The Applicability of Wav2Vec2 and Whisper for Low-Resource Maltese ASR

Aiden Williams, Andrea Demarco, Claudia Borg

University of Malta, Malta

aiden@um.edu.mt, andrea.demarco@um.edu.mt, claudia.borg@um.edu.mt

Abstract

Maltese is a low-resource language with limited digital tools, including automatic speech recognition. With very limited datasets of Maltese speech available, a recent project, MASRI, developed further speech datasets and produced an initial prototype trained using the Jasper architecture. The best system achieved 55.05% WER on the MASRI test set. Our work builds upon this, producing a further two-and-a-half-hour annotated speech corpus from a domain in which no data was previously available (Parliament of Malta). Moreover, we experiment with existing pre-trained self-supervised models (Wav2Vec2.0 and Whisper) and further fine-tune these models on Maltese annotated data. A total of 30 Maltese ASR models are trained and evaluated using the WER and the CER. The results indicate that the performance of the models scales with the quantity of data, although not linearly. The best model achieves state-of-the-art results of 8.53% WER and 1.93% CER on a test set extracted from the CommonVoice project and 24.98% WER and 8.37% CER on the MASRI test set.

Index Terms: automatic speech recognition, low-resource languages

1. Introduction

Due to the lack of annotated speech data, Automatic Speech Recognition (ASR) for Maltese is considered low-resource, making it a challenge to work with from a technical perspective. Efforts have been made to source new annotated data, but the available data is still very limited. This means that a data-efficient method is imperative to create a Maltese ASR system.

Pre-trained self-supervised models, fine-tuned in low-resource scenarios outside the ASR field, perform better than if they were only trained on the available labelled data [1, 2]. Schneider et al. claim to be the first to implement a self-supervised English ASR system [3]. They presented the Wav2Vec model, which is pre-trained and then fine-tuned on an English speech dataset of 1,040 annotated hours. When evaluated, Wav2Vec surpassed the state-of-the-art at the time while requiring 11 times less annotated data. Conneau et al. present the Wav2Vec system in a multilingual pre-training approach [4]. They show that Cross-Lingual Speech Representations (XLSR) can be learnt and used by monolingual models. This is further developed by Babu et al. with the XLS-R models [5]. These models are pre-trained on more than 436,000 hours of unannotated speech data, sourced from 128 languages, including Maltese. Most recently, OpenAI unveiled their latest ASR model Whisper, pre-trained on 680,000 hours [6].

Following this trend, it is clear that training an ASR system for a low-resource language such as Maltese should leverage these larger models. The main aim of this research is to explore fine-tuning Wav2Vec XLS-R variants (300m & 2B) and Whisper variants (tiny, small & large-v2) for the purposes of creating a Maltese ASR. Specifically, we investigate the differences in performance in fine-tuning Wav2Vec 2.0 XLS-R and Whisper on a number of different speech datasets for Maltese. Moreover, we also present a further two-and-a-half hours of annotated spoken data in Maltese.

Our results demonstrate that all XLS-R models show improvements as more data is used for fine-tuning. However, diminishing returns start to appear at larger scales. On the other hand, diminishing returns on Whisper were not observed. The resulting ASR model has been made available on Huggingface.

2. Related Work

2.1. Maltese ASR

Maltese is the only Semitic language written in the Latin alphabet with an additional four letters: ħ, ċ, ġ, ż. The alphabet is represented in UTF-8 and consists of 24 consonants and 6 vowels (a, e, i, o, u, and ie). When the final syllable is stressed in certain Maltese words, grave accent vowels such as 'è' are used. Non-alphabetic characters such as '-' *(dash)* and ' *(apostrophe)* are also important in Maltese orthography.

Recent efforts were made by the Maltese Speech Recognition (MASRI)[1] project, which developed a number of speech corpora. These include the MASRI-Headset and other corpora including MEP, Merlin, and Tube [7]. Mena et al. train a Jasper model [8] on the MASRI-Headset, achieving 63.82% WER [9]. The inclusion of a 100-hour English dataset in the pre-training phase improved WER performance to 55.05%.

2.2. Wav2Vec for Self-Supervised ASR

The most recent Wav2Vec architecture is introduced by Baevski et al. with Wav2Vec 2.0 [10]. The novel architecture consists of three modules: a feature encoder and a vector quantization module similar to previous work, and a new Transformer module implemented within the architecture. A method was developed [11] in which the Transformer model takes a masked version of the latent speech representation Z as input, while ignoring the quantized representation Q, to create the context representation C. Self-supervised learning is done by combining the contrastive loss on C, and a weighted diversity loss on Q. This is done so that both modules are trained efficiently and so, help the entire model learn speech representations. To create a final ASR model, a fully connected layer is appended to the model allowing for Connectionist Temporal Classification (CTC) decoding. Therefore, the final training process is split into two phases while maintaining the one-model architecture.

[1] https://www.um.edu.mt/projects/masri/

During pre-training, the model learns to represent the discrete phonemes via self-supervision. Once this is done, the fully connected random initialised layer is appended to the trained model and fine-tuned using CTC on the annotated dataset [11].

An English dataset of unannotated data consisting of roughly 54000 hours is used for pre-training. The 1000 hour annotated Librispeech corpus [12] is used for fine-tuning. Subsets were created to observe the model performance in a pseudo-low-resource setting. Furthermore, the sizes and complexity of each module within the architecture are changed and experimented with. For the case of the Transformer module, the smaller BASE model has 12 layers, while the larger LARGE model has 24 layers. Including more layers means more trainable parameters. To train these additional parameters, the LARGE model is pre-trained on the entire dataset, while the BASE model is pre-trained on Librispeech only. Table 1 shows the performance of both models when evaluated on English.

Table 1: *Wav2Vec 2.0 WER results on the test-other Librispeech English set, based on fine-tuning set size and model parameter size [10].*

	1h	10h	100h
Wav2Vec 2.0 BASE	11.3	9.5	8.0
Wav2Vec 2.0 LARGE	5.8	4.9	4.0

Conneau et al. found that due to the robust architectural design of Wav2Vec 2.0, models are able to learn cross-lingual speech representations while pre-training on massive amounts of data [4]. This is put into practice with the XLSR models, which are pre-trained on up to 53 different languages. The largest model was subsequently pre-trained on a total of 56 thousand hours of speech data. To test out the XLSR approach, several Wav2Vec BASE models are pre-trained either monolingually or multilingually. Monolingual models follow the process previously taken. This process is changed slightly for multilingual models which are pre-trained on ten languages, then at the fine-tuning stage, a model is fine-tuned for each language. The experiment also included the pre-training of the Wav2Vec LARGE XLSR-53 model, which was pre-trained on the entire dataset of unannotated data, and just like the multilingual models, a separate model is then created for each language it was evaluated on during fine-tuning.

The work on the XLSR approach is continued with the release of the XLS-R models [5], which saw an increase in both the size of the unannotated data and the languages included. A total of 436 thousand unannotated hours are used for pre-training. 9000 hours of unannotated Maltese speech, from the Voxpopuli corpus [13], are included in this corpus. The model parameters were also increased with the smallest model using the Wav2Vec LARGE model, and the largest models increased both the number of Transformer blocks from 24 to 48 as well as the size of the feedforward layers to 1920 from 1024. The performance of the different approaches is evaluated on four languages: Assamese, Tagalog, Swahili, and Georgian are shown in Table 2. In these languages, the multilingual models, outperform the monolingual model.

2.3. Whisper for ASR

The trend of large pre-trained transformer-based models for ASR continues with OpenAI's Whisper model. Trained on 680,000 hours, we can assume that the same 9000 Maltese unannotated hours of the Voxpopuli corpus [13] have been used as the corpus is mentioned several times in their paper [6]. The model architecture is based on the transformer with 2 1D convolutional layers placed at the front. Substantial work and effort were made to make the model multi-modal, as Whisper is not only intended to be used for ASR. The authors present Whisper as a ready-to-use model for any language included in its training set, in the case of Maltese they report roughly 80% WER when trained on just one hour.

Table 2: *XLSR Wav2Vec 2.0 performance on low-resource settings when evaluated using WER. Assamese (AS), Tagalog (TL), Swahili (SW), and Georgian (KA) are the languages presented.*

Language	AS	TL	SW	KA
Annotated Data (h)	55	76	30	46
XLSR-10	44.9	37.3	35.5	-
XLSR-53	44.1	33.2	36.5	31.1
XLS-R (0.3B)	42.9	33.2	24.3	28.0
XLS-R (1B)	40.4	30.6	21.2	25.1
XLS-R (2B)	39.0	29.3	21.0	24.3

3. Data & Data Processing

Literature shows that transformer-based models are the current state-of-the-art system in low-resource ASR. The system can be implemented in two different methods. In the first option, a model is pre-trained from scratch on massive amounts of unannotated Maltese speech data and then fine-tuned on a smaller annotated dataset. Otherwise, an already pre-trained model, such as XLS-R or Whisper, is chosen and fine-tuned for Maltese ASR. While some efforts were made to collect unannotated Maltese speech, the amount of data collected so far is not enough for pre-training. For this reason, the latter mode of operation was chosen, and the XLS-R and Whisper models were fine-tuned on various subsets of Maltese speech.

3.1. Annotated Speech Corpora

Supervised deep learning techniques train models by minimising the loss function of a model. To match the output of a model with the actual label of a sample, the individual weights and biases of the layers within a neural network are modified via the backpropagation algorithm. XLS-R uses CTC decoding to predict the transcription of an audio file. This means that samples in the dataset that were used for fine-tuning require only the inclusion of the audio signal and the corresponding transcription.

Local efforts have been made to collect high-quality speech corpora. MASRI is a research project led by the University of Malta. It has spearheaded research and development in the field of speech related to the Maltese language. Several speech corpora developed by the MASRI project [7] has also been used in this project, namely the MASRI-Headset, MEP, Merlin, and Tube speech corpora. In total, these corpora amount to 40 hours and 24 minutes of annotated Maltese speech. A test and validation set are also included, each totalling 1 hour.

A 2-hour 30-minute corpus from Maltese parliamentary sessions was annotated for this project. Speeches from 23 male and 3 female parliament members were used, sourced from the parliament website. Audio files were loaded into Audacity, an audio editing software, to identify, clip, and export phrases or sentences of up to 10 seconds. The path and transcription of each new file were recorded in a CSV file. Unofficial sitting

transcriptions provided guidance and annotations were made by the authors. Audio clips were mostly sampled from the longer "Second Reading" stage of a bill reading. Annotated speeches were primarily read, not conversational or spontaneous, excluding chaotic or fast-paced speeches for ease of annotation.

CommonVoice (CV) is a Mozilla-led initiative with the aim of creating a diverse speech dataset that spans multiple languages [14]. The dataset is powered by global volunteers, who have enabled the inclusion of 80 languages. This project uses version 8 released in January 2022. Of the 80 languages, the Maltese subset consists of 8 hours and 53 minutes of verified annotated data. An additional 8 hours of unvalidated recorded speech are also included. 4794 unique prompts are included. Mozilla recommends having around 14,000 prompts to collect an audio dataset of 16 hours. However, only 4794 unique prompts are included in the Maltese dataset. This means that there are a number of audio samples repeating the same prompt, which limits textual variety. It is important to note here that, while prompts might be repeated, it is doubtful that contributors will record the same prompt more than once. A test set including 670 unique utterances, totalling 53 minutes in length, is used. A two-hour validation set was extracted from the validated set. Of the 8 unvalidated hours, an additional three hours are used, named CV Other.

3.2. Data Processing

Both audio and text processing is done either on acquisition or, in the case of the Parlament Speech corpus while annotating. The Wav2Vec system requires audio to be encoded as WAV files at 16KHz. To preserve space, only one channel (mono) was used. Text pre-processing was more complex. Character cases are lowered. Non-alphabetic characters are removed, except for the dash and the apostrophe. Accented letters were also kept in order to make the model learn the difference in pronunciation. Numeric characters are converted to Maltese numbers using Korpus Malti resources. Finally, the Maltese corpora are joined and exported as a MsgPack file. The order in which data is joined in matters [15]. Thus, while joining, the corpora were amended one after the other in the following manner; Headset, CV Validated, Parliament, MEP, Tube, Merlin, CV Other.

The authors behind the XLSR and XLS-R models claim that the models are capable of generating good transcriptions after fine-tuning with as little as 10 minutes of audio [4, 5]. In their work, the authors use splits of 10 minutes, 1 hour, 10 hours, and 100 hours to prove these claims. Therefore, the Maltese dataset is split into seven subsets: 10 minutes, 30 minutes, 1 hour, 10 hours, 20 hours, and 50 hours to validate these claims. Additional subsets are included to monitor the effect of increasing the corpus size. The ordered join was done so that the higher-quality Headset and CV-validated corpora are included in the smaller subsets.

3.3. Fine-tuning the models for Maltese ASR

Three versions of the XLS-R Wav2Vec model were produced. These include 300 million (300M), 1 billion (1B), and 2 billion (2B) parameter models [5]. Of these three, the 300M and 2B variants are used to develop 16 Maltese ASR models. To make the final trained models comparable, all models are trained using the same hyperparameters. Each model was trained for 30 epochs, and the AdamW criterion was used with a starting learning rate of 3e-4, where the first 500 training steps are used as warm-up steps. Depending on the size of the model, the batch size was changed to maximise the use of the available GPU VRAM. To hasten training, the gradient was accumulated every four steps which helped us reduce our VRAM usage. Six versions of Whisper were produced. These include the Tiny, Base, Small, Medium, Large and Large-v2 models. They have 39, 74, 244, 769, and 1550 million parameters respectively (Large and Large-v2 have the same parameter count) and so the Small and Large-v2 models are chosen as they roughly correspond to the Wav2Vec 2.0 models chosen. The Tiny Whisper model is also fine-tuned as it is interesting to see the performance of a smaller model, which would ultimately be more cost-efficient if deployed and used. Similar to how we trained the Wav2Vec 2.0 models, all models are trained using the same hyperparameters. Each model was trained for 30 epochs, and the AdamW criterion was used with a starting learning rate of 1e-5 where the first 500 training steps are used as warm-up steps. As the Whisper models are smaller than the Wav2Vec 2.0 models, we were able to make use of a much larger batch size in general, although this was still a variable dependent on the size of the model. We still used gradient accumulation when we fine-tuned the Large-v2 variant.

4. Evaluation

Two test sets are used, one as provided by MASRI and the other made up of validated CommonVoice audio. As mentioned previously, the MASRI test set includes several non-standard Maltese speech utterances. This set includes speech excerpts from rural speakers, political debates, and news broadcasts. In contrast, most of the utterances recorded for the CommonVoice project read Maltese in a standard way.

The utterances spoken by speakers with a heavy rural accent, usually elderly people, included a number of character errors, while the spoken words could still be inferred from the transcription output. Maltese speakers have a high tendency to code-switch to English. There is no attempt by the speakers to change pronunciation and so the models find it difficult to correctly transcribe English words spoken by Maltese speakers. Examples range from numbers such as "for" (four) and "wan" (one) and common food items such as the "big mek" (big mac). While grammatically incorrect these transcriptions are phonetically plausible, which also means that the model is correctly learning the correct Maltese character pronunciation.

4.1. MASRI Test Results

The results for the MASRI test set are shown in Figures 1 and 2, which present a worse performance than the results for the CommonVoice test set, shown in Figures 3 and 4. This is attributed to the fact that the MASRI test set was created with an 'in the wild' approach, where performance on the set would be more comparable to realistic use cases of the model rather than to the controlled environments used for data collection. The XLS-R 2B model attains the best performance of all the models when trained on 50 hours of Maltese speech. In general, Whisper did not perform as well as Wav2Vec 2.0 with Whisper-large model performance close to the Wav2Vec 2.0 however the smaller models lag behind. This is especially the case for the Whisper-tiny model which remained at 100% WER for all experiments.

4.2. CommonVoice Test Results

Looking at Figures 3 and 4 we can see that the XLS-R models perform very well with both the 300M and 2B 50-Hour models achieving results near 2% CER. WER is also considerably

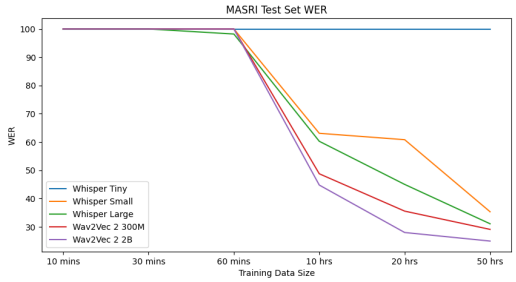

Figure 1: *XLS-R & Whisper WER Results on the MASRI Test Set.*

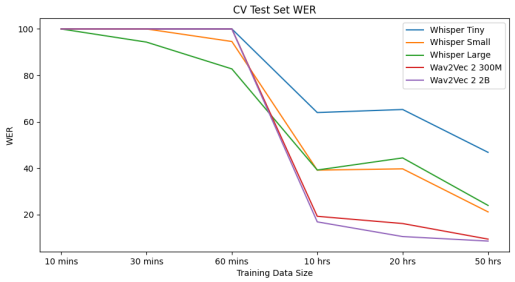

Figure 3: *XLS-R & Whisper WER Results on the CommonVoice Test Set.*

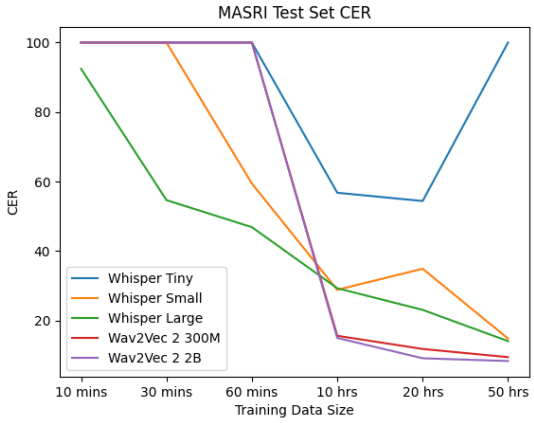

Figure 2: *XLS-R & Whisper CER Results on the MASRI Test Set.*

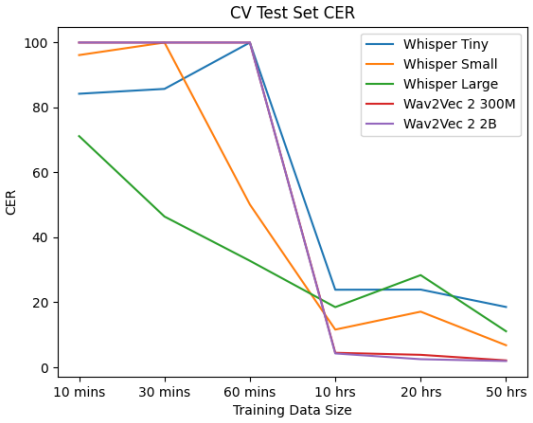

Figure 4: *XLS-R & Whisper CER Results on the CommonVoice Test Set.*

smaller, at 8.53%, the best recorded WER result for a Maltese ASR system so far. Note that the CommonVoice dataset includes more standard and clean speech audio. In this setting Whisper-large and Whisper-small performed very close to each other, meaning in a real-world use case it would be more cost-efficient to train a Whisper-small model and use it, rather than the bulkier Whisper-large.

5. Conclusion

The main aim of this project was to develop a Maltese ASR model using the Wav2Vec 2.0 and Whisper models. This was achieved, and the current state-of-the-art for Maltese ASR is presented in this paper along with the process behind it.

The effect investigated was that of scaling the training data for the fine-tuning task. All Maltese XLS-R models show improvement with more data. Although diminishing returns do start to appear, especially at larger scales, these are not at a point where further data collection would be futile. This is in contrast to the experiments made with the Whisper models where diminishing returns were not observed. In fact, the results indicate that we should revisit both models and train further.

Further improvements can be made to address model performance for in-the-wild scenarios. We can reinvigorate the data collection that has happened in the past by making use of the XLS-R-2B 50 model. At the same time, we would need to keep in mind that some audio, such as Maltese-English, is not correctly transcribed by this model. The issues the models experienced with Maltese-English require us to push towards collecting audio for that particular case, without the help of the trained models.

In conclusion of this project, the Wav2Vec 2.0 system has been successfully applied for Maltese ASR with the fine-tuning of the 2B XLS-R Model with 50 hours of annotated Maltese Speech Data. The effects of data scaling have been noted for future work to follow. The groundwork for further data annotation as well as model development has been made.

6. Acknowledgements

We acknowledge LT-Bridge Project (GA 952194) and DFKI for access to the Virtual Laboratory.

7. References

[1] P. Swietojanski, A. Ghoshal, and S. Renals, "Unsupervised cross-lingual knowledge transfer in dnn-based lvcsr," in *2012 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2012, pp. 246–251.

[2] S. Thomas, M. L. Seltzer, K. Church, and H. Hermansky, "Deep neural network features and semi-supervised training for low resource speech recognition," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013, pp. 6704–6708.

[3] S. Schneider, A. Baevski, R. Collobert, and M. Auli, "wav2vec: Unsupervised Pre-Training for Speech Recognition," in *Proc. Interspeech 2019*, 2019, pp. 3465–3469.

[4] A. Conneau, A. Baevski, R. Collobert, A. Mohamed, and M. Auli, "Unsupervised Cross-Lingual Representation Learning for Speech Recognition," in *Proc. Interspeech 2021*, 2021, pp. 2426–2430.

[5] A. Babu, C. Wang, A. Tjandra, K. Lakhotia, Q. Xu, N. Goyal, K. Singh, P. von Platen, Y. Saraf, J. Pino, A. Baevski, A. Conneau, and M. Auli, "XLS-R: Self-supervised Cross-lingual Speech Representation Learning at Scale," in *Proc. Interspeech 2022*, 2022, pp. 2278–2282.

[6] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," 2022.

[7] C. D. Hernandez Mena, A. Gatt, A. DeMarco, C. Borg, L. van der Plas, A. Muscat, and I. Padovani, "MASRI-HEADSET: A Maltese corpus for speech recognition," in *Proceedings of the 12th Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, May 2020, pp. 6381–6388. [Online]. Available: https://aclanthology.org/2020.lrec-1.784

[8] J. Li, V. Lavrukhin, B. Ginsburg, R. Leary, O. Kuchaiev, J. M. Cohen, H. Nguyen, and R. T. Gadde, "Jasper: An End-to-End Convolutional Neural Acoustic Model," in *Proc. Interspeech 2019*, 2019, pp. 71–75.

[9] C. D. H. Mena, A. DeMarco, C. Borg, L. van der Plas, and A. Gatt, "Data augmentation for speech recognition in maltese: A low-resource perspective," *CoRR*, vol. abs/2111.07793, 2021. [Online]. Available: https://arxiv.org/abs/2111.07793

[10] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in Neural Information Processing Systems*, vol. 33, pp. 12 449–12 460, 2020.

[11] A. Baevski and A. Mohamed, "Effectiveness of self-supervised pre-training for asr," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 7694–7698.

[12] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An asr corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 5206–5210.

[13] C. Wang, M. Riviere, A. Lee, A. Wu, C. Talnikar, D. Haziza, M. Williamson, J. Pino, and E. Dupoux, "VoxPopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Online: Association for Computational Linguistics, Aug. 2021, pp. 993–1003. [Online]. Available: https://aclanthology.org/2021.acl-long.80

[14] R. Ardila, M. Branson, K. Davis, M. Kohler, J. Meyer, M. Henretty, R. Morais, L. Saunders, F. Tyers, and G. Weber, "Common voice: A massively-multilingual speech corpus," in *Proceedings of the 12th Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, May 2020, pp. 4218–4222. [Online]. Available: https://aclanthology.org/2020.lrec-1.520

[15] A. Agarwal and T. Zesch, "German end-to-end speech recognition based on deepspeech," in *Proceedings of the 15th Conference on Natural Language Processing (KONVENS 2019): Long Papers*. Erlangen, Germany: German Society for Computational Linguistics & Language Technology, 2019, pp. 111–119.