INTL JOURNAL OF ELECTRONICS AND TELECOMMUNICATIONS, 2024, VOL. 70, NO. 2, PP. 315–321
Manuscript received February 14, 2024; revised June, 2024. doi: 10.24425/ijet.2024.149547

End-To-End deep neural models for Automatic Speech Recognition for Polish Language

Karolina Pondel-Sycz, Agnieszka Paula Pietrzak, and Julia Szymla

***Abstract*—This article concerns research on deep learning models (DNN) used for automatic speech recognition (ASR). In such systems, recognition is based on Mel Frequency Cepstral Coefficients (MFCC) acoustic features and spectrograms. The latest ASR technologies are based on convolutional neural networks (CNNs), recurrent neural networks (RNNs) and Transformers. The article presents an analysis of modern artificial intelligence algorithms adapted for automatic recognition of the Polish language. The differences between conventional architectures and ASR DNN End-To-End (E2E) models are discussed. Preliminary tests of five selected models (QuartzNet, FastConformer, Wav2Vec 2.0 XLSR, Whisper and ESPnet Model Zoo) on Mozilla Common Voice, Multilingual LibriSpeech and VoxPopuli databases are demonstrated. Tests were conducted for clean audio signal, signal with bandwidth limitation and degraded. The tested models were evaluated on the basis of Word Error Rate (WER).**

***Keywords*—Automatic Speech Recognition; Deep Neural Networks; End-To-End; Polish Language**

I. Introduction

THE ability to convert spoken language into written text, known as Automatic Speech Recognition (ASR), is an important element of developing human-computer interaction. ASR systems have become integral components in numerous applications, from voice assistants to transcription services, and have found applications in a wide array of languages. However, for Polish language, because of its' complicated structure and limited data resources, automatic speech recognition still requires research. In recent years, the emergence of End-To-End (E2E) approaches based on deep neural networks (DNNs) has accelerated ASR research, and E2E DNN systems show promise in processing Polish speech.

This paper presents preliminary tests of five E2E ASR models adapted for Polish language recognition, conducted on the Mozilla Common Voice [1] (MCV), Multilingual LibriSpeech [2] (MLS) and VoxPopuli [3] (VP) databases. Models adapted for automatic speech recognition in Polish were tested, two models available in the NVIDIA NeMo toolkit [4]: QuartzNet [5] and FastConformer Transducer-CTC [6]), Whisper [7] (developed by OpenAI), Model Zoo [8] from ESPnet toolkit and a Wav2Vec 2.0 XLSR-53 [9] model (developed by MetaAI). Whisper and ESPnet Model Zoo are multilingual, general-purpose models with language detection stage before recognition; the Wav2Vec 2.0 XLSR-53 version used is an additionally fine-tuned version of the multilingual model (which originally also covered Polish), which can be used as an ASR model for Polish (without the language detection stage); the described versions of the QuartzNet and FastConformer models have been fine-tuned for Polish based on pre-trained English models.

Section II describes conventional and E2E approaches for building ASR systems. Section III details the used databases and the architecture of the tested models. Section IV covers a description of the performed experiments. The results of the tests conducted are in Section V, and the conclusions are in VI.

Karolina Pondel-Sycz, Agnieszka Paula Pietrzak and Julia Szymla are with Faculty of Electronics and Information Technology, Warsaw University of Technology, Warsaw, Poland (e-mail: karolina.sycz@pw.edu.pl, agnieszka.pietrzak@pw.edu.pl, julia.szymla.stud@pw.edu.pl).

II. Automatic Speech Recognition (ASR)

In the field of automatic speech recognition, there are two main approaches: Conventional (Fig. 1), based on three models: acoustic (AM), linguistic (LM) and pronunciation (PM) and End-To-End (E2E, Fig. 2), based on a integrated deep neural model (DNN). In both solutions the input is an audio speech signal, and the output results in a transcription, i.e. a textual notation of the content contained in the input signal. The both approaches are distinguished by the models used, the preparation of training data, training proces and post-processing.

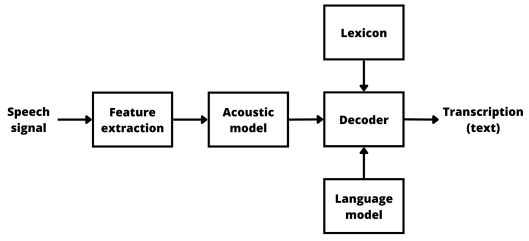

Fig. 1. Conventional ASR workflow.

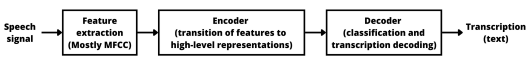

Fig. 2. End-To-End Encoder-Decoder ASR workflow.

A. Conventional ASR

In the Conventional ASR (Fig. 1) approach, models are trained separately and exchange information with each other. Typically, such systems are based on HMM (Hidden Markov Models [10]) AM and n-gram [11] LM and PM as a lexicon/pronunciation dictionary. The process of preparing training data for conventional ASR systems is therefore complex and time-consuming.

HMM AM consist of hidden markov chains and observed variables. In ASR, hidden states correspond to phonemes (graphical representation of speech sound) and observed variables represent to sound frames (acoustic features). Using HMM requires segmenting the speech signal into smaller parts and assigning them corresponding graphical representations (using forced alignment) [12]. In addition, in Conventional ASR AM use a wide range of acoustic features, such as formant frequencies [13], Perceptual Prediction Coefficients (PLP) [14], power-normalized cepstral coefficients (PNCC) [15] or Mel Frequency Cepstrum Coefficient (MFCC) [16]. This means that to train AM in Conventional ASR requires additional steps (in addition to the above, analysis and appropriate selection of acoustic features) - steps eliminated or partially eliminated in the E2E DNN approach.

Subsequent words are predicted based on previous words, and the meaning of the utterance is discovered based on the local context [17]. Statistical approaches are appropriate for English, which is positional and has a specific sentence formation (most commonly subject-verb-object (SVO) [18]). For Polish, which is inflected and has almost arbitrary sentence formation, coverage of both local and global context is required.

B. End-To-End ASR

The requirement of Polish to simultaneously capture the local and global demands a different approach to building ASR systems. Research is considering E2E ASR (Fig. 2) architectures based on an integrated deep neural network model. Such an architecture requires no or no extensive forced alignment and is simpler to train. The weakness of such an architecture is the huge computational power requirements, which are constantly increasing with the development of larger models and new types of neural networks.

One of the earliest deep neural networks used in the E2E ASR approach was Recurrent Neural Networks [19] (RNNs). RNNs are suitable for the analysis of sequential data (such as speech), and their prediction is based not only on the input at a given time, but is also updated with previous predictions [20]. However, they require a data pre-segmentation. In RNNs, the network has access to the entire previous sequence, but covers more local context - for the last parts of the input sequence and the prediction [21].

Other deep neural networks used in E2E ASR architectures are convolutional networks [22] (CNNs). CNNs were originally applied to image recognition but can be implemented for speech recognition in combination with its graphical representation (spectrograms [23]). Context of the time-dependency of speech is covered at the relevant depth of the CNNs [24]. It doesn't precise require pre-segmentation of data (on phonemes and phoneme labelling during model fine-tuning). The CNN network filter moves across the mel-spectrogram of whole utterance (e.g a word), so it doesn't require to be divided into smaller segments.

The latest deep neural networks with immense potential in ASR systems are Transformers [25]. It uses a Multi-Headed Self-Attention mechanism [26] covering both global and local contexts. They were originally adapted for the task of natural language processing (NLP), i.e. texts. Architectures of E2E ASR models using Transformers networks are enriched with CNN layers, which process features from spectrograms, and a CTC loss function or Transducer, which have their origin in RNNs [25]. The most effective E2E ASR models, therefore, seem to be those built from layers of all the above-mentioned types of deep neural networks.

III. Models Architecture and Datasets

For testing the E2E ASR architecture, suitable databases had to be prepared. For this purpose, 3 multilingual databases containing the Polish language and open source were selected. Five deep E2E ASR models based on both RNNs, CNNs and Transformers were selected for testing. This Section describes the chosen data resources and models.

A. Data

Databases containing both recordings and reference texts are required to test selected E2E ASR models. The recordings are used for transcription, and the reference texts for evaluation of the models (selected evaluation methods are described in Section III-C). For the experiments, Polish parts from 3 multilingual open-source databases were selected: Mozilla Common Voice [1] (MCV), Multilingual LibriSpeech (MLS) [2] and VoxPopuli (VP) [3].

1) ***Mozilla Common Voice*** *(MCV):* database provided for 112 languages (including Polish) and covers MP3 recordings of speech, corresponding transcriptions (text) and metadata on age, gender and accent. The datasets for each language are divided into training (train), development (dev) and test sets. The Polish portion of MCV is constantly being expanded and currently contains 173 hours of speech recordings (163 hours verified) in 15.0 version [27]. The MCV version available at [28] was sampled for the research.

2) ***Multilingual LibriSpeech*** *(MLS):* database is a multilingual version of the LibriSpeech [29] database (only for English). MLS includes, in addition to English, 7 other languages (including Polish). It contains speech read from publicly available LibriVox [30] audiobooks and Project Gutenberg text data [31] (44,500 hours in English and a total of 6,000 hours in other languages). The dataset is divided into training, development and testing sets. The Polish set contains recordings in subsets of respectively: 103.65, 2.08 and 2.14 hours [32].

*3) **VoxPopuli** (VP):* database is a multilingual speech corpus (includes 23 languages, including Polish) that contains recordings from the European Parliament from 2009 to 2020. The database includes unlabeled (400,000 hours) and transcribed (1,800 hours for 16 languages, with Polish); speech-to-speech interpretation data and transcribed accented speech data [33]. The Polish part contains 21.2 thousand hours of non-transcribed recordings, including 111 hours for which transcription is available.

B. Models Architectures

Five open-source, DNNs models were selected for the experiment, which are adapted to Polish speech recognition: QuartzNet [5], FastConformer [6], Wav2Vec 2.0 XLSR [9], Whisper [7] and ESPnet Model Zoo [8]. The QuartzNet model is fully convolutional with a BxR block architecture. Other models have an Encoder-Decoder architecture, in which the Encoder converts the input audio signal into high-level representations, and then the Decoder decodes the content and returns the most probable transcription.

The Whisper and Wav2Vec 2.0 XLSR models are OpenAI and MetaAI (formerly Facebook AI) projects, respectively. The Whisper model is available on the Github [34] platform, and the Wav2Vec 2.0 XLS adapted to Polish language originates from the Hugging Face [35] platform (an AI community sharing database and model knowledge). QuartzNet and FastConformer models are available in the NVIDIA NeMo Toolkit [4], and Model Zoo in the ESPnet toolkit [36]. The Whisper and ESPnet Model Zoo are general-purpose multilingual models which include Polish among their known languages. The described versions of the QuartzNet, FastConformer and Wav2Vec 2.0 XLSR-53 models are fine-tiuned to Polish based on pre-trained models recognizing speech in English. This section describes the characteristics of selected models.

*1) **QuartzNet**:* The QuartzNet [37] model is a fully convolutional model (CNN) based on the Jasper [38] model architecture - a deep time-delay neural network [39] (TDNN) consisting of blocks of layers of 1D CNNs. The model has a BxR architecture, where B is the number of blocks and R is the number of convolutional sub-blocks in a block. The QuartzNet model is distinguished from the pure Jasper version by separable splices and larger filters, making it perform close to Jasper, with an order of magnitude fewer parameters. The model chosen for the experiment is a QuartzNet version fine-tuned to Polish, trained on MCV 6.0.

*2) **FastConformer**:* Conformer [40] is a convolution-augmented Transformer. In this model, the CNN layers are not just the initial feature processing layers (as in Speech-Transformer), but the Transformer block has been replaced by a Conformer with additional CNN layers behind the Multi-Head Self Attention. FastConformer is an optimized version of the Conformer model. The result of data encoding can be decoded using either RNN-T or CTC loss (RNN-T by default) [41]. The model was trained on the MCV 12.0, MLS and VP databases.

*3) **Wav2Vec 2.0 XLSR-53**:* The Wav2Wec 2.0 [42], [43] model is a model developed by MetaAI. It is based on CNNs and Transformers networks. The model is self-supervised, learning speech structure from raw audio. XLSR-53 [44] is a cross-linguistic approach to teach speech units common to several languages. The model does not require data labeling, forced alignment or segmentation. The large Wav2Vec 2.0 XLSR-53 model has been trained for 53 languages on the MLS, MCV and BABEL [45] databases (a total of 56,000 hours of speech data, all of which include Polish). The model used in the experiment is an additional fine-tuned version of Wav2Vec 2.0 XLSR-53 for Polish using MCV 6.1 training and validation parts [9].

*4) **Whisper**:* The above-described models were originally developed for English and then adapted for Polish speech recognition. Using models adapted for Polish doesn't require a language identification step, because it is default for these models. Whisper [7] is a general-purpose speech recognition model, has been trained on a large set of varied audio data, and can perform multilingual speech recognition, speech translation and language identification. Using comand line, it is possible to determine the recognition language. When used in a Python implementation, the entire model is loaded first (e.g., large version) and the first stage of recognition is language identification, followed by transcription. Whisper architecture based on Speech-Transformer ideas. Whisper model documentation includes results of evaluation based on Fluers [46] dataset, for 57 languages, including Polish (WER = 5.4%).

*5) **ESPnet Model Zoo**:* Model Zoo [8] originates from ESPnet2 [47] - an updated version of the ESPnet toolkit. ESPnet2 includes changes and updates based on experience and feedback from users of the toolkit. In ESPnet2, audio data is directly entered into the model as in all of the above models. Model Zoo is a population of models trained on multiple datasets, both uni- and multilingual available for public use. Model Zoo implemented in ESPnet2 consists of over 164 models, in which over 20 are intended solely for ASR purpose. Whereas most of those models were designed for English language recognition only, there is a limited choice of models that can be used for other languages. Considering that Polish is a low-resource language, only "open_li52" corpus [48] from ESPnet toolkit (containing 52 languages, including the MCV database) could actually be applied here. Model Zoo is Transformer based and was trained on multilingual dataset containing tokens with Polish diacritic signs - letters with dashes, overdots and tails.

C. Evaulation

The Word Error Rate [49] (WER) was chosen to evaluate the E2E ASR models. This rate reflects the number of errors made in the recognition process. A low WER indicates a low number of errors produced by the model in the speech recognition process, and vice versa. The WER is commonly used to evaluate conventional and E2E ASR. It will enable comparison of the tested models with each other and to literature

sources. The formula for calculating the WER reveals the Eq. 1.

$$WER = \frac{I + D + S}{N} \cdot 100 \quad (1)$$

Where: S - number of substitutions, D - number of deletions, I - number of insertions, N - number of words in the reference text.

IV. Experiment

In this section, we present the WER values of the tested models. We performed the tests for five samples from each of the MCV, MLS and VP databases (test sets). We degraded the selected samples, with two types of degradation: by limiting the bandwidth and by applying an equalizer to the selected audio frequency. We investigated the effects of these types of signal degradation on recognition quality. All tests were performed in the Google Colab [50] engine using the Python 3 execution environment and the T4 GPU hardware accelerator. Due to the limitations of the budget allocated for the Colab environment in the study, the available disk memory and computing power were limited, so the tests were conducted for 5 randomly selected samples from the test section from each database (the same samples for each model).

A. Data preparation

The MCV, MLS and VP databases were sampled along with metadata. A csv file was created for each database containing the file name and corresponding reference transcription. The references were used in the model evaluation stage. Three types of data were used in the study: clean (data taken from a database), bandwidth limitation and degraded. Due to the requirements of the models derived from the NeMo Toolkit (QuartzNet and FastConformer) and the effective bandwidth limitation, a conversion to mono-channel, a sampling rate of 16 kHz and bits per sample equal to 16 was performed for all data used.

*1) **Bandwidth limitation**:* To test the effect of bandwidth limitation on recognition quality, limiting the bandwidth to the range 300 Hz - 3 kHz was applied.

*2) **Degraded**:* The audio_degrader tool [51] was used to degrade the data. Audio amplitude normalization to the range (-1.0, 1.0) was applied to the samples; the filter gain was set to 6 dB, the dynamic range compression (compression ratio) was set to 3 (hard), and a two-pole peaking equalisation (EQ) filter with the parameters: filter center frequency 500 Hz, filter bandwidth 10 Hz, and filter gain 30 dB.

B. Usage of models

All models were tested according to the following scheme:

- loading the model into the Colab notebook,
- feeding audio samples to the model input,
- for Whisper and Model Zoo models language detection, for other models this step was skipped,
- performing the automatic speech recognition process,
- saving the obtained results to a csv file,
- calculating the WER for each sample and the average.

All results were saved to csv files. assigned to each model. All models except Whisper are capable of recognizing speech from sound with no limit on its duration. Whisper is only capable of recognizing recordings with a maximum duration of 30 seconds (if the sound is longer, Whisper recognizes only the first 30 seconds and the transcription is cut off). The duration of all samples used in the tests did not exceed 30 seconds, so there was no need for additional data segmentation for Whisper.

Punctuation marks have been removed from both references and transcriptions and case has been omitted (all tests converted to lower case). The evaluation was carried out using the wer() method from the JiWER [52] library by comparing the preference with the recognition result obtained in the tests and calculating the WER. WERs obtained for all models, databases and degrees of audio degradation, are summarized in Tables 1-3 in Section V.

V. Results

The result of the tests is the summary attached in this Section. The results apply to all tested models, databases and degrees of audio degradation. Table I shows the WER values for the clean audio files, Table II for the signal with bandwidth limitation, and Table III for the degenerate audio.

TABLE I
WER [%] (CLEAN AUDIO SIGNAL).

	MCV	MLS	VP
FastConformer	0.00	3.78	3.80
QuartzNet	4.44	36.20	50.00
Wav2Vec	15.55	8.52	27.55
Whisper	8.89	2.91	6.65
ESPnet	47.20	63.77	83.15

TABLE II
WER [%] (AUDIO SIGNAL WITH BANDWIDTH LIMITATION).

	MCV	MLS	VP
FastConformer	0.00	3.78	3.80
QuartzNet	6.67	41.27	54.55
Wav2Vec	26.67	8.91	29.85
Whisper	6.67	3.94	5.65
ESPnet	69.91	84.21	81.04

In the results obtained, the influence of the training data on the WER achieved is evident. The FastConformer model was trained on all of the databases used in the study, and although the training data is from a different part of the databases than the test data, its influence on the quality of the model is visible (Table I). This may indicate an overfitting of the model to the data, which may be a source of comparison non-objectivity.

TABLE III
WER [%] (DEGRADED AUDIO SIGNAL).

	MCV	MLS	VP
FastConformer	0.00	6.71	1.80
QuartzNet	8.89	44.25	64.30
Wav2Vec	30.00	17.15	40.05
Whisper	28.89	5.29	4.65
ESPnet	48.15	68.41	87.44

The QuartzNet model was adapted to the Polish language on the basis of the MVC database as reflected in the results obtained. The low WER for this database and model is also due to over-fitting of the model to the data. The WER for clean audio (Table I) and signal with bandwidth limitation (Table II), for the MVC database known to the model, is approx. 10 times lower than for samples unknown to the model (MLS and VP). For the degraded signal (Table III), the WER for the MCV database is approx. 5 times smaller than for the MLS and VP databases.

The Wav2Vec 2.0 XLSR-53 model was adapted to Polish on basis of MCV 6.1, but it achieved the best recognition results for the MLS database (in all Tables), and did not exhibit an over-fitting to the MCV database, for which it obtained a relatively high WER. This is a deviation relative to all other models, and the WER obtained for MLS is approx. 2 times smaller than for MVC, for which the fine-tuning was performed. However, according to [44], the original Wav2Vec XLSR-52 model (before additional fine-tuning to Polish with MCV 6.1) was trained as a large model on 53 languages on the MLS, MCV and BABEL [45] databases (containing Polish). Therefore, it can be assumed that the model fitted most closely to the MLS database.

Information on the exact content of the Whisper model training data is not available. It is therefore not possible to conclude conclusively whether the training data had an impact on its over-fitting. If the range of training data for this model was sufficiently large and diverse, this effect on recognition can be disregarded. The Whisper model gets similar results for all tested databases (Tables I and II), which could indicate that the training database contained data from all considered databases, or the model generalizes well.

In the information available in the ESPnet2 toolkit documentation, the only corpus that may contain Polish (open_li52, which contains 52 languages, there is no information as to which languages these are, but the other databases do not contain Polish, so one might suspect that Polish is included in this database) contains samples from the MCV database. The Model Zoo, like the previous models, recognised the samples best, from the database known to it (in all Tables). However, in the case of this model, the WER differences between the data were not as large as for the other models, although overall the WER for the Model Zoo is high in all performed tests.

For the QuartzNet and Wav2Vec 2.0 XLSR-53 models, the WER value increased with the degree of audio signal degradation. For Whisper, such a relationship is only apparent for the MLS (Table III). For the MCV, signal bandwidth limitation (Table II), Whisper's WER decreased, while for a degraded signal (Table III) it increased significantly. For the VP, the WER decreased as the degree of data degradation increased. For Model Zoo, the highest WER results were obtained for signal with bandwidth limitation (Table II), followed by the degraded signal (Table III), and the lowest WER for the clean audio (Table I).

As Model Zoo and Whisper are multilingual and general-purpose, the recognition task can be divided into two important parts: the actual detection of language and the actual recognition of words. In both cases there was incorrect language detection, with this occurring more frequently for Model Zoo. Whisper confused the language for only one sample, which came from the MCV degraded signal, and instead of Polish, it detected Ukrainian, and recognised the phrase "byle tylko jechać" (just to go) as rider in Ukraininan. In the case of Model Zoo, incorrect language detection has a major impact on the prediction results. Polish, being a slavic family language, would often be mistaken for Russian. MCV clean audio were recognised most successfully with every sample identified as a polish language, leaving MLS and VP behind with samples mistaken for Russian and Italian. In the case of a degraded signal none of the subset samples were fully recognised as polish, being mistaken for Russian and Portuguese. For signal with bandwidth limitation some samples were falsely recognised as Russian, English and Spanish. Although Ukrainian and Russian come from the same family of inflectional languages as Polish and the pronunciation of some words with the same meaning may be similar, the languages are distinguished from Polish by their writing - Ukrainian and Russian use the Cyrillic alphabet, while Polish uses the Latin alphabet. In this case, the WER could not be calculated correctly even for words with similar sounds in both languages (WER=100% or above, which contributes to such poor performance of Model Zoo).

VI. CONCLUSIONS

Although the study was limited in terms of the amount of data analyzed, it uncovered the following weaknesses in the models tested: excessive fit to the data and enormous computing power requirements. In addition, in the case of Whisper and ESPnet Model Zoo, the extra stage of language detection, appeared to be the source of the decline in the performance of these models (with this having a much greater impact on ESPnet Model Zoo). Since the tests performed showed an over-fitting of the models to the used databases, the impact of signal degradation on the recognition process cannot be clearly assessed. The conclusions indicate the need for further testing, on a database that is known not to have been presented to the tested models during training and fine-tuning. Since the contents of the Whisper and Model Zoo training databases are not fully known, this may require the preparation of a self-created database. Due to the limitations of the budget allocated for the Colab environment, limited tests could be performed.

For more accurate results, on a larger number of samples, there is a need to increase the budget for computing power and disk capacity. The requirement for computing power of E2E ASR models is huge, so there is also a need to develop methods to limit this demand.

ACKNOWLEDGMENT

We would like to express our deep gratitude to the head of the Department of Electroacoustics at the Warsaw University of Technology, Jan Żera, Prof. D.Sc. Eng., for the opportunity to conduct this research. Karolina Pondel-Sycz would also like to sincerely thanks her supervisor Piotr Bilski, Assoc. Prof. D.Sc. Eng.

REFERENCES

[1] R. Ardila *et al.*, "Common Voice: a Massively-Multilingual Speech Corpus," *Proc. of the 12th Conf. on Language Resources and Evaluation (LREC 2020)*, pp. 4218–4222, 11–16 May 2020. [Online]. Available: https://www.aclweb.org/anthology/2020.lrec-1.520.pdf

[2] J. Mahaveer *et al.*, "Mls: A large-scale multilingual dataset for speech research," in *INTERSPEECH 2020*. Shanghai, China: ISCA Speech, 25-29 Oct. 2020, pp. 2757–2761. [Online]. Available: https://doi.org/10.21437/Interspeech.2020-2826

[3] C. Wang *et al.*, "VoxPopuli: A Large-Scale Multilingual Speech Corpus for Representation Learning, Semi-Supervised Learning and Interpretation," *arXiv (Cornell University)*, Jan. 2021. [Online]. Available: https://doi.org/10.48550/arXiv.2101.00390

[4] "NVIDIA NeMo conversational AI toolki github repository." [Online]. Available: https://github.com/NVIDIA/NeMo

[5] "STT Pl Quartznet15x5 model." [Online]. Available: https://catalog.ngc.nvidia.com/orgs/nvidia/teams/nemo/models/stt_pl_quartznet15x5

[6] "NVIDIA FastConformer-Hybrid Large (pl) model." [Online]. Available: https://catalog.ngc.nvidia.com/orgs/nvidia/teams/nemo/models/stt_pl_fastconformer_hybrid_large_pc

[7] "Whisper Github repository." [Online]. Available: https://github.com/openai/whisper

[8] "Zoo espnet multi-purpose pre-trained model." [Online]. Available: https://github.com/espnet/espnet_model_zoo

[9] "Fine-tuned wav2vec2-xlsr-53 large model for speech recognition in polish." [Online]. Available: https://huggingface.co/jonatasgrosman/wav2vec2-large-xlsr-53-polish

[10] L. Rabiner and B. Juang, "An introduction to hidden Markov models," *IEEE ASSP magazine*, vol. 3, no. 1, pp. 4–16, Jan. 1986. [Online]. Available: https://doi.org/10.1109/massp.1986.1165342

[11] W. Cavnar and J. Trenkle, "N-gram-based text categorization," *Proc. of the 3rd Annu. Symp. on Document Analysis and Information Retrieval*, pp. 161–175, 11-13 Apr 1994. [Online]. Available: https://www.let.rug.nl/vannoord/TextCat/textcat.pdf

[12] M. J. F. Gales and S. Young, "The application of hidden Markov models in speech recognition," *Foundations and Trends in Signal Processing*, vol. 1, no. 3, pp. 195–304, Jan. 2007. [Online]. Available: https://doi.org/10.1561/2000000004

[13] J. Holmes, W. Holmes, and P. Garner, "Using formant frequencies in speech recognition," in *Proc. 5th European Conf. on Speech Communication and Technology (Eurospeech 1997)*. Rhodes, Greece: ISCA Speech, 22-25 Sep. 1997, pp. 2083–2086. [Online]. Available: https://doi.org/10.21437/Eurospeech.1997-551

[14] F. Honig, G. Stemmer, C. Hacker, and F. Brugnara, "Revising Perceptual Linear Prediction (PLP)," in *INTERSPEECH 2016*. San Francisco, CA, USA: ISCA Speech, 8–12 Sep. 2016, pp. 410–414. [Online]. Available: https://doi.org/10.21437/Interspeech.2016-1446

[15] C. Kim and R. M. Stern, "Power-Normalized Cepstral Coefficients (PNCC) for Robust Speech Recognition," *IEEE/ACM Transactions on Audio Speech and Language Processing*, vol. 24, no. 7, pp. 1315 – 1329, Jul. 2016. [Online]. Available: https://doi.org/10.1109/TASLP.2016.2545928

[16] F. Zheng, G. Zhang, and Z. Song, "Comparison of different implementations of MFCC," *Journal of Computer Science and Technology*, vol. 16, no. 6, pp. 582–589, 11 2001. [Online]. Available: https://doi.org/10.1007/bf02943243

[17] A. Pohl and B. Ziółko, *Using part of speech N-Grams for improving automatic speech recognition of Polish*, Jan. 2013. [Online]. Available: https://doi.org/10.1007/978-3-642-39712-7_38

[18] K. Marasek *et al.*, "Spoken language translation for polish," *arXiv (Cornell University)*, Nov. 2015. [Online]. Available: https://doi.org/10.48550/arXiv.1511.07788

[19] A. Graves and N. Jaitly, "Towards End-To-End Speech Recognition with Recurrent Neural Networks," *Proc. of 31st Int. Conf. on Machine Learning*, pp. 1764–1772, 21–26 Jun. 2014. [Online]. Available: http://proceedings.mlr.press/v32/graves14.pdf

[20] O. Vinyals, S. V. Ravuri, and D. Povey, "Revisiting recurrent neural networks for robust asr," in *2012 IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*. Kyoto, Japan: IEEE, 25 - 30 Mar. 2012, pp. 4085–4088. [Online]. Available: https://doi.org/10.1109/ICASSP.2012.6288816

[21] D. Jurafsky and J. H. Martin, "Speech and Language Processing (3rd ed. draft)." [Online]. Available: https://web.stanford.edu/~jurafsky/slp3/

[22] Y. Zhang *et al.*, "Towards End-to-End Speech Recognition with Deep Convolutional Neural Networks," in *INTERSPEECH 2005 - Eurospeech, 9th European Conf. on Speech Communication and Technology*. Lisbon, Portugal: ISCA Speech, 4-8 Sep. 2005, pp. 2997–3000. [Online]. Available: https://doi.org/10.21437/Interspeech.2005-138

[23] A. V. Oppenheim, "Speech spectrograms using the fast Fourier transform," *IEEE Spectrum*, vol. 7, no. 8, pp. 57–62, 8 1970. [Online]. Available: https://doi.org/10.1109/mspec.1970.5213512

[24] S. William and C. Jim, "End-to-end deep neural network for automatic speech recognition," *Stanford CS224D Reports*, 2015. [Online]. Available: http://cs224d.stanford.edu/reports/SongWilliam.pdf

[25] B. X. Linhao Dong, Shuang Xu, "Speech-transformer: A no-recurrence sequence-to-sequence model for speech recognition," in *2018 IEEE 40th Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 15-20 Apr. 2018, pp. 5884–5888. [Online]. Available: https://doi.org/10.1109/ICASSP.2018.8462506

[26] A. Vaswani *et al.*, "Attention is all you need," in *Advances in Neural Information Processing Systems 30 (NIPS 2017)*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[27] "Common Voice open source, multi-language dataset of voices." [Online]. Available: https://commonvoice.mozilla.org/en/datasets

[28] "Dataset Card for CommonVoice Hugging Face." [Online]. Available: https://huggingface.co/datasets/common_voice

[29] V. Panayotov *et al.*, "Librispeech: an asr corpus based on public domain audio books," in *2015 IEEE 40th Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 19-24 Apr. 2015, pp. 5206–5210. [Online]. Available: https://doi.org/10.1109/ICASSP.2015.7178964

[30] "LibriVox free public domain audiobooks." [Online]. Available: https://librivox.org/

[31] "Project Gutenberg free eBooks." [Online]. Available: https://www.gutenberg.org/

[32] "Multilingual LibriSpeech (MLS) Website." [Online]. Available: https://www.openslr.org/94/

[33] "VoxPopuli Github repository." [Online]. Available: https://github.com/facebookresearch/voxpopuli

[34] "Github The AI-powered developer platform to build, scale, and deliver secure software." [Online]. Available: https://github.com/

[35] "Hugging Face The AI community building the future." [Online]. Available: https://huggingface.co/

[36] S. Watanabe *et al.*, "Espnet: End-to-end speech processing toolkit," in *INTERSPEECH 2018*. ISCA Speech, 2-6 Sep. 2018, pp. 2207–2211. [Online]. Available: https://doi.org/10.21437/Interspeech.2018-1456

[37] S. Kriman *et al.*, "Quartznet: Deep automatic speech recognition with 1d time-channel separable convolutions," in *2020 IEEE International Conf. on Acoustics, Speech and Signal Processing (ICASSP)*. Barcelona, Spain: IEEE, 04–08 May 2020, p. 6124–6128. [Online]. Available: https://doi.org/10.1109/ICASSP40776.2020.9053889

[38] J. Li *et al.*, "Jasper: An end-to-end convolutional neural acoustic model," *arXiv (Cornell University)*, Apr. 2019. [Online]. Available: https://doi.org/10.48550/arXiv.1904.03288

[39] A. Waibel *et al.*, "Phoneme recognition using time-delay neural networks," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 37, no. 3, pp. 328–339, Mar. 1989. [Online]. Available: https://doi.org/10.1109/29.21701

[40] A. Gulati *et al.*, "Conformer: Convolution-augmented Transformer for Speech Recognition," in *INTERSPEECH 2020*. Shanghai, China:

ISCA Speech, 25-29 Oct. 2020, pp. 5036–5040. [Online]. Available: https://doi.org/10.21437/Interspeech.2020-3015

[41] "NVIDIA Models - NeMo Core." [Online]. Available: https://docs.nvidia.com/deeplearning/nemo/user-guide/docs/en/main/asr/models.html#hybrid-transducer-ctc

[42] A. Baevski *et al.*, "wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations," *Neural Information Processing Systems*, vol. 33, pp. 12 449–12 460, 6 2020. [Online]. Available: https://proceedings.neurips.cc/paper/2020/file/92d1e1eb1cd6f9fba3227870bb6d7f07-Paper.pdf

[43] "MetaAI Wav2vec 2.0: Learning the structure of speech from raw audio." [Online]. Available: https://ai.meta.com/blog/wav2vec-20-learning-the-structure-of-speech-from-raw-audio/

[44] C. Alexis *et al.*, "Unsupervised cross-lingual representation learning for speech recognition," in *INTERSPEECH 2021*. Brno, Czechia: ISCA Speech, 30 Aug. - 3 Sep. 2021, pp. 2426–2430. [Online]. Available: https://doi.org/10.21437/Interspeech.2021-329

[45] P. Roach *et al.*, "Babel: An eastern european multi-language database," in *Proc. of 4th International Conf. on Spoken Language Processing. ICSLP '96*. IEEE, 3-6 Oct. 1996. [Online]. Available: https://doi.org/10.1109/ICSLP.1996.608002

[46] A. Conneau *et al.*, "FLEURS: FEW-Shot Learning Evaluation of Universal Representations of Speech," *2022 IEEE Spoken Language Technology Workshop (SLT)*, 1 2023. [Online]. Available: https://doi.org/10.1109/slt54892.2023.10023141

[47] "Espnet2 Espnet major update." [Online]. Available: https://espnet.github.io/espnet/espnet2_tutorial.html

[48] "List of ESPnet2 corpora." [Online]. Available: https://github.com/espnet/espnet/blob/master/egs2/README.md

[49] A. Ahmed and R. Steve, "Word Error Rate Estimation for Speech Recognition: e-WER," in *Proc. of the 56th Annu. Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Melbourne, Australia: -, Association for Computational Linguistics 2018, pp. 20—-24. [Online]. Available: https://aclanthology.org/P18-2004/

[50] "Google Colaboratory hosted Jupyter Notebook service." [Online]. Available: https://colab.google/

[51] "Audio degradation toolbox in python, with a command-line tool. github repository." [Online]. Available: https://github.com/emilio-molina/audio_degrader

[52] "Jiwer ASR evaluation library." [Online]. Available: https://jitsi.github.io/jiwer/