Phonologically-aware Automatic Speech Recognition Evaluation of Low-Resource Languages: The Case of Basque Dialects

Christoforos Souganidis, Asier Herranz, Ibon Saratxaga, Eva Navas, Inma Hernáez
HiTZ Basque Center for Language Technology - Aholab Signal Processing Laboratory,
University of the Basque Country UPV/EHU, Spain
{christoforos.souganidis, asier.herranz, ibon.saratxaga, eva.navas, inma.hernaez}@ehu.eus

Abstract

Automatic speech recognition models are typically trained with data of standard languages. However, their performance degrades when dealing with non-standard dialectal speech. In this paper, we present the first evaluation of an automatic speech recognition system for Basque, a low-resource language, based on spontaneous broadcast speech with high representation of dialectal speech. It relies on a 140-h manually annotated proprietary corpus of television programs in Basque, including dialect-level labels, as well as standardized and pseudo-phonetic transcriptions. We find that recognition performance significantly degrades for dialectal compared to standard speech, for all dialects present in our corpus. Subsequently, we provide a quantitative analysis of phonological phenomena based on single-word substitution errors, and identify 52 recurrent phenomena, grouped into sound deletions, epentheses, and substitutions. We further show a modest but statistically significant correlation between the number of phonological phenomena in an utterance and its recognition error rate. Our findings highlight the limitations of dialect-agnostic evaluation and motivate linguistically informed, dialect-aware strategies for automatic speech recognition in low-resource and typologically diverse languages.

Keywords: automatic speech recognition, dialectal speech evaluation, Basque dialects

1. Introduction

The vast majority of language technologies have been trained with standardized varieties of languages. However, dialectal variation has gained increasing attention in natural language processing (NLP), as existing systems frequently struggle to handle socially conditioned linguistic variation such as dialects, underscoring the need to systematically integrate social factors into NLP models and applications (Hovy and Yang, 2021; Zampieri et al., 2020). However, a recent survey on dialects in NLP (Joshi et al., 2025) emphasizes that despite the social and ethical implications of dialectal NLP, most of the applications focus on high-resource languages with many speakers, such as English or Spanish (Zampieri et al., 2024). A great recent advance towards the incorporation of language varieties in NLP was the creation of DIALECTBENCH, the first large-scale benchmark for dialectal NLP (Faisal et al., 2024), and the dialectal evaluation of machine translation benchmark CODET (Alam et al., 2024).

In addition to affecting NLP, social factors, such as race or dialectal variation, are found to affect the performance of automatic speech recognition (ASR) models. The impact of accent on ASR performance has troubled the field for over two decades (e.g. Huang et al., 2004), with phonology playing a crucial role (Yang et al., 2018). Various studies show that, up to date, ASR performance degrades for speech of non-white speakers or speakers of non-standard varieties (Koenecke et al., 2020; Tatman, 2017; Wassink et al., 2022; Chang et al., 2024; Vakirtzian et al., 2025; Martin and Tang, 2020). As in the case of NLP, ASR advances in dialectal speech focus mainly on English (Hinsvark et al., 2021), leaving other languages underexplored, although recent benchmarking initiatives, namely ML-SUPERB 2.0 (Chen et al., 2025; Shi et al., 2024), have started to integrate dialectal speech.

1.1. Basque dialectal landscape

Basque is a language isolate spoken in the Autonomous Community of the Basque Country and Navarre in Spain, and in the Northern Basque Country in France. Although it has approximately only 1 million speakers (Basque Government, 2016), there is a great deal of dialectal variability in terms of phonetics and phonology (Hualde, 1991; Hualde and Ortiz de Urbina, 2003). Various proposals have been made regarding the classification of Basque dialects, ranging from 3 to 8, many of which contain subdialectal divisions (Zuazo, 2008; Aurrekoetxea et al., 2017, 2019). In this paper, we adopt a dialectal classification based on the division of Basque-speaking regions into provinces: Biscayan (biz), Gipuzkoan (gip), Upper Navarrese (naf), Lapurdian (lap), Lower Navarrese (naf-low), and Souletin. The Royal Academy of the Basque language delevoped Standard Basque (eus), the standardized version of the Basque language, based on central dialects, mainly on Gipuzkoan, incorporating features of

Proceedings of Workshop on Dialects in NLP — A Resource Perspective (DialRes) , pages 48–57
16 May 2026.

other dialects such as Lapurdian and Upper Navarrese (Hualde and Zuazo, 2007). This is also reflected in recent NLP studies, showing lower Levenshtein distance between Standard Basque and dialects closer to Standard Basque (i.e., Gipuzkoan and Upper Navarrese), than between the standard and more peripheral dialects such as Biscayan (Bengoetxea et al., 2025).

Regarding phonetics and phonology in Basque dialects, we can classify the phenomena into three major categories: deletions, epentheses and substitutions. Deletions of sounds can occur in various word positions, such as between vowels (intervocalic deletions), for example, <egun> ('day') > <eun>, or in the final position of a syllable (coda deletions), as in the case of <usain> ('smell') > <usai> (Hualde et al., 2010). Consonant (C) epentheses are typically attested in an intervocalic position, for example, <mendia> ('mountain') > <mendixa> (Hualde et al., 2010). Substitutions can vary in form, such as substitution of a vowel (V) with another vowel or vowels, substitution of consonants including palatalization, and voicing assimilation. Basque dialects show a wide variety of vocalic interactions, for example <neska> ('girl') that can be pronounced either as <neske>, <neskia> or <neskie> (Hualde and Gaminde, 1998; deCastro Arrazola et al., 2015). Some Basque dialects maintain the voiceless sibilant distinction between apico-alveolar <s> and laminal <z>, while others merge the two sounds (Hualde and Gaminde, 1998; Larraza et al., 2017). Palatalization in Basque orthography is denoted by the use of a double consonant, typically <tt> or <dd> (palatalized versions of <t> and <d>, respectively), as in the case of <polita> ('pretty') > <politta> (Hualde et al., 2010). Voicing assimilation of voiceless consonants is common when the preceding sound is voiced, for example, in the expression <lan eta lan> ('working non-stop') which is pronounced <lan da lan>, after the common deletion of <e> in <eta> ('and') in spoken language.

As illustrated by the phonological phenomena described above, the linguistic landscape of Basque dialects is highly complex. This is particularly relevant for computational resources of dialects of an already low-resource language, as the lack of specific subdialectal orthography makes it difficult for the representation of those dialects in NLP and speech technologies (e.g. Vakirtzian et al., 2025).

1.2. Basque dialectal resources

Despite Basque being considered a low-resource language, recent efforts are starting to include dialectal variation, reflecting the growing interest in Basque NLP and speech technologies.

Regarding NLP resources of Basque dialects, there have been a handful of contributions: a historical corpus of Basque dialects (Estarrona et al., 2020), a corpus of syntactic variation (Uria and Etxepare, 2012), the XNLI-var dataset, including Biscayan, Gipuzkoan, and Upper Navarrese dialects (Bengoetxea et al., 2025), and the recent BasPhyCo$_{west}$ for commonsense reasoning in Biscayan (Bengoetxea et al., 2026). Although not specifically targeted as a resource for dialectal NLP, the corpus of tweets by Fernandez de Landa (2019) includes dialectal material from Biscayan and Gipuzkoan varieties, which was later integrated as part of the Diverse Corpora corpus as an NLP resource aimed at capturing language variation (Azurmendi et al., 2025).

Compared to the increasing availability of dialectal resources for NLP for Basque, dialectal speech processing resources remain relatively underexplored, with a couple of notable exceptions. The Speech_Dat (II) telephone speech database (Hernaez et al., 2003) contains untranscribed recordings of Standard and Biscayan Basque speakers from 23 regions. The Basque Speecon-like and Basque SpeechDat MDB-600 databases (Odriozola et al., 2014) contain manually transcribed spontaneous and read speech by 832 speakers from the Autonomous Community of the Basque Country and Navarre in Spain. Nevertheless, in these datasets there is an under-representation of spontaneous speech, and, as in the case of Speech_Dat (II), include only Basque speakers from Spain. In a study comparing performance of ASR in Basque and Spanish, it is found that performance for Basque is noticeably worse than for Spanish, and although it is not systematically investigated, dialectal variability is likely a contributing factor for this discrepancy (Peñagarikano et al., 2023).

1.3. Aims of this work

Despite recent advances in Basque dialect resources for NLP, there is still a lack of studies looking at the performance of ASR systems for Basque dialects. In particular, we hypothesize that the variable phonological phenomena attested in Basque dialects could be responsible for the behavior of ASR systems.

In this study, our contributions are twofold:

- a dialect-aware evaluation of Basque ASR of spontaneous speech on real broadcast data, and
- a systematic documentation of phonological phenomena in various Basque dialects.

This work provides empirical evidence on ASR performance across Basque dialects, and documents relevant phonological phenomena in spoken Basque, contributing to both dialect-aware ASR research and linguistic studies of dialectal variation.

Dialect	$N_{speakers}$	$N_{utterances}$	N_{words}	H
eus	924	63565	489620	59
lap	27	812	7905	1
gip	547	49103	296365	33
biz	335	29791	170043	19
naf	42	2073	15290	2
naf-low	1	8	28	<1

Table 1: Number of speakers, number of utterances, number of words, and number of hours, for each dialect.

2. Data

In all experiments carried out in this paper, we used a proprietary dataset of TV shows in Basque broadcast by Basque Radio Television (Euskal Irrati Telebista, EITB), the Autonomous Community of the Basque Country's public broadcast service. The dataset consists of 140 hours of manually annotated and transcribed audio from speakers of various Basque dialects.

It contains two types of transcription: a standardized transcription and a pseudo-phonetic transcription. The standardized transcription corresponds to the written form of each utterance, either in Standard or Dialect-Specific Basque. However, due to the aforementioned dialectal diversity even within each dialect, the limits of what constitutes a standardized transcription are sometimes difficult to define. The pseudo-phonetic transcription corresponds to the orthographic representation of what was pronounced by each speaker. This type of transcription includes phonological phenomena such as across-word or within-word co-articulation.

The audio data was annotated for speaker turns, speaker identity (each speaker was assigned a unique label), and speaker dialect (one dialect label per speaker). Table 1 shows the number of speakers per dialect. It should be noted that the dialect annotation denoted the dialect identified for each speaker, regardless of the content of each utterance. This means that even if an utterance is pronounced by a speaker of a dialect and a speaker of Standard Basque in exactly the same way, the former is annotated as dialectal, while the latter is annotated as standard. The speakers whose dialect could not be identified were annotated as speakers of Standard Basque. This corpus contains the following dialects: Biscayan, Gipuzkoan, Upper Navarrese, Lapurdian and Lower Navarrese.

3. Analyses and results

The statistical analyses described in this section were carried out in JASP v0.19.1.0 (JASP Team, 2026). A non-parametric Kruskal–Wallis test was conducted, with the evaluation metric as the dependent variable and *Dialect* as the independent variable. Since the independent variable had more than two levels, we conducted Dunn's post-hoc pairwise comparisons using Holm correction of *p*-values.

3.1. Evaluation of ASR

For the experiments conducted in this study, we used a Conformer-Transducer ASR model developed within the NVIDIA NeMo[1] framework. The specific model used was `stt_eu_conformer_transducer_large_v2`, which is publicly available on the Hugging Face Hub[2]. The standardized transcriptions were used as reference for the ASR evaluation.

The model was trained on a total of 772 hours of Basque speech. The training corpus includes a diverse collection of speech data, with 95.75 hours of spontaneous speech sourced from broadcast radio programs, distinct from the programs in our proprietary dataset. Notably, this training set incorporates specific dialectal representation, including 19 hours of Biscayan, 11.7 hours of Gipuzkoan, 3.45 hours of Lapurdian and 1 hour of Upper Navarrese dialects. This inclusion of dialect-rich broadcast data allows the model to better generalize across the phonological variations characteristic of the Basque linguistic landscape described in Section 1.1.

The descriptive statistics (Table 2) show that the ASR system had better performance in terms of word error rate (WER) for standard compared to dialectal speech. The non-parametric Kruskal-Wallis test with *WER* as the dependent variable and *Dialect* as the independent variable showed a significant effect of *Dialect* on *WER* (*H(5)* = 6630.784, $p < .001$). Pairwise comparisons (Table 3) revealed that performance for Standard Basque was significantly better than for all other dialects ($p = .005$ for Lower Navarrese, $p < .001$ for the rest of the dialects). The only dialect that was significantly more challenging for ASR compared to all other dialects except Lower Navarrese was Gipuzkoan. It should be noted that although Lower Navarrese had a numerically higher WER, the limited amount of observations as shown in Table 2 was not enough for a statistically reliable difference.

3.1.1. Substitution patterns

In this section, we focus on the single-word substitutions produced by the ASR model described above, as our goal is to explore the effect of dialectal speech on ASR performance. Since we are aiming to investigate the phonological phenomena that

[1]NeMo: https://github.com/NVIDIA/NeMo

[2]Model: https://huggingface.co/HiTZ/stt_eu_conformer_transducer_large_v2

Dialect	$\text{WER}_{\text{utt}} \pm \text{SD}$	$\text{WER}_{\text{global}}$
eus	0.192 ± 0.324	0.126
lap	0.245 ± 0.283	0.213
gip	0.350 ± 0.407	0.270
biz	0.332 ± 0.406	0.256
naf	0.318 ± 0.413	0.227
naf-low	0.619 ± 0.301	0.643

Table 2: Mean WER and standard deviation per utterance, and global WER, per dialect.

Comparison	z	p_{holm}
eus − lap	−9.406	< 0.001
eus − gip	−76.136	< 0.001
eus − biz	−54.375	< 0.001
eus − naf	−16.544	< 0.001
eus − naf-low	−3.422	0.005
lap − gip	−3.525	0.004
lap − biz	−1.382	0.501
lap − naf	−0.887	0.750
lap − naf-low	−2.496	0.095
gip − biz	10.302	< 0.001
gip − naf	3.930	< 0.001
gip − naf-low	−2.128	0.133
biz − naf	0.549	0.750
biz − naf-low	−2.342	0.106
naf − naf-low	−2.373	0.106

Table 3: Dunn's post-hoc pairwise comparisons for *Dialect*, reporting *z* statistics and Holm-corrected p-values.

might affect ASR, the focus of the following evaluation shifts to the character error rate metric (CER). We believe that CER reflects more accurately than WER the direct effect of dialectal variation on ASR performance, as WER could be influenced by other factors such as the number of dialect-specific words per utterance.

To specifically analyze substitution patterns across dialects, we focused the analysis on single-word substitution errors. First, we filtered out all the correctly recognized utterances by our ASR model, as well as the word deletions and word insertions. Then, for the remaining data, we kept only single-word substitutions that differed between the manual transcriptions and the automatic transcriptions and calculated the CER, considering the manual transcriptions as the ground truth.

Table 4 displays the number of single-word substitutions, as well as the mean CER and standard deviation per dialect of these substitutions. In this table, we can observe that the CER in Standard Basque is the lowest compared to dialectal speech. The non-parametric Kruskal-Wallis test with *CER* as the dependent variable and *Dialect* as the independent variable showed a significant effect of *Dialect* on *CER* (*H(5)* = 652.503, $p < .001$). As

Dialect	N	$\text{M} \pm \text{SD}$
eus	22993	0.344 ± 0.300
lap	498	0.361 ± 0.281
gip	26293	0.385 ± 0.330
biz	14020	0.394 ± 0.321
naf	1033	0.378 ± 0.341
naf-low	6	0.372 ± 0.338

Table 4: Number of utterances, mean CER and standard deviation of single-word substitutions per dialect.

Comparison	z	p_{holm}
eus − lap	−2.799	0.056
eus − gip	−21.293	< 0.001
eus − biz	−22.158	< 0.001
eus − naf	−4.862	< 0.001
eus − naf-low	−0.084	1.000
lap − gip	−1.448	1.000
lap − biz	−2.427	0.137
lap − naf	−0.511	1.000
lap − naf-low	0.225	1.000
gip − biz	−4.319	< 0.001
gip − naf	1.186	1.000
gip − naf-low	0.387	1.000
biz − naf	2.568	0.102
biz − naf-low	0.497	1.000
naf − naf-low	0.294	1.000

Table 5: Dunn's post-hoc pairwise comparisons for *Dialect*, reporting *z* statistics and Holm-corrected *p*-values.

shown in Table 5, CER was significantly lower in Standard Basque compared to most dialects ($p < .001$), namely Gipuzkoan, Biscayan and Upper Navarrese, which were the dialects with the highest amount of data in our corpus. Between dialects, the only significant difference found was between gip and biz, with the latter showing a significantly higher CER ($p < .001$). Our results are aligned with the distance of each dialect from the standard, with Upper Navarrese being closer, followed by Gipuzkoan, and Biscayan being the most different (Bengoetxea et al., 2025).

3.2. Phonological analysis

In order to investigate the phonological phenomena that could affect ASR performance, we compared the standardized and the pseudo-phonetic transcriptions available in this corpus. First, we identified the single words in the corpus that had a different standardized transcription compared to the pseudo-phonetic one. Then, we kept only unique transcription pairs and calculated the frequency of occurrence of each pair. We identified 53541 pairs, 13412 of which were unique. We manually anno-

tated the phonological phenomena of 367 unique pairs that occurred more than 15 times. These manual annotations correspond to a total of 31343 occurrences, approximately 58,54% of all pair instances.

We identified 52 phenomena (see Table 7), which can be grouped into three: deletions ($n_{deletions} = 6$), epentheses ($n_{epentheses} = 6$), and substitutions ($n_{substitutions} = 40$).

Deletions. We identified two types of deletion, namely consonant (C) deletions and vowel (V) deletions. We further classified C-deletions depending on the position of the deleted sound, either in syllabic coda (without sound specification) or in intervocalic position (including sound specification). For example, if the word ‹ezagutu› ("to get to know"), was pronounced ‹ezautu›, the pair was annotated with the phenomenon *intervocalic g deletion*.

Epentheses. We identified two types of epentheses, i.e., single C and syllabic. Single C epentheses included the specification of the sound, while syllabic ones did not. For example, if the word ‹seiehun› ("six hundred"), was pronounced ‹seirehun›, the pair was annotated with the phenomenon *intervocalic r epenthesis*.

Substitutions. Three types of substitutions were identified: CV-to-CV, C-to-C, V-to-V. For C-to-C and V-to-V substitutions, the C/V could be either a single C/V or two C/V. In cases where only one C/V was involved in the substitution but was adjacent to another C/V which did not change, both C or V letters were annotated as part of the phenomenon. As an example, the word ‹orduan› ("then") pronounced as ‹orduen›, the pair was annotated as *ua-to-ue substitution*. It should be noted that, despite the fact that we annotated the phenomena involving two C/V as substitutions, these could also be viewed as cases of deletions or epentheses. Regarding C-to-C substitutions, two of the phenomena identified were palatalization and voicing of a voiceless stop sound. Although some C-to-C substitutions could also be viewed as palatalizations (e.g., *s-to-x substitution*), we annotated as *palatalization* only cases where a C is written as a double C, which in Basque are palatalized.

In some pairs, more than one phenomenon was identified, which were either considered independent or dependent. In the latter, we considered that some phonological phenomena were applied recursively, particularly where a deletion of a sound could have triggered a subsequent substitution. For instance, if the word ‹lehenago› ("earlier") was pronounced ‹lehenau›, the pair was annotated with the phenomenon *intervocalic g deletion*, which triggered the phenomenon *ao-to-au substitution*.

We identified 51 phenomena in Gipuzkoan, 50 in Biscayan, 31 in Upper Navarrese, 15 in Lapurdian, while Lower Navarrese was not present in our annotation due to insufficient data. Of all the identified phenomena, only one was unique to a single dialect: *oa-to-ue substitution* was only present in Biscayan. Finally, we explored the co-occurrence of phenomena that were unique to a specific dialect: The co-occurrence of *e-to-i substitution* and *intervocalic t deletion* was unique to Biscayan, while *intervocalic g deletion* and *intervocalic r deletion* co-occurrence and the co-occurrence of *intervocalic g deletion*, *ea-to-ia substitution*, and *ei-to-i substitution* were unique to Gipuzkoan.

Dialect	M ± SD
lap	1.000 ± 0.000
gip	1.161 ± 0.395
biz	1.362 ± 0.498
naf	1.200 ± 0.404

Table 6: Mean number of phenomena per word (M) and standard deviation (SD), per dialect.

Last, we explored whether the number of attested phenomena differs between dialects. Table 6 displays the descriptive statistics. A non-parametric Kruskal-Wallis test with the Number of phenomena as the dependent variable and Dialect as the independent variable showed a significant effect of Dialect (H(3)= 101.029, p < .001). As Dialect had four levels, we conducted Dunn's post-hoc pairwise comparisons using Holm correction of p-values. Biscayan showed significantly more phenomena compared to all other dialects, i.e., Gipuzkoan (z = 9.829 , p < .001), Upper Navarrese (z = 2.603, p = .037), and Lapurdian (z = 3.685, p = .001).

3.3. Linking phonological variations and ASR errors

In this section, we bring together the results from the ASR evaluation regarding Basque dialects and our phonological analyses. We investigated if ASR performance in terms of CER is mediated by the number of identified phenomena in utterances with single word substitutions. For this reason, we conducted a Spearman's correlation analysis between the total CER and the number of phenomena in the utterance. Our results suggest a modest but significant correlation (r = .216, p < .001).

4. Discussion

This study provides, to our knowledge, the first dialect-aware analysis of Basque ASR performance on spontaneous broadcast speech. The results consistently show that recognition errors are not uniformly distributed across varieties, as Standard Basque shows significantly better performance in terms of WER and CER compared to

Group	Phenomenon	Dialects attested	Example		
			Standardized transcription	Pseudophonetic transcription	Dialect
deletion	coda	biz, gip, naf	bat	ba	biz
deletion	e	biz, gip, lap, naf	honetan	hontan	gip
deletion	intervocalic g	biz, gip, lap, naf	egin	ein	biz
deletion	intervocalic r	biz, gip, lap, naf	gero	geo	gip
deletion	intervocalic t	biz, gip, naf	dute	due	gip
deletion	intervocalic d	biz, gip, lap, naf	didazu	diazu	gip
epenthesis	d	biz, gip	genuen	gendun	gip
epenthesis	n	biz, gip, naf	hola	holan	biz
epenthesis	r	biz, gip, naf	seiehun	seirehun	biz
epenthesis	syllable	gip, naf	izan	izandu	naf
epenthesis	x	biz, gip	egia	egixa	biz
epenthesis	z	biz, gip	dira	diez	biz
substitution	a-to-ai	biz, gip, lap	zergatik	zergaitik	biz
substitution	a-to-au	biz, gip, lap, naf	handia	haundia	gip
substitution	a-to-e	biz, gip, lap, naf	behala	bezela	gip
substitution	a-to-o	biz, gip	daukat	dakot	biz
substitution	ao-to-au	biz, gip	dago	dau	gip
substitution	ao-to-o	biz, gip, lap, naf	lasaiago	lasaio	gip
substitution	au-to-a	biz, gip, lap, naf	daukazue	dakazue	gip
substitution	da-to-te	biz, gip	dodala	dotela	biz
substitution	e-to-a	biz, gip, lap, naf	hemezortzi	hamazortzi	biz
substitution	e-to-i	biz, gip, naf	behar	bihar	gip
substitution	ea-to-e	biz, gip	behar	ber	gip
substitution	ea-to-i	biz, gip	azkenean	azkenin	gip
substitution	ea-to-ia	biz, gip, lap, naf	batean	batian	gip
substitution	ea-to-ie	biz, gip, naf	azkenean	azkenien	biz
substitution	ei-to-e	biz, gip, naf	hogeita	hogeta	gip
substitution	ei-to-i	biz, gip, naf	begira	bira	gip
substitution	eo-to-o	biz, gip	edo	o	gip
substitution	ia-to-i	biz, gip	guztian	guztin	gip
substitution	ia-to-ie	biz, gip	egia	egie	gip
substitution	i-to-e	biz, gip, lap, naf	nirekin	nerekin	gip
substitution	ie-to-i	gip, naf	horiek	hoik	gip
substitution	kl-to-k	biz, gip	klaro	karo	gip
substitution	n-to-l	biz, gip	lehenengo	lelengo	gip
substitution	o-to-u	biz, gip, naf	non	nun	gip
substitution	oa-to-o	biz, gip	orain	oin	biz
substitution	oa-to-u	biz, gip, naf	joatea	jutea	gip
substitution	oa-to-ua	biz, gip, naf	goazen	guazen	gip
substitution	oa-to-ue	biz	gaurkoan	gaurkuen	biz
substitution	oe-to-o	biz, gip, lap, naf	dagoen	dagon	biz
substitution	ou-to-u	biz, gip	moduz	muz	gip
substitution	palatalization	biz, gip, naf	ditu	dittu	gip
substitution	s-to-x	biz, gip	samarra	xamarra	biz
substitution	t voicing	biz, gip, naf	eta	da	gip
substitution	u-to-eu	biz, gip	nuke	neuke	gip
substitution	u-to-o	biz, gip, naf	justu	justo	biz
substitution	ua-to-u	biz, gip, lap, naf	orduan	ordun	gip
substitution	ua-to-ue	biz, gip, naf	orduan	orduen	biz
substitution	ue-to-u	biz, gip, lap, naf	genituen	genitun	gip
substitution	z-to-s	biz, gip	ezta	esta	biz
substitution	z-to-x	biz, gip, naf	dezente	dexente	gip

Table 7: Examples of dialectal phonological phenomena with the dialect of the example and the dialects where each phenomenon is attested.

most dialectal varieties. This finding aligns with earlier observations in other languages, where ASR systems trained predominantly on standardized varieties struggle to generalize to socially and regionally marked speech (Koenecke et al., 2020; Tatman, 2017; Wassink et al., 2022; Chang et al., 2024; Martin and Tang, 2020; Vakirtzian et al., 2025).

A particularly relevant outcome is the systematic disadvantage observed for Biscayan and Gipuzkoan, which are both well represented in our dataset, being the two dialects with the highest number of utterances. The fact that Biscayan exhibits significantly higher CER than Gipuzkoan, despite high data availability, suggests that the observed differences cannot be explained by data volume alone. Instead, our phonological analysis supports the interpretation that the linguistic distance to the standard variety plays a central role. This interpretation is further reinforced by the finding that Biscayan utterances display a significantly higher number of phonological phenomena per word than the other dialects considered, consistent with previous research (Bengoetxea et al., 2025).

The significant correlation between the number of phonological phenomena in an utterance and its CER provides evidence that phonological variability contributes to recognition errors. Although the correlation is not strong, this is expected given that ASR errors are influenced by multiple interacting factors, including background noise and speaker characteristics. Nevertheless, our results indicate that the accumulation of phonological processes is a relevant predictor of ASR performance in dialectal Basque.

Importantly, our qualitative analysis reveals that most of the attested phenomena are not dialect-exclusive. Only one phenomenon (*oa-to-ue substitution*) was unique to a single dialect (i.e., Biscayan), and even co-occurrence patterns were limited to a small number of cases, one pattern for Biscayan (i.e., *e-to-i substitution* and *intervocalic t deletion*) and two for Gipuzkoan (*intervocalic g deletion* and *intervocalic r deletion*, and *intervocalic g deletion* and *ea-to-ia substitution*, and *ei-to-i substitution*). From the perspective of ASR, this implies that dialect-aware modeling strategies based solely on dialect labels are unlikely to fully capture the relevant variability. Instead, approaches that explicitly model systematic phonological alternations such as deletion, epenthesis, and substitution patterns may offer a more robust way of improving recognition for non-standard speech. Alternatively, since most ASR systems are trained with data of standard forms of a language, universal speech models such as OpenAI Whisper (Radford et al., 2023) might be more suitable for dialectal speech when dialectal transcriptions are needed, as the transcription of dialectal speech is closer to dialectal transcriptions than standard transcriptions.

The analysis of substitutions further highlights an important methodological issue. Many of the errors contributing to CER arise from vowel and vowel-sequence substitutions (e.g., *ea-to-ia, ua-to-u, oa-to-o*), which are known to be frequent and productive in dialectal Basque (deCastro Arrazola et al., 2015). These alternations often preserve lexical identity for human listeners but create orthographic mismatches for ASR systems evaluated against standardized references. This observation resonates with recent findings in dialectal ASR research showing that the lack of standardized or pronunciation-oriented representations for dialectal speech is a major source of evaluation and modeling difficulties (Vakirtzian et al., 2025). In this respect, our use of pseudo-phonetic transcriptions proves particularly valuable, as it allows us to disentangle errors due to pronunciation variation from those related to lexical errors.

Another relevant aspect of our results concerns the imbalance between dialects. Although Lapurdian, Lower Navarrese and Upper Navarrese are included in the corpus, their limited representation prevents strong statistical conclusions about these varieties, particularly for Lower Navarrese which did not have enough data to be included in our phonological analyses or Souletin that did not have any representation in our dataset. This reflects a broader structural limitation in Basque speech technology resources, where northern dialects remain underrepresented. Consequently, future efforts should prioritize targeted data collection for northern and less represented dialects, as well as for informal and highly spontaneous speech, in order to obtain a more balanced view of dialectal ASR performance.

However, our study has certain limitations that should be taken into account in future research. First, although the standardized transcriptions in our dataset were designed to reflect the orthography of the major dialects included, the considerable degree of subdialectal variation within each Basque dialect, together with the absence of standardized orthography for many varieties, poses challenges for both database annotation and ASR evaluation. Second, the grouping of dialects adopted in this work constitutes another limitation, as there is no broad consensus regarding the number and boundaries of Basque dialects and subdialects (Zuazo, 2008; Aurrekoetxea et al., 2017, 2019). Finally, speech not explicitly identified as dialectal was labeled as Standard Basque, and dialect annotation was performed at the speaker level, even though speakers did not necessarily use dialectal forms consistently. Therefore, we should be parsimonious in the interpretation of our results comparing standard to dialectal Basque.

5. Conclusions

Overall, this study demonstrates that phonological variability in Basque dialects has measurable and systematic effects on ASR performance. The results of our fine-grained phonological analyses provide empirical evidence that supports the need for linguistically informed, dialect-aware modeling strategies in low-resource and typologically distinct languages such as Basque.

6. Acknowledgments

This work is funded by the Ministerio para la Transformación Digital y de la Función Pública and Plan de Recuperación, Transformación y Resiliencia - Funded by EU – NextGenerationEU within the framework of the project ILENIA with reference 2022/TL22/00215335 and the project ALIA (Plan Nacional de Tecnologías de Lenguaje-ENIA 2024, SEDIA 19.08.2024), and by a grant from the Department of Culture and Language Policy of the Basque Government (IKER-GAITU project).

7. Bibliographical References

Md Mahfuz Ibn Alam, Sina Ahmadi, and Antonios Anastasopoulos. 2024. CODET: A Benchmark for Contrastive Dialectal Evaluation of Machine Translation. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1790–1859, St. Julian's, Malta. Association for Computational Linguistics.

Gotzon Aurrekoetxea, Ane Arandia, Malalen Camino, and Xantiana Etchebest. 2017. Dialectología perceptual del vasco: aplicación del Little Arrow Method. In Florentino Paredes García, Ana M. Cestero Mancera, and Isabel Molina Martos, editors, *Investigaciones actuales en Lingüística. Vol. V: Sobre variación geolectal y sociolingüística*, pages 51–66. Universidad de Alcalá, Servicio de Publicaciones, Alcalá de Henares, Spain.

Gotzon Aurrekoetxea, Iñaki Gaminde, José Luis Ormaetxea, and Xarles Videgain. 2019. *Euskalkien Sailkapen Berria [New Classification of Basque Dialects]*. UPV/EHU, Bilbao, Spain.

Ekhi Azurmendi, Joseba Fernandez de Landa, Jaione Bengoetxea, Maite Heredia, Julen Etxaniz, Mikel Zubillaga, and Ander Soraluze. 2025. BERnaT: Basque Encoders for Representing Natural Textual Diversity. *arXiv preprint*.

Basque Government. 2016. VI. Inkesta Soziolinguistikoa Euskal Autonomia Erkidegoa [6. The Sociolinguistic Survey of the Basque Autonomous Community]. Technical report, Basque Government, Vitoria-Gasteiz, Spain. Accessed: 2026-02-24.

Jaione Bengoetxea, Itziar Gonzalez-Dios, and Rodrigo Agerri. 2025. Lost in variation? evaluating NLI performance in Basque and Spanish geographical variants. In *Proceedings of the 29th Conference on Computational Natural Language Learning*, pages 452–468, Vienna, Austria. Association for Computational Linguistics.

Jaione Bengoetxea, Itziar Gonzalez-Dios, and Rodrigo Agerri. 2026. Physical Commonsense Reasoning for Lower-Resourced Languages and Dialects: a Study on Basque. *arXiv preprint*. ArXiv:2602.14812 [cs.CL].

Kalvin Chang, Yi-Hui Chou, Jiatong Shi, Hsuan-Ming Chen, Nicole Holliday, Odette Scharenborg, and David R. Mortensen. 2024. Self-supervised Speech Representations Still Struggle with African American Vernacular English. In *Interspeech 2024*, pages 4643–4647.

William Chen, Chutong Meng, Jiatong Shi, Martijn Bartelds, Shih-Heng Wang, Hsiu-Hsuan Wang, Rafael Mosquera, Sara Hincapie, Dan Jurafsky, Antonis Anastasopoulos, Hung-yi Lee, Karen Livescu, and Shinji Watanabe. 2025. The ML-SUPERB 2.0 Challenge: Towards Inclusive ASR Benchmarking for All Language Varieties. In *Proceedings of Interspeech 2025*, pages 2093–2097.

Varun deCastro Arrazola, Edoardo Cavirani, Kathrin Linke, and Francesc Torres-Tamarit. 2015. A typological study of vowel interactions in Basque. *Isogloss*, 2015(1/2):147–177.

Ainara Estarrona, Izaskun Etxeberria, Ricardo Etxepare, Manuel Padilla-Moyano, and Ander Soraluze. 2020. Dealing with dialectal variation in the construction of the Basque historical corpus. In *Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 79–89, Barcelona, Spain. International Committee on Computational Linguistics (ICCL).

Fahim Faisal, Orevaoghene Ahia, Aarohi Srivastava, Kabir Ahuja, David Chiang, Yulia Tsvetkov, and Antonios Anastasopoulos. 2024. DIALECT-BENCH: A NLP Benchmark for Dialects, Varieties, and Closely-Related Languages. *arXiv preprint*.

Joseba Fernandez de Landa. 2019. Large Scale Linguistic Processing of Tweets to Understand Social Interactions among Speakers of Less Resourced Languages: The Basque Case. *Information*, 10(6).

Inma Hernaez, Iker Luengo, Eva Navas, Maren Zubizarreta, Iñaki Gaminde, and Jon Sanchez. 2003. The basque speech_dat (II) database: a description and first test recognition results. In *8th European Conference on Speech Communication and Technology (Eurospeech 2003)*, pages 1549–1552.

Arthur Hinsvark, Natalie Delworth, Miguel Del Rio, Quinten McNamara, Joshua Dong, Ryan Westerman, Ilya Pirkin, and Nishchal Bhandari. 2021. Accented Speech Recognition: A Survey. *arXiv preprint*.

Dirk Hovy and Diyi Yang. 2021. The Importance of Modeling Social Factors of Language: Theory and Practice. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 588–602, Online. Association for Computational Linguistics.

José Ignacio Hualde. 1991. *Basque Phonology*. Routledge.

José Ignacio Hualde and Iñaki Gaminde. 1998. On the phonology and morphology of Basque dialects. *Anuario del Seminario de Filología Vasca "Julio de Urquijo"*, 32:71–108.

José Ignacio Hualde, Oihana Lujanbio, and Juan Joxe Zubiri. 2010. Goizueta Basque. *Journal of the International Phonetic Association*, 40(1):113–127.

José Ignacio Hualde and Jon Ortiz de Urbina. 2003. *A Grammar of Basque*. Mouton de Gruyter.

José Ignacio Hualde and Koldo Zuazo. 2007. The standardization of the Basque language. *Language Problems and Language Planning*, 31(2):143–168.

Xuedong Huang, Federico Alleva, Hsiao-Wuen Hon, Wei Hwang, and Edward Reichenbach. 2004. Spoken language processing: A guide to theory, algorithm, and system development. *International Journal of Speech Technology*, 7(2):111–126.

JASP Team. 2026. JASP (Version 0.19.1.0) [Computer software]. Accessed: 2026-02-24.

Aditya Joshi, Raj Dabre, Diptesh Kanojia, Zhuang Li, Haolan Zhan, Gholamreza Haffari, and Doris Dippold. 2025. Natural Language Processing for Dialects of a Language: A survey. *ACM Computing Surveys*, 57(6):149:1–149:37.

Allison Koenecke, Andrew Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zion Mengesha, Connor Toups, John R. Rickford, Dan Jurafsky, and Sharad Goel. 2020. Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences of the United States of America*, 117(14):7684–7689.

Saioa Larraza, Arthur G. Samuel, and Miren Lourdes Oñederra. 2017. Where do dialectal effects on speech processing come from? Evidence from a cross-dialect investigation. *Quarterly Journal of Experimental Psychology*, 70(1):92–108.

Joshua L. Martin and Kevin Tang. 2020. Understanding Racial Disparities in Automatic Speech Recognition: The Case of Habitual "be". In *Interspeech 2020*, pages 626–630.

Igor Odriozola, Inma Hernaez, María Inés Torres, Luis Javier Rodriguez-Fuentes, Mikel Peñagarikano, and Eva Navas. 2014. Basque speecon-like and Basque SpeechDat MDB-600: speech databases for the development of ASR technology for Basque. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 2658–2665, Reykjavik, Iceland. European Language Resources Association (ELRA).

Mikel Peñagarikano, Amparo Varona, Germán Bordél, and Luis Javier Rodriguez-Fuentes. 2023. Semisupervised Speech Data Extraction from Basque Parliament Sessions and Validation on Fully Bilingual Basque–Spanish ASR. *Applied Sciences*, 13(14):848492.

Alec Radford, Jong Wook Wu, Dario Amodei, et al. 2023. Whisper: Robust Speech Recognition via Large-Scale Weak Supervision. arXiv preprint. ArXiv:2212.04356 [cs.CL].

Jiatong Shi, Shih-Heng Wang, William Chen, Martijn Bartelds, Vanya Bannihatti Kumar, Jinchuan Tian, Xuankai Chang, Dan Jurafsky, Karen Livescu, Hung yi Lee, and Shinji Watanabe. 2024. ML-SUPERB 2.0: Benchmarking Multilingual Speech Models Across Modeling Constraints, Languages, and Datasets. In *Interspeech 2024*, pages 1230–1234.

Rachael Tatman. 2017. "oh, I've Heard That Before": Modelling Own-Dialect Bias After Perceptual Learning by Weighting Training Data. In *Proceedings of the 7th Workshop on Cognitive Modeling and Computational Linguistics (CMCL 2017)*, pages 29–34, Valencia, Spain. Association for Computational Linguistics.

Larraitz Uria and Ricardo Etxepare. 2012. Hizkeren arteko aldakortasun sintaktikoa aztertzeko metodologiaren nondik norakoak: BASYQUE aplikazioa. *Lapurdum. Euskal ikerketen aldizkaria | Revue d'études basques | Revista de estudios vascos | Basque studies review*, (16):117–135.

Socrates Vakirtzian, Vivian Stamou, Yannis Kazos, and Stella Markantonatou. 2025. Dialectal treebanks and their relation with the standard variety: The case of East Cretan and Standard Modern Greek. In *Proceedings of the Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies (NoDaLiDa/Baltic-HLT 2025)*, pages 776–784, Tallinn, Estonia. University of Tartu Library.

Alicia Beckford Wassink, Cady Gansen, and Isabel Bartholomew. 2022. Uneven success: automatic speech recognition and ethnicity-related dialects. *Speech Communication*, 140:50–70.

Xuesong Yang, Kartik Audhkhasi, Andrew Rosenberg, Samuel Thomas, Bhuvana Ramabhadran, and Mark Hasegawa-Johnson. 2018. Joint Modeling of Accents and Acoustics for Multi-Accent Speech Recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, page 1–5. IEEE Press.

Marcos Zampieri, Preslav Nakov, and Yves Scherrer. 2020. Natural language processing for similar languages, varieties, and dialects: A survey. *Natural Language Engineering*, 26(6):595–612.

Marcos Zampieri, Preslav Nakov, and Yves Scherrer. 2024. Natural language processing for similar languages, varieties, and dialects: A survey. *Natural Language Engineering*, 26(6):595–614.

Koldo Zuazo. 2008. *Euskalkiak. Euskararen dialektoak*. Elkar.