Procesamiento del Lenguaje Natural, Revista nº 70, marzo de 2023, pp. 135-144 recibido 15-12-2022 revisado 31-01-2023 aceptado 10-02-2023

The state of end-to-end systems for Mexican Spanish speech recognition

El estado de los sistemas end-to-end *para el reconocimiento de voz del Español de México*

Carlos Daniel Hernández-Mena[1], **Ivan Vladimir Meza Ruiz**[2]
[1]Language and Voice Laboratory, Reykjavík University.
[2]Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas,
Universidad Nacional Autónoma de México.
carlosm@ru.is, ivanvladimir@turing.iimas.unam.mx

Abstract: Current end-to-end speech recognizer systems report an excellent performance for Spanish. However, this is not reported for specific variants. Moreover, it is unclear if there would be a benefit in creating a fine-tuned version for a particular variant. To investigate these aspects, particularly for Mexican Spanish, we evaluate four different of-the-shelf speech recognizers (one commercial and three open-source); additionally, we fine-tune two systems for Mexican Spanish. We evaluate read and spontaneous speech, present an error analysis and show that fine-tuning for a variant decreases the error rate. As a result of our experimentation, we build two new systems available to the community.
Keywords: speech recognition, acoustic models, mexican spanish.

Resumen: El desempeño actual de los reconocedores de voz se reporta como notablemente bueno para el español, sin embargo, no se especifica el desempeño para variantes especificas, y sobretodo no se establece si existe un beneficio de crear una versión ajustada explicitamente a una variante particular. Para investigar estos aspectos, y especificamente para el español de México, nuestro trabajo evalua el desempeño de cuatro sistemas de reconocimiento de voz (uno comercial y tres de código abierto); adicionalmente creamos dos versiones especificas al español de México mediante la técnica de *fine-tuning*. Se evaluan los sistemas en voz leída y espontanea, presentamos un análisis de error y mostramos que ajustando los sistemas actuales con la variante todavía se puede reducir el error. Como resultado de la experimentación se obtuvieron dos nuevos sistemas que se hacen disponibles a la comunidad.
Palabras clave: reconocimiento de voz, modelos acústicos, español de México.

1 Introduction

Recent progress in end-to-end speech recognition has shown that the performance of Spanish is among the best ones (Radford et al., 2022). In addition, current practices in sharing models and their associated systems have made this technology accessible to a larger pool of potential users. However, it is crucial to notice that this advancement has been reached through the extensive use of private resources and the surge of multilingual and self-supervision settings. This situation makes it difficult to understand the nuances of the field's current state, particularly for individual languages and their variants. It also complicates researching possible improvements to the existing approaches. In this work, we shed light on the current state of Mexican Spanish speech recognition. We use several available language resources and fine-tune well-established speech recognition models to understand the current state better.

We focus on Mexican Spanish, one of the language's most spoken variants, which is spoken predominately in Mexico (Hernández-Mena et al., 2017; Pineda et al., 2010b) and extensively in the United States. It makes

ISSN 1135-5948 DOI 10.26342/2023-70-11

use of 24 phonemes. Table 1 shows the phonetic repertoire of Mexican Spanish in terms of the points of articulation for consonant and vowel sounds. In particular, two phonemes in Mexican Spanish are characteristic of the variant. They are related to the influence of the Nahuatl language (indigenous language of Central Mexico), and they are used in everyday speech (Hernández-Mena and Herrera-Camacho, 2014; Cuétara Priede and others, 2004; Hernández-Mena, 2019): /ʃ/ as in "Xolos" (or "shadow" in English), /t͡l/ at the end of words as "Popocatépetl" (with no counterpart in English). In comparison, the Spain variant of the Spanish language also uses 24 phonemes. The two phonemes that are different are /θ/ and /ʎ/ (for more details see (Quilis, 1993)).

The contribution of this research is to present an evaluation of four different of-the-shelve systems and two fine-tuned versions to quantify the current speech recognition performance for the Mexican Spanish variant. These fine-tuned systems are made freely available. Our evaluation focuses on reading and spontaneous speech; for a general evaluation of Spanish, we include other variants, which include Latin-American, Spain and US-based dialects. First, section 2 presents the work done for Mexican Spanish speech resources and the current rise of end-to-end Speech recognition. Next, in section 3 we present the corpora used in fine-tuning and our evaluation. In section 4, we present the experimentation, results and error analysis to highlight the different system behaviors. Finally, section 5 presents our main findings.

2 Previous work

Since the second half of the nineties, there have been corpora that included or consisted of Mexican Spanish speech recordings. These resources have been commonly used to model the acoustic properties of speech; the resulting model is frequently referred to as *acoustic model*. There have been three primary strategies behind the efforts to collect the data for the acoustic models:

1. To explicitly record samples of Latin-American speakers for speaker identification or speech recognition.
2. To focus only on Mexican speakers.
3. Projects that collected spoken audio in Spanish and later were processed to work as data for speech recognition.

Table 2 list the corpora that was reported and available originally in an academic context: HUB4-NE, a Spanish broadcast news corpus (Consortium, 1997; Fiscus et al., 2001); VoxForge, a corpus of reading speech collected through Internet volunteers (Voxforge.org, 2006); DIMEx100, a phonetic balanced reading speech corpus by Mexican Speakers (Pineda et al., 2010a); DIMEx100 niños, is a version of DIMEx100 where the speakers are children (Moya et al., 2011); Golem-Universum, contains spontaneous interactions of children with a dialogue system system (Venegas-Brione, Meza-Ruiz, and Pineda, 2011); LATINO-40, a corpus of Spanish reading news (Bernstein et al., 1995); West Point Heroico, a corpus of spontaneous speech from Mexican and non-native Spanish speakers (Morgan, 2006); Fisher Spanish, a collection of spontaneous telephone calls (Graff et al., 2010); Hispanic, a collection of reading recordings (Byrne et al., 2014); CIEMPIESS, a spontaneous Mexican Spanish Corpus (Hernández-Mena and Herrera, 2015); CIEMPIES Light, an updated version of CIEMPIESS corpus (Hernández-Mena and Herrera, 2017); CIEMPIESS Balance, a corpus to gender balance CIEMPIESS (Hernández-Mena, 2018); CIEMPIESS experimentation, a version of CIEMPIESS to develop speech recognition systems, it includes CIEMPIES Test for testing Mexican Spanish spontaneous speech (Hernández-Mena, 2019a); LibriVox[1], corpus based on book readings (Hernández-Mena, 2020);Wikipedia grabada[2], corpus of readings of Wikipedia articles (Hernández-Mena and Ruiz, 2021); TEDx, collection of TED talks in Spanish (Hernández-Mena, 2019b). Table 2 summarizes the sizes and availability of the different corpora[3].

Acoustic resources, speech recordings and transcriptions became more relevant with the advent of end-to-end systems, which avoided two traditional sources of informa-

[1]LibriVox website `https://librivox.org/` (last visited April 2022).

[2]Wikipedia *grabada* website `https://es.wikipedia.org/wiki/Wikiproyecto:Wikipedia_grabada` (last visited April 2022).

[3]For further detail about these corpora see (Hernández-Mena et al., 2017) and (Mena and Meza-Ruiz, 2022).

	Points of articulation						
Manners of articulation	**Consonants**	**Labial**	**Labiodental**	**Dental**	**Alveolar**	**Palatal**	**Velar**
	Voiceless Stop	p		t			k
	Voiced Stop	b		d			g
	Voiceless Affricate					t͡ʃ	
	Voiceless Fricative		f		s	ʃ	x
	Voiced Fricative					ʝ	
	Nasal	m			n	ɲ	
	Rhotic				ɾ/ r		
	Lateral				l	t͡l	
	Vowels				**Front**	**Central**	**Back**
	Close				i		u
	Mid				e		o
	Open					a	

Table 1: Phonetic repertoire of Mexican Spanish (Hernández-Mena et al., 2014).

Corpora	**Hours**	**Year**	**Av.**	**Modality**	**Variants**
LATINO-40	6.8h	1995	Cost	Read	Latin-American
HUB4-NE	31h	1997	Cost	Spontaneous	US
CALLHOME Spanish	13*h*	1997	Cost	Spontaneous	Latin-American
DIMEx100	6.1h	2004	Req.	Read	Mexican
VoxForge	50h	2006	Free	Read	Mix
West Point Heroico	16.6h	2006	Cost	Both	North-American
Fisher	163h	2010	Cost	Spontaneous	Latin-American
DIMEx100 niños	8h	2011	Unk.	Read	Mexican
Golem-Universum	0.2h	2011	Unk.	Read	Mexican
CIEMPIESS	17h	2015	Free	Spontaneous	Mexican
CHM150	1.6h	2016	Free	Spontaneous	Mexican
CIEMPIESS Light	18h	2017	Free	Spontaneous	Mexican
CIEMPIESS Balance	18h	2018	Free	Spontaneous	Mexican
CIEMPIESS Experimentation	40h	2019	Free	Spontaneous	Mexican
TEDx Spanish	24h	2019	Free	Spontaneous	Mix
LibriVox Spanish	73h	2020	Free	Read	Mix
Wikipedia Spanish	25h	2021	Free	Read	Mix
Mozilla Common Voice Spanish	320h	2022	Free	Read	Mix

Table 2: Corpora that include Mexican Spanish for the development of speech recognizers (in bold, those that only focus on Mexican Spanish; Av., Availability; Unk., unknown; Req., by request).

tion required by previous approaches: pronunciation dictionaries and language models. Pronunciation dictionaries are a list of words with other corresponding computer-based phonetic transcription; language models are probabilistic models that determine the probability of a sequence of words. Although these last ones nowadays are heavily used to re-score the output of end-to-end systems (Wang, Wang, and Lv, 2019). In particular, end-to-end speech recognition relies on deep neural networks to relate segments of the acoustic signal with the character sequence of transcription (Hannun et al., 2014) and on the CTC loss function (Graves et al., 2006), which collapse the sequence of characters and compare it to the correct transcription during training.

Another recent advancement in the field consisted of using self-supervision settings to train models that rely on the acoustic signal to create better sound representations than what traditional acoustic models can reach (Schneider et al., 2019). In addition, self-supervision allows reaching good performance in multilingual settings. In this setting, first, a model is self-trained with speech recordings from multiple languages without the need for transcriptions; for instance, it is trained to predict the next segment of the

audio signal; later, this model gets fine-tuned using an end-to-end setting to perform speech recognition (Conneau et al., 2020). This arrangement was the backbone during the creation of the Whisper system, which became state-of-the-art in the field (Radford et al., 2022). As a result, Whisper reaches a 5.4% word error rate performance for Spanish, the best-reported performance for the language up to this moment.

3 Systems and datasets

For our experiments, we selected four out-of-the-shelf systems:

Google speech recognizer[4] This is a commercial system widely adopted in Mexico which supports Latin America variants. However, its parts and training are not publicly shared.

Quarztnet[5]: It is an implementation of a Quarztnet architecture (Kriman et al., 2020) on the NeMo platform developed by NVIDIA (Kuchaiev et al., 2019; Fidjeland et al., 2009). This model was first trained with several English corpora to be later fine-tuned using the Spanish Common Voice Mozilla corpus. This is an example of transfer learning using a pre-trained model.

Wav2vec[6]: Model based on the XLSR-53 system (Conneau et al., 2020) train in a multilingual setting. In particular, this model was also fine-tuned with the Spanish Common Voice Mozilla corpus.

Whisper[7]: Model trained on 680,000 hours of several languages recordings, including Spanish (Radford et al., 2022). There are five pre-trained versions of this system which vary in size: *tiny*, *base*, *small*, *medium* and *large*.

In addition to the pre-trained systems, we fine-tuned two new models:

Quartznet fine-tuned[8] based on a Spanish Quarztnet model described above[9].

Wav2vec fine-tuned[10] based on a XLSR wav2vec large model[11].

Both fine-tuned systems were further trained with 944h of predominantly Mexican Spanish. For the Mexican Spanish, we use the corpora: CIEMPIESS Light, CIEMPIESS Balance, CIEMPIES FEM, CHM150, TEDx Spanish, DIMEX100, DIMEX100 niños, Golem-Universum, VoxForge, LIBRIVOX Spanish, WIKIPEDIA Spanish, Spanish Mozilla Common Voice 10.0, West Point Heroico, LATINO-40, CALLHOME Spanish, HUB4NE Spanish, FISHER Spanish. Additionally, we also incorporate two private collections called *Tele con Ciencia* (28h16m) and extra recordings from another private collection of Mexican recordings (118h22m), which can not be shared given their copyright status. At the fine-tuning stage, we also added Spanish variants corpora: the Spanish portion of MediaSpeech (10h) citemediaspeech2021, Spanish form Spain (6h40m) (Garrido et al., 2013), Chilean (7h08m), Colombian (7h34m), Peruvian (9h13m), Argentinian (8h01m) and Puerto Rican (1h00m) all of these corpora are part of the project *Crowdsourcing Latin American Spanish* (Guevara-Rukoz et al., 2020).

During testing we use Mozilla Common Voice Speech (MCVS) since given its accessibility, results on MCVS are commonly reported; however, since their modality is read speech, the performance tends to be better than expected for spontaneous speech. To contrast, we have evaluated performance in three spontaneous speech corpora: the HUB4-NE, CALLHOME, and CIEMPIESS. We also isolated the Mexican speakers from the MCVS and evaluated performance on this segment of this corpus. We also report

[4]Website describing system: `https://cloud.google.com/speech-to-text/` (last visited November 2022).

[5]Website for the pre-trained STT Es Quartznet15x5 model: `https://catalog.ngc.nvidia.com/orgs/nvidia/teams/nemo/models/stt_es_quartznet15x5` (last visited November 2022).

[6]Website for Fine-tuned XLSR-53 large model for speech recognition in Spanish: `https://huggingface.co/jonatasgrosman/wav2vec2-large-xlsr-53-spanish` (last visited November 2022).

[7]Website for whisper model: `https://github.com/openai/whisper` (last visited November 2022).

[8]Fine-tuned model available at: `https://huggingface.co/carlosdanielhernandezmena/stt_es_quartznet15x5_ft_ep53_944h` (last visited November 2022).

[9]Description of the model: `https://catalog.ngc.nvidia.com/orgs/nvidia/teams/nemo/models/stt_es_quartznet15x5` (last visited November 2022).

[10]Fine-tuned model available at: `https://huggingface.co/carlosdanielhernandezmena/wav2vec2-large-xlsr-53-spanish-ep5-944h` (last visited November 2022).

[11]Description of the model: `https://huggingface.co/facebook/wav2vec2-large-xlsr-53`

our development results on the MCVS and CALLHOME which include this partition.

4 Experiments and results

Table 4 shows the Word Error Rate (*WER*) results of ten versions of the four systems (the lower the result, the better). As it can be seen, the *wav2vec* system performs the best in six of the eight testing scenarios, while *Whisper large* performs the best in the rest (two). Notice that both versions of *wav2vec* are fine-tuned, *W2V* originally was fine-tuned using the Mozilla Spanish Corpus (includes read Mexican Spanish). In contrast, our fine-tuned version was trained with the collection of predominately Mexican Spanish corpora (spontaneous Mexican Spanish). The benefit of this fine-tuning can be appreciated within the results for the CIEMPIESS test corpus, which consists of Mexican Spanish recordings; this system reaches the best performance for spontaneous speech: 11.17 of WER. Also, it reaches a new best score for the HUB4-NE with 7.48. On the other hand, the *Whisper* system consistently gets a competitive performance and scores two of the best performances for the Mozilla Common Voice Speech test corpora. Remember, this is a large model which relies on more than half a million training hours in a multilingual setting, including Spanish speech.

Another aspect to consider with the new *Whisper* system is the longer time it takes to transcribe. As expected, the larger the model, the more parameters and time to transcribe. This can be seen in Table 3, which records more than 7 hours for the large version of the system. As a point of comparison, the other systems took no longer than 30 minutes on the same amount of data. Their performance was so consistent among themselves that we did record them. However, when Whisper took too long, we started to record it.

In Table 4, we also notice some drawbacks when fine-tuning a model. While it would be logical that fine-tuning the model using data closest to the target will improve the performance, this does not always happen since the performance of the original model could be degraded. For instance, *Quartznet* had an excellent performance for the Mozilla Common Voice, but this became worst with the fine-tuning. This effect has been noticed previously (Huang et al., 2020). Sometimes, the fine-tuning degrades the performance from a previously scored performance. However, the positive impact can be noticed with the rest of the testing corpora in which the fine-tuned version produces fewer mistakes. We hypothesize that this is related to the "closeness" of the variant. By fine-tuning, it stops being close to reading Mexican speech and gets closer to the spontaneous version.

Another interesting aspect is the difficulty associated with the CALLHOME corpus (associated with the worst performances). Our experience points out that this is a problematic corpus. A preliminary analysis of the transcriptions shows a challenging setting where it is common to find overlapping speech among speakers. Additionally to this, there is no suitable transcription protocol for such cases.

4.1 Error analysis

Word Error Rate (WER)is based on edit-distance operations: *insert*, *delete*, and *replace*. WER quantifies the percentage of operation to transform the expected output (reference transcription) into the system's transcription (hypothesis). Figure 1 shows the percentage of *insertions* per word and normalized by the occurrence of such term in the system transcription. These percentages were ranked from larger to lower. A sound system should start transitioning from words for which all occurrences were inserted (1.0) to terms for which a low percentage of the occurrences were inserted. The *wav2vec* system has fewer insert operations in this figure (green line); it is followed by Whisper medium and large (grey and light pink lines). This can be interpreted as these systems being less eager to propose words. Insertions could be viewed as an acoustic hallucination (the system "listen" to a word which is not there). Table 5 shows some examples of hallucinated words and their frequencies (between parentheses). In total hallucinations are in the hundreds, but most words get inserted only once. One source of error is the insertion of single letters, which is expected partly because these systems are end-to-end and allow bits of the signal to relate to a bit of transcription, even though it does not relate to a word.

On the other hand, Figure 2 shows the ranking of the *deletions* per word and normalized by the reference corpus. Similarly to

Corpus	Tiny	Base	Small	Medium	Large
MCVS dev	1h47m	1h41m	2h48m	5h23m	7h51m
MCVS test	1h47m	1h46m	2h40m	5h25m	8h2m

Table 3: Whisper time to transcribe 15k audios from the Mozilla Common Voice Speech corpora *test* and *dev* portions. The runs were done using NVidia GeForce GTX Titan X GPU.

System	Go.	QN	W2V	Whisper					fine-tuned	
Corpora				Tn.	Bs.	Sm.	Med.	Lg.	QN	W2V
HUB4-NE	17.79	22.87	12.84	29.27	22.84	15.63	11.92	10.82	14.48	**7.48**
MCVS dev	17.68	12.85	**4.12**	33.75	20.89	10.27	6.49	5.86	15.97	8.02
MCVS Mex dev	17.81	11.72	**4.69**	32.0	20.32	9.90	6.66	5.87	14.96	7.59
MCVS test	19.59	14.89	8.70	37.02	23.32	11.73	7.58	**6.80**	17.99	9.20
MCVS Mex test	21.84	14.29	8.04	33.82	21.75	11.7	7.92	**6.89**	16.33	8.93
CALLHOME dev	52.93	78.68	61.45	91.76	74.52	53.37	44.42	41.44	56.34	**40.39**
CALLHOME test	51.92	78.07	60.29	86.43	70.18	50.32	41.91	39.25	55,43	**39.12**
CIEMPIESS test	18.19	36.69	23.16	28.59	22.17	18.28	15.10	15.25	18.57	**11.17**

Table 4: Word error rate for evaluation corpora (lower the better; Go., Google QN, Quartznet; W2V, wav2vec; Tn, tiny;Bs, base; Sm., small; Med., medium; Lg., Large).

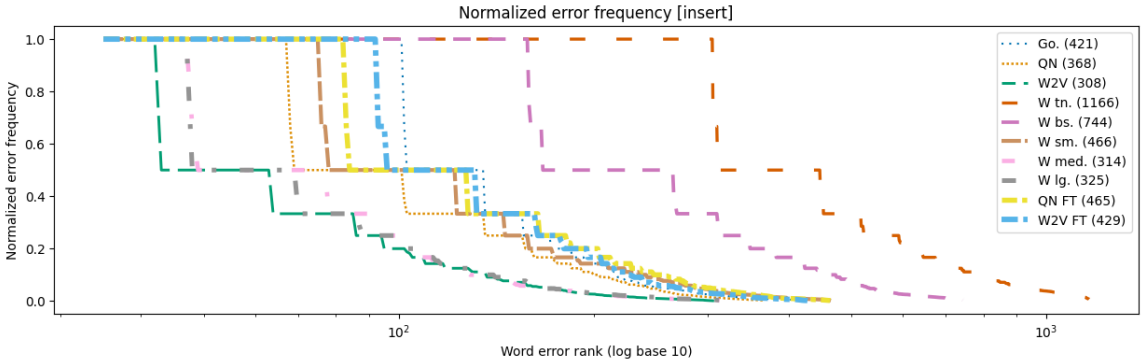

Figure 1: Ranked normalized *insertions* frequency per word (lower the better; Go., Google QN, Quartznet; W2V, wav2vec; FT, fine-tuned; Tn, tiny;Bs, base; Sm., small; Med., medium; Lg., Large).

the *insertions*, the faster the system transitions from a high percentage (1.0) to lower, the better performance. Here we can see that both fine-tuned systems have the fewer deletions, *wav2vec* (blue line) and *QuartzNet* yellow line.

Figure 3 shows the normalized frequencies per word replaced. This operation is harder to normalize because it can be interpreted as a combination of a deletion and an insertion and is anchored to two words, the deleted and the inserted. To normalize, we use the higher count of any of the words: inserted or deleted. This is related to replacement operations being higher than insertions and deletions. In order to gain further understanding regarding the different systems, we calculate the number of operations per word; this is a proxy to learn how different the words in the hypothesis transcription are compared to the references. Figure 4 shows the histograms of several edits on each word to replace. As can be noticed, most terms need to be replaced by a word very close in spelling and only different in one edit. The *W. Lg.* is the system that better performs with 7,735 replacements. However, remember this system implies more significant transcription times, so it might not be a reasonable cost-effective compromise.

Reflecting on the performance of the systems, the best ones *Whisper Large* and *W2V fine-tuned* performance are very similar, but they have different behaviours. *W2V fine-tuned* takes more risks proposing words (higher insertion error), but the proposed ones are usually correct (lowest deletion error). On the other hand, the *Whisper Large*

System	Total	Exaples
Go.	90	post(2), auto(2), 16(2), refuerzo(2), 70(2), reorganizar(1), pos(1),...
QN	127	l(5), n(3), s(3), auto(3), digamo(2), dy(2), qu(2), tam(2), d(2), ...
W2V	86	l(6), n(4), mas(2), pos(2), d(2), puras(2), fracean(1), your(1), metodo(1), ...
W Tn.	565	os(6), auto(4), p(4), estero(4), gabriel(2), s(2), tr(2), l(2) ...
W Bs.	287	auto(3), transcribe(3), estimar(2), os(2), juris(2), pura(2), tango(2), s(2), ...
W Sm.	180	qte(5), auto(3), post(2), agarró(2), pura(2), mecánicas(2), extra(2), eh(2), ...
W Med.	72	high(3), pura(2), post(1), método(1), lacktut(1), delan(1), ow(1), ...
W Lg.	110	agarró(2), pura(2), manny(2), auto(1), método(1), hertz(1), papito(1), ...
QN FT	164	n(8), s(4), piro(4), l(4), d(3), g(2), bum(2), payz(2), epaíses(1), escribier(1), ...
W2V FT	116	s(8), h(8), n(5), pos(3), eh(3), l(3), pura(2), d(2), método(1), cl(1), seso(1), ...

Table 5: Example of words transcribed by the systems but not present in the reference transcription.

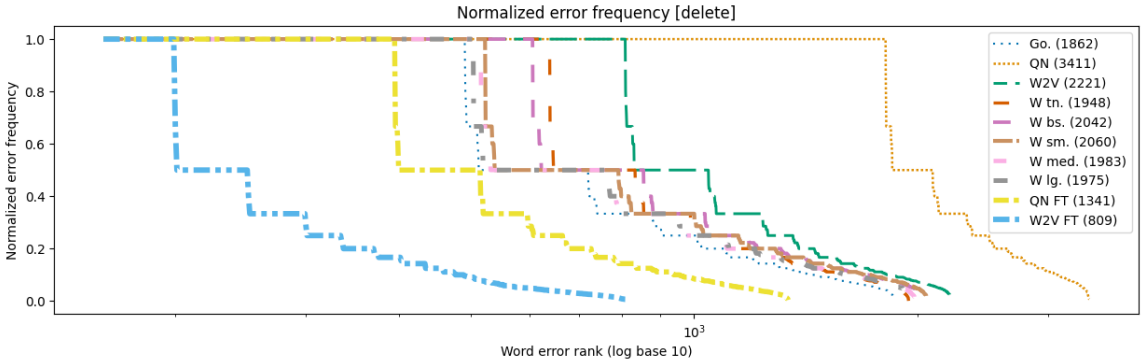

Figure 2: Ranked normalized *deletions* frequency per word (lower the better; Go., Google QN, Quartznet; W2V, wav2vec; FT, fine-tuned; Tn, tiny;Bs, base; Sm., small; Med., medium; Lg., Large).

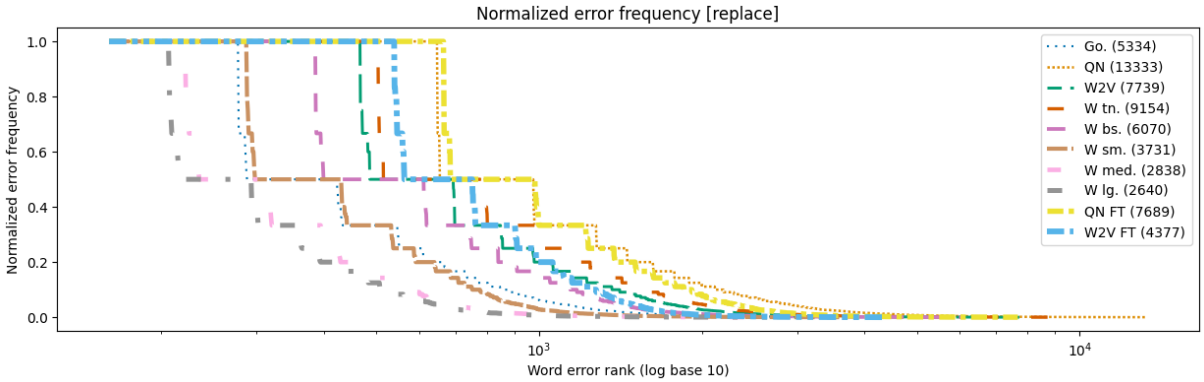

Figure 3: Ranked normalized *replace* frequency per word (lower the better; Go., Google QN, Quartznet; W2V, wav2vec; FT, fine-tuned; Tn, tiny;Bs, base; Sm., small; Med., medium; Lg., Large).

is shier when proposing words, so it omits several words (higher deletion error, but lowest insertion error). This characteristic of taking more risks when proposing words is also observed in the replacement, on which *W2V fine-tuned* also gets a higher rate of errors. However, it seems to be a good strategy for a speech recognizer since it performs better. The code for the analysis is freely available[12].

5 *Conclusion*

We have presented an evaluation of speech recognisers for the Mexican variant of Spanish. For several decades there has been

[12]Code for the error analysis in speech transcription: `https://github.com/ivanvladimir/speech_transcriptions_analysis` (last visited March 2023)

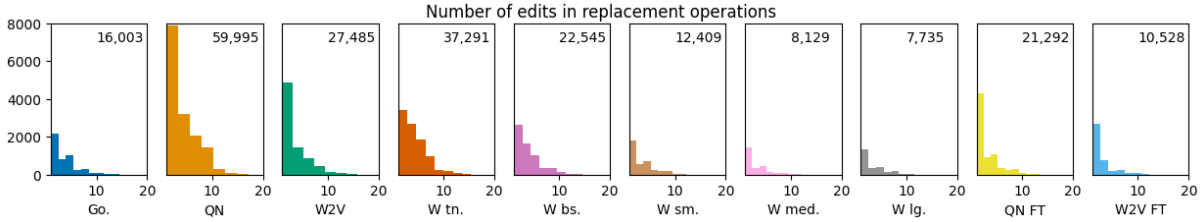

Figure 4: Histograms of Levenshtein distances between replaced words (the lower, the better; a sum of all the distances is in the upper right corner; it follows the same colours than 3).

an effort to support the creation of language resources focused on this variant. Although new methods such as end-to-end speech recognition facilitate the construction of multilingual settings and have helped increase the performance of the current systems, in this work, we show that fine-tuning for a specific variant still has its benefits. In our experience, we will continue to recommend the collection of language resources for specific variants and the fine-tuning based on pre-trained models.

On the other hand, there are still open questions regarding these adaptations. First, our observations must be confirmed on new multilingual general models recently released, such as *Wav2vec XLR* or new *Whisper* versions, which unfortunately rely on much more computer power. Second, we believe it would be important to the development of speech technology to have more diversity of variants with their corresponding resources. However, at this moment, there is no clear answer on how to mix these variants to reach a good performance for most speakers without losing performance during the fine-tuning process. We also would like to create fine-tuned versions of specific variants without including other ones to quantify the effect of variants and sub-variants and the support the rest of the variants can provide.

Acknowledgments

Authors thank Mónica Alejandra Ruiz López for verifying and correcting the transcriptions of the CIEMPIESS Test corpus. Carlos Hernández-Mena thanks the support from Language and Voice Laboratory from Reykjavik University in the realization of this manuscript.

References

Bernstein, J., B. Grundy, E. Rosenfeld, A. Najmi, and P. Mankoski. 1995. Latino-40 spanish read news ldc95s28. CD.

Byrne, W., E. Knodt, J. Bernstein, and F. Emami. 2014. Hispanic-english database ldc2014s05. CD.

Conneau, A., A. Baevski, R. Collobert, A. Mohamed, and M. Auli. 2020. Unsupervised cross-lingual representation learning for speech recognition. *arXiv preprint arXiv:2006.13979*.

Consortium, L. D. 1997. 1997 spanish broadcast news speech (hub4-ne) ldc98s74. Web download.

Cuétara Priede, J. O. et al. 2004. *Fonética de la ciudad de México. Aportaciones desde las tecnologías del habla*. Tesis-UNAM.

Fidjeland, A. K., E. B. Roesch, M. P. Shanahan, and W. Luk. 2009. Nemo: a platform for neural modelling of spiking neurons using gpus. In *2009 20th IEEE international conference on application-specific systems, architectures and processors*, pages 137–144. IEEE.

Fiscus, J. G., J. S. Garofolo, M. Przybocki, W. Fisher, and D. Pallett. 2001. 1997 hub4 broadcast news evaluation non-english test material ldc2001s91. Web download.

Garrido, J. M., D. Escudero, L. Aguilar, V. Cardeñoso, E. Rodero, C. De-La-Mota, C. González, C. Vivaracho, S. Rustullet, O. Larrea, et al. 2013. Glissando: a corpus for multidisciplinary prosodic studies in spanish and catalan. *Language resources and evaluation*, 47(4):945–971.

Graff, D., S. Huang, I. Cartagena, K. Walker, and C. Cieri. 2010. Fisher spanish speech ldc2010s01. CD.

Graves, A., S. Fernández, F. Gomez, and J. 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.

Guevara-Rukoz, A., I. Demirsahin, F. He, S.-H. C. Chu, S. Sarin, K. Pipatsrisawat, A. Gutkin, A. Butryna, and O. Kjartansson. 2020. Crowdsourcing Latin American Spanish for Low-Resource Text-to-Speech. In *Proceedings of The 12th Language Resources and Evaluation Conference (LREC)*, pages 6504–6513, Marseille, France, May. European Language Resources Association (ELRA).

Hannun, A., C. Case, J. Casper, B. Catanzaro, G. Diamos, E. Elsen, R. Prenger, S. Satheesh, S. Sengupta, A. Coates, and A. Y. Ng. 2014. Deep speech: Scaling up end-to-end speech recognition.

Hernández-Mena, C. D. 2019. The CIEMPIESS proper-names pronouncing dictionary. In *Corpus presented at OpenCor 2019 Conference, Guanajuato City, Mexico. Available online at https://opencor.gitlab.io/corpora-list/*, page 1.

Hernández-Mena, C. D. and J.-A. Herrera-Camacho. 2014. Ciempiess: A new open-sourced mexican spanish radio corpus. In *LREC*, volume 14, pages 371–375.

Hernández-Mena, C. D., I. Meza-Ruiz, J. Herrera-Camacho, et al. 2017. Automatic speech recognizers for Mexican Spanish and its open resources. *Journal of Applied Research and Technology*, 15(3):259–270.

Hernández-Mena, C. D. 2018. Ciempiess balance ldc2018s11.

Hernández-Mena, C. D. 2019a. Ciempiess experimentation ldc2019s07.

Hernández-Mena, C. D. 2019b. TEDx Spanish Corpus. Audio and transcripts in Spanish taken from the TEDx Talks; shared under the CC BY-NC-ND 4.0 license. Web Download.

Hernández-Mena, C. D. 2020. Librivox spanish ldc2020s01.

Hernández-Mena, C. D. and A. Herrera. 2015. Ciempiess ldc2015s07.

Hernández-Mena, C. D. and A. Herrera. 2017. Ciempiess light ldc2017s23.

Hernández-Mena, C. D. and I. V. M. Ruiz. 2021. Wikipedia spanish speech and transcripts ldc2021s07.

Huang, J., O. Kuchaiev, P. O'Neill, V. Lavrukhin, J. Li, A. Flores, G. Kucsko, and B. Ginsburg. 2020. Cross-language transfer learning, continuous learning, and domain adaptation for end-to-end automatic speech recognition. *arXiv preprint arXiv:2005.04290*.

Kriman, S., S. Beliaev, B. Ginsburg, J. Huang, O. Kuchaiev, V. Lavrukhin, R. Leary, J. Li, and Y. Zhang. 2020. Quartznet: Deep automatic speech recognition with 1d time-channel separable convolutions. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6124–6128. IEEE.

Kuchaiev, O., J. Li, H. Nguyen, O. Hrinchuk, R. Leary, B. Ginsburg, S. Kriman, S. Beliaev, V. Lavrukhin, J. Cook, et al. 2019. Nemo: a toolkit for building ai applications using neural modules. *arXiv preprint arXiv:1909.09577*.

Mena, C. D. H. and I. Meza-Ruiz. 2022. Creating mexican spanish language resources through the social service program. In *Proceedings of the 2nd Workshop on Novel Incentives in Data Collection from People: models, implementations, challenges and results within LREC 2022*, pages 20–24.

Morgan, J. 2006. West point heroico spanish speech ldc2006s37. Web Download.

Moya, E., M. Hernandez, L. Pineda, and I. Meza. 2011. Speech recognition with limited resources for children and adult speakers. In *2011 10th Mexican International Conference on Artificial Intelligence*, pages 57–62. IEEE.

Pineda, L. A., H. Castellanos, J. Cuétara, L. Galescu, J. Juárez, J. Llisterri, P. Pérez, and L. Villaseñor. 2010a. The corpus dimex100: transcription and evaluation. *Language Resources and Evaluation*, 44(4):347–370.

Pineda, L. A., H. Castellanos, J. Cuétara, L. Galescu, J. Juárez, J. Llisterri, P. Pérez, and L. Villaseñor. 2010b. The Corpus DIMEx100: transcription and evaluation. *Language Resources and Evaluation*, 44(4):347–370.

Quilis, A. 1993. *Tratado de fonología y fonética españolas*, volume 2. Gredos Madrid.

Radford, A., J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever. 2022. Robust speech recognition via large-scale weak supervision. *OpenAI Blog*.

Schneider, S., A. Baevski, R. Collobert, and M. Auli. 2019. wav2vec: Unsupervised pre-training for speech recognition. In *INTERSPEECH*.

Venegas-Brione, E., I. Meza-Ruiz, and L. A. Pineda. 2011. Evaluation of a dialogue system for children based on an interaction-oriented cognitive architecture. *Procesamiento del lenguaje natural*, pages 113–120.

Voxforge.org. 2006. Free speech... recognition (linux, windows and mac) - voxforge.org. `http://www.voxforge.org/`.

Wang, D., X. Wang, and S. Lv. 2019. An overview of end-to-end automatic speech recognition. *Symmetry*, 11(8):1018.