Towards an ASR System for Documenting Endangered Languages: A Preliminary Study on Sardinian

Ilaria Chizzoni[1], Alessandro Vietti[1]

[1]Free University of Bozen-Bolzano

Abstract

Speech recognition systems are still highly dependent on textual orthographic resources, posing a challenge for low-resource languages. Recent research leverages self-supervised learning of unlabeled data or employs multilingual models pre-trained on high resource languages for fine-tuning on the target low-resource language. These are effective approaches when the target language has a shared writing tradition, but when we are confronted with mainly spoken languages, being them endangered minority languages, dialects, or regional varieties, other than labeled data, we lack a shared metric to assess speech recognition performance. We first provide a research background on ASR for low-resource languages and describe the specific linguistic situation of Campidanese Sardinian, we then evaluate five multilingual ASR models using traditional evaluation metrics and an exploratory linguistic analysis. The paper addresses key challenges in developing a tool for researchers to document and analyze the phonetics and phonology of spoken (endangered) languages.

Keywords

Speech recognition, Campidanese Sardinian, Resource and evaluation, Spoken language documentation

1. Introduction

The growing interest in understudied languages has led to categorizing them on the basis of resource availability, defining them as high, low, or zero-resource languages. In the narrowest sense, zero and low-resource languages are those lacking sufficient data to train statistical and machine learning models [1] [2] [3]. However, such a technical definition is not adequate to account for the different linguistic scenarios of world languages. As a matter of fact, in the literature, the term low and zero resource languages is still used inconsistently. Sometimes, it is used to describe standard, widely spoken languages with a shared orthography, that cannot rely on many hours of transcribed or annotated speech, see Afrikaans, Icelandic, and Swahili in [4]. Sometimes, it is used for non-standard, widely spoken languages, lacking a shared orthography (no orthography or multiple proposed orthographies) as for Swiss German dialects [5] or Nasal and Besemah [6]. And sometimes to refer to non-standard, endangered languages lacking a shared orthography, like Bribri, Mi'kmaq and Veps [3].

These scenarios are mainly being addressed with two approaches: The first leverages self-supervised learning, and uses unlabeled data from the target language to learn linguistic structures [7]. Self-supervised learning is an optimal choice in low-resource settings because only requires to gather more audio data. However, it seems costly and prone to catastrophic forgetting [6] [4]. The second approach involves training a multilingual model on labeled data from highly-resourced languages and then applying the trained model to transcribe unseen target languages. This includes the benefits of a supervised learning setting and proved to be effective [8]. Pre-trained multilingual models can then be fine-tuned on just a smaller dataset of labeled data in the target language. Since fine-tuning is a straightforward, efficient approach, it is the preferred one to address the problem of low-resource languages [6]. However, the success of this approach still depends on the amount of available labeled data in the target language or whether or not it is possible to generate more, e.g., via data augmentation.

Several data augmentation approaches for low-resource languages are currently being explored, including self-learning [6], text-to-speech (TTS) [6] or optimized dataset creation approaches [9]. Bartelds and colleagues [6] propose data augmentation techniques to develop ASR for minority languages, regional languages or dialects. They employ a self-training method on Besemah and Nasal two Austronesian languages spoken in Indonesia. In self-training, a teacher XLS-R model is fine-tuned on manually transcribed data, the teacher model is used to transcribe unlabeled speech and then a student model is fine-tuned on the combined datasets of manually and automatically transcribed data. Since the collected 4 hours of manually transcribed speech for Besemah and Nasal followed different orthography conventions, the transcriptions were first normalized to working orthographies and then used for fine-tuning. In the same framework, they leveraged a pre-existing

CLiC-it 2024: Tenth Italian Conference on Computational Linguistics, Dec 04 — 06, 2024, Pisa, Italy

*Corresponding author.

†These authors contributed equally.

ilaria.chizzoni@unibz.it (I. Chizzoni); alessandro.vietti@unibz.it (A. Vietti)

0009-0009-9936-1220 (I. Chizzoni); 0000-0002-4166-540X (A. Vietti)

TTS system available for Gronings, a Low-Saxon language variant spoken in the province of Groningen in the Netherlands, to generate more synthetic training data from textual sources and they achieved great results [6].

While fine-tuning paired with data augmentation techniques works for low-resource, widely-spoken languages, developing a speech recognition system for endangered spoken languages also involves ethical considerations towards the local community. More participatory research is required to understand the native speakers' relationship with the written form of their language, as well as with language technologies. In their position paper [3] Liu and colleagues emphasize the importance of creating language technologies in consultation with speakers, activists, and community language workers. They present a case study on Cayuga, an endangered indigenous language of Canada with approximately 50 native elder speakers and an increasing number of young L2 speakers. After gaining insights from the community, they began collaborating on a morphological parser. This tool aids teachers and young L2 students in language learning while gradually providing morphological annotations and segmentations useful for developing ASR systems for researchers. Blaschke and colleagues [10] surveyed over 327 native speakers of German dialects and regional varieties, finding that respondents prefer tools that process speech over text and favor language technology that handles dialect speech input rather than output. Understanding the needs of the speech community and differentiating them from those of linguistic researchers can guide research more effectively.

This paper outlines the first steps towards a speech recognition system for researchers to aid the systematic analysis of the phonetics and phonology of Campidanese, an endangered language spoken in southern Sardinia. To achieve this goal, we first describe the situation of the speech community of the target language, we then select five speech recognition multilingual and ready for inference models and evaluate them on Campidanese Sardinian. When multilingual models were not available for speech recognition task, we chose multilingual models fine-tuned on Italian, which we assume to be a relatively close language both genealogically and structurally. We assess the goodness of the models' inferences, first by computing the traditional evaluation metrics, i.e., average Word Error Rate (WER) and Character Error Rate (CER), and then carrying out a qualitative linguistic analysis to have better insights of which model best meets the needs for language documentation and research. This work is part of "New Perspectives on Diphthong Dynamics (DID)", a joint project between the University of Bozen and the Ludwig-Maximilians-Universität München, focusing on the study of diphthongs dynamics in two understudied languages, i.e., Campidanese Sardinian and Tyrolean and aims to build a corpus for the linguistic documentation of these two languages.

2. Campidanese Sardinian

Sardinian is a Romance language spoken on the Sardinia island in Italy [11]; it is considered an official minority language and is protected by National Law n.482/1999 and Regional Law n.26/1997 but does not have a written standard [12]. Sardinia has a high internal linguistic diversity but the two main macro varieties are Logudorese (ISO code 639-3 src), spoken in the northern sub-region and Campidanese (ISO code 639-3-sro), spoken in the southern sub-region of Sardinia [12]. To date, there are no quantitative studies on the real number of Sardinian speakers. The first sociolinguistic survey [13] carried out by Regione Sardegna in 2007 on 2437 speakers states that 68.4% of the respondents claim to know and speak a variety of the local languages. However, the survey was based on the speakers' self-assessment. As far as Campidanese Sardinian is concerned, Ethnologue lists it as an endangered indigenous language [14] and research [12] claims it is used as a first language just by some elder adults in the ethnic community, and not taught to children anymore. In 2017, Rattu [15] carried out a sociolinguistic survey on 310 Cagliari speakers, where a self-assessment questionnaire was followed by a language test (mostly translation tasks from Italian to Sardinian) and only a minority of respondents over the age of 45 achieved good or excellent results.

The Sardinian Regional Administration presented two proposals for an official standard language: the first in 2001, presented as a linguistic compromise but actually over representative of Logudorese (Limba Sarda Unificada LSU), and the second in 2006, mainly based on the central regional variety (Limba Sarda Comuna LSC) [12]. The latter remains the one used for communication by the Regional Administration, while in the Cagliari Province a proposal of orthographic rules for Campidanese called *Sa Norma Campidanesa* has been put forward in 2009 by the *Comitau Scientìficu po sa normalisadura de sa bariedadi campidanesa de sa lìngua sarda* [16]. Without discussing the issue of the orthographic norm, which is inherently political, we would like to point out that these proposals do not seem to have become part of everyday language use by the speech community [17]. This is primarily because they were not based on any official data regarding the linguistic and sociolinguistic situation or language use [18]. Therefore, these standards remained limited to administrative communications.

Some tendencies in the speakers' linguistic attitudes emerged from the DID project data collection fieldwork conducted in 2023 in the city of Sinnai. Native speakers of Campidanese are often unfamiliar with the written version of their language. Elder native speakers had

no way or need to write the language, except in the last decade through social networks. Whereas, the few young people who use the language even in its written version to communicate with friends and family via message service apps, do not use *Sa Norma Campidanesa*, but rather use a transcription that intuitively approximates their pronunciation.

3. Experiments

3.1. Campidanese Sardinian dataset

We decided to evaluate the speech recognition models on a small sample of highly controlled Sardinian data, in order to carry out a qualitative linguistic analysis of the output transcription. The dataset includes short audios of read speech with an average length of 3.5 seconds (read_short), long audios of read speech with an average length of 23 seconds (read_long), and short audios of spontaneous speech with an average length of 5.3 seconds (spontaneous). Read speech is a subset of the corpus gathered during the DID project fieldwork in Sinnai. For the read_short, participants were asked to read aloud short sentences developed by the research group, using an orthography close to *Sa Norma Campidanesa.* In particular, twenty audio clips of four native speakers (2F and 2M) were selected. Two longer audio clips were selected from the same corpus: one of a female speaker reading an autograph poem, and another of a male speaker reading an excerpt of an autograph story. To have speech style variability, chunks of spontaneous speech from ethnographic interviews collected by Mereu [19] in Cagliari in 2016 were included. Twelve audio chunks were extracted from two of the interviews conducted with two male native speakers of Campidanese. The orthographic transcripts followed different Campidanese conventions either being written or validated by native speakers.

3.2. Methods

From HuggingFace's Open ASR Leader board [20], ready-to-test models with low Real-Time-Factor (RTF) values were selected. Out of the five tested models, two are multilingual models containing at least one Romance language in their training dataset i.e., `whisper-large-v2` and `multilingual-fastconformer-hybrid-large`; and three were multilingual models fine-tuned on Italian datasets and ready for inference, this is the case for `it-fastconformer-hybrid-large` from NVIDIA and `wav2vec2-large-xlsr-53-italian` and `wav2vec2-xlsr-53-espeak-cv-ft` from Facebook.

Open AI Whisper is a Transformer sequence-to-sequence multilingual and multitask model trained on performing multilingual speech recognition, speech translation, spoken language identification, and voice activity detection [21]. We tested it without passing a specific language.

The multilingual FastConformer Hybrid Transducer-CTC model is a model developed by NVIDIA, combining the FastConformer architecture with a hybrid Transducer-CTC approach [22]. NVIDIA FastConformers come across as very competitive for their efficiency and computational speed. We tested both the multilingual model version 1.20.0, trained on Belarusian, German, English, Spanish, French, Croatian, Italian, Polish, Russian, and Ukrainian [22], and the Italian model version 1.20.0 trained specifically on Italian (Mozilla Common Voice 12, Multilingual LibriSpeech and VoxPopuli) [23].

By Facebook we chose Wav2Vec 2.0 XLSR, a model that learns cross-lingual speech representations from the raw waveform of speech in multiple languages during pre-training [24]. We use `wav2vec2-large-xlsr-53-italian`, the Wav2Vec 2.0 model pre-trained on multilingual data from Multilingual LibriSpeech, Mozilla Common Voice and BABEL and fine-tuned on Italian [25]. To attempt an automatic phonetic transcription we used `wav2vec2-xlsr-53-espeak-cv-ft`, the same Wav2Vec 2.0 Large XLSR model, fine-tuned on multilingual Common Voice dataset to recognize phonetic labels [8].

In order to have a standard reference, traditional evaluation metrics for speech recognition systems like WER and CER were computed via the `evaluate` HuggingFace library [26]. Since the output text was normalized differently by the different models, a text normalization was done on both reference and hypothesis transcriptions, removing every special characters (non-alphanumeric characters) before computing WER and removing special characters and spaces (tabs, spaces and new lines) before computing CER. We made no additional changes to the inferences, and no default parameters of the models were modified. All tests were run locally to respect data privacy policies.

3.3. Models evaluation

Regarding the WER metric, we assume models to perform possible word recognition based on the inventory of multilingual or Italian tokens, since the model has not been trained or fine-tuned on any Sardinian data. This is why in our case average WER is poorly significant. We therefore evaluate performance mainly by looking at CER.

In Table 1 we can see there is little difference in the performance between Whisper medium and large-v2. Surprisingly, however, Whisper medium performs better on long read-speech data, reaching a CER of 0.22 versus Whisper large-v2 only achieving 0.36. This could be due to a better performance of the translation task in Whisper large-v2. However, the larger model performs better on spontaneous speech (CER 0.39) then the medium model

Table 1
Whisper Models

Model	Style	Length (s)	CER	WER
large-v2	read_short	3.5	0.69	1.02
large-v2	read_long	23.5	**0.36**	**0.76**
large-v2	spontaneous	5.3	0.39	0.90
medium	read_short	3.5	0.70	1.00
medium	read_long	23.5	**0.22**	**0.79**
medium	spontaneous	5.3	0.52	1.12

Table 2
FastConformer NVIDIA Models

Model	Style	Length (s)	CER	WER
FC-ML	read_short	3.5	0.69	1.00
FC-ML	read_long	23.5	**0.22**	**0.79**
FC-ML	spontaneous	5.3	0.34	0.88
FC-IT	read_short	3.5	0.69	1.00
FC-IT	read_long	23.5	**0.28**	**0.83**
FC-IT	spontaneous	5.3	0.41	0.97

Table 3
Wav2Vec XLSR Italian

Model	Style	Length (s)	CER	WER
W2V-IT	read_short	3.5	0.68	1.00
W2V-IT	read_long	23.5	**0.25**	**0.81**
W2V-IT	spontaneous	5.3	0.36	0.90

(CER 0.52). As shown in Table 2, both NVIDIA Fast Conformer models achieve low values on long audios of read speech. While multilingual FastConformer reaches the best values overall, Wav2Vec XLSR fine-tuned on Italian performs better than the multilingual FastConformer fine-tuned on Italian (see Table 3).

Overall, CER is relatively low on long read speech, which is intuitively understandable, considering the selected models have all been trained mainly on read speech (Mozilla Common Voice data and audio books). Poor performance on short audios was also expected, since all the tested models where pre-trained on longer audio chunks, ranging from 20 to 30 seconds [27] [21] [7]. Given the similar average length of the audio inputs, it is surprising that every model performs better on short spontaneous speech than on short read speech.

The relatively low CER values suggest promising potential, particularly for the multilingual models. Therefore, we decided to get more phonetically informative outputs to evaluate how well these models generalize beyond word boundaries and language-specific spelling conventions. We select `wav2vec2-xlsr-53-espeak-cv-ft`, a Wav2Vec 2.0 XLSR model fine-tuned on multilingual Common Voice dataset to recognize phonetic labels [28]. While using the exact same architecture as Wav2Vec2, Wav2Vec2Phoneme maps phonemes of the training languages to the target language using articulatory features [8]. Since the model outputs a string of tab-separated phonetic labels, we computed the CER metric only. As a reference, we used the story *Sa tramuntana e su soli* which was phonemically and phonetically transcription provided by Mereu [12]. The input file is a single 43-second audio of a young female native speaker of Campidanese Sardinian. When comparing the Wav2VecPhoneme predictions with the human phonemic transcription we get a Phoneme Error Rate (PER) of 0.28, while when comparing it with the phonetic human transcription, PER decreases to 0.23. This results suggest that an automatic transcription into phonemes rather than characters would be a path worth exploring, allowing a systematic description of the phonetics and phonology of endangered spoken languages, while bypassing the orthography issue. These results align with recent work on cross-lingual transfer [29] proposing a very similar solution to develop a multilingual phoneme recognizer.

4. Exploratory Linguistic Analysis

In this section, we present an exploratory linguistic analysis to evaluate to what extent the orthographic transcriptions from the tested ASR models capture the phonetic events present in the speech signal. The analysis is based on the inventory of phonological phenomena described for Campidanese Sardinian spoken in Cagliari [12].

In multilingual FastConformer's predictions some known phonological processes of Campidanese can be recognized. For instance, in Campidanese Sardininan the alveolar tap [ɾ] is an allophone of /r/ in word-medial intervocalic position and a sociophonetic variant of /t/ and /d/ in the Cagliari variety [12]. In examples 1 and 4, the intervocalic /t/ across word boundaries (*si lui* and *ma lui*) is transcribed as /l/, which can be considered a good orthographic approximation to an alveolar tap. Following a process of lenition of voiceless plosives and fricatives, the intervocalic labiodental fricatives /f/ across word boundaries are also consistently transcribed as their voiced counterpart /v/, see example 1 *asivato*, example 4 *con savorza* and *deno vusti*. Voiceless plosives /p/, /t/, and /k/ in word-medial intervocalic positions are expected to be realized with a long duration, in the predictions are recognized as geminate sounds, see example 5 in *deppidi* and *mascetti*, yet not always, see example 1 *depidi*. We also notice the insertion of paragogic vowels, which in Campidanese are inserted after a final consonant to avoid consonant in word-final coda position [12], as in example 1 *depidi* and *zinotenesi* or *a rosasa* in example 3. Except for *esaminat* in exemple 1 where it was expected and actually produced in the audio.

Although this model seems to propose an orthographic transcription close enough to the phonetic one, it sometimes makes systematic choices that are unfaithful to the acoustic signal. We provide an example where /u/ both in word medial and final position is generally transcribed as /o/, not only when there is an Italian equivalent or phonetically close lexical item e.g, antunietta>*antonietta*; coru>*coro*; su>*suo*; cun>*con*, but also when the item is unknown to the model ollastu>*ollasto*; dentradura>*dentradora*, giving reason to believe that the model might have information about the phonotactic constraints in Italian, e.g. no [u] in word final position.

1. esaminat si tui as fatu su percursu cumenti si depit [1]
   examina si lui asivato subercurso come zi depidi
2. e si non tenis atrus problemas in sa vida in foras [2]
   e zinotenesi a tus problema in savira in forez
3. sa vida no es stettia tuttu arrosas [3]
   savidano e stetti a dotto a rosasa
4. ma tui con sa forza de unu fusti di ollastu [4]
   ma lui con savorza deno vusti di ollasto
5. no si deppiti imperai ma sceti castiai [5]
   nosi deppidi imperai mascetti gastiai

Regarding Whisper large-v2, we notice in some cases a near-perfect Italian translation of the Sardinian input audios, see example 5 and 6 below; in others cases, a poorer Italian translation with the deletion of repetitions, as in 7. Surprisingly, in example 8 and 9 we see how the tentative translations (or identifications with the phonetically most similar lexical items in a known language) also happens to Portuguese. Similar behavior is observed in Whisper medium: tentative Italian and Portuguese translations, and hallucinations both in spontaneous and read short input audios.

5. esaminat si tui as fatu su percursu cumenti si depit
   esamina se lui ha fatto il suo percorso come si deve
6. e si non tenis atrus problemas in sa vida in foras
   se non ha altri problemi in vita in forza
7. chi est o de un annu o de duus annus eccetera eccetera chi depis chi depis [6]
   chi e di un anno o di due anni chi deve essere
8. in su mesi e friaxu si cumentzat a fai su casu [7]
   em cima das evriagens o segundo mes ate faz sucesso
9. sanguidda si cuat in mesu e su ludu [8]
   sanguidas igual em mesa sulado açuludo

Similarly to multilingual FastConformer, Wav2Vec XLSR accounts for many of the phonological phenomena of Campidanese. The voiceless plosives /k/ and /p/, lenited to voiced fricatives [ɣ] and [β] when found in intervocalic environment across word boundaries [12], are transcribed as /g/ and /v/ in *gusta vingiara* and *sugauli* in example 13. While in Wav2Vec model the alveolar tap [ɾ] is rendered as /r/ instead of /l/ see *sirui* in example 10.

10. esaminat si tui as fatu su percursu cumenti si depit
    einasidu sirui ha sivato su bercursu come zi deperi
11. e si non tenis atrus problemas in sa vida in foras
    esino tenesi atosproblema sainsavvira in forese
12. su boi est un animali de meda importantzia [9]
    su boe e un animale de meda importanza
13. su cauli coit mellus in custa pingiada [10]
    sugauli coi melusu in gusta vingiara
14. ma tui con sa forza de unu fusti di ollastu
    madoi con savorza de unovusti diolastu

Unlike Whisper large-v2, Wav2Vec XLSR never performs translations and, unlike the FastConformer fine-tuned on Italian, does not seem to respect the Italian phonotactic constraints, see *diolastu* in example 14.

5. Conclusions and Future steps

The preliminary analysis carried out in this paper provided insight into how various speech recognition models transcribe data in a Romance language not encountered in the model training. All evaluated models improve their performance as the audio length increases. Best CER values are achieved on audio of read speech longer than 20 seconds. However, short audios of spontaneous speech with an average length of 5.3 seconds achieved a remarkably low CER, meaning better precision compared to the similarly short (3.5 seconds) read speech chunks. These results suggest that speech style might also play a role. To investigate whether the models are sensitive to speech style, other linguistic, speaker-specific, or technical variables, such as the topic, age, gender of the speaker, or the acoustic quality of the audio data, should be taken into account. For example, both datasets of spontaneous speech are produced by males over 45, and models might be biased toward an adult male speaker profile. For the time being, we attribute it to the poor representativeness of the dataset and will investigate it in future work.

[1] [He/she] makes sure you have done the proper training.
[2] And if you have no other problems in your life in general.
[3] Life has not been all roses.
[4] Yet you, with the strength of a wild olive trunk.
[5] It is not to be used but only looked at.
[6] That it is either one or two years long, and so on and so forth – that it has to – that it has to
[7] February sees the start of cheese making.
[8] The eel hides in the mud.
[9] The ox is a very important animal.
[10] The cabbage cooks best in this pan.

A controlled yet diverse dataset facilitated a qualitative linguistic analysis of the predictions. Interestingly, some models seem to follow the phonotactic constraints of the languages they have been trained on, but at the same time they generalize well to unfamiliar languages, providing quite accurate phonetically-like orthographic transcription of Campidanese Sardinian. These initial considerations should be validated with tests on a larger corpus to eliminate data bias and a more systematic linguistic analysis to avoid cherry-picking. We also plan to look in detail at the speech recognition models' architectures in order to make a informed choice at the fine-tuning phase.

In conclusion, it seems that state-of-the-art transcription models, especially multilingual ones, produce a phonetically accurate orthographic transcription of Campidanese Sardinian and thus provide a promising basis for fine-tuning. Specifically, `Wav2Vec2 large XLSR-53` and `STT Multilingual FastConformer Hybrid` proved to be the best models according to the evaluation metrics and preliminary linguistic analysis. `STT Multilingual FastConformer Hybrid` was the best and most efficient in terms of computational resources, which makes it our first choice for further testing and fine-tuning. However, it is worth noticing, speech recognition systems with orthographic output can be costly in terms of human and computational resources, poorly informative for speech researchers and uninteresting to native speakers; whereas recent work on multilingual automatic phonemic recognition seems a viable alternative worth exploring for documenting endangered spoken languages.

Acknowledgments

Work funded by the New Perspectives on Diphthong Dynamics (DID) project #I83C22000390005.

We would like to extend our gratitude to Daniela Mereu for providing the essential data for this research and for her invaluable perspective. We also thank Loredana Schettino and Aleese Block for their support and helpful insights.

References

[1] A. Magueresse, V. Carles, E. Heetderks, Low-resource languages: A review of past work and future challenges, arXiv (2020). URL: https://arxiv.org/abs/2006.07264. arXiv:2006.07264.

[2] P. Joshi, S. Santy, A. Budhiraja, K. Bali, M. Choudhury, The state and fate of linguistic diversity and inclusion in the NLP world, CoRR abs/2004.09095 (2020). URL: https://arxiv.org/abs/2004.09095. arXiv:2004.09095.

[3] Z. Liu, C. Richardson, R. J. Hatcher, E. T. Prudhommeaux, Not always about you: Prioritizing community needs when developing endangered language technology, in: Annual Meeting of the Association for Computational Linguistics, 2022. URL: https://api.semanticscholar.org/CorpusID:248118721.

[4] Y. Liu, X. Yang, D. Qu, Exploration of whisper fine-tuning strategies for low-resource asr, EURASIP Journal on Audio, Speech, and Music Processing 2024 (2024) 29. URL: https://doi.org/10.1186/s13636-024-00349-3. doi:10.1186/s13636-024-00349-3.

[5] C. Sicard, K. Pyszkowski, V. Gillioz, Spaiche: Extending state-of-the-art asr models to swiss german dialects, in: Swiss Text Analytics Conference, 2023. URL: https://arxiv.org/abs/2304.11075. doi:10.48550/arXiv.2304.11075. arXiv:2304.11075.

[6] M. Bartelds, N. San, B. McDonnell, D. Jurafsky, M. B. Wieling, Making more of little data: Improving low-resource automatic speech recognition using data augmentation, in: Annual Meeting of the Association for Computational Linguistics, 2023. URL: https://api.semanticscholar.org/CorpusID:258762740. doi:10.48550/arXiv.2305.10951.

[7] A. Baevski, Y. Zhou, A. Mohamed, M. Auli, wav2vec 2.0: A framework for self-supervised learning of speech representations, Advances in neural information processing systems 33 (2020) 12449–12460. doi:10.48550/arXiv.2006.11477.

[8] Q. Xu, A. Baevski, M. Auli, Simple and effective zero-shot cross-lingual phoneme recognition, in: Interspeech, 2021. URL: https://arxiv.org/abs/2109.11680. doi:10.21437/interspeech.2022-60.

[9] A. Yeroyan, N. Karpov, Enabling asr for low-resource languages: A comprehensive dataset creation approach, arXiv preprint arXiv:2406.01446 (2024). URL: https://arxiv.org/abs/2406.01446. arXiv:2406.01446.

[10] V. Blaschke, C. Purschke, H. Schütze, B. Plank, What do dialect speakers want? a survey of attitudes towards language technology for german dialects, arXiv preprint arXiv:2402.11968 (2024). doi:10.48550/arXiv.2402.11968.

[11] G. Mensching, E.-M. Remberger, 270sardinian, in: The Oxford Guide to the Romance Languages, Oxford University Press, 2016, p. 270–291. URL: https://doi.org/10.1093/acprof:oso/9780199677108.003.0017. doi:10.1093/acprof:oso/9780199677108.003.0017.

[12] D. Mereu, Cagliari sardinian, Journal of the International Phonetic Association 50 (2020) 389–405. doi:10.1017/S0025100318000385.

[13] A. Oppo, Le lingue dei sardi. una ricerca sociolinguistica (2007).

[14] Ethnologue, Sardinian, campidanese, 2024. URL: https://www.ethnologue.com/language/sro/.

[15] R. Rattu, Repertorio Plurilingue e Variazione Linguistica a Cagliari: I Quartieri di Castello, Marina, Villanova, Stampace, Bonaria e Monte Urpinu, Master's thesis, Università degli Studi di Cagliari, 2017.

[16] B. F. Eduardo, C. Amos, C. Stefano, D. Nicola, M. Massimo, M. Michele, M. Francesco, M. Ivo, P. Pietro, P. Oreste, R. Antonella, S. Paola, S. Marco, Z. Paolo, Arrègulas po ortografia, fonètica, morfologia e fueddàriu de sa Norma Campidanesa de sa Lìngua Sarda, ALFA EDITRICE, 2009.

[17] D. Mereu, Efforts to standardise minority languages: The case of sardinian, Europäisches Journal für Minderheitenfragen. European Journal of Minority Studies (2021) 76–95. doi:10.35998/ejm-2021-0004.

[18] S. Gunsch, La distribuzione delle parti del discorso nel parlato e nello scritto campidanese e fenomeni del parlato in una lingua minoritaria di contatto, Master's thesis, Free University of Bozen-Bolzano, 2022.

[19] D. Mereu, Il sardo parlato a Cagliari: una ricerca sociofonetica., FrancoAngeli., Milano, 2019.

[20] V. Srivastav, S. Majumdar, N. Koluguri, A. Moumen, S. Gandhi, et al., Open automatic speech recognition leaderboard, https://huggingface.co/spaces/hf-audio/open_asr_leaderboard, 2023.

[21] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, I. Sutskever, Robust speech recognition via large-scale weak supervision, in: International conference on machine learning, PMLR, 2023, pp. 28492–28518. doi:10.48550/arXiv.2212.04356.

[22] NVIDIA, Stt multilingual fastconformer hybrid large pc, 2023. URL: https://catalog.ngc.nvidia.com/orgs/nvidia/teams/nemo/models/stt_multilingual_fastconformer_hybrid_large_pc.

[23] NVIDIA, Stt it fastconformer hybrid large pc, 2023. URL: https://catalog.ngc.nvidia.com/orgs/nvidia/teams/nemo/models/stt_it_fastconformer_hybrid_large_pc.

[24] H. Face, Xls-r wav2vec2 model documentation, 2024. URL: https://huggingface.co/docs/transformers/en/model_doc/xlsr_wav2vec2.

[25] H. Face, wav2vec2-large-xlsr-53-italian, 2021. URL: https://huggingface.co/facebook/wav2vec2-large-xlsr-53-italian.

[26] H. Face, Evaluate: A library for evaluation in machine learning, 2024. URL: https://github.com/huggingface/evaluate.

[27] D. Rekesh, S. Kriman, S. Majumdar, V. Noroozi, H. Juang, O. Hrinchuk, A. Kumar, B. Ginsburg, Fast conformer with linearly scalable attention for efficient speech recognition, 2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU) (2023) 1–8. URL: https://api.semanticscholar.org/CorpusID:258564901.

[28] H. Face, wav2vec2-xlsr-53-espeak-cv-ft, 2021. URL: https://huggingface.co/facebook/wav2vec2-xlsr-53-espeak-cv-ft.

[29] K. Glocker, A. Herygers, M. Georges, Allophant: Cross-lingual phoneme recognition with articulatory attributes, in: Proceedings of Interspeech, 2023. URL: http://dx.doi.org/10.21437/Interspeech.2023-772. doi:10.21437/interspeech.2023-772.