



Gram Vaani ASR Challenge on spontaneous telephone speech recordings in regional variations of Hindi

Anish Bhanushali¹, Grant Bridgman², Deekshitha G³, Prasanta Ghosh³, Pratik Kumar¹, Saurabh Kumar³, Adithya Raj Kolladath¹, Nithya Ravi¹, Aaditeshwar Seth⁴, Ashish Seth¹, Abhayjeet Singh³, Vrunda N. Sukhadia¹, Umesh S¹, Sathvik Udupa³ and Lodagala V. S. V. Durga Prasad¹

¹Department of Electrical Engineering, Indian Institute of Technology, Madras-600036, India.

²Uliza CEO, Co-founder, City of Cape Town, Western Cape, South Africa.

³Electrical Engineering Department, Indian Institute of Science (IISc), Bangalore-560012, India.

⁴Khosla School of Information Technology, Indian Institute of Technology, New Delhi-110016, India.

gramvaani.challenge.2022@gmail.com

Abstract

This paper describes the corpus and baseline systems for the Gram Vaani Automatic Speech Recognition (ASR) challenge in regional variations of Hindi. The corpus for this challenge comprises the spontaneous telephone speech recordings collected by a social technology enterprise, *Gram Vaani*. The regional variations of Hindi together with spontaneity of speech, natural background and transcriptions with variable accuracy due to crowdsourcing make it a unique corpus for ASR on spontaneous telephonic speech. Around, 1108 hours of real-world spontaneous speech recordings, including 1000 hours of unlabelled training data, 100 hours of labelled training data, 5 hours of development data and 3 hours of evaluation data, have been released as a part of the challenge. The efficacy of both training and test sets are validated on different ASR systems in both traditional time-delay neural network-hidden Markov model (TDNN-HMM) frameworks and fully-neural end-to-end (E2E) setup. The word error rate (WER) and character error rate (CER) on eval set for a TDNN model trained on 100 hours of labelled data are 29.7% and 15.1%, respectively. While, in E2E setup, WER and CER on eval set for a conformer model trained on 100 hours of data are 32.9% and 19.0%, respectively.

Index Terms: Real-world ASR challenge, spontaneous telephone speech data, Gram Vaani, Hindi speech data.

1. Introduction

Humanizing speech technology for a wide range of population is challenging due to language differences arising from regional variations in a language. This is particularly true in a country like India, where the official language, Hindi, spoken by 528 million people, has 56+ dialectal variations across multiple states including Delhi, Rajasthan, Uttar Pradesh, Madhya Pradesh, Bihar, Jharkhand, and other states [Census 2011] [1]. These variations pose a problem for running socially beneficial services such as agricultural and health advisory, and voice-based interactions for call centre automation, for which speech recognition can potentially improve the efficiency and usability of the services, as well as make them more equitably accessible to less-literate populations. *Gram Vaani*¹ (GV) is a social enterprise operating in rural India that provides such services to support socio-economic development through a voice-based par-

ticipatory media platform called *Mobile Vaani*² [2]. Users can call the Mobile Vaani Interactive Voice Response (IVR) system, and through keypress-based navigation they can record voice messages, or listen to voice messages recorded by other users. The use of IVR makes the Mobile Vaani platform accessible through feature phones, without requiring any Internet access, and serves as a voice forum for rural communities to share local news, agriculture related queries and discussions, report about problems they face with accessing welfare schemes, and record cultural expressions in the form of songs and poems in their local languages. Over 25 district-level instances of Mobile Vaani have been operational for more than seven years. Together, these instances serve more than 100,000 monthly unique users, and can benefit significantly from automatic speech recognition (ASR) tools to automate moderation of the voice reports [3], improve data collection through voice-surveys [4], and provide automated question-answering services [5]. Several other similar voice-forum platforms operating around the world can also benefit from ASR improvements [6, 7, 8, 9].

The paper presents a real-world ASR challenge by sharing spontaneous telephone speech recordings obtained through the Mobile Vaani platform operating in North India, with several regional variations of Hindi. The recordings are similar to call centre data with natural noisy backgrounds - street settings, crowded areas, outdoor nature, etc. For the purpose of this ASR challenge, the recordings are accompanied by their corresponding transcriptions generated by crowd workers recruited via the Uliza platform³, along with a variety of metadata for the recordings including the location and gender. The recordings come from more than four states (Bihar, Jharkhand, Madhya Pradesh, Uttar Pradesh) and most recordings have district level information as well.

From Table 1 it is clear that, even though similar huge datasets have been released for the development of ASR systems in other well-resourced languages such as English [18, 19, 20], Russian, [21], the low-resource regional variations of Hindi, together with the spontaneity of telephone speech, natural background noises, and transcriptions with varying degrees of accuracy due to crowdsourcing, make this a unique corpus for the ASR challenge. First 8 rows of Table 1 lists some challenges and corpora where Hindi speech data is available and last 3 rows lists the spontaneous telephonic corpora. The GV chal-

¹<http://gramvaani.org/>

²<http://mobilevaani.in/vaani/#/1/home>

³<https://www.uliza.org/>

Table 1: *Prior initiatives in Hindi and spontaneous data collection*

Name	Language	Size (Hindi data)	Type	Ref/Link
Hindi Raw Speech Corpus	Hindi	121 h		4
Hindi Speech Database	Hindi	500 sent	Read speech	[10]
MUCS 2021	Hindi	106.09 h	Read speech	[11]
Hindi-Tamil-English ASR challenge	Hindi	188.1 h	Read and conversational speech	5
IITKGP-MLILSC Speech Database	Hindi	134.7 h	Television broadcast	[12]
IIITH-ILSC Speech Database	Hindi	4.5 h		[13]
MeitY	Hindi	6 h	Read, conversation,& lecture	
CALLFRIEND Hindi	Hindi	60 files ranging from 5-30 mnt	Telephone conversations	[14]
Bangla telephonic speech corpus	Bangla	0 h	Telephonic speech	[15]
Spoken Corpora	15 languages	0 h	Telephone conversations	6
Iraqi Arabic CTS	Arabic	0 h	Telephone conversations	[16]
Gulf Arabic CTS	Gulf Arabic	0 h	Telephone conversations	[17]

h: hours; sent: sentence; mnt: minutes

lenge provides 1000 hours of unlabelled recordings, as well as 100 hours of labelled recordings, split into accurate transcriptions and noisy transcriptions. Three tracks have been created for the challenge: 1) Closed set, 2) Self-supervised, and 3) Open set⁷. Results from several baseline systems are provided against which participants can compare their solutions. Through the challenge, we will release the results of a held-out blind test set, on which the submitted systems will be evaluated.

The remainder of the paper is summarized as follows: Section 2 provides an overview of the data collection, text preparation, statistical analysis of the entire data released through this challenge. Section 3 presents the experimental setup and baseline results. Finally, Section 4 concludes the paper.

2. Corpus and Challenge Definition

Through the GV challenge, 1000 hours of unlabelled *Mobile Vaani* recordings, together with 100 hours of labelled data (80 hours with noisy transcriptions, and 20 hours with reliable transcriptions) have been released. 5 hours of dev set and 2.8 hours of eval set having spontaneous natural noisy telephone recordings have also been released, on which our baseline numbers have been reported. The dev set is similar to the eval set and can be used to fine-tune the hyperparameters of the model.

2.1. Corpus Preparation

As described, Mobile Vaani, the voice-based participatory media platform, uses an IVR system as the primary channel for interaction. People call a unique district-specific phone number that is publicized by the Gram Vaani field teams, the IVR server then cuts the call and automatically calls the person back, thus making the call free for people. Over this call, users can record a voice message which they want to share, or listen to voice messages left by other users. These voice messages range across a number of domains: hyperlocal news reported by citizen journalists, questions on agriculture or health, grievances related to access to social entitlements, and also folk songs and poems. When a voice message is received, Gram Vaani’s content moderators review the message, and if deemed acceptable then the message is published on the platform and can be heard by other users, who can add comments or replies or contribute their own messages.

⁷More details about the GV ASR challenge can be found here: <https://sites.google.com/view/gramvaaniasrchallenge/home>. The datasets released through this challenge can be found on the following OpenSLR page: <https://www.openslr.org/118/>

2.1.1. Stage 1: Audio Collection

Voice recordings shared by Mobile Vaani users on about 25 district-specific instances comprise the ASR dataset. Only those voice recordings have been provided that were accepted for publication by the moderators. Recordings with very poor audio quality due to phone disturbances, or inarticulate message recordings, or objectionable content, were not included in the dataset.

It is well-known that automatic recognition of telephonic speech poses a unique set of challenges due to usual telephony hazards, such as channel distortions, clipping, speech truncation, and audio jump, present in the speech [15, 22, 23]. Therefore, crowd workers of the Uliza were asked to generate a variety of metadata for the recordings. A sample of the labels generated for each audio is listed in Table 2. Metadata about the district location is provided for around 65% of the unlabelled audio recordings. For most of the remaining voice recordings, the state location is provided. Metadata about the gender of the speaker (male/female) is also available for most of the labelled recordings. Other metadata includes whether the recording has multiple speakers, such as in the case of a citizen journalist interviewing another person, and occasional audio disturbances such as coughing or recording problems.

Table 2: *Metadata available in the corpus*

Metadata	Labels
Gender	#(Female), #(Male), #(Gender unknown)
Speaker age	#(Child), #(Adult), #(Senior), #(Age unknown)
Location	#(State name), #(District name)
Sentiment	#(Happy), #(Normal), #(Sad), #(Serious)
Accent/Dialect	#(Bihari), #(Bhojpuri), #(Accent unknown)
Background	#(Music), #(No speaking), #(Noise)

2.1.2. Stage 2: Text Preparation

Uliza recruits crowd workers from developing and low-income regions, and assigns micro-tasks such as translation and transcription to provide them with an income stream. 24 native Hindi-speaking crowd workers were recruited by Uliza from the

⁵<https://data.ldcil.org/hindi-raw-speech-corpus>

⁶<https://sites.google.com/view/indian-language-asrchallenge>

⁷<https://www.clarin.eu/resource-families/spoken-corpora>

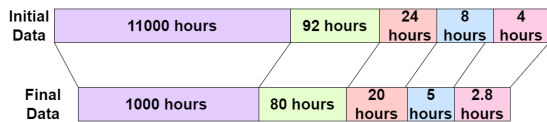


Figure 1: *Preprocessing of challenge dataset.*

same regions where Mobile Vaani is operating. 128 hours of voice recordings were labelled through the Uliza crowdsourcing platform. For each audio file, a single transcription file has been created. 5% of the transcriptions were randomly selected for verification by an Uliza expert and found to be of satisfactory quality, with noise largely arising from sporadic audio issues that made it hard to understand some words. Subsequently, 28% of the audio and their corresponding transcriptions, constituting of two subsets of 12 hours and 24 hours of speech data randomly selected from the entire 128 hours of labelled data were reviewed by a Gram Vaani team and corrected to provide a subset of accurate transcriptions. The 12 hours subset was then further rechecked multiple times to create dev and eval sets.

2.2. Challenge Dataset

As a part of the challenge, 1108 hours of real-world, spontaneous telephone speech recordings in local dialects of Hindi have been released. It includes 1000 hours of unlabelled training data, 100 hours of labelled training data, 5 hours of development data and 3 hours of blind test data. Figure 1 details the effort taken to prepare the challenge dataset from the original GV recordings.

Training sets: The training set consists of two type of datasets: labelled and unlabelled. For the labelled training set, initially, approximately 116 hours of speech data, including the 24 hours subset manually checked by GV team, have been selected. Then, Connectionist Temporal Classification (CTC) [24] alignments have been computed on the entire data. Based on the CTC alignments, all sentences with log likelihood value less than -2.5 have been selected. The final labelled training set consists of approximately 100 hours of speech data, out of which around 20 hours of data are from the 24 hours subset manually checked and corrected by GV team. On the other hand, the unlabelled training set consists of 1000 hours of speech data randomly selected from a total of approximately 11000 hours of unlabelled speech collected by the Gram Vaani team. We have made sure that sampling rate distributions in both labelled and unlabelled training sets remains similar to some extent.

Development set: The development set (or dev set) comprises 5 hours of labelled speech data. To ensure the accuracy of the text data, the dev data set has been gone through two rounds of manual check and subsequent programmatic checks as well. With respect to the original text (v0), the WER/CER (in %) of dev set after each round of manual check are 0.078/0.045 (v1) and 2.258/1.552 (v2), respectively.

Evaluation set: The evaluation Set (or eval set) consists of 2.81 hours of labelled speech data. Just like dev set, eval data has been finalized after several rounds of programmatic and manual checks. WERs/CERs (in %) of each round text (v1–v5) w.r.t the original text (v0): 2.507/1.965 (v1), 2.515/1.965 (v2), 3.811/3.221 (v3), 5.889/4.993 (v4), and 7.021/6.126 (v5).

2.3. Challenge Types

Three tracks have been created for the challenge: 1) Closed set: where the participants can use only the provided 100 hours of labelled train dataset to build ASR systems, 2) Self-supervised: where the participants can use only the Gram Vaani 1000 hours (unlabelled) and 100 hours labelled Train dataset, and 3) Open set: where the participants can use other resources in addition to the data provided in this challenge. Results from several baseline systems are provided against which participants can compare their solutions.

2.4. Corpus Analysis

As mentioned earlier, the entire audio data have been collected from different locations and environments. Therefore, it has some inconsistencies, such as handset variability, channel variability, and different sampling rates. Table 3 shows the sampling rate wise distribution of the entire dataset that have been released through this challenge. As in Table 3, the datasets includes audio with a mix of sampling rates ranging from 8 kHz to 48 kHz for both labelled and unlabelled data.

3. Experimental evaluations

In this section, experimental studies performed to validate the efficacy of training and test data are presented in detail.

3.1. Experimental setup

Hybrid TDNN-HMM: Kaldi ASR toolkit [25] is used to build three different ASR models with time-delay neural network (TDNN) architectures. The lattice free maximum mutual information (LF-MMI) objective function was used to train all three models [26]. The most of the parameters reported in Kaldi example recipes were used without adjustments. The model architectures of all three ASR systems are as follows:

TDNN-LSTM⁸: Consists of 10 TDNN layers of dimension 1260 stacked with 3 long-short term memory projected (LSTMP) [27] layers of dimension 1536.

CNN-TDNN-F⁹: It consists of six one-dimensional convolutional layers with ReLu batch normalization followed by 12 factored TDNN (TDNN-F) layers [28] of dimension 1536 with linear bottleneck dimension of 160.

TDNN-F¹⁰: It consists of 12 TDNN-F layers of dimension 1024 and a linear bottleneck dimension of 128.

Lexicon: A sequitur grapheme-to-phoneme (G2P) model [29] has been trained using a dictionary containing 30198 words, shared through a Hindi ASR challenge¹¹. Pronunciation of 19435 out-of-vocabulary (OOV) words present in the 100 hours labelled training data and 5 hours dev set have been generated using the trained G2P model. Since the transcript of 100 hrs labelled data is noisy, it is expected to have some erroneous words. Finally, pronunciations generated for OOVs are merged with the existing dictionary. It is to note that even after the addition of OOV pronunciations in the existing dictionary, the final

⁸https://github.com/kaldi-asr/kaldi/blob/master/egs/librispeech/s5/local/chain/tuning/run_tdnn_lstm_lb.sh

⁹https://github.com/kaldi-asr/kaldi/blob/master/egs/librispeech/s5/local/chain/tuning/run_cnn_tdnn_1a.sh

¹⁰https://github.com/kaldi-asr/kaldi/blob/master/egs/wsjs/s5/local/chain2/tuning/run_tdnn_1i.sh

¹¹<https://sites.google.com/view/indian-language-asrchallenge>

Table 3: Sample rate and duration statistics of the challenge dataset

SR (Hz)	Training data (Labelled)			Training data (Unlabelled)			Dev			Eval		
	#Utt.	Dur (h)	Avg Dur (s)	#Utt.	Dur (h)	Avg Dur (s)	#Utt.	Dur (h)	Avg Dur (s)	#Utt.	Dur (h)	Avg Dur (s)
8000	22781	61.76	9.76	47068	483.00	36.94	1169	3.07	9.45	587	1.64	10.06
16000	300	0.82	9.81	466	17.17	132.62	20	0.06	9.96	16	0.05	10.33
22050	0	0.00	0.00	2101	55.99	95.94	2	0.00	7.98	0	0.00	0.00
24000	0	0.00	0.00	60	1.85	111.15	0	0.00	0.00	0	0.00	0.00
32000	97	0.24	8.80	183	2.06	40.48	0	0.00	0.00	0	0.00	0.00
44100	12662	34.76	9.88	17712	400.03	81.31	623	1.70	9.84	371	0.97	9.43
48000	1312	3.32	9.11	2001	39.92	71.82	71	0.18	9.33	58	0.14	8.50
Total	37152	100.89	9.78	69591	1000.01	51.73	1885	5.02	9.58	1032	2.80	9.75

SR: Sampling Rate; Dur: Duration; Avg: Average; #Utt.: Number of utterance; Hz: Hertz; h: hours; s: seconds

eval set has an OOV rate of 8.2%.

Language model (LM): A 3-gram LM trained on 100 hours of labelled data with perplexity of 243.66 on dev set was built using SRILM toolkit [30].

End-to-end (E2E) ASR: For developing baselines for E2E ASR, ESPnet [31] and wav2vec 2.0 [32] toolkit are used.

ESPnet: A transformer [33] and a conformer [34] architecture based models are built. Both models are trained on a joint loss comprised of ctc-attention based multitask loss [35]. Both models have 12 layers of encoder and 6 layers of decoder blocks, each with 8 attention heads. The model configuration files (yaml files) and pre-trained model checkpoints are available on GitHub repository¹² of the GV challenge.

wav2vec 2.0: Fairseq [36] library is used to build wav2vec 2.0 based ASR systems. The pre-trained wav2vec model is trained (with contrastive loss and codebook diversity loss) on 1000 hours of unlabelled data, and then fine-tuned with CTC loss on 100 hours of labelled data. We have used wav2vec2 base and large models for fine-tuning.

3.2. Results and Discussion

Table 4 shows the WERs and CERs on both dev and eval sets obtained from ESPnet and Kaldi based ASR systems. For both transformer and conformer models, LM is not used. As evident from the Table 4, conformer performed better than transformer model on both dev and eval sets.

On the other hand, among the three HMM based models, TDNN-LSTM model is the largest in terms of the number of parameters, while TDNN-F being the smallest. But, as per the Table 4, TDNN-F outperforms the other two models. Since a large part of both the training audio data and the corresponding transcript are very noisy, overfitting seems a possible explanation for the smaller model performing better, but a more comprehensive study is required to make any solid claim.

Finally, the WERs obtained on dev set for wav2vec 2.0 based models have been reported in Table 5. The first two rows show the results on wav2vec 2.0 base model and a large model when 1000 hrs of unlabelled data are used for pretraining, and they are further fine-tuned using 100 hrs of labelled data. The last two rows, in Table 5, show the WERs on the dev set for two pretrained models from Open-Speech-EkStep¹³ and AI4Bharat¹⁴, respectively. Both pretrained models are fine-tuned using 100 hours of labelled data.

¹²https://github.com/anish9208/gramvaani_hindi_asr

¹³<https://github.com/Open-Speech-EkStep>

¹⁴<https://ai4bharat.org/>

Table 4: WER/CER values with respect to the different ASR systems trained on 100 hrs labelled data only

Model	WER / CER (%) on Test Data	
	Dev	Eval
Transformer (ESPnet)	37.3 / 19.6	35.6 / 20.8
Conformer (ESPnet)	33.4 / 17.5	32.9 / 19.0
TDNN-LSTM (Kaldi)	35.9 / 21.1	35.5 / 20.0
CNN-TDNN-F (Kaldi)	30.9 / 17.0	30.6 / 16.2
TDNN-F (Kaldi)	30.1 / 16.1	29.7 / 15.1

Table 5: WER obtained on dev set for wav2vec 2.0 ASR systems fine-tuned on 100 hours of labelled data

Model type	Pretraining data	WER (in %) on Dev set
Base	1000h (unlabelled)	35.9
Large	1000h (unlabelled)	35.8
Base	pretrained (OSE)	34.3
Base	pretrained (AI4B)	33.3

OSE: Open-Speech-EkStep; AI4B: AI4Bharat

4. Conclusion

This paper presents dataset and baseline system details of the Gram Vaani ASR challenge on a real-world spontaneous telephonic speech in regional variations of Hindi. The audio data for this challenge have been collected through the Mobile Vaani platform having users from all across India and, hence, it includes regional/dialectal variations of Hindi. A part of recordings are accompanied by their corresponding transcriptions done by crowd workers, along with a variety of metadata for the recordings including location, dialect, recording environments, and gender. Approximately 1000 hours of unlabelled data and 108 hours of labelled spontaneous speech data have been released through the challenge. Experimental evaluations reported in this paper validates the efficacy of the labelled and unlabelled speech data in the traditional TDNN-HMM as well as end-to-end ASR frameworks.

5. Acknowledgements

Special thanks to Sangeeta Saini, Paramita Panjal, Sayonee Chatterjee, Deepak Kumar, Rimjhim Kumari, Adwitee Verma, and Gram Vaani content moderators, for their invaluable contributions for data checking and correction. Mittul Singh for constructing a dictionary to convert transliterated Hindi words written to Roman text to Devanagari text.

6. References

- [1] "Census of India 2011," https://censusindia.gov.in/2011Census/C-16_25062018_NEW.pdf, Last accessed 07 October 2021.
- [2] A. Moitra, V. Das, G. Vaani, A. Kumar, and A. Seth, "Design lessons from creating a mobile-based community media platform in rural India," 2016.
- [3] A. Khullar, P. Panjal, R. Pandey, and et al., "Experiences with the introduction of ai-based tools for moderation automation of voice-based participatory media forums," 2021.
- [4] A. Khullar, P. Hitesh, S. Rahman, and et al., "Costs and benefits of conducting voice-based surveys versus keypress-based surveys on interactive voice response systems," 2021.
- [5] A. Khullar, M. Santosh, P. Kumar, and et al., "Early results from automating voice-based question-answering services among low-income populations in india," 2021.
- [6] P. Mudliar, J. Donner, and W. Thies, "Emergent practices around egnat swara, a voice forum for citizen journalism in rural india," vol. 9, no. 2, 2013.
- [7] N. Patel, D. Chittamuru, A. Jain, P. Dave, and T. Parikh, "Avaaj Otalo - a field study of an interactive voice forum for small farmers in rural India," 2010.
- [8] A. Vashistha, E. Cutrell, G. Borriello, and B. Thies, "Sangeet Swara: A Community-Moderated Voice Forum in Rural India," 2015.
- [9] S. Randhawa, T. Ahmad, J. Chen, and A. Raza, "Karamad: A voice-based crowdsourcing platform for underserved populations," 2021.
- [10] K. Samudravijaya, P. V. S. Rao, and S. S. Agrawal, "Hindi speech database," in *INTERSPEECH*, 2000.
- [11] A. Diwan, R. Vaideeswaran, S. Shah, and et al., "MUCS 2021: Multilingual and Code-Switching ASR Challenges for Low Resource Indian Languages," in *Proc. Interspeech 2021*, 2021, pp. 2446–2450.
- [12] S. Maity, A. K. V., K. S. Rao, and D. Nandi, "Iitkgp-mlilsc speech database for language identification," in *National Conference on Communications (NCC)*, 2012, pp. 1–5.
- [13] R. Kumar Vuddagiri, K. Gurugubelli, P. Jain, and et al., "IIITH-ILSC Speech Database for Indian Language Identification," in *Proc. 6th Workshop on Spoken Language Technologies for Under-Resourced Languages (SLTU 2018)*, 2018, pp. 56–60.
- [14] A. Canavan and G. Zipperlen, "CALLFRIEND Hindi LDC96s52," 1996. [Online]. Available: <https://catalog ldc.upenn.edu/LDC96S52>
- [15] J. Basu, M. S. Bepari, R. Roy, and S. Khan, "Real time challenges to handle the telephonic speech recognition system," in *Proc. of the 4th Int. Conf. on Signal and Image Processing 2012 (ICSIP 2012)*, M. S and S. S. Kumar, Eds. India: Springer India, 2013, pp. 395–408.
- [16] Appen Pty Ltd, Sydney, and Australia, "Iraqi Arabic conversational telephone speech LDC2006s45," 2006. [Online]. Available: <https://catalog ldc.upenn.edu/LDC2006S45>
- [17] Appen Pty Ltd, Sydney, and Australia, "Gulf Arabic conversational telephone speech LDC2006s43," 2006. [Online]. Available: <https://catalog ldc.upenn.edu/LDC2006S43>
- [18] G. Chen, S. Chai, G. Wang, and et al., "GigaSpeech: An evolving, multi-domain ASR corpus with 10,000 hours of transcribed audio," 6 2021. [Online]. Available: <http://arxiv.org/abs/2106.06909>
- [19] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 5206–5210.
- [20] P. K. O'Neill, V. Lavrukhin, S. Majumdar, and et al., "SPGISpeech: 5,000 hours of transcribed financial audio for fully formatted end-to-end speech recognition," 4 2021. [Online]. Available: <http://arxiv.org/abs/2104.02014>
- [21] N. Karpov, A. Denisenko, and F. Minkin, "Golos: Russian dataset for speech research," 6 2021. [Online]. Available: <https://arxiv.org/abs/2106.10161>
- [22] G. Saon, G. Kurata, T. Sercu, and et al., "English conversational telephone speech recognition by humans and machines," *CoRR*, vol. abs/1703.02136, 2017. [Online]. Available: <http://arxiv.org/abs/1703.02136>
- [23] J. Gauvain, L. Lamel, H. Schwenk, and et al., "Conversational telephone speech recognition," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing, (ICASSP '03)*, vol. 1, 2003, pp. I–I.
- [24] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks," in *Proc. of 23rd Int. Conf. on Machine Learning*, ser. ICML '06. New York, NY, USA: Association for Computing Machinery, 2006, p. 369–376. [Online]. Available: <https://doi.org/10.1145/1143844.1143891>
- [25] D. Povey, A. Ghoshal, G. Boulianne, and et al., "The Kaldi Speech recognition toolkit," in *Proc. ASRU*, December 2011.
- [26] D. Povey, V. P., D. Galvez, and et al., "Purely Sequence-Trained Neural Networks for ASR Based on Lattice-Free MMI," in *Proc. Interspeech 2016*, 2016, pp. 2751–2755.
- [27] H. Sak, A. W. Senior, and F. Beaufays, "Long short-term memory based recurrent neural network architectures for large vocabulary speech recognition," *CoRR*, vol. abs/1402.1128, 2014. [Online]. Available: <http://arxiv.org/abs/1402.1128>
- [28] D. Povey, G. Cheng, Y. Wang, and et al., "Semi-Orthogonal Low-Rank Matrix Factorization for Deep Neural Networks," in *Proc. Interspeech 2018*, 2018, pp. 3743–3747.
- [29] M. Bisani and H. Ney, "Joint-sequence models for grapheme-to-phoneme conversion," *Speech Communication*, vol. 50, no. 5, pp. 434–451, 2008. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0167639308000046>
- [30] A. Stolcke, "SRILM - An extensible language modeling toolkit," in *INTERSPEECH*, 2002.
- [31] S. Watanabe, T. Hori, S. Karita, T. Hayashi, J. Nishitoba, Y. Unno, N. E. Y. Soplin, J. Heymann, M. Wiesner, N. Chen, A. Renduchintala, and T. Ochiai, "Espnet: End-to-end speech processing toolkit," *CoRR*, vol. abs/1804.00015, 2018. [Online]. Available: <http://arxiv.org/abs/1804.00015>
- [32] A. Baevski, H. Zhou, A. M., and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *CoRR*, vol. abs/2006.11477, 2020. [Online]. Available: <https://arxiv.org/abs/2006.11477>
- [33] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *CoRR*, vol. abs/1706.03762, 2017. [Online]. Available: <http://arxiv.org/abs/1706.03762>
- [34] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu et al., "Conformer: Convolution-augmented transformer for speech recognition," *arXiv preprint arXiv:2005.08100*, 2020.
- [35] S. Kim, T. Hori, and S. Watanabe, "Joint ctc-attention based end-to-end speech recognition using multi-task learning," *CoRR*, vol. abs/1609.06773, 2016. [Online]. Available: <http://arxiv.org/abs/1609.06773>
- [36] M. Ott, S. Edunov, A. Baevski, and et al., "fairseq: A fast, extensible toolkit for sequence modeling," *CoRR*, vol. abs/1904.01038, 2019. [Online]. Available: <http://arxiv.org/abs/1904.01038>