The Impact of Dialect Variation on Robust Automatic Speech Recognition for Catalan

Zachary Hopton and **Eleanor Chodroff**
{zacharywilliam.hopton,eleanor.chodroff}@uzh.ch
University of Zurich

Abstract

To accurately transcribe a speech signal, automatic speech recognition (ASR) systems must show robustness to a wide range of task-independent variation, such as speaker factors, recording quality, or even "adversarial noise" designed to disrupt performance.We manipulated the dialect composition of fine-tuning data for ASR to study whether balancing the relative proportion of dialects had an impact on models' robustness to two such sources of variation: dialect variation and adversarial perturbations. We fine-tuned XLSR-53 for Catalan ASR using four different dialect compositions, each containing the Central Catalan dialect. These were defined as 100%, 80%, 50%, and 20% Central Catalan, with the remaining portions split evenly between four other Catalan dialects. While increasing the relative proportion of dialect variants improved models' dialect robustness, this did not have a meaningful impact on adversarial robustness. These findings suggest that while improvements to ASR can be made by diversifying the training data, such changes do not sufficiently counteract adversarial attacks, leaving the technology open to security threats.

https://github.com/zhopto3/DialAttack

1 Introduction

Effectively handling dialect variation is an important attribute of a high-performing automatic speech recognition system. While incorporating dialect variation into a model's training data may benefit this robustness (Jie et al., 2024; Dan et al., 2022; Lonergan et al., 2023), the relative lack of speech data with clean accent or dialect labels poses a challenge for this line of research. Moreover, the exact approach to incorporating data from other dialects could have consequences beyond just the model's ASR performance. Research on the robustness of neural networks to adversarial attacks has indicated that unbalanced training datasets may result in models that are more susceptible to adversarial attacks (Wu et al., 2021; Richards et al., 2023). When evaluating the susceptibility of Open AI's Whisper to adversarial noise, Olivier and Raj (2023) found that attacks on the model's language identification token were more effective when the attack's language was less present in the model's training data. For models trained on data from mutually intelligible dialects, it is an open question as to whether an unbalanced dataset would increase susceptibility to adversarial attacks. Training on multiple dialects may even confer a robustness to adversarial noise, analogous to training models on geometrically transformed or adversarial examples (Silva and Najafirad, 2020).

Catalan's well-documented dialect variation makes it a good candidate for studying questions of multi-dialect ASR (Veny, 2015, 1982; Calvo and Segura-Llopes, 2022). A great deal of research has focused on the compilation of Catalan corpora (Kjartansson et al., 2020; Kulebi et al., 2022; Boleda et al., 2006; Ljubešić and Toral, 2014). Catalan also has a substantial presence in the Common Voice corpus, for which diverse speakers of the language write, record, and validate data points on a voluntary basis (Ardila et al., 2020). Catalan's presence in Common Voice has grown substantially in recent releases of the corpus thanks to data management and campaigning efforts from a number of bodies (Armentano-Oller et al., 2024). Notable among these efforts is the encouragement of Catalan speakers with various accents to contribute to the corpus, and to include their accent in the metadata of the recording.

Here we use data from five Catalan dialects to answer two questions about multi-dialect ASR: First, to what extent is balancing the quantity of data from different dialects necessary when training models meant to accurately transcribe multiple varieties of a language? And second, does training an ASR

Proceedings of the The 22nd SIGMORPHON workshop on Computational Morphology, Phonology, and Phonetics, pages 23–33

model on an unbalanced dataset with multiple dialects impact the model's security at inference?

2 Methods

2.1 Multi-dialect Catalan Speech Recognition

The first step of our experiments consisted of fine-tuning XLSR-53 for Catalan ASR (Conneau et al., 2021). We used XLSR-53 instead of a later version of XLSR since Catalan was not part of the pretraining data for this model, eliminating a potential confound in the manipulation of the dialect composition in the fine-tuning data[1].

Our data source was the validated portion of Catalan Common Voice 18.0[2] (Ardila et al., 2020). Similar to Armentano-Oller et al. (2024), we first mapped each unique accent label for a given audio file to one of five Catalan macro-dialects: Balearic, Central, Nord, Nord-Occidental, and Valencià. Audio files from second language learners were excluded, as were underspecified accent responses such as "Catalan" or "normative."

We then sampled four training and development datasets with 100%, 80%, 50%, and 20% Central Catalan; the remaining portion of each set was split evenly between the other four dialects. This meant that the 20% Central train and development datasets were perfectly balanced with respect to the five macro-dialects studied. We randomly sampled a single test set from the remaining data with an equal number of samples from each macro-dialect. Figure 2 shows the final proportion of each dialect in the four models' training data. All four models were fine-tuned on 152 hours of data with 19 hours of development data. This training set size is comparable to previous work that has used XLSR for multi-dialect speech processing (Zuluaga-Gomez et al., 2023; Lonergan et al., 2023). See Appendix A for fine-tuning details.

2.2 Robustness to Adversarial Noise

After training our ASR models, we randomly selected 50 audio files from each dialect's evaluation set to train a total of 250 adversarial attacks on each of the four models. Put generally, we aimed to add noise to our input audio files that resulted in the model outputting an adversarial target—"Porta'm a un lloc web malvat," or "Take me to an evil website"—despite the perceivable audio input saying something else. More specifically, we trained targeted, adversarial noise δ for a given audio file x as in Carlini and Wagner (2018). In such Carlini-and-Wagner (CW) attacks, the objective is also to make the noise relatively imperceptible. As in Olivier and Raj (2023), we judged the perceptibility of the noise relative to the signal in decibels using the signal-to-noise ratio (SNR), but using the L_∞ metric, similar to Carlini and Wagner (2018):

$$\mathrm{SNR}(x, \delta) = 20(\log |x|_\infty - \log |\delta|_\infty) \quad (1)$$

For details on the algorithm and hyperparameters used to train the adversarial noise δ, see Appendix B.

2.3 Evaluation

Following fine-tuning, we ran inference over the withheld test set of Catalan data. Each ASR model was evaluated on the same 19-hour evaluation dataset with equal representation from each dialect (2486 audio files per dialect).

To assess variation in WER and CER, a generalized linear model for a gamma-distributed dependent variable was implemented in R, using the identity link function. The model included main effects of the model dialect composition (100%, 80%, 50%, or 20% Central Catalan), the speech input dialect (Balear, Central, Nord, Nord-Occidental, or Valencià), and the interaction between the model dialect composition and speech dialect. A gamma distribution was chosen given that the WER and CER distributions have a strong positive skew and cannot be negative.[3] Each predictor was sum-coded with the held-out levels first set to the 100% Central model and the Catalan test files, and then rotated to test each main effect and interaction against the average performance.

To evaluate the effectiveness of the CW attacks on each fine-tuned model, we primarily use the percentage of successful attacks. An attack was considered successful if—at any SNR—the WER of the model output compared to the adversarial target ("Porta'm a un lloc web malvat.") was 0. To assess the influences on a successful or unsuccessful attack, we implemented a binomial logistic regression model with fixed effects of dialect, the fine-tuning data composition of the model, and

[1] https://github.com/facebookresearch/fairseq/tree/main/examples/wav2vec

[2] https://commonvoice.mozilla.org/en/datasets

[3] As the gamma prediction can only predict positive values greater than 0, we transformed any WER or CER of 0 in our data to 0.00001.

	WER					CER				
	bal	cen	nor	no-oc	val	bal	cen	nor	no-oc	val
100%	0.280	0.152	0.191	0.244	0.234	0.079	0.038	0.049	0.064	0.056
80%	0.224	0.146	0.160	0.196	0.165	0.060	0.037	0.040	0.050	0.039
50%	**0.181**	**0.140**	**0.133**	0.164	0.126	**0.047**	**0.035**	**0.033**	**0.041**	**0.029**
20%	0.189	0.164	0.144	**0.161**	**0.125**	0.050	0.043	0.036	0.042	0.035

Table 1: Percentages in the far left column represent the percent Central Catalan data used in fine-tuning; bal: Balearic, cen: Central, nor: Nord, no-oc: Nord-Occidental, val: Valencià; Bold: lowest WER/CER for each dialect.

their interaction. Categorical predictors were sum-coded.

3 Results

3.1 ASR

In terms of WER and CER, ASR for all dialects improved with more dialect-balanced data (Table 1). The WER model revealed significant influences of model composition and dialect input: relative to average, the 20% and 50% Central models had significantly lower WERs, while the 100% Central model had significantly higher WERs across all dialects. The average performance across models was significantly better on Central, Nord and Valencià, but worse on Balear and Nord-Occidental. The main effects were significantly tempered in several interactions; a significant interaction can be interpreted as a significant modulation from the expected WER performance based on the main effects alone. Beyond the main effects, the 20% Central model performed significantly better on Balear, Nord-Occidental, and Valencià, but worse on Central and Nord. The 50% Central model also performed significantly better on Balear and Valencià, but the main effect of the 50% Central model was significantly tempered for the Central test files: While the Central WER of the 50% Central model was numerically lowest across model types, the improvement was not as great as expected based on the main effects alone. The 100% Central model had significantly improved performance on Central and Nord, but significantly worse performance on Balear, Nord-Occidental, and Valencià. For the full model results, see Table 5 in Appendix C).

For the CER model, the same pattern of significance emerged for the main effects, but the interactions differed slightly. While the 20% Central model still had significantly higher performance on Balear and Nord-Occidental, its performance was significantly worse on Central test files. In addition, while the 50% model still had higher performance on Balear and Valencià, it performed significantly worse on Central and Nord. The full results can be found in Table 6 in Appendix D.

Measured with WER and CER, the 50% Central model had a consistently strong performance across dialects, followed closely by the 20% Central model. Performance was generally higher for Central, Nord and Valencian dialects, but lower for the Balearic and Nord-Occidental dialects.

3.2 CW Attacks

We obtained a high percentage of successful attacks in all conditions (see Figure 1). The attacks' high average SNR implies that successful attacks were relatively imperceptible (see Tables 3 and 4). Given that our models had relatively low WER and CER on the non-adversarial test set, these results are in line with the common finding in the adversarial attack literature that even high-performing models are susceptible to adversarial perturbations. The logistic regression yielded no significant main effects and only one significant interaction, indicating that adversarial attacks using Central Catalan audio were significantly less successful in the 80% Central model relative to main effects alone. For the full model results, see Table 7 in Appendix E).

4 Discussion

In the present study, we manipulated the balance of five different dialects of Catalan in a dataset that we used to fine-tune XLSR-53 for automatic speech recognition. We tested how biasing a dataset toward one variety (Central Catalan) would affect the robustness of the model to both dialect variation and targeted adversarial noise at inference.

With respect to ASR performance on multiple dialects, we found that including larger portions of different dialects in fine-tuning data does make for a model that is more robust to dialect variation at test time. However, it is not necessary for a model to be perfectly balanced with respect to dialect composition to obtain maximal gains in performance. Other researchers have studied how to make models that are more robust to dialect variation at test time, for instance focusing on the config-

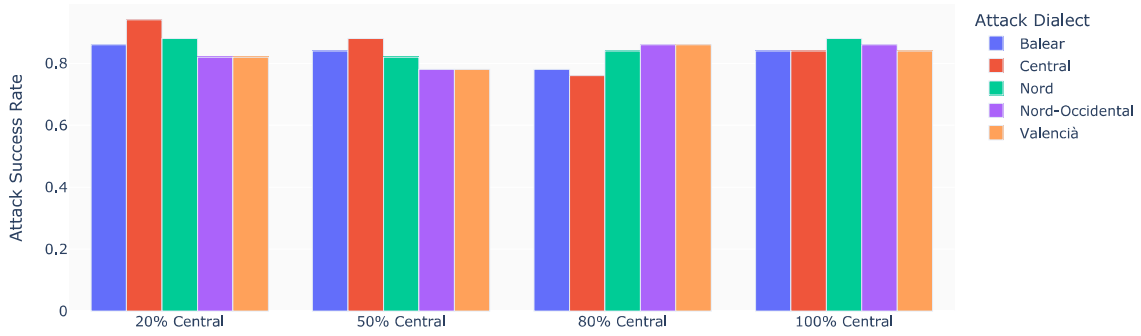

Figure 1: Attack success rate broken down by the dialect of the attack's audio (bar color) and the proportion of Central Catalan in the models' fine-tuning data (bar group).

uration of language models used during decoding (Yadavalli et al., 2022), the usefulness of normalizing dialects' scripts in the training data (Nigmatulina et al., 2020), or the utility of jointly learning to identify the variety and perform ASR for it (Waheed et al., 2023). While our experiments required relatively large amounts of data for which we knew the dialect of origin, we were encouraged about the implications of the results for low-resource dialect settings, as we see that even if half of the fine-tuning data used is from a single dialect, we can still observe substantial ASR improvements in performance for various dialects. Similar to a related study on the impact of balanced corpora on multi-dialect Irish ASR, we found that a perfectly balanced fine-tuning corpus does not lead to equivalent performance across dialects (Lonergan et al., 2023). Indeed, in our perfectly balanced model, the Balearic dialect stands out as having a relatively high WER (though the CER scores indicate closer performance across dialects). As suggested by Lonergan et al. (2023), this implies that the Balearic dialect may need to be up-sampled in future corpus compilation efforts.

As for adversarial noise, our results indicate that systematically adding dialect variation to fine-tuning data for ASR is neither helpful nor hurtful in the case of XLSR-53. Though CW attacks have previously been shown to be powerful against ASR systems, it is still interesting to consider if there were any parts of our experiments that made the models particularly vulnerable to the attacks. One potential susceptibility may be the presence of noisy labels in the training data. Paleka and Sanyal (2023) demonstrate that mislabeled images in training data can result in a loss of adversarial robustness for image classification models. We restricted our datasets to the validated text-audio pairs of Catalan data in Common Voice 18.0, and sentences contributed to the dataset for Catalan have undergone substantial validation (Armentano-Oller et al., 2024). Still, there is no guarantee that the dialect of the sentence matches the accent with which the reader speaks Catalan in the dataset. For instance, in our test set, we see a reader who speaks Valencià was assigned a sentence containing the feminine, third-person possessive pronoun from Central Catalan, "la seva." In Valencià, however, it is written and spoken as "la seua" (Calvo and Segura-Llopes, 2022), which is what our 100% Central Catalan model transcribes for this data point. Thus, it is feasible that there is inconsistency in our data's labels (the text) and what is actually spoken in the audio. A study of how languages' orthographic transparency impacts adversarial vulnerability in ASR models would be an interesting means of exploring the impact of such noisy labels.

5 Conclusion

Taking Catalan as the language of study given the large amounts of available annotated data, this study demonstrated that a more balanced dialect composition indeed confers robustness to dialect variation in test data. However, dialect composition of the training data had little influence on adversarial robustness. We hope that these findings will motivate the consideration of datasets' dialect composition in the development of ASR systems in the

future. Indeed, our findings suggest that even if the dialects present in training are not perfectly balanced, including such variation to some degree is beneficial for *all* dialects represented in the training data (even for Central Catalan, in this case). As for adversarial robustness, we encourage further work on multilingual and multi-dialect speech processing models to assess specific vulnerabilities that might come from unbalanced datasets or mismatched labels and audio that spur from orthographic depth or dialect variation.

Limitations

Given that languages' dialects can differ in their mutual intelligibility, an important limitation arises in the use of only Catalan's varieties. It may be the case that for a language with less mutually intelligible dialects, less cross-dialect transfer is possible. Though we predict this would make for worse performance on lower resource dialects in the biased dataset conditions, more work on a larger sample of languages is needed. Collecting such multilingual, transcribed speech datasets with accent annotations presents a limitation in itself to this line of work. However, this paper and Zuluaga-Gomez et al. (2023)—who use Common Voice to create such a multiilingual dataset with accent labels—demonstrate that in some cases, existing datasets can be repurposed to study multi-dialect speech processing.

As we worked with limited computational resources, we were only able to fine-tune one time per data composition. Ideally, we would repeat the fine-tuning with several random samples and report the average results, but this was not feasible here. We encourage repetition of our experiments using other ASR architectures, including larger versions of XLSR, and with other languages and their dialects.

Ethical Considerations

This work studies adversarial attacks on automatic speech recognition, which could potentially be used to alter the behavior of ASR models with malicious intentions. We do not introduce any new algorithms for attacking models, and conducted the study with the intent of studying if multi-dialect speech processing models are more or less susceptible to existing attacks. In doing so, we hoped to assess not just the quality, but also the trustworthiness of speech recognition models that could potentially be used by speakers of lower-resource language varieties.

Acknowledgements

This research was supported by SNSF Grant PR00P1_208460 to EC.

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 Twelfth Language Resources and Evaluation Conference*, pages 4218–4222, Marseille, France. European Language Resources Association.

Carme Armentano-Oller, Montserrat Marimon, and Marta Villegas. 2024. Becoming a high-resource language in speech: The Catalan case in the Common Voice corpus. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2142–2148.

Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick Platen, Yatharth Saraf, Juan Pino, Alexei Baevski, Alexis Conneau, and Michael Auli. 2022. XLS-R: Self-supervised cross-lingual speech representation learning at scale. In *Interspeech, 2022*, pages 2278–2282.

Gemma Boleda, Stefan Bott, Rodrigo Meza, Carlos Castillo, Toni Badia, and Vicente López. 2006. CUCWeb: A Catalan corpus built from the web. In *Proceedings of the 2nd International Workshop on Web as Corpus*.

Vicent Beltran Calvo and Carles Segura-Llopes. 2022. *Els parlars valencians (actualitzada)*, volume 34. Universitat de València.

Nicholas Carlini and David Wagner. 2018. Audio adversarial examples: Targeted attacks on speech-to-text. In *2018 IEEE Security and Privacy Workshops (SPW)*, pages 1–7. IEEE.

Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli. 2021. Unsupervised cross-lingual representation learning for speech recognition. In *Interspeech, 2021*.

Zhengjia Dan, Yue Zhao, Xiaojun Bi, Licheng Wu, and Qiang Ji. 2022. Multi-task transformer with adaptive cross-entropy loss for multi-dialect speech recognition. *Entropy*, 24(10):1429.

Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. 2006. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the*

23rd international conference on Machine learning, pages 369–376.

Zhou Jie, Gao Shengxiang, Yu Zhengtao, Dong Ling, and Wang Wenjun. 2024. DialectMoE: An end-to-end multi-dialect speech recognition model with mixture-of-experts. In *Proceedings of the 23rd Chinese National Conference on Computational Linguistics (Volume 1: Main Conference)*, pages 1148–1159.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Oddur Kjartansson, Alexander Gutkin, Alena Butryna, Isin Demirsahin, and Clara Rivera. 2020. Open-source high quality speech datasets for Basque, Catalan and Galician. In *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)*, pages 21–27.

Baybars Kulebi, Carme Armentano-Oller, Carlos Rodriguez-Penagos, and Marta Villegas. 2022. ParlamentParla: A speech corpus of Catalan parliamentary sessions. In *Proceedings of the Workshop ParlaCLARIN III within the 13th Language Resources and Evaluation Conference*, pages 125–130, Marseille, France. European Language Resources Association.

Nikola Ljubešić and Antonio Toral. 2014. caWaC – a web corpus of Catalan and its application to language modeling and machine translation. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1728–1732, Reykjavik, Iceland. European Language Resources Association (ELRA).

Liam Lonergan, Mengjie Qian, Neasa Ní Chiaráin, Christer Gobl, and Ailbhe Ní Chasaide. 2023. Towards dialect-inclusive recognition in a low-resource language: Are balanced corpora the answer? In *Interspeech 2023*, pages 5082–5086.

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).

Raphaël Olivier and Bhiksha Raj. 2023. There is more than one kind of robustness: Fooling Whisper with adversarial examples. In *24th Annual Conference of the International Speech Communication Association, Interspeech 2023, Dublin, Ireland, August 20-24, 2023*, pages 4394–4398.

Daniel Paleka and Amartya Sanyal. 2023. A law of adversarial risk, interpolation, and label noise. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*.

Luke E Richards, Edward Raff, and Cynthia Matuszek. 2023. Measuring equality in machine learning security defenses: A case study in speech recognition. In *Proceedings of the 16th ACM Workshop on Artificial Intelligence and Security*, pages 161–171.

Samuel Henrique Silva and Peyman Najafirad. 2020. Opportunities and challenges in deep learning adversarial robustness: A survey. *CoRR*.

Joan Veny. 1982. *Els parlars catalans: síntesi de dialectologia*. Biblioteca 'Raixa'. Moll.

Joan Veny. 2015. Català occidental/català oriental, encara. *Estudis Romànics*, 37:31–65.

Abdul Waheed, Bashar Talafha, Peter Sullivan, Abdelrahim Elmadany, and Muhammad Abdul-Mageed. 2023. VoxArabica: A robust dialect-aware arabic speech recognition system. In *Proceedings of Arabic-NLP 2023, Singapore (Hybrid), December 7, 2023*, pages 441–449. Association for Computational Linguistics.

Tong Wu, Ziwei Liu, Qingqiu Huang, Yu Wang, and Dahua Lin. 2021. Adversarial robustness under long-tailed distribution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8659–8668.

Aditya Yadavalli, Ganesh Sai Mirishkar, and Anil Vuppala. 2022. Exploring the effect of dialect mismatched language models in Telugu automatic speech recognition. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Student Research Workshop*, pages 292–301, Hybrid: Seattle, Washington + Online. Association for Computational Linguistics.

Juan Zuluaga-Gomez, Sara Ahmed, Danielius Visockas, and Cem Subakan. 2023. Commonaccent: Exploring large acoustic pretrained models for accent classification based on common voice. *Interspeech 2023*.

A Fine-Tuning Details

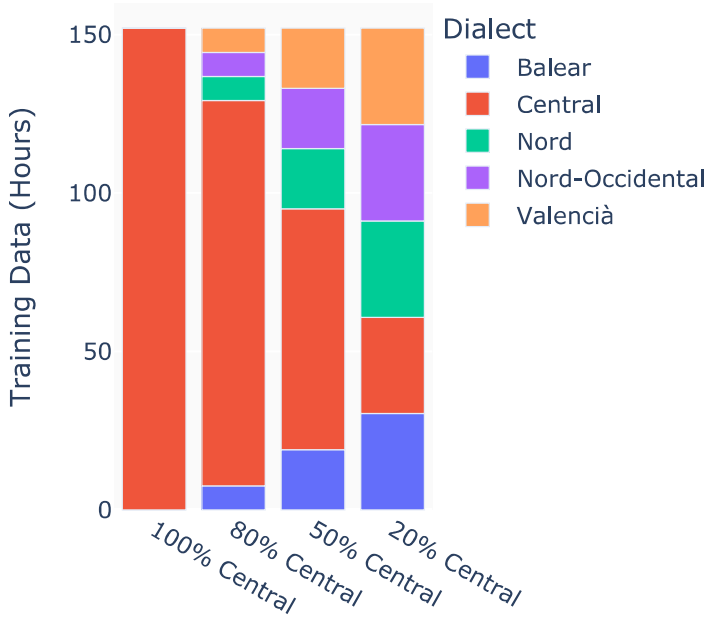

Figure 2: The proportion of five Catalan macro-dialects present in the training split of fine-tuning data for all four conditions.

To fine-tune XLSR-53 for Catalan ASR, we added a fully connected linear layer to the model that output logits over the vocabulary for each time step of the input audio. Similar to Conneau et al. (2021) and Babu et al. (2022), we used the connectionist temporal classification (CTC) loss function during training (Graves et al., 2006) and froze the weights in the model's CNN feature extractors. The audio files were resampled to the model sample rate and the text targets were tokenized into characters after decasing and removing punctuation. We retained any diacritics that are phonetically meaningful in Catalan and performed Unicode normalization so diacritics were represented consistently.

We trained with batch sizes of 16 but accumulated gradients for 16 batches before updating weights. Weight updates were made with the Adam optimizer (Kingma and Ba, 2015) and training continued until improvements to the development set loss were less than 0.05 for three epochs in a row. We found a stable learning rate of $9e^{-4}$ worked well across models. During inference, we used greedy decoding to transcribe input audio.

B CW Adversarial Attacks

B.1 Attack Training Algorithm

While training the adversarial noise, we set some upper limit ϵ to the value $|\delta_\infty|$. We then sought the perturbation δ for a given audio file x that minimized the CTC loss with respect to an adversarial target y while also keeping the noise's L_2 norm small. Thus, we aimed to minimize the following objective function:

$$\ell(x+\delta, y) + c|\delta|_2^2 \tag{2}$$

The constant c controls the relative importance of the regularizing term and therefore the noise's perceptibility. We iteratively calculated the objective in Equation 2 and updated the adversarial perturbation using the Adam optimizer. After every update to the noise, it was clamped such that $|\delta|_\infty \leq \epsilon$. Once a δ was found that successfully outputs the adversarial target y (as measured by a word error rate of 0 between y and the model output), ϵ was multiplied by a term α which is smaller than one to reduce the search radius and look for a quieter perturbation that successfully attacked the model. This continued until the search space was reduced k times or until a maximum number of updates n was carried out. See Table 2 for a summary of the hyperparameters we used while fitting attacks. We largely based our values off of those by Olivier and Raj (2023), though we found a higher learning rate worked better for our models.

Initial ϵ	0.10
c	0.25
α	0.70
learning rate	0.10
k	8
n	2000

Table 2: Hyperparameters for training CW attacks.

B.2 Detailed CW Attack Results

	Success Rate	Average SNR
100% Central	0.852	35.62
80% Central	0.820	35.44
50% Central	0.820	34.21
20% Central	0.864	35.80

Table 3: Attack success rate and average signal-to-noise ratio among successful CW attacks, broken down by the proportion of Central Catalan in each model's fine-tuning data. A higher average SNR of the attack indicates that among the successful attacks, quieter perturbations could be used to attain a successful attack.

	Success Rate	Average SNR
bal	0.830	36.73
cen	0.855	35.03
nor	0.855	34.27
no-oc	0.830	35.97
val	0.825	34.85

Table 4: Attack success rate and average signal-to-noise ratio among successful CW attacks, broken down by dialect of the attack's audio. A higher average SNR of the attack indicates that among the successful attacks, quieter perturbations could be used to attain a successful attack; bal: Balearic, cen: Central, nor: Nord, no-oc: Nord-Occidental, val: Valencià.

C Gamma Regression Output for WER

	Estimate	Standard Error	p-value
Intercept	0.176	8.84e-04	<2e-16***
20% Central	-0.019	0.001	<2e-16***
50% Central	-0.027	0.001	<2e-16***
80% Central	0.002	0.002	0.121
100% Central	0.044	0.002	<2e-16***
Balearic	0.042	0.002	<2e-16***
Central	-0.025	0.002	<2e-16***
Nord	-0.019	0.002	<2e-16***
Nord-Occidental	0.015	0.002	3.00e-16***
Valencià	-0.013	0.002	1.18e-15***
20% Central × Balearic	-0.010	0.003	0.002**
50% Central × Balearic	-0.010	0.003	0.001**
80% Central × Balearic	0.003	0.004	0.339
100% Central × Balearic	0.017	0.004	4.44e-05***
20% Central × Central	0.033	0.003	<2e-16***
50% Central × Central	0.017	0.003	1.70e-11***
80% Central × Central	-0.006	0.003	0.016*
100% Central × Central	-0.043	0.003	<2e-16***
20% Central × Nord	0.007	0.003	0.009**
50% Central × Nord	0.004	0.002	0.158
80% Central × Nord	1.66e-04	0.003	0.953
100% Central × Nord	-0.010	0.003	0.001**
20% Central × Nord-Occidental	-0.011	0.003	1.38e-04***
50% Central × Nord-Occidental	-2.82e-04	0.003	0.923
80% Central × Nord-Occidental	0.002	0.003	0.468
100% Central × Nord-Occidental	0.009	0.003	0.018*
20% Central × Valencià	-0.018	0.002	4.39e-13***
50% Central × Valencià	-0.010	0.002	1.20e-04***
80% Central × Valencià	4.23e-04	0.003	0.883
100% Central × Valencià	0.027	0.004	5.27e-14***

Table 5: β estimates, standard errors, and p-values of the gamma regression predicting WER. Factors are sum-coded. *: significant at threshold 0.05; **: significant at threshold 0.01; ***: significant at threshold 0.001.

D Gamma Regression Output for CER

	Estimate	Standard Error	p-value
Intercept	0.045	2.81e-04	<2e-16***
20% Central	-0.004	4.51e-04	<2e-16***
50% Central	-0.008	4.26e-04	<2e-16***
80% Central	-1.14e-05	4.81e-04	0.981
100% Central	0.012	5.74e-04	<2e-16***
Balearic	0.014	6.87e-04	<2e-16***
Central	-0.007	4.86e-04	<2e-16***
Nord	-0.005	5.04e-04	<2e-16***
Nord-Occidental	0.004	5.94e-04	3.18e-13***
Valencià	-0.006	5.10e-04	<2e-16***
20% Central × Balearic	-0.005	0.001	4.75e-07***
50% Central × Balearic	-0.004	0.001	1.17e-04***
80% Central × Balearic	0.001	0.001	0.215
100% Central × Balearic	0.008	0.001	5.40e-08***
20% Central × Central	0.009	8.69e-04	<2e-16***
50% Central × Central	0.005	7.79e-04	8.61e-12***
80% Central × Central	-0.001	8.24e-04	0.090
100% Central × Central	-0.013	8.90e-04	<2e-16***
20% Central × Nord	6.69e-04	8.16e-04	0.413
50% Central × Nord	0.002	7.73e-04	0.040*
80% Central × Nord	5.00e-04	8.71e-04	0.566
100% Central × Nord	-0.003	0.001	0.007**
20% Central × Nord-Occidental	-0.003	9.31e-04	9.82e-04***
50% Central × Nord-Occidental	-9.45e-05	9.08e-04	0.917
80% Central × Nord-Occidental	3.55e-04	0.001	0.728
100% Central × Nord-Occidental	0.003	0.001	0.022*
20% Central × Valencià	-9.33e-04	8.02e-04	0.244
50% Central × Valencià	-0.003	7.34e-04	9.39e-05***
80% Central × Valencià	-9.45e-04	8.57e-04	0.270
100% Central × Valencià	0.005	0.001	1.47e-05***

Table 6: β estimates, standard errors, and p-values of the gamma regression predicting CER. Factors are sum-coded. *: significant at threshold 0.05; **: significant at threshold 0.01; ***: significant at threshold 0.001.

E Logistic Regression Output for CW Attacks

	Estimate	Standard Error	p-value
Intercept	1.689	0.089	<2e-16***
20% Central	0.229	0.166	0.168
50% Central	-0.149	0.149	0.316
80% Central	-0.148	0.149	0.321
100% Central	0.067	0.155	0.663
Balearic	-0.09	0.172	0.603
Central	0.200	0.195	0.306
Nord	0.101	0.181	0.578
Nord-Occidental	-0.086	0.173	0.620
Valencià	-0.125	0.171	0.465
20% Central × Balearic	-0.013	0.315	0.966
50% Central × Balearic	0.208	0.298	0.484
80% Central × Balearic	-0.186	0.281	0.508
100% Central × Balearic	-0.008	0.301	0.977
20% Central × Central	0.633	0.405	0.118
50% Central × Central	0.253	0.330	0.444
80% Central × Central	-0.588	0.292	0.044*
100% Central × Central	-0.298	0.314	0.343
20% Central × Nord	-0.027	0.331	0.935
50% Central × Nord	-0.124	0.296	0.675
80% Central × Nord	0.016	0.303	0.958
100% Central × Nord	0.135	0.325	0.678
20% Central × Nord-Occidental	-0.316	0.300	0.292
50% Central × Nord-Occidental	-0.188	0.281	0.503
80% Central × Nord-Occidental	0.360	0.306	0.240
100% Central × Nord-Occidental	0.145	0.309	0.640
20% Central × Valencià	-0.277	0.299	0.355
50% Central × Valencià	-0.149	0.280	0.595
80% Central × Valencià	0.399	0.305	0.191
100% Central × Valencià	0.027	0.300	0.929

Table 7: β estimates, standard errors, p-values of a logistic regression predicting adversarial attack success. Factors are sum-coded. *: significant at threshold 0.05; ***: significant at threshold 0.001.