou Abdelrahman Mohamed Michael Auli

XLS-R: SELF-SUPERVISED CROSS-LINGUAL SPEECH REPRESENTATION LEARNING AT SCALE

Arun Babu^{△*}, Chaghan Wang^{△*}, Andros Tjandra[△], Kushal Lakhota^{◇†}, Qiantong Xu[△],
Naman Goyal[△], Kritika Singh[△], Patrick von Platen[✱], Yatharth Saraf[△], Juan Pino[△],
Alexei Baevski[△], Alexis Conneau^{□‡}, Michael Auli^{△‡}

Robust Speech Recognition via Large-Scale Weak Supervision

Alec Radford^{*1} Jong Wook Kim^{*1} Tao Xu¹ Greg Brockman¹ Christine McLeavey¹ Ilya Sutskever¹

UNSUPERVISED CROSS-LINGUAL REPRESENTATION LEARNING FOR SPEECH RECOGNITION

Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, Michael Auli
Facebook AI

Understanding how limited training data affects model performance paves the way for creating more inclusive speech technology

Can standard data augmentation approaches
improve low-resource speech recognition
performance using real-world data?

Can standard data augmentation approaches
improve low-resource speech recognition
performance using real-world data?

Self-training and TTS-generated speech

XLST: Cross-lingual Self-training to Learn Multilingual Representation for Low Resource Speech Recognition

Zi-Qiang Zhang, Yan Song, Ming-Hui Wu, Xin Fang, Li-Rong Dai

GENERATING SYNTHETIC AUDIO DATA FOR ATTENTION-BASED SPEECH RECOGNITION SYSTEMS

Nick Rossenbach, Albert Zeyer, Ralf Schlüter, Hermann Ney

Human Language Technology and Pattern Recognition, Computer Science Department,
RWTH Aachen University, 52074 Aachen, Germany
AppTek GmbH, 52062 Aachen, Germany
<surname>@i6.informatik.rwth-aachen.de

SPEAKER AUGMENTATION FOR LOW RESOURCE SPEECH RECOGNITION

Chenpeng Du, Kai Yu

MoE Key Lab of Artificial Intelligence
SpeechLab, Department of Computer Science and Engineering
Shanghai Jiao Tong University, China
{duchenpeng, kai.yu}@sjtu.edu.cn

SELF-TRAINING FOR END-TO-END SPEECH RECOGNITION

Jacob Kahn, Ann Lee, Awni Hannun

Facebook AI Research

Published as a conference paper at ICLR 2023

CONTINUOUS PSEUDO-LABELING FROM THE START

Dan Berrebbi*

Carnegie Mellon University
dberrebb@andrew.cmu.edu

**Ronan Collobert, Samy Bengio,
Navdeep Jaitly, Tatiana Likhomanenko**

Apple
{collobert,bengio,njaitly,antares}@apple.com

PSEUDO-LABELING FOR MASSIVELY MULTILINGUAL SPEECH RECOGNITION

Loren Lugosch^{1}, Tatiana Likhomanenko^{2†}, Gabriel Synnaeve², Ronan Collobert^{2†}*

¹McGill University / Mila, ²Facebook AI Research

SPEECH RECOGNITION WITH AUGMENTED SYNTHESIZED SPEECH

Andrew Rosenberg, Yu Zhang, Bhuvana Ramabhadran, Ye Jia, Pedro Moreno, Yonghui Wu, Zelin Wu

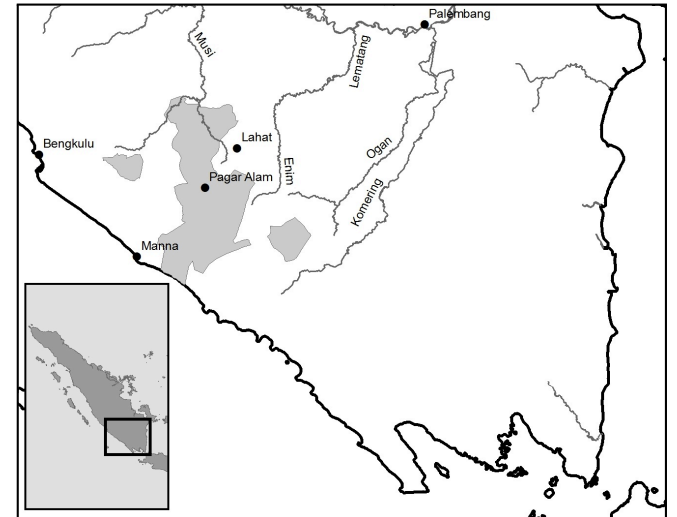
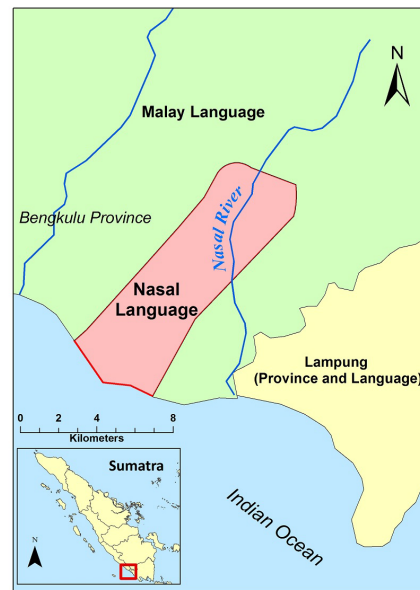
Google
{rosenberg, ngyuzh, bhuv, jiaye, pedro, yonghui, zelinwu}@google.com

MAGIC DUST FOR CROSS-LINGUAL ADAPTATION OF MONOLINGUAL WAV2VEC-2.0

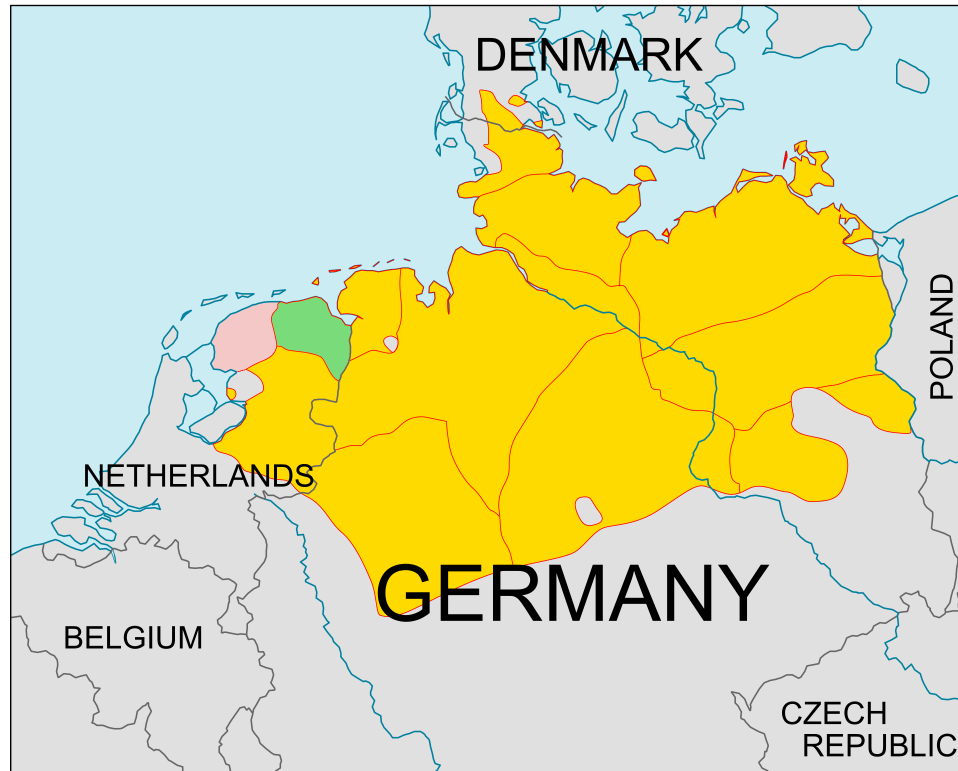
Sameer Khurana¹, Antoine Laurent², James Glass¹

¹MIT Computer Science and Artificial Intelligence Laboratory, Cambridge, MA, USA

²LIUM - Le Mans University, France

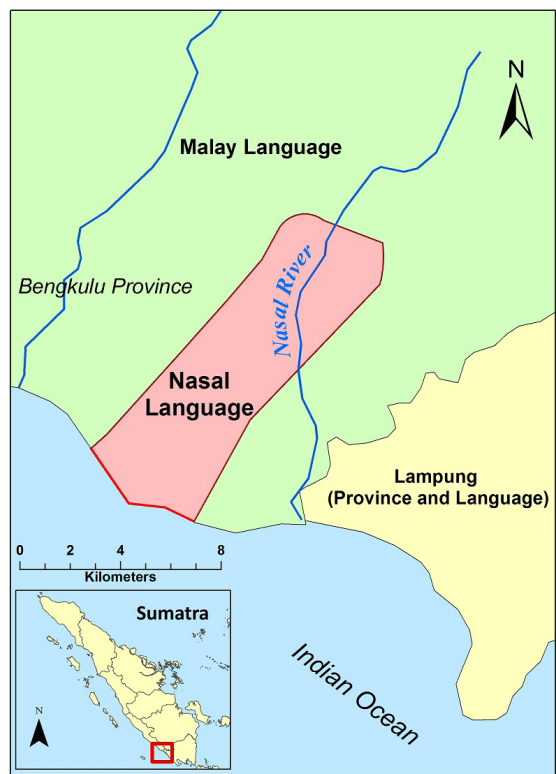


Images: https://en.wikipedia.org/wiki/Low_German (modified), Anderbeck, K., & Aprilani, H. (2013). The improbable language: Survey report on the Nasal language of Bengkulu, Sumatra. *SIL Electronic Survey Report*, 12, McDonnell, B. J. (2016). *Symmetrical voice constructions in Besemah: A usage-based approach*. University of California, Santa Barbara.

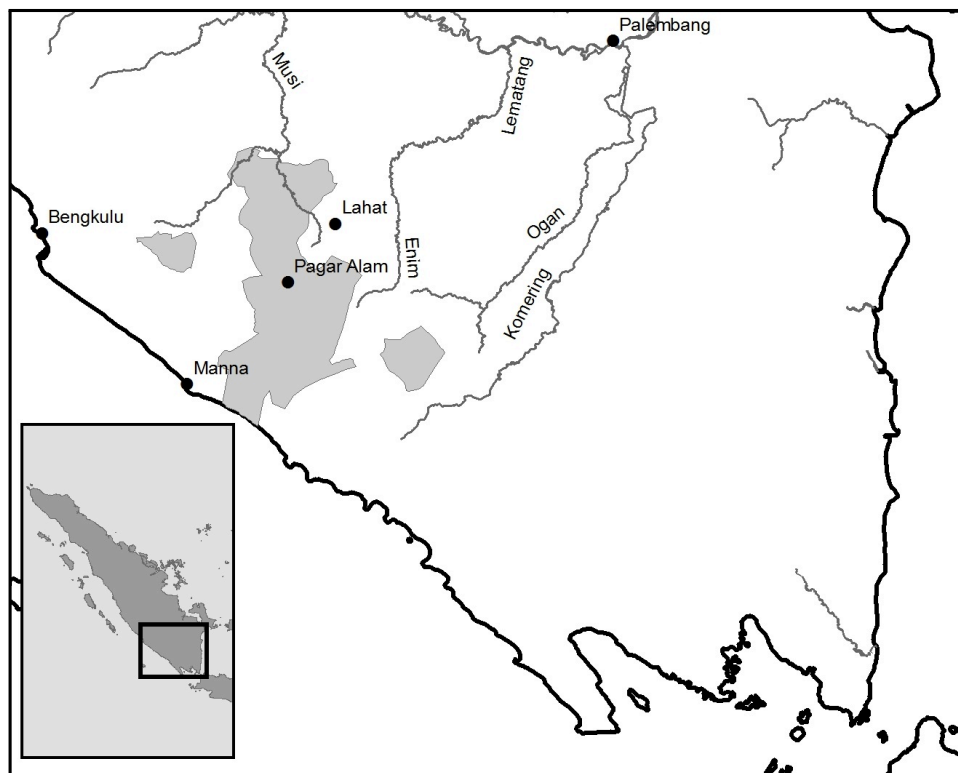


■ West-Frisian: ± 875,000 ■ Gronings: ± 260,000

Images: https://en.wikipedia.org/wiki/Low_German (modified), Anderbeck, K., & Aprilani, H. (2013). The improbable language: Survey report on the Nasal language of Bengkulu, Sumatra. *SIL Electronic Survey Report*, 12, McDonnell, B. J. (2016). *Symmetrical voice constructions in Besemah: A usage-based approach*. University of California, Santa Barbara.

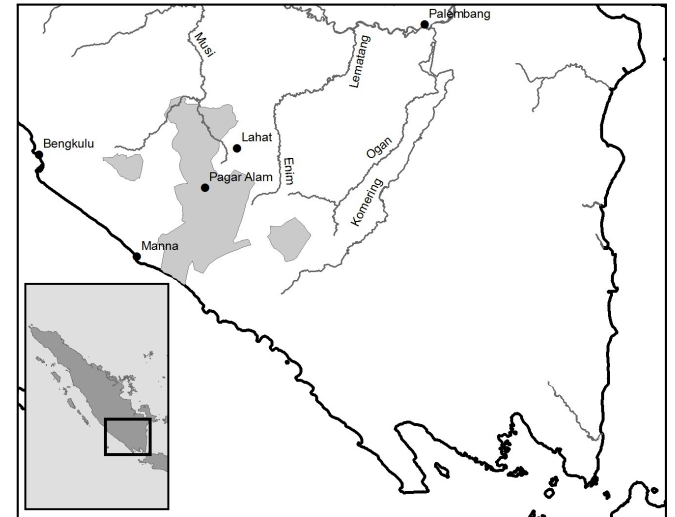
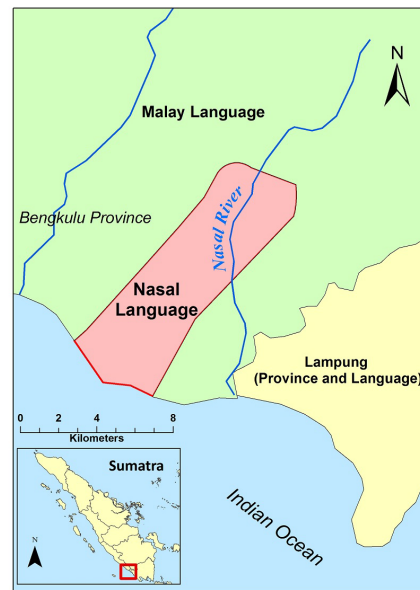


Nasal: ± 3,000



Besemah: ± 500,000

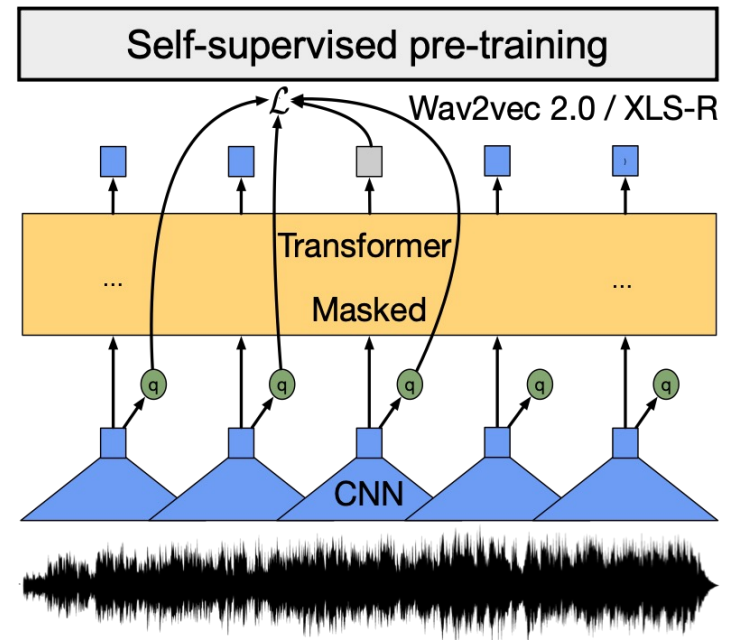
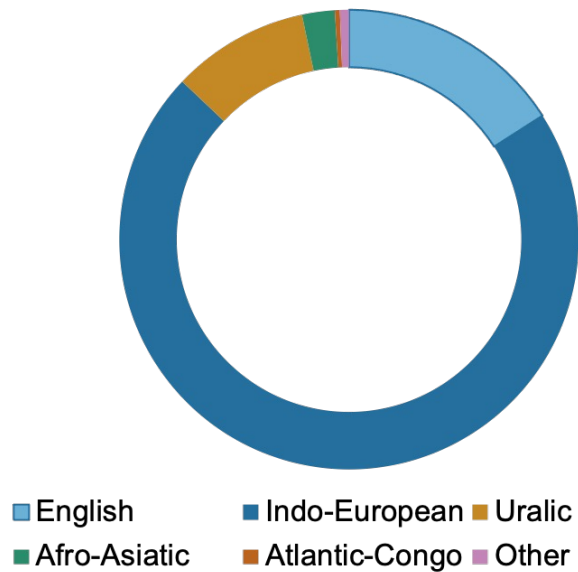
Images: https://en.wikipedia.org/wiki/Low_German (modified), Anderbeck, K., & Aprilani, H. (2013). The improbable language: Survey report on the Nasal language of Bengkulu, Sumatra. *SIL Electronic Survey Report*, 12, McDonnell, B. J. (2016). *Symmetrical voice constructions in Besemah: A usage-based approach*. University of California, Santa Barbara.



Images: https://en.wikipedia.org/wiki/Low_German (modified), Anderbeck, K., & Aprilani, H. (2013). The improbable language: Survey report on the Nasal language of Bengkulu, Sumatra. *SIL Electronic Survey Report*, 12, McDonnell, B. J. (2016). *Symmetrical voice constructions in Besemah: A usage-based approach*. University of California, Santa Barbara.

XLS-R

436,000 hours in 128 languages



Effect of data size

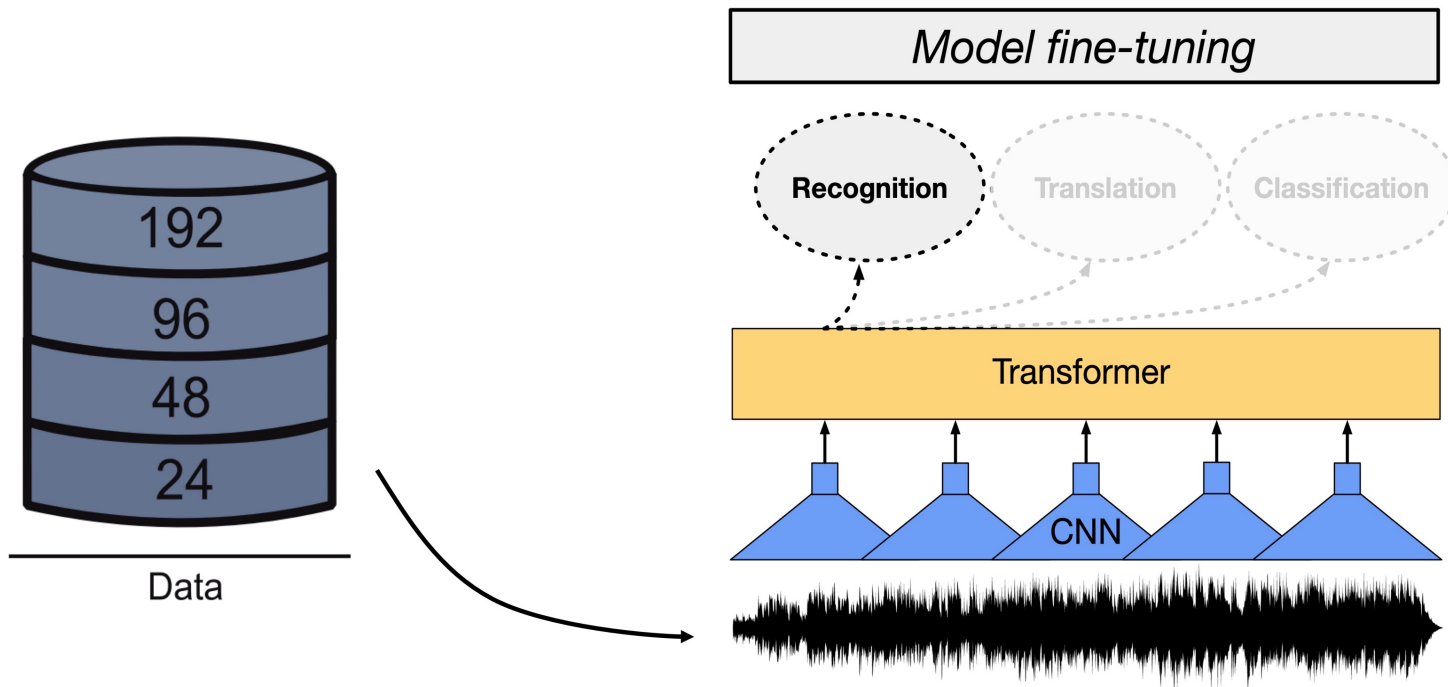
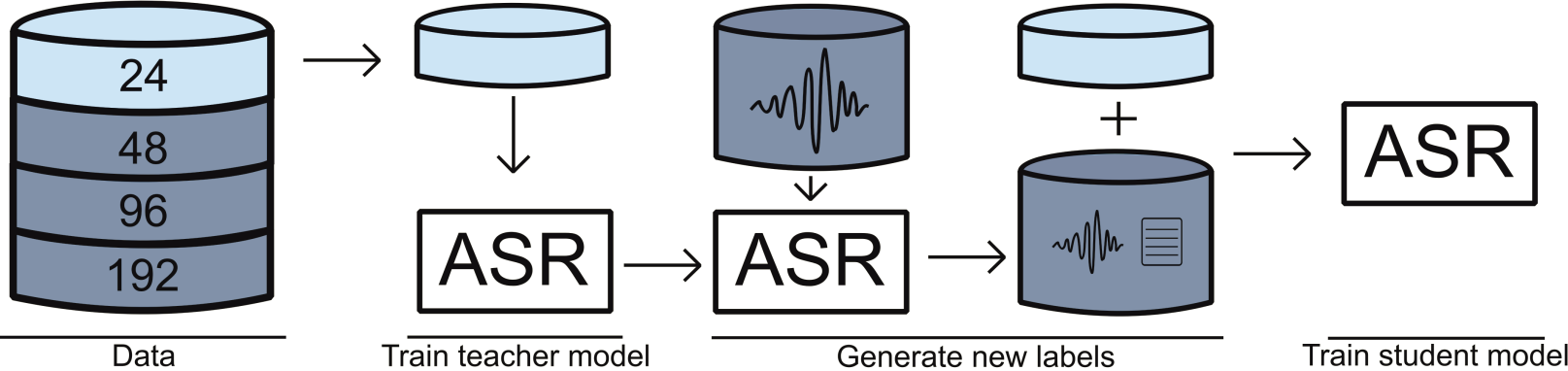
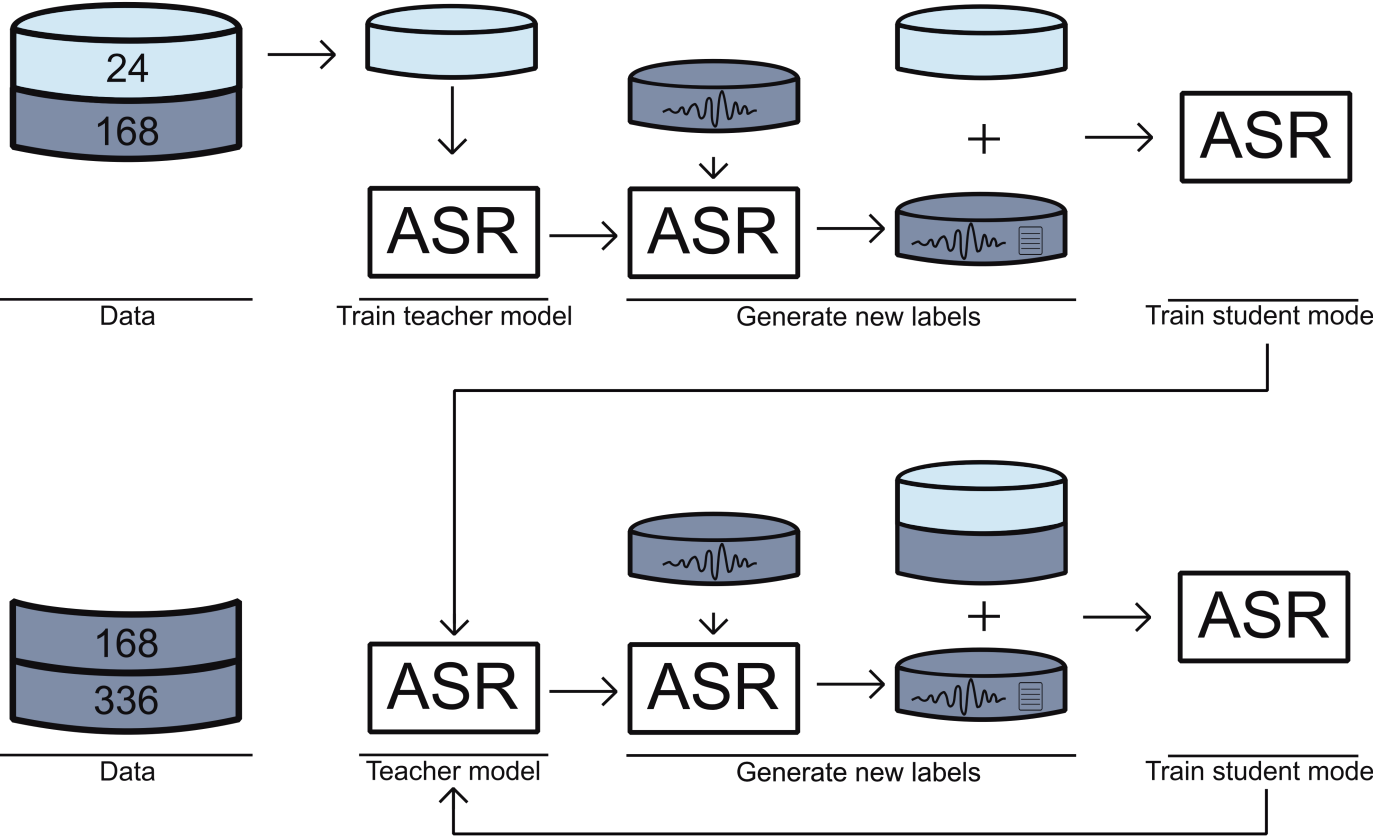


Image (modified): Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, Alexei Baevski, Alexis Conneau, and Michael Auli. 2021. XLS-R: Self-supervised Cross-lingual Speech Representation Learning at Scale.

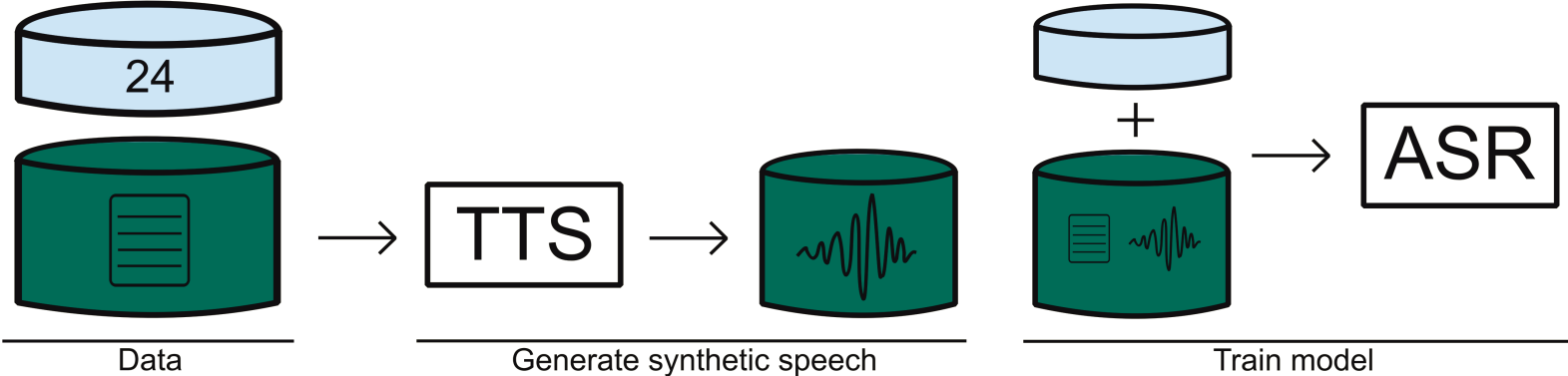
Self-training

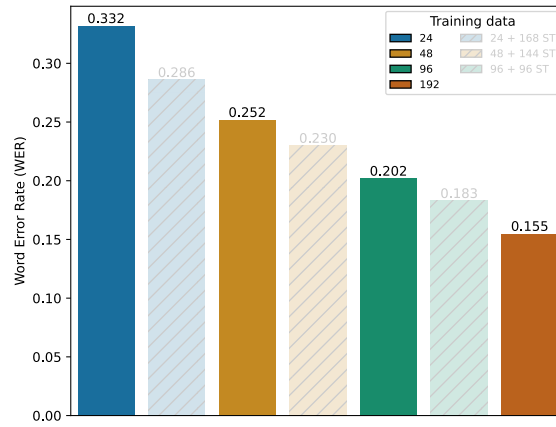


Self-training on Gronings

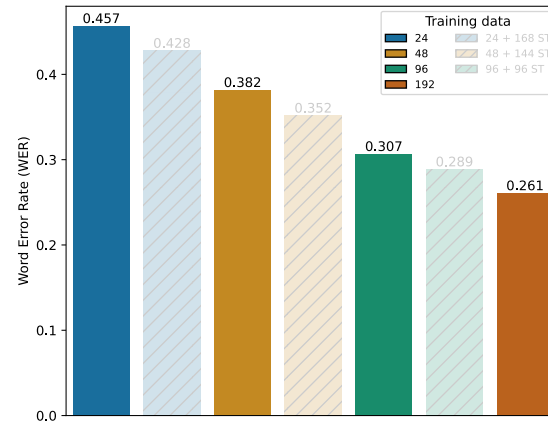


Gronings synthetic speech

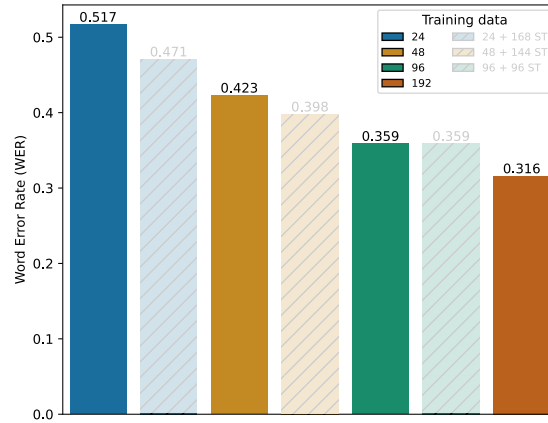




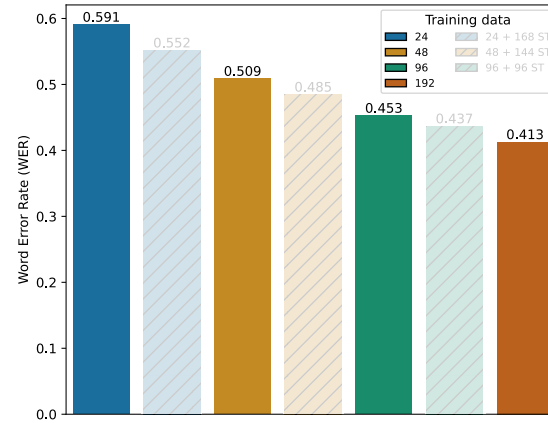
(a) Results for the Gronings test set.



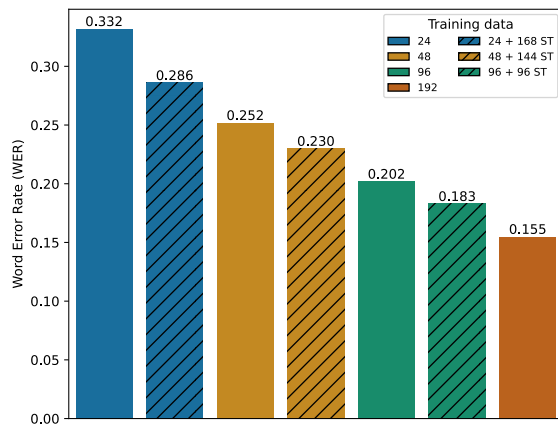
(b) Results for the West-Frisian test set.



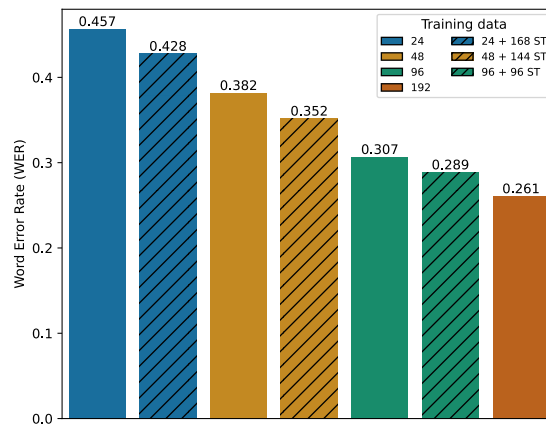
(c) Results for the Besemah test set.



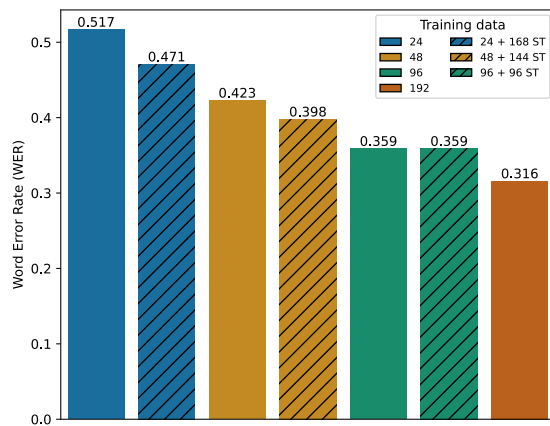
(d) Results for the Nasal test set.



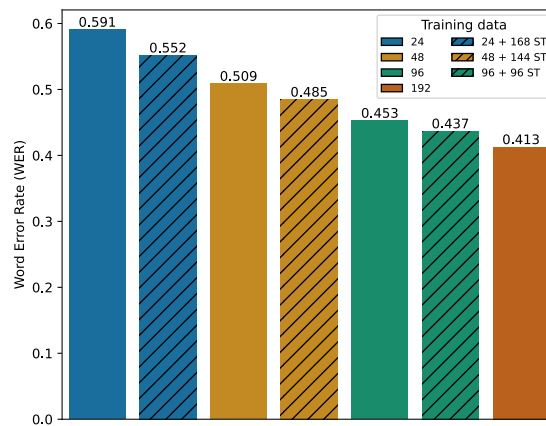
(a) Results for the Gronings test set.



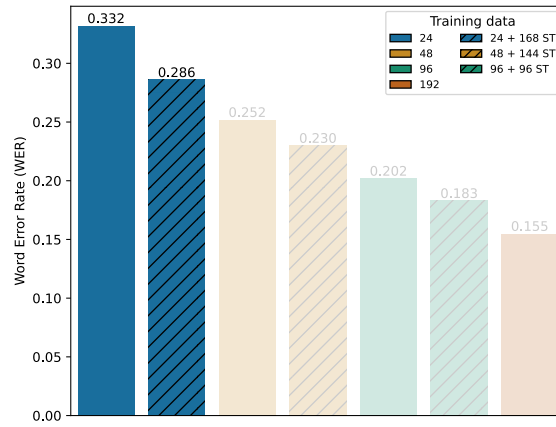
(b) Results for the West-Frisian test set.



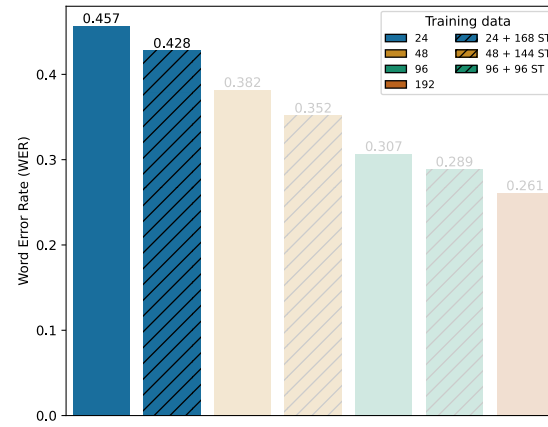
(c) Results for the Besemah test set.



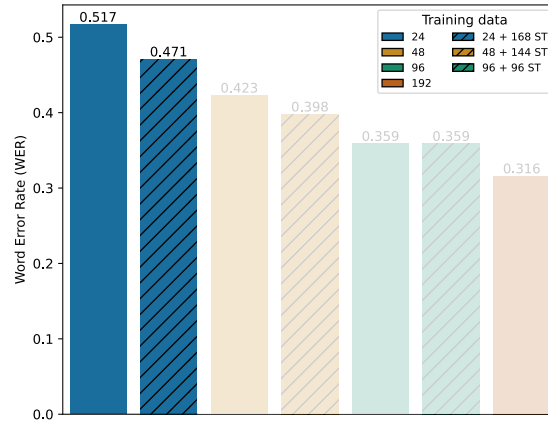
(d) Results for the Nasal test set.



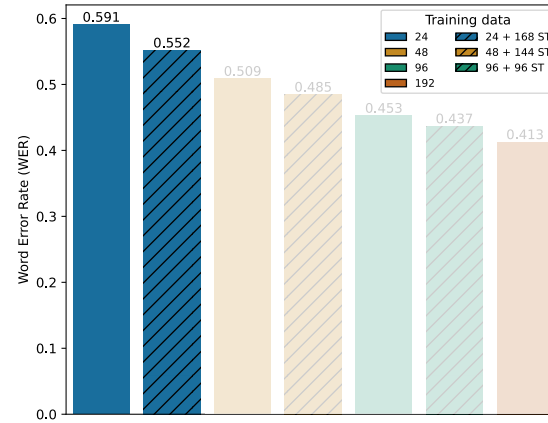
(a) Results for the Gronings test set.



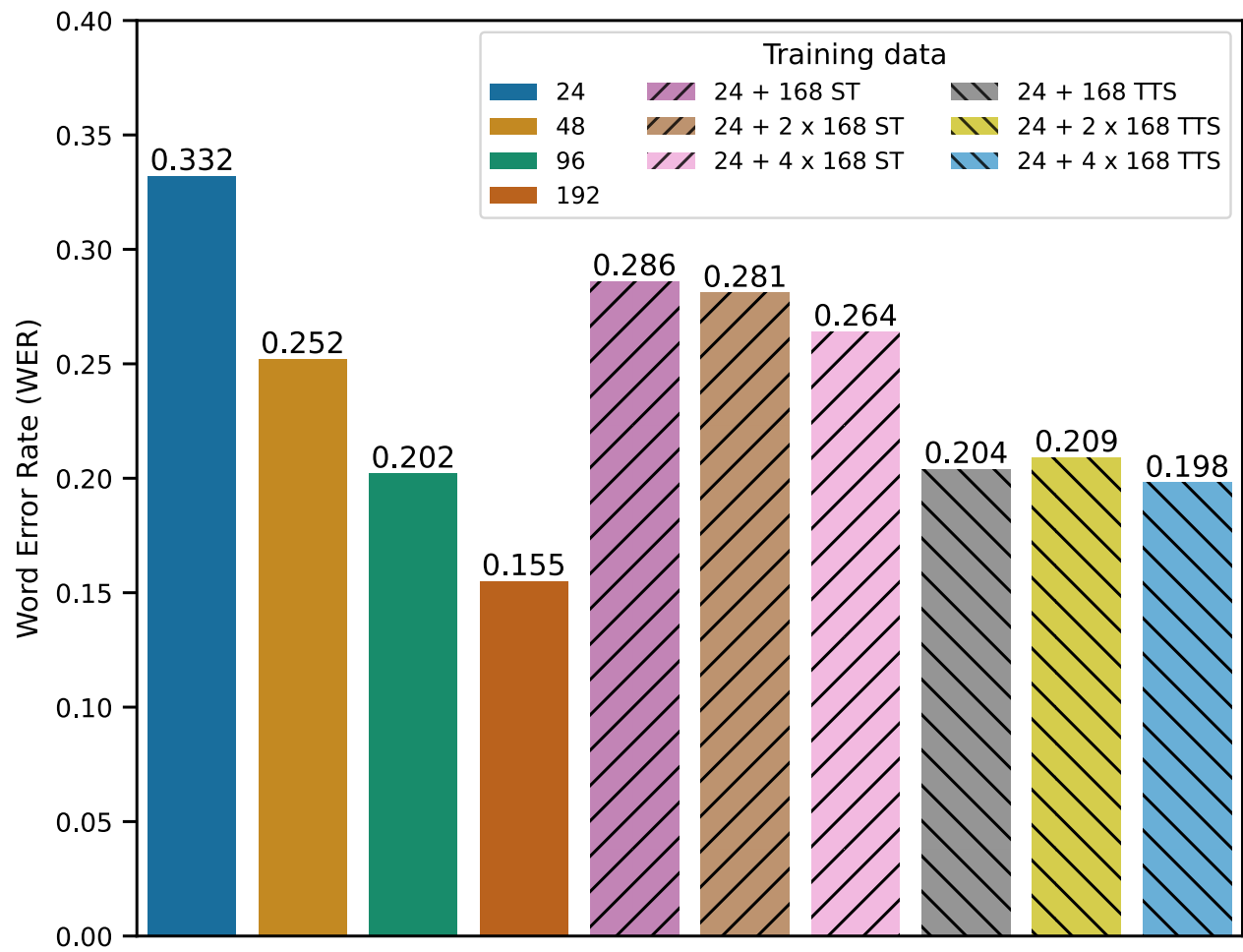
(b) Results for the West-Frisian test set.

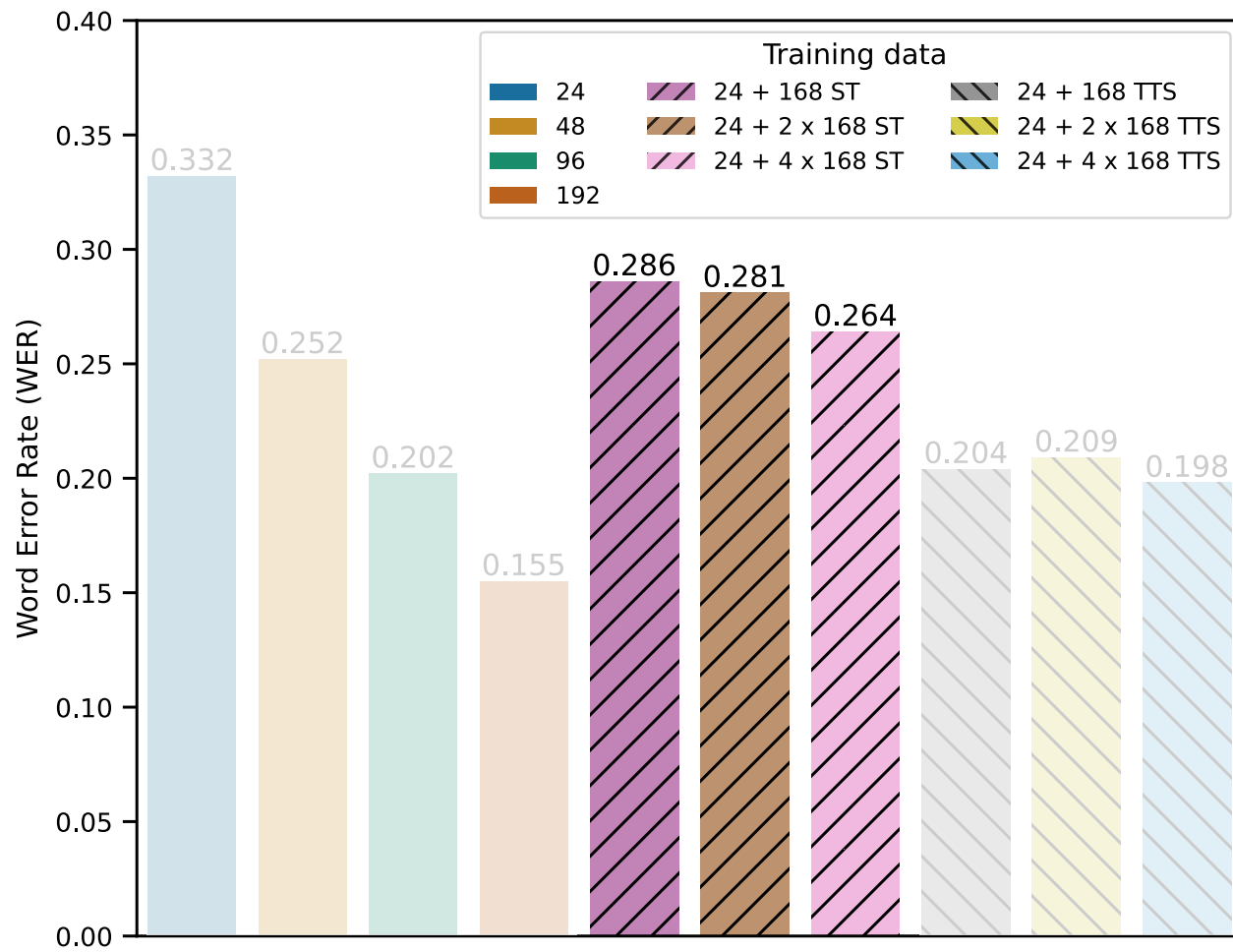


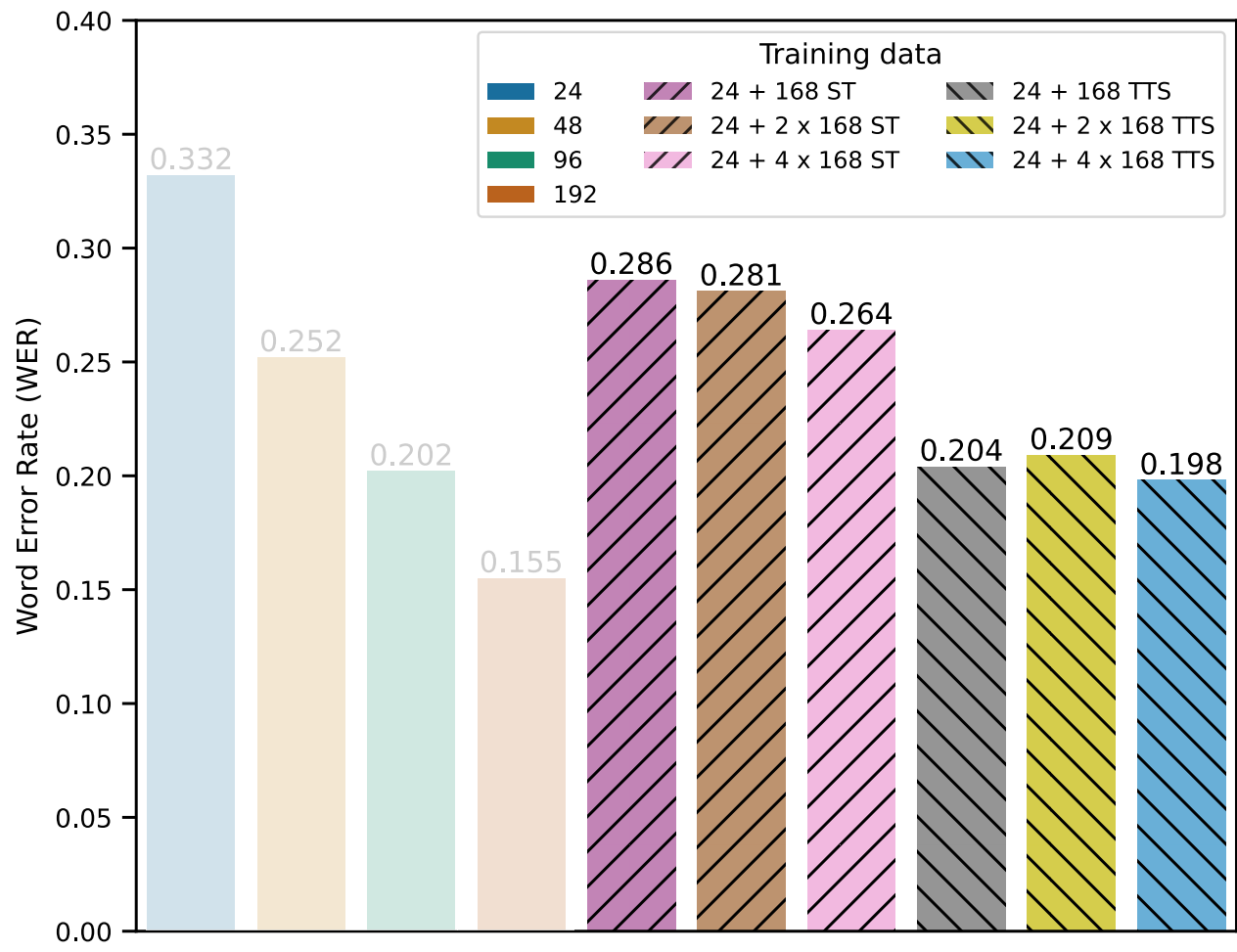
(c) Results for the Besemah test set.

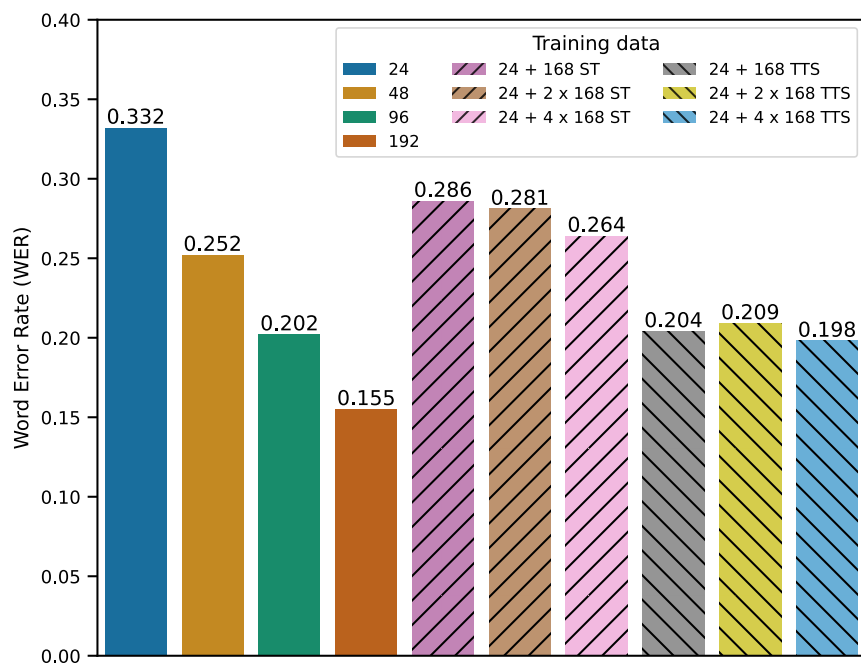


(d) Results for the Nasal test set.

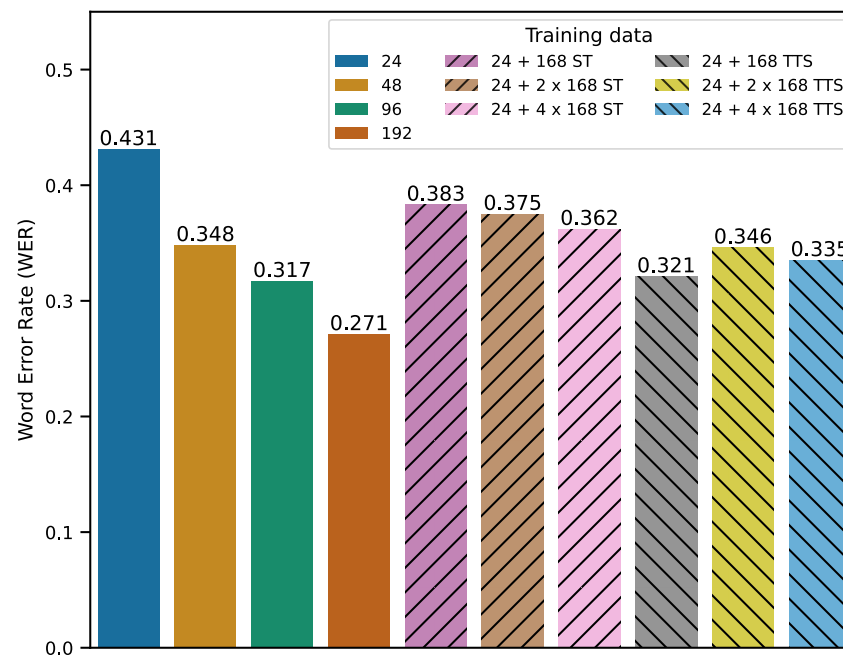




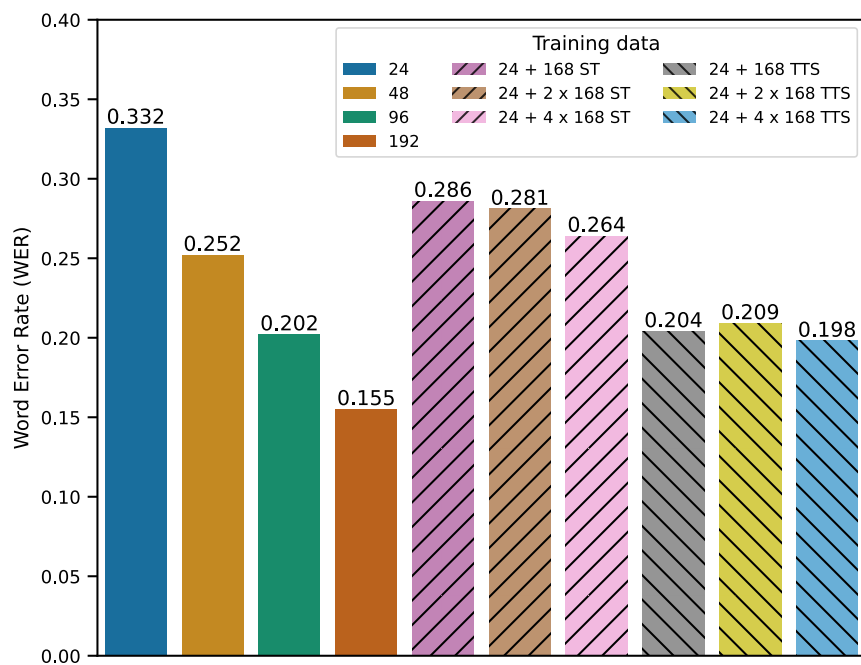




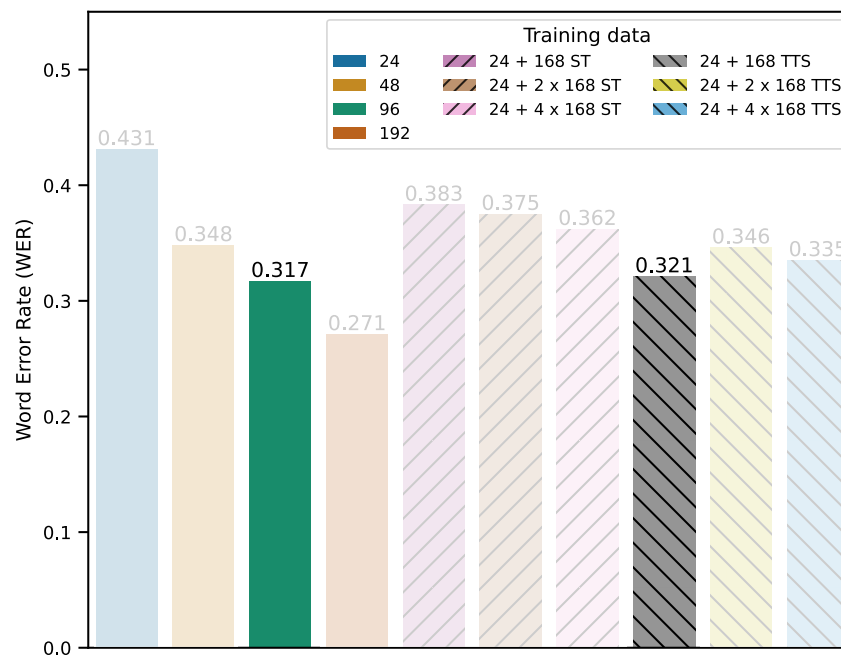
(a) Results for the regular Gronings test set.



(b) Results for the out-of-domain Gronings test set.



(a) Results for the regular Gronings test set.



(b) Results for the out-of-domain Gronings test set.

Summary

We found that:

Data augmentation techniques may serve as a **cost-effective** way to improve low-resource ASR performance in a real-world setting;

Summary

We found that:

Data augmentation techniques may serve as a **cost-effective** way to improve low-resource ASR performance in a real-world setting;

The largest performance gains were observed when increasing the amounts of manually transcribed data;

Summary

We found that:

Data augmentation techniques may serve as a **cost-effective** way to improve low-resource ASR performance in a real-world setting;

The largest performance gains were observed when increasing the amounts of manually transcribed data;


We hope our experiments help further the development of more inclusive speech technology.


Summary

To help researchers to include real-world data into their analysis and applications, we publicly release our datasets in addition to our code and models.

Summary

 github.com/Bartelds/neural-acoustic-distance (Journal of Phonetics 2022)

 github.com/Bartelds/language-variation (NAACL 2022)

 github.com/Bartelds/asr-augmentation (ACL 2023)

 bartelds@stanford.edu |  [@barteldsmartijn](https://twitter.com/barteldsmartijn)