applied sciences

MDPI

Article

Improving Aphasic Speech Recognition by Using Novel Semi-Supervised Learning Methods on AphasiaBank for English and Spanish

Iván G. Torre *, Mónica Romero and Aitor Álvarez *

Vicomtech Foundation, Basque Research and Technology Alliance (BRTA), Mikeletegi 57, 20009 Donostia, Spain; mromero@vicomtech.org

* Correspondence: igonzalez@vicomtech.org (I.G.T.); aalvarez@vicomtech.org (A.Á.)

Abstract: Automatic speech recognition in patients with aphasia is a challenging task for which studies have been published in a few languages. Reasonably, the systems reported in the literature within this field show significantly lower performance than those focused on transcribing non-pathological clean speech. It is mainly due to the difficulty of recognizing a more unintelligible voice, as well as due to the scarcity of annotated aphasic data. This work is mainly focused on applying novel semi-supervised learning methods to the AphasiaBank dataset in order to deal with these two major issues, reporting improvements for the English language and providing the first benchmark for the Spanish language for which less than one hour of transcribed aphasic speech was used for training. In addition, the influence of reinforcing the training and decoding processes with out-of-domain acoustic and text data is described by using different strategies and configurations to fine-tune the hyperparameters and the final recognition systems. The interesting results obtained encourage extending this technological approach to other languages and scenarios where the scarcity of annotated data to train recognition models is a challenging reality.

Keywords: aphasia; speech recognition; wav2vec2.0; semi-supervised learning; aphasiabank; low-resource

Citation: Torre, I.G.; Romero, M.; Álvarez, A. Improving Aphasic Speech Recognition through Novel Semi-Supervised Learning Methods on AphasiaBank for English and Spanish. *Appl. Sci.* **2021**, *11*, 8872. https://doi.org/10.3390/app11198872

Academic Editors: Inma Hernaez Rioja, José A. González-López and Heidi Christensen

Received: 31 July 2021
Accepted: 20 September 2021
Published: 24 September 2021

1. Introduction

Aphasia is a language disorder that causes impairments in dimensions including speech, writing, interaction or communication. People with aphasia (PWA) mainly acquire this disorder after suffering a stroke, a traumatic brain injury, a tumoral brain or any other affection in some specific areas of the brain that are related to language. Particularly, aphasia is more likely to be developed when the affected areas are located in the left hemisphere [1]. Every year, millions of people worldwide acquire aphasia through one of these issues and its prevalence on the full population ranges between 6 and 62 people per 100.000 inhabitants depending on the region and country [2–4]. These values may increase even up to 30–60% in people who have survived a stroke, which is the second cause of death globally [4–6].

PWA may acquire communication impairments that affect their daily life in different grades depending on the severity of the disorder [7]. Usually, these impediments are classified with the scale proposed by the Western Aphasia Battery (WBA) [8] ranging from mild to very severe depending on the performance on several tasks that include reading, speech or writing, among others [8]. On the other hand, aphasia disorders can also be distinguished by a combination of symptoms and the affected physical areas [7]. The most extended classification uses the Wernicke–Lichtheim model, which associates communication capabilities with different brain regions [9,10], differentiating three main types of aphasia depending on the area damaged: Broca, Wernick and Anomic. Nevertheless, language comprehension and production are not isolated at the specific brain areas

considered by this model [11], and more modern and complete theories, e.g., dual stream model [12] consider that language capabilities are organized in a distributed system in different cortical regions, emphasizing the connections between them [13–15]. However, cortical damages that causes aphasic impairment have barely been mapped using these new theories; therefore, the Wernicke–Lichtheim model is still the most widely used method in clinical assessment [11].

Intensive speech therapy conducted by interdisciplinary groups of clinical experts has a fundamental role in recovering the communication abilities of PWA [16]. During the last years, intense research carried out in speech recognition technology promises to support the work of these clinical experts by automating processes and improving access to therapy related to isolated areas and/or less favored socioeconomic environments and collectives. In this sense, some applications such as *Constant Therapy* [17], *Lingraphica* [18] and *Tactus Therapy* [19], for which their usefulness has been recognized by the National Aphasia Association of United States (accessed on 15 July 2021) https://www.aphasia.org/, provide exercises to practice speech, language and cognitive tasks by customizing the PWA progress. These applications have been proven to reinforce the therapy, achieving marked goals in less time [20], especially in rural areas [21]. Other technological applications focus on the adaptation of standard cognitive tests [22] or on the automatic quantitative analysis of aphasia severity through speech [23]. Taken together, these new techniques and solutions promise to enhance face-to-face therapy, to extend the treatment to more patients and, therefore, to improve the quality of life of PWA.

Nonetheless, there are still challenges related to automatic speech recognition (ASR) that must be solved worldwide in order to extend these therapy applications, since they basically depend on adequate engines that should properly recognize aphasic speech. ASR systems are usually trained with the voices of people without any speech pathology, and their performance degrades when they are applied to aphasic speech [23–27]. Furthermore, ASR systems are usually language-dependent and have to be trained with hundreds or thousands of hours of transcribed speech. This idiosyncrasy avoids, in many cases, extending their use to the thousands of languages currently spoken in the world and, particularly, to the use case of aphasic speech recognition due to the lack of so many annotated data for training recognition models following the more traditional supervised learning methods.

In this work, we explore the application of novel semi-supervised end-to-end (E2E) learning methods on ASR to perform aphasic speech recognition in English and Spanish in a very challenging scenario with few annotated data. More specifically, we make use of the *wav2vec2.0* architecture [28], building models adapted to aphasic speech for English and Spanish and comparing the results with previous fully supervised technological approaches presented in the literature. In particular, we achieved a relative error reduction in Word Error Rate (WER) for the English test set by ∼25% when comparing with previous published results. In addition, we demonstrate that this technological approach can be extended to perform aphasic speech recognition with few annotated data. To this end, we built the first Spanish E2E model adapted to aphasic speech recognition with less than one hour of data from PWA and report the first results in the literature for this language and domain.

The rest of the paper is organized as follows: Section 2 introduces previous work in aphasic speech recognition. Section 3 details the process performed over the main corpora used for the experiments in addition to the creation and compositions of the train, validation and test partitions. In Section 4, the speech recognition architectures and constructions are explained, whilst the evaluation results obtained over different configurations of the systems are presented in Section 5 for English and Spanish. Finally, Section 6 concludes the paper and presents future work.

2. Related Work in Aphasic Speech Recognition

ASR is a technological field that has remarkably evolved over the last years from the hand of new methods and architectures based on Deep Neural Networks (DNNs), which are closer to reaching human-like performance in controlled acoustic environments [28–32]. These improvements have great potential to impact new ASR clinical applications and to develop new e-health solutions [33–35]. Particularly, ASR technology applied to disordered voices brings the opportunity to implement new assisted and personalized therapies, generate automatic cognitive tests or to develop adapted applications for people with impairments.

The first ASR systems for aphasic speech recognition found in the literature were focused on recognizing isolated words within small vocabularies for English [36] and Portuguese [24]. More recently, thanks to the advancements in deep learning speech recognition technologies, new studies achieved up to 90% accuracy on assessing *correct* versus *incorrect* naming attempts in controlled utterance verification systems [37]. However, the biggest challenge in the field nowadays is to improve the performance of the continuous recognition of aphasic speech in large vocabularies. To the best of our knowledge, the published works in the task of aphasic continuous speech recognition of large vocabularies only consider English [23,38,39] and Cantonese [40] to date. In this sense, the performance and results for these systems widely oscillate depending on the severity level of aphasia, ranging WER from 33 on mildest cases to more than 60 on very severe cases. All these studies employ the same AphasiaBank database [41] as the main corpus for training and evaluation, but they usually differ on the train-test-validation partitions and on the evaluation metrics employed, given that some studies used the Phoneme Error Rate (PER) as its main metric and others employed the Character Error Rate (CER). This decision strongly depends on the configuration and the basic modeling unit used to train their systems (phonemes or characters). Hence, a fair and balanced comparison between systems and technological approaches cannot always be guaranteed. Nonetheless, in some cases, notable improvements can be appreciated between the 52.3 of PER in moderate aphasia test group presented in [25] and the more recent 41.7 of PER reported in [39]. These results seems to be in line with the 38.3 global Syllable Error Rate (SER) reported for the full test set in Cantonese [40], where more than 60% of the test set was composed of mild severity speech data.

Regarding technological approaches, previous works focused on developing ASR technology for aphasic speech considering architectures based on hybrid Acoustic Models (AMs) such as Deep Neural Networks and Hidden Markov Models (DNN-HMM) [25], Bidirectional Long Short-Term Memory and Recurrent Neural Models (BLSTM-RNN) [23], and solutions based on Mixture of Experts (MoEs) [39]. More specifically, in the work presented in [38], the authors established the first large-vocabulary continuous speech recognition baseline for English built on the AphasiaBank dataset using a DNN-HMM hybrid AM trained on unseen train-validation-test partitions and by distinguishing performances depending on aphasia severity. They reached PER metrics between 47.41 for mild severity test and 75.81 for very severe test set and reported that appending utterance fixed-length speaker identity vectors (i-vectors) to frame-level acoustic features resulted in PER reductions specially in speakers with more severe levels of aphasia. These results were then improved by using an acoustic modeling method based on a BLSTM-RNN architecture enriched with a trigram language model (LM) estimated on the transcripts of the training audios [23]. In this case, the training of the AM was reinforced with transcribed data from healthy speakers, achieving an improved WER ranging from 33.68 on mild test set to 53.17 on very severe test set. In the work described in [39], an AM based on a MoE of DNN models was proposed, where each expert in the model was specialized on specific aphasia severity. Additionally, an Speech Intelligibility Detector (SID) composed of two hidden layers and a final softmax function was trained to detect the Aphasia Quotient (AQ) severity level of a given speech frame by using the acoustic features and utterance-level speaker embeddings. At inference time, the contribution of each expert was decided by the

SID module. Once again, the train-validation-test partitions were randomly generated, and they achieved PER values ranging from 33.37 on mild test set to 61.41 on severe test set.

Finally, the first ASR system for Cantonese continuous aphasic speech was described in the work presented in [40]. They used a Time Delay Neural Network (TDNN) combined with a BLSTM model as the main AM, which was trained with both in-domain and out-of-domain speech data and a syllable-based trigram LM. The performance of the system was evaluated at the syllable level by using the SER metric. In this work, any distinctions between aphasia severities, yielding an overall SER of 38.77 for aphasic speech and 15.07 of SER for the healthy speakers, were not reported.

As it can be concluded, over the last years, the speech recognition of aphasic voices has benefited from the latest improvements in the ASR based on fully supervised learning methods, gradually enhancing its performance and, thus, allowing its application in real clinical and therapists tools. In this work, we show that semi-supervised learning methods have great potential in this particular domain, reporting interesting WER improvements for English and competitive results for Spanish considering the scarcity of annotated PWA data (less than 1 h) for this language.

3. AphasiaBank Dataset Description and Processing

3.1. General Description

In this work, transcribed speech data from the AphasiaBank dataset [41] were used as the main corpus. The AphasiaBank corresponded to a computerised database of interviews between PWA and clinicians. The interviews are presented in recorded video format, and they were transcribed and transformed into CHAT file format following a protocol designed by a table of experts based on previous successful experiences [42]. This protocol mainly consisted of narrative and procedural discourse in order to maximize task comparability across participants [41].

The contents in the original AphasiaBank dataset are organized by the severity of the aphasia impairment for the English language. This measurement was performed with the standardized comprehensive assessment by using the WAB scale and yielding an AQ value which ranged from 0 to 100. Lower AQ value meant a higher degree of aphasia severity. The AQ score served as a threshold to classify patients into four aphasic levels, including mild ($AQ < 75$), moderate ($50 < AQ \leq 75$), severe ($25 < AQ \leq 50$) and very severe ($0 < AQ \leq 25$) [41].

Regarding the amount of data, at the time the authors accessed the database, the full English subpart of the AphasiaBank dataset included 116 h and 54.9 h of transcribed speech from 435 PWA and healthy control speakers, respectively, collected at various sites across the United States and Canada [41]. The PWA speakers were organized by their severity of the aphasia impairment. By contrast, for the case of Spanish, the available data only included chunks from 4 PWA collected at four different sites across the United States, summarizing a total of 1.2 h of transcribed speech [41]. In this case, with the aim of adding contents from healthy people, 1 h (700 speech utterances) from the Spanish Mozilla Common Voice corpus [43] was selected in order to reinforce the training of the Spanish AM. It should be noted that no information about the aphasia severity of the Spanish PWA patients was reported in the original database.

3.2. Data Processing

The original data from the AphasiaBank dataset were processed at different acoustic and text levels in order to generate suitable corpora to build the E2E AMs for English and Spanish. The audio files were first extracted from the video recordings and converted to PCM WAV 16 kHz 16-bit format using the open sourced FFmpeg tool [44]. Since the time-codes were provided at the sentence level, the audio was split into correctly aligned audio chunks by using the SoX [45] tookit in order to manage shorter segments for the training of the neural models. In this respect, audio chunks shorter than 0.3 s were discarded to avoid future problems when computing Fourier transform for the spectrograms generation

or during the CTC layer alignment in the neural network. Furthermore, audio chunks longer than 30 s were not included in our corpus, with the aim of avoiding memory issues during training.

Concerning text transcriptions, they originally contained enriched information including not only literally transcribed words and phenomena such as repetitions, sound fragments and phonological transcription but also artifacts such as misalignments or phoneme omissions. In the latter cases, different criteria were applied in order to maintain or definitively discard these phenomena. In the cases where some phonemes were missing but the full word was intelligible, we chose to maintain the entire word, although some of its phonemes may not have been properly pronounced. Moreover, the repetitions of words and semantic mismatches that may occur during the speech were also preserved, since replacing them would not reflect the real speech patterns of the PWA collective. Additionally, it should be remarked that transcriptions also included special symbols representing isolated noises interjections or fillers, including *um*, *uh*, *uhuh* or *huh*, among others. These symbols included *(FLR)* to represent fillers; *(SPN)* for spoken noises; *(BRTH)* as breathing sounds; and *(LAU)* for laughter. These special symbols were included for training and considered as individual words and characters in the acoustic E2E model. Moreover, contents with empty or mismatched transcriptions were discarded. We illustrated in Table 1 this methodology showing a real example that includes the original and processed transcription from an audio chunk performed by a female moderate non-fluent Broca English speaker.

Table 1. Example of original and processed transcription from an audio chunk performed by a female moderate non-fluent Broca English speaker.

Original transcript	&=sighs very ʃɑrmIŋ@u [:charming] [*p:n] [//] &-uh Cinderella ɑrmIŋ@u [:charming] [*p:n] &-uh
Processed transcript	F B very charming F cinderella charming F

Once the cleaning up process was performed, the English corpus included 89.9 h of PWA patients and 51.3 h of healthy controls, whilst the Spanish corpus summed up a total of 1.2 h of PWA speakers and 1 h of healthy controls.

Since standard partitions for train, validation and test are not provided in the original AphasiaBank dataset, we applied the following criteria to split the processed data.

For the English corpus, we randomly selected 25% of PWA speakers from each severity level for the test partition, 19% of PWA speakers for the validation test set and the remaining 56% for the training set to create an unseen train/test/validation set. This train partition was called *PWA acoustic set*. In addition, we also created a second training set, which we called *Mixed acoustic set*, by adding data from healthy controls. The configuration of the train/test/validation partitions was mainly thought so that speakers cannot appear simultaneously in more than one subset while the data remained balanced throughout the aphasia severities. Moreover, both validation and test sets were composed only with data from PWA. In this manner, we could compare two different train sets to investigate the usefulness of adding healthy control data in order to improve the performance of the ASR model. Detailed information of the constructed English corpus can be found in Table 2, including the number of subjects, the amount of hours per partition and the levels of aphasia considered.

3.3. Experimental Setup

Given that the original Spanish corpus from the AphasiaBank dataset was composed by only 4 PWA participants without information about their aphasia severity level, a different configuration was followed for this language but maintaining the same partition percentages. In this case, 56%, 19% and 25% of the audio chunks were randomly selected from each PWA speaker to form the train, validation and test set, respectively. As in the case of the English language, two train sets were also created for Spanish; the *PWA acoustic*

set including only PWA data for training and the *Mixed acoustic set*, which added data from healthy controls. In the case of the *PWA acoustic set*, its configuration allow the authors to explore the ability to train an ASR system with an extremely small number of data using semi-supervised learning methods. Detailed information for each Spanish partition including the total number of speakers and hours is summarized in Table 3.

Table 2. Train, validation and test partitions of the English corpus.

English	Train		Validation		Test		Total	
	Speakers	**Hours**	**Speakers**	**Hours**	**Speakers**	**Hours**	**Speakers**	**Hours**
Mild	105	26	36	9.9	48	12.8	189	48.7
Moderate	79	18.7	27	6.1	36	8	142	32.8
Severe	22	3.9	8	1.3	10	1.6	40	6.8
Very severe	8	0.5	3	0.4	4	0.7	15	1.6
Total PWA	214	49.1	74	17.7	98	23.1	386	89.9
Controls	277	51.3	-	-	-	-	277	51.3
Total (PWA + Controls)	491	100.4	74	17.7	98	23.1	663	141.2

Table 3. Train, validation and test partitions of the Spanish corpus.

Spanish	Train		Validation		Test		Total	
	Speakers	**Hours**	**Speakers**	**Hours**	**Speakers**	**Hours**	**Speakers**	**Hours**
Total PWA	4	0.69	4	0.23	4	0.28	4	1.2
Controls	415	1	-	-	-	-	415	1
Total (PWA + Controls)	419	1.69	4	0.23	4	0.28	419	2.2

4. Semi-Supervised Learning Based System

In this section, the ASR architecture based on semi-supervised learning techniques used during this research is described, providing details on the strategies employed to find the best hyperparameters and the fine-tuning techniques implemented. Finally, the two decoding strategies used to generate the recognition hypothesis are described as well.

4.1. Main Architecture

The main ASR architecture used in this work is based on the unsupervised E2E model *wav*2*vec*2.0 proposed by Facebook AI [28], which is schematically represented in Figure 1. The *wav*2*vec*2.0 model maps speech audio through a multi-layer convolutional feature encoder $f : \chi \rightarrow Z$ to latent speech representations $z_1, ...z_T$, which are fed into a Transformer network $g : Z \rightarrow C$ to output context representations $c_1, ...c_T$. These context representations are then quantized to $q_1...q_T$ in order to represent the targets in the self-supervised learning objective [28,46]. The feature encoder contains seven blocks, and the temporal convolutions in each block include 512 channels with strides $(5, 2, 2, 2, 2, 2, 2)$ and kernel widths $(10, 3, 3, 3, 3, 2, 2)$. The transformer used had 24 blocks, a model dimension of 1024, an inner dimension of 4096 and a total of 16 attention heads. The model was pretrained by solving a contrastive task over masked feature encoder outputs. Afterwards, it was fine-tuned relative to the aphasia domain by adding a randomly initialized linear projection on top of the context network into C classes representing the vocabulary of the task [47] and optimized by using a Connectionist Temporal Classification (CTC) layer [28,46,48].

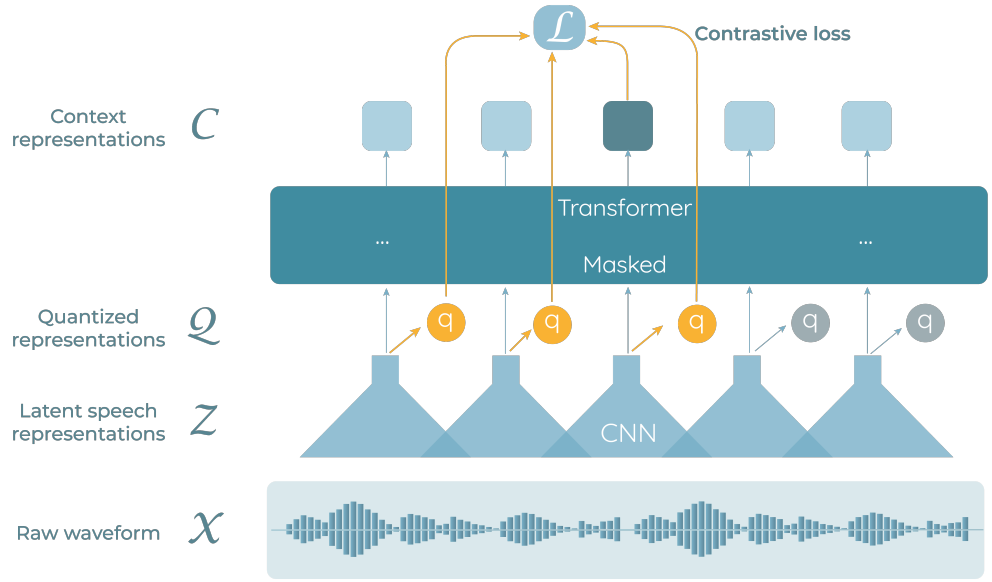

Figure 1. Main architecture of the ASR model based on the wav2vec2.0 representation. Raw waveform is mapped to speech representations that are fed into a transformer network to output context representations. Context representations are then quantized to represent targets in the self-supervised task.

The pretrained task was based on the *XLSR-53* [46] model, which was originally trained with 56,000 h of nontranscribed speech data in 53 different languages, including English and Spanish. These data were composed of audio from the CommonVoice [43], Babel [49] and Multilingual Librispeech (MLS) [50] datasets. The unsupervised task learns a set of quantized latent speech representations shared across languages that are later combined together on the supervised training to identify the phonemes or characters to decode. The speech audio representations are learned by solving a contrastive task, which requires identifying the true quantized latent speech representation for a masked time step within a set of distractors [28]. This strategy has been shown to be capable of learning non-language-dependent universal quantized representations of speech that can then be combined to train specific phonemes and sounds of each language [46].

4.2. Supervised Fine-Tuning Phase

The fine-tuning phase of the pre-trained *XLSR-53* model corresponded to the supervised training where quantized representations of speech are mapped into the output vocabulary by using Connectionist Temporal Classification (CTC) loss [48]. The last layer corresponded to the vocabulary set, and it was composed of 35 characters for the case of English and 38 for the case of the Spanish language.

In a first step, we performed grid search hyperparameter tuning on the validation set, training models with small subsets of the train partition by using the Weights & Biases tools [51]. Using this information, we set the learning rate to 2×10^{-5} using a warm-up during the first 10% of updates and then using a linear decay learning rate scheduler. Additionally, the feature and layer dropouts were set to 0.05 and 0.02, respectively, whilst the accumulation steps was set to 3, the mask time to 0.057 and the activation and attentions dropouts were established as 0.03 and 0.036, respectively.

In addition, we also applied a masking strategy to the feature encoder outputs similar to the *SpecAugment* technique presented in [30], and mask embeddings were randomly applied, as explained in [28]. Previous research studies reported weight update optimal values between 16 k and 300 k during training, depending on training corpus size, training batch-size and number of GPU (Graphics Processing Unit) cards employed [46]. Following

these recommendations and considering our hardware resources, we used a batch size of 6 during training and performed finetuning during 10 epochs on English (∼21 k updates on the *PWA acoustic set* and ∼50 k on the *Mixed acoustic set*). For the Spanish dataset, our best results were achieved by finetuning the model during 100 epochs when using the *PWA acoustic set* (∼2 k updates) and 200 epochs when using the *Mixed acoustic set* (∼ 13 k updates).

4.3. Decoding Strategies and External LMs

Two different decoding strategies were applied during the experiments for both languages. The first decoding strategy was based on a greedy-search approximation, which selected the most likely character at each step in the output sequence. Although this approach had the benefit of being very fast, its performance strongly depends on the robustness of the E2E AM and the quality of the final output sequences may not be the most optimal.

As the second decoding strategy, a beam-search approximation was applied by using external LMs for rescoring the initial hypothesis of the E2E AM. Different external LMs were built and constructed for the experiments. For the case of the English language, three LMs were trained and tested: (i) a model trained only with the transcriptions of the audio of the *PWA acoustic set* called *In-domain LM*; (ii) a second LM model using the transcriptions of the audio from the *Mixed acoustic set* called *Mixed LM*, which mixed audio of the *PWA acoustic set* and healthy controls; and (iii) a final large LM model, called *Large LM*, which includes the transcriptions of the above two acoustic sets plus texts from the Librispeech [52] and CommonVoice [43] public datasets. Each LM was trained with 250 k words, 600 K words and 813.2 million words, respectively. With respect to Spanish language, given the low amount of texts from the *PWA acoustic set* and *Mixed acoustic set*, only one external LM was trained, including the transcriptions of the audios from the *PWA acoustic set* and *Mixed acoustic set*, in addition to texts extracted from the public CommonVoice dataset (1.8 million words) and generic news extracted from Spanish digital newspapers (25.2 million words). The model was identified as *Large LM*. In total, the Spanish text corpus contained 27.1 million words.

The LMs were built through the KenLM toolkit [53] in which modified Kneser–Ney smoothed 3-gram models were estimated. Beam-search decoding was performed with a beam-width value of 10 in all experiments, whilst the LM weight parameters *alpha* and the insertion weight *beta* were tuned with the validation dataset for each language. In this manner, for English, an *alpha* value of 0.8 and a *beta* value of 0 were used on*In-domain LM* and *Mixed LM*, while *alpha* value of 1.4 and a *beta* value of 0 were used on the *Large LM*, whilst the *alpha* and *beta* parameters for Spanish were set to 1.4 and 0, respectively.

5. Evaluation Results and Discussion

In this section, the evaluation results for English and Spanish are reported, together with the results obtained by the reference ASR systems of the literature, which are shown in Table 4. All the evaluations were performed following the experimental setup, neural acoustic and language models and decoding strategies detailed in Sections 3 and 4. In addition, a discussion of the results achieved is provided as well.

Table 4. Reference ASR baselines' performance on the AphasiaBank English dataset. The results are organized by the AM used in the ASR system and the aphasia severity.

PER (Phoneme Error Rate)				
AM	**Mild**	**Moderate**	**Severe**	**Very Severe**
DNN-HMM [38]	47.41	52.79	61.00	75.81
MoE-DNN and SID [39]	33.37	41.69	61.41	-
WER (Word Error Rate)				
AM	**Mild**	**Moderate**	**Severe**	**Very Severe**
BLSTM-RNN [23]	33.68	41.11	49.21	63.17

5.1. Semi-Supervised ASR Performance for English

The performances of the different ASR systems developed in this work for aphasic speech recognition in English are reported in Tables 5 and 6 for the CER and WER metrics, respectively. The results are organized by the AM of the ASR system, the acoustic data used to finetune the pre-trained *XLSR-53-wav2vec2.0* model, the decoding type, the external LM used for rescoring the initial lattices and the aphasia severity level.

Table 5. CER results on the English corpus of AphasiaBank detailed by severity level of aphasia: mild, moderate, severe and very severe. The *PWA acoustic set* is only composed by PWA patients, and *Mixed acoustic set* combines PWA and healthy controls. *In-domain LM* was trained by using transcriptions from the *PWA acoustic set*, *Mixed LM* was trained with the transcriptions from the audio of the *Mixed acoustic set* and the *Large LM* by using the transcriptions from the above acoustic sets and texts from Librispeech and CommonVoice datasets.

CER (Character Error Rate)							
AM	**Acoustic Data**	**Decoding**	**LM**	**Mild**	**Moderate**	**Severe**	**Very Severe**
XLSR-53-wav2vec2.0	PWA acoustic set	Greedy	-	14.1	**23.0**	23.5	49.0
XLSR-53-wav2vec2.0	Mixed acoustic set	Greedy	-	**13.4**	23.5	**22.5**	46.7
XLSR-53-wav2vec2.0	Mixed acoustic set	Beam	In-domain	14.1	24.3	23.0	**46.0**
XLSR-53-wav2vec2.0	Mixed acoustic set	Beam	Mixed	14.6	24.4	23.4	46.2
XLSR-53-wav2vec2.0	Mixed acoustic set	Beam	Large	17.4	27.4	23.3	47.2

Table 6. WER results on English corpus of AphasiaBank detailed by severity level of aphasia: mild, moderate, severe and very severe.

WER (Word Error Rate)							
AM	**Acoustic Data**	**Decoding**	**LM**	**Mild**	**Moderate**	**Severe**	**Very Severe**
XLSR-53-wav2vec2.0	PWA acoustic set	Greedy	-	25.1	36.2	39.0	62.5
XLSR-53-wav2vec2.0	Mixed acoustic set	Greedy	-	23.6	36.8	36.4	59.1
XLSR-53-wav2vec2.0	Mixed acoustic set	Beam	In-domain	23.2	35.2	35.2	55.8
XLSR-53-wav2vec2.0	Mixed acoustic set	Beam	Mixed	**22.3**	**35.1**	**34.1**	**55.5**
XLSR-53-wav2vec2.0	Mixed acoustic set	Beam	Large	26.9	39.4	34.1	62.0

As it was expected, audio contents from more severe levels of PWA are more challenging to transcribe, whilst the speech segments from mild severity cases are recognized with lower error rates on CER and WER values. The differences between the performance in the different groups that establish the degree of the aphasia severity are quite significant, obtaining up to $2x$ error on the most severe groups when comparing with mild cases. These

big differences between AQ level groups are in line with previous publications [23,38,39], which PER and WER results are summarized in Table 4.

At acoustic levels, the best performance was obtained when finetuning the *XLSR-53* pre-trained model with data from the *Mixed acoustic set*, which included audio content from PWA and healthy controls. In this sense, we report CER and WER reductions of almost a ∼5% when adding the healthy controls in comparison with using only audios of PWA for training. It implies that the impact of the scarcity of annotated aphasic speech can be partially reduced by incorporating speech from healthy speakers and domains. This finding was explored and applied later on the Spanish dataset.

Regarding the beam-search decoding using external LMs for rescoring the initial lattices, it was demonstrated that this strategy clearly improves the performance of the speech recognition systems, showing different results depending on the level of severity of aphasia and the type of LM employed. At this point, it is worth remarking that the *Large LM* does not enhance overall results when comparing with the other LMs, even if it includes more than 803 million extra words, and the special symbols were ignored in order to compute metrics. It suggests that, in this case, the texts from the Librispeech and CommonVoice datasets used for training the LM are too far from the domain sentences of the AphasiaBank dataset. In this manner, the best results are achieved using the *Mixed LM* model, reaching a 22.3 WER on the mild severity level group, a 35.1 WER over the moderate subset, a 34.1 WER for severe PWA and 55.5 WER on very severe cases. Overall, this LM reported improvements of ∼2% in comparison with using the *In-domain LM* and ∼7% when comparing to greedy decoding.

The results obtained show that, despite the great differences in the quality of pronunciation in speakers from mild to very severe groups, the semi-supervised learning method applied in this work is able to generalize the learning of contextualized speech representations of a very diverse type of speech, improving the ASR performance for all cases. This strategy is again demonstrated in Section 5.2 for the Spanish language. Finally, although a fair and well-balanced comparison of these results cannot be fully established with the ones published in previous studies (see Table 4) considering the differences in the modeling units (character versus phoneme) and the possible mismatch in data partitions, the results provided in this work for the English language (Tables 5 and 6) constitute a significant improvement in the quality of aphasic speech recognition systems tested to date on the AphasiaBank dataset.

5.2. Semi-Supervised ASR Performance for Spanish

The evaluation results achieved for the Spanish language are summarized in Table 7 at CER and WER levels. Firstly, it is worth noting that, even when we used less than one hour of PWA transcribed speech, we were able to achieve performances of 25.8 of CER and 49.8 of WER on the test set using the most simple greedy search decoding. These results were further improved by integrating audio from healthy control speakers and the *Large LM* trained with million of words to rescore and enhance the initial recognition hypothesis. If we consider the challenge of the task and the previous benchmarks of English and Cantonese ASR systems, which were trained with up to 50x more hours of transcribed speech, these results can be considered very competitive and promising. Moreover, these results are, to the best of our knowledge, the first benchmark of aphasic speech recognition published for Spanish.

Table 7. CER and WER metrics on Spanish test set of AphasiaBank where there was no clinical information on the severity of aphasia of participants. The *Mixed acoustic set* combines data from PWA and one hour of clean speech from the CommonVoice dataset. The *Large LM* was trained with texts from the Common Voice dataset and digital news from the generic domain. They do not include special symbols.

AM	Acoustic Data	Decoding	LM	CER	WER
XLSR-53-wav2vec2.0	PWA acoustic set	Greedy	-	25.8	49.8
XLSR-53-wav2vec2.0	Mixed acoustic set	Greedy	-	**24.1**	45.3
XLSR-53-wav2vec2.0	Mixed acoustic set	Beam	Large	24.8	**42.8**

The best initial results with the Spanish AM models trained with the *PWA acoustic set* were reached by fine-tuning the pre-trained model for 100 epochs, achieving a CER 25.8 and a WER of 49.8. However, previous results in English demonstrated that augmenting the training dataset with data from healthy controls improved the overall ASR performance. In this manner, the Spanish model trained with the *Mixed acoustic set* improved the WER performance at around 10% when finetuning the pre-trained model for 200 epochs. Once again, this approach showed that using semi-supervised methods on clinical data scarcity domains together with non-pathological data augmentation results in a very promising and interesting strategy.

Finally, the best performance for this language was achieved through a beam search decoding with the external *Large LM* model. Once again, the special symbols *FLR*, *SPN*, *BRTH* and *LAU* were discarded during the evaluation since these symbols were not covered in the generic texts. Following this strategy, we achieved a 24.8 of CER and a 42.8 of WER on the test set. These results differs with the English subset where the external *Large LM* did not improve the results at all. This may be due to the fact that the Spanish AM, fine-tuned with much fewer data, did not learn special symbols properly. As a result, they could be removed during evaluation without a negative impact on the performance.

6. Conclusions and Future Work

In this work, we show that semi-supervised learning methods applied to the ASR are promising solutions for improving the performance on aphasic speech recognition. Moreover, we set new benchmarks for the English AphasiaBank dataset, and we performed the first study for the Spanish language. The acoustic data for training were augmented using a mix of data from PWA and healthy controls, demonstrating that this strategy considerably improves the performance. This benefit was boosted for the case of Spanish, which included less than one hour of available aphasic speech data. These results open the door to improve ASR systems for people with aphasia and other clinical speech pathologies, or even simply to make speech recognition engines available for those languages with few annotated and available data.

As future work, it would be interesting to check if the performance of the systems could be improve by considering some other learning rate schedulers, by tuning the SpecAugment parameters or by considering other hyperparameters configurations. Moreover,whether the results could be enhanced by fine-tuning specific models for each level of aphasia severity should be evaluated, as speakers in each group probably perform similar speech and acoustic patterns. Another strategy worth studying would be to train AMs by directly removing the special symbols and then rescoring with an external *Large LM*. In any case, this point should be considered depending on the application, since special symbol information can be important for clinical practice but irrelevant for voice assistants. Furthermore, AMs may even be finetuned relative to individual patient speech by using Federated Learning approaches [54]. Finally, future studies should be also focused on extending this semisupervised learning method to other languages where no benchmarks on aphasic speech recognition voices has been reported, probably due to the scarcity of annotated data.

In addition, this technology should be tested in clinical practice, as well as in real medical environments and applications.

Author Contributions: Conceptualization, I.G.T. and A.Á.; methodology, I.G.T., M.R. and A.Á.; software, I.G.T. and M.R.; validation, I.G.T. and M.R.; formal analysis, I.G.T. and A.Á.; investigation, I.G.T., M.R. and A.Á.; resources, I.G.T. and M.R.; data curation, M.R.; writing—original draft preparation, I.G.T.; writing—review and editing, I.G.T. and A.Á.; visualization, A.Á.; supervision, I.G.T. and A.Á.; project administration, I.G.T.; funding acquisition, A.Á. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Acknowledgments: The authors would like to acknowledge AphasiaBank and, especially, to the people who contributed to it.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses or interpretation of data; in the writing of the manuscript or in the decision to publish the results.

Abbreviation

The following abbreviations are used in this manuscript:

PWA	Person with aphasia;
WBA	Western Aphasia Battery;
ASR	Automatic Speech Recognition;
E2E	End-to-end;
WER	Word Error Rate;
DNN	Deep Neural Network;
PER	Phoneme Error Rate;
CER	Character Error Rate;
SER	Syllable Error Rate;
AM	Acoustic Model;
HMM	Hidden Markov Model;
BLSTM-RNN	Bidirectional Long Short-Term Memory Recurrent Neural Network;
MoE	Mixture of experts;
LM	Language Model;
SID	Speech Intelligibility Detector;
AQ	Aphasia Quotient;
TDNN	Time-delay neural network;
MLS	Multilingual Librispeech;
CTC	Connectionist Temporal Classification;
GPU	Graphics Processing Unit.

References

1. Engelter, S.T.; Gostynski, M.; Papa, S.; Frei, M.; Born, C.; Ajdacic-Gross, V.; Gutzwiller, F.; Lyrer, P.A. Epidemiology of aphasia attributable to first ischemic stroke: Incidence, severity, fluency, etiology, and thrombolysis. *Stroke* **2006**, *37*, 1379–1384. [CrossRef]
2. Ellis, C.; Urban, S. Age and aphasia: A review of presence, type, recovery and clinical outcomes. *Top. Stroke Rehabil.* **2016**, *23*, 430–439. [CrossRef]
3. Martinez, E.O.; Saborit, A.R.; Carbonell, L.B.T.; Contreras, R.M.D. Epidemiología de la afasia en Santiago de Cuba. *Neurol. Argent.* **2014**, *6*, 77–82. [CrossRef]
4. Pedersen, P.M.; Stig Jørgensen, H.; Nakayama, H.; Raaschou, H.O.; Olsen, T.S. Aphasia in acute stroke: Incidence, determinants, and recovery. *Ann. Neurol. Off. J. Am. Neurol. Assoc. Child Neurol. Soc.* **1995**, *38*, 659–666. [CrossRef] [PubMed]
5. Mitchell, C.; Gittins, M.; Tyson, S.; Vail, A.; Conroy, P.; Paley, L.; Bowen, A. Prevalence of aphasia and dysarthria among inpatient stroke survivors: Describing the population, therapy provision and outcomes on discharge. *Aphasiology* **2021**, *35*, 950–960. [CrossRef]
6. Scarpa, M.; Colombo, A.; Sorgato, P.; De Renzi, E. The incidence of aphasia and global aphasia in left brain-damaged patients. *Cortex* **1987**, *23*, 331–336. [CrossRef]
7. Goodglass, H. *Understanding Aphasia*; Academic Press: Cambridge, MA, USA, 1993.
8. Shewan, C.M.; Kertesz, A. Reliability and validity characteristics of the Western Aphasia Battery (WAB). *J. Speech Hear. Disord.* **1980**, *45*, 308–324. [CrossRef] [PubMed]

9. Wernicke, C. *Der Aphasische Symptomencomplex: Eine Psychologische Studie auf Anatomischer Basis*; Cohn & Weigert: Breslau, Poland, 1874.
10. Lichtheim, L. On aphasia. *Brain* **1885**, *7*, 433–484. [CrossRef]
11. Fridriksson, J.; den Ouden, D.B.; Hillis, A.E.; Hickok, G.; Rorden, C.; Basilakos, A.; Yourganov, G.; Bonilha, L. Anatomy of aphasia revisited. *Brain* **2018**, *141*, 848–862. [CrossRef]
12. Hickok, G.; Poeppel, D. Dorsal and ventral streams: A framework for understanding aspects of the functional anatomy of language. *Cognition* **2004**, *92*, 67–99. [CrossRef]
13. Hickok, G.; Poeppel, D. The cortical organization of speech processing. *Nat. Rev. Neurosci.* **2007**, *8*, 393–402. [CrossRef] [PubMed]
14. Papo, D.; Buldú, J.M.; Boccaletti, S.; Bullmore, E.T. Complex network theory and the brain. *Philos. Trans. R. Soc. Biol. Sci.* **2014**, *369*. [CrossRef] [PubMed]
15. Fornito, A.; Zalesky, A.; Bullmore, E. *Fundamentals of Brain Network Analysis*; Academic Press: Cambridge, MA, USA, 2016.
16. Bhogal, S.K.; Teasell, R.; Speechley, M. Intensity of aphasia therapy, impact on recovery. In *Database of Abstracts of Reviews of Effects (DARE): Quality-assessed Reviews [Internet]*; Centre for Reviews and Dissemination: York, UK, 2003.
17. Kiran, S.; Des Roches, C.; Balachandran, I.; Ascenso, E. Development of an impairment-based individualized treatment workflow using an iPad-based software platform. In *Seminars in Speech and Language*; Thieme Medical Publishers: New York, NY, USA, 2014; Volume 35, pp. 038–050.
18. Lingraphica: AAC Devices for Communication. Available online: https://www.aphasia.com (accessed on 28 July 2021).
19. Orr, J. The Effectiveness of Tactus Therapy for Individuals with Apraxia of Speech. of the University. Ph.D. Thesis, Indiana University, Bloomington, IN, USA, 2020.
20. Godlove, J.; Anantha, V.; Advani, M.; Des Roches, C.; Kiran, S. Comparison of therapy practice at home and in the clinic: A retrospective analysis of the constant therapy platform data set. *Front. Neurol.* **2019**, *10*, 140. [CrossRef]
21. Munsell, M.; De Oliveira, E.; Saxena, S.; Godlove, J.; Kiran, S. Closing the digital divide in speech, language, and cognitive therapy: Cohort study of the factors associated with technology usage for rehabilitation. *J. Med. Internet Res.* **2020**, *22*, e16286. [CrossRef]
22. Wall, K.J.; Cumming, T.B.; Koenig, S.T.; Pelecanos, A.M.; Copland, D.A. Using technology to overcome the language barrier: The Cognitive Assessment for Aphasia App. *Disabil. Rehabil.* **2018**, *40*, 1333–1344. [CrossRef]
23. Le, D.; Licata, K.; Provost, E.M. Automatic quantitative analysis of spontaneous aphasic speech. *Speech Commun.* **2018**, *100*, 1–12. [CrossRef]
24. Abad, A.; Pompili, A.; Costa, A.; Trancoso, I.; Fonseca, J.; Leal, G.; Farrajota, L.; Martins, I.P. Automatic word naming recognition for an on-line aphasia treatment system. *Comput. Speech Lang.* **2013**, *27*, 1235–1248. [CrossRef]
25. Le, D.; Licata, K.; Persad, C.; Provost, E.M. Automatic assessment of speech intelligibility for individuals with aphasia. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2016**, *24*, 2187–2199. [CrossRef]
26. Jamal, N.; Shanta, S.; Mahmud, F.; Sha'abani, M. Automatic speech recognition (ASR) based approach for speech therapy of aphasic patients: A review. In *AIP Conference Proceedings*; AIP Publishing LLC.: Melville, NY, USA, 2017; Volume 1883, p. 020028.
27. Ballard, K.J.; Etter, N.M.; Shen, S.; Monroe, P.; Tien Tan, C. Feasibility of automatic speech recognition for providing feedback during tablet-based treatment for apraxia of speech plus aphasia. *Am. J. Speech Lang. Pathol.* **2019**, *28*, 818–834. [CrossRef]
28. Baevski, A.; Zhou, H.; Mohamed, A.; Auli, M. wav2vec 2.0: A framework for self-supervised learning of speech representations. *arXiv* **2020**, arXiv:2006.11477.
29. Wang, D.; Wang, X.; Lv, S. An overview of end-to-end automatic speech recognition. *Symmetry* **2019**, *11*, 1018. [CrossRef]
30. Park, D.S.; Chan, W.; Zhang, Y.; Chiu, C.C.; Zoph, B.; Cubuk, E.D.; Le, Q.V. Specaugment: A simple data augmentation method for automatic speech recognition. *arXiv* **2019**, arXiv:1904.08779.
31. Park, D.S.; Zhang, Y.; Jia, Y.; Han, W.; Chiu, C.C.; Li, B.; Wu, Y.; Le, Q.V. Improved noisy student training for automatic speech recognition. *arXiv* **2020**, arXiv:2005.09629.
32. Baevski, A.; Hsu, W.N.; Conneau, A.; Auli, M. Unsupervised Speech Recognition. *arXiv* **2021**, arXiv:2105.11084.
33. Fong, R.; Tsai, C.F.; Yiu, O.Y. The implementation of telepractice in speech language pathology in Hong Kong during the COVID-19 pandemic. *Telemed. E-Health* **2021**, *27*, 30–38. [CrossRef] [PubMed]
34. Muhetaer, P.; Ayifu, M.; Dawa, I.; Silamu, W. A Multi-Lingual and Text-Speech Dialog Support System for e-health. In *Journal of Physics: Conference Series*; IOP Publishing: Bristol, UK, 2020; Volume 1549, p. 052029.
35. Blackley, S.V.; Huynh, J.; Wang, L.; Korach, Z.; Zhou, L. Speech recognition for clinical documentation from 1990 to 2018: A systematic review. *J. Am. Med. Inform. Assoc.* **2019**, *26*, 324–338. [CrossRef] [PubMed]
36. Wade, B.P.R.; Cain, J. Voice recognition and aphasia: Can computers understand aphasic speech? *Disabil. Rehabil.* **2001**, *23*, 604–613. [CrossRef] [PubMed]
37. Barbera, D.S.; Huckvale, M.; Fleming, V.; Upton, E.; Coley-Fisher, H.; Doogan, C.; Shaw, I.; Latham, W.; Leff, A.P.; Crinion, J. NUVA: A Naming Utterance Verifier for Aphasia Treatment. *Comput. Speech Lang.* **2021**, *69*, 101221. [CrossRef] [PubMed]
38. Le, D.; Provost, E.M. *Improving Automatic Recognition of Aphasic Speech with AphasiaBank*; Interspeech: San Francisco, CA, USA, 2016; pp. 2681–2685.
39. Perez, M.; Aldeneh, Z.; Provost, E.M. Aphasic Speech Recognition using a Mixture of Speech Intelligibility Experts. *arXiv* **2020**, arXiv:2008.10788.

40. Qin, Y.; Lee, T.; Kong, A.P.H. Automatic Assessment of Speech Impairment in Cantonese-Speaking People with Aphasia. *IEEE J. Sel. Top. Signal Process.* **2019**, *14*, 331–345. [CrossRef]
41. MacWhinney, B.; Fromm, D.; Forbes, M.; Holland, A. AphasiaBank: Methods for studying discourse. *Aphasiology* **2011**, *25*, 1286–1307. [CrossRef] [PubMed]
42. MacWhinney, B. *The CHILDES Project: Tools for Analyzing Talk Transcription Format and Programs*; Psychology Press: London, UK, 2000; Volume 1.
43. Ardila, R.; Branson, M.; Davis, K.; Henretty, M.; Kohler, M.; Meyer, J.; Morais, R.; Saunders, L.; Tyers, F.M.; Weber, G. Common voice: A massively-multilingual speech corpus. *arXiv* **2019**, arXiv:1912.06670.
44. Tomar, S. Converting video formats with FFmpeg. *Linux J.* **2006**, *2006*, 10.
45. Sox Sound eXchange. Available online: http://sox.sourceforge.net/ (accessed on 1 July 2021).
46. Conneau, A.; Baevski, A.; Collobert, R.; Mohamed, A.; Auli, M. Unsupervised cross-lingual representation learning for speech recognition. *arXiv* **2020**, arXiv:2006.13979.
47. Baevski, A.; Auli, M.; Mohamed, A. Effectiveness of self-supervised pre-training for speech recognition. *arXiv* **2019**, arXiv:1911.03912.
48. Graves, A.; Fernández, S.; Gomez, F.; Schmidhuber, J. Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. In Proceedings of the 23rd International Conference on Machine Learning, Pittsburgh, PA, USA, 25–29 June 2006; pp. 369–376.
49. Gales, M.J.; Knill, K.M.; Ragni, A.; Rath, S.P. Speech recognition and keyword spotting for low-resource languages: BABEL project research at CUED. In Proceedings of the Fourth International Workshop on Spoken Language Technologies for Under-Resourced Languages (SLTU-2014), St. Petersburg, Russia, 14–16 May 2014; pp. 16–23.
50. Pratap, V.; Xu, Q.; Sriram, A.; Synnaeve, G.; Collobert, R. MLS: A Large-Scale Multilingual Dataset for Speech Research. *arXiv* **2020**, arXiv:2012.03411.
51. Biewald, L. Experiment Tracking with Weights and Biases. 2020. Available online: wandb.com (accessed on 1 July 2021).
52. Panayotov, V.; Chen, G.; Povey, D.; Khudanpur, S. Librispeech: An asr corpus based on public domain audio books. In Proceedings of the 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), South Brisbane, QLD, Australia, 19–24 April 2015; pp. 5206–5210.
53. Heafield, K. KenLM: Faster and smaller language model queries. In Proceedings of the Sixth Workshop on Statistical Machine Translation, Edinburgh, UK, 30–31 July 2011; pp. 187–197.
54. Bonawitz, K.; Eichner, H.; Grieskamp, W.; Huba, D.; Ingerman, A.; Ivanov, V.; Kiddon, C.; Konečnỳ, J.; Mazzocchi, S.; McMahan, H.B.; et al. Towards federated learning at scale: System design. *arXiv* **2019**, arXiv:1902.01046.