OpenASR20: An Open Challenge for Automatic Speech Recognition of Conversational Telephone Speech in Low-Resource Languages

Kay Peterson[1]*, Audrey Tong*[1]*, Yan Yu*[2]

[1]National Institute of Standards and Technology, USA
[2]Dakota Consulting Inc., USA

kay.peterson@nist.gov, audrey.tong@nist.gov, yan.yu@nist.gov

Abstract

In 2020, the National Institute of Standards and Technology (NIST), in cooperation with the Intelligence Advanced Research Project Activity (IARPA), conducted an open challenge on automatic speech recognition (ASR) technology for low-resource languages on a challenging data type - conversational telephone speech. The OpenASR20 Challenge was offered for ten low-resource languages - Amharic, Cantonese, Guarani, Javanese, Kurmanji Kurdish, Mongolian, Pashto, Somali, Tamil, and Vietnamese. A total of nine teams from five countries fully participated, and 128 valid submissions were scored. This paper gives an overview of the challenge setup and procedures, as well as a summary of the results. The results show overall high word error rate (WER), with the best results on a severely constrained training data condition ranging from 0.4 to 0.65, depending on the language. ASR with such limited resources remains a challenging problem. Providing a computing platform may be a way to level the playing field and encourage wider participation in challenges like OpenASR.

Index Terms: automatic speech recognition, evaluation, low-resource language, conversational telephone speech, IARPA MATERIAL, Amharic, Cantonese, Guarani, Javanese, Kurmanji Kurdish, Mongolian, Pashto, Somali, Tamil, Vietnamese

1. Introduction

The performance of ASR technologies has been under investigation for decades. NIST started conducting benchmark ASR tests in the 1980s, beginning with English read speech in limited domains. In collaboration with the Spoken Language Program of Defense Advanced Research Projects Agency (DARPA), a series of Large Vocabulary Continuous Speech Recognition (LVCSR) tests took place in the 1990s, over time adding more data and data genres such as broadcast news and conversational speech, as well as other high-resource languages such as Arabic and Spanish. Overviews of the progression of these tests can be found in [1], [2]. The DARPA Effective, Affordable, Reusable Speech-to-Text (EARS) Program that ran from 2002 to 2004 also marked the beginning of the Rich Transcription (RT) evaluation series that ran from 2002 to 2009[3]. From 2006 to 2011, the DARPA Global Autonomous Language Extraction (GALE) program, while not having an ASR main focus, had ASR as a component and continued to further speech recognition evaluation. Performance improved over years of repeated testing, with some languages and genres reaching a similar performance as human transcription of speech.

As with human language technologies in general, ASR presents a bigger and different challenge if less is data available for training and development. There are over 7000 languages spoken in the world; the vast majority of them are considered low-resource. At the same time, improved performance for such languages is becoming increasingly more important, with more widespread use of devices and technologies in more languages for which technologies such as ASR are an integral part of the user experience. The performance of ASR also affects the quality of downstream applications such as machine translation. As human language technologies are maturing, the challenges of low-resource language technologies have come more into focus as a subject of increased research interest in the 2000s, as surveyed e.g. in [4]. The Workshop on Spoken Language Technologies for Under-Resourced Languages, held biannually since 2008 and most recently as the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages and Collaboration and Computing for Under-Resourced Languages (CCURL) in 2020, has been a venue for continued research into low-resource speech technologies including speech recognition.[5] The IARPA Babel program, which ran from 2012 to 2016, tested rapid development of ASR and keyword search technologies for conversational telephone speech in languages with little transcribed data available.[6]

Against this background, the OpenASR Challenge[7] hosted by NIST is an open evaluation challenge created as a spin-off of the IARPA Machine Translation for English Retrieval of Information in Any Language (MATERIAL) program[8], which encompasses more tasks, including cross-language information retrieval, language and domain identification, and summarization. For every year of MATERIAL, NIST supports a simplified, smaller scale evaluation open to all, focusing on a particular technology aspect of MATERIAL. The capabilities tested in the open challenges are expected to ultimately support the MATERIAL task of effective triage and analysis of large volumes of text and audio content in a variety of less-studied languages.

In 2020, the focus was on assessing the state of the art of ASR for low-resource languages. Ten low-resource languages were offered. The task was to perform ASR on speech data in these languages, producing written text output. OpenASR20 was implemented as a track of NIST's Open Speech Analytic Technologies (OpenSAT) evaluation series.[9] The evaluation made use of existing data, thus offering a low-cost option to perform an evaluation on a multitude of languages.

2. Challenge Setup

2.1. Languages

The OpenASR20 Challenge was offered for these ten low-resource languages (shorthand used in results in parentheses): Amharic (AMH), Cantonese (CAN), Guarani (GUA), Javanese (JAV), Kurmanji Kurdish (KUR), Mongolian (MON), Pashto (PAS), Somali (SOM), Tamil (TAM), and Vietnamese (VIE).

Apart from being considered low-resource, several of these

languages exhibit additional challenges that often come with low-resource status, such as inconsistent spelling conventions, underspecified orthographies, potential code-switching, and dialectal variations. Participants could attempt as many of the ten languages as they wished.

2.2. Training Conditions

The OpenASR20 Challenge offered two different training conditions, Constrained (CONSTR) and Unconstrained (UNCONSTR). For any language processed, participants were required to make a submission for the Constrained Training condition. The Unconstrained Training condition was optional, but encouraged.

As the name implies, the *Constrained* training condition limited training data resources to allow better cross-team comparisons. The constraint was quite severe; the only speech data permissible for training under this condition was a 10-hour training set specified by NIST in the language being processed. Additional text data in any language, either from the provided training set or publicly available resources, was permissible for training in the Constrained Training condition. Any such additional text training data had to be specified in sufficient detail in the system description. Under the *Unconstrained* training condition, participants were allowed to use speech data outside of the provided 10-hour training set, as well as additional publicly available speech and text training data from any language. This condition allowed for gauging performance gain from additional training data. Any such additional training data had to be specified in the system description.

Participants were not allowed to hire native speakers for data acquisition, system development, or analysis for either training condition.

2.3. Data

The data used in the challenge consisted of conversational telephone speech between two individuals speaking on a topic of their choosing from a list of suggested topics, for approximately ten minutes. Separate datasets for system training (BUILD), development (DEV), and evaluation (EVAL) were provided for each of the languages. The conversations were provided as separate channels for each speaker. For a few cases, only one side of the conversation was available due to not passing quality control checks. The data were sampled at 8kHz, 44.1kHz, or 48kHz and provided in .sph or .wav format, depending on the language, and were marked for gender, age, dialect, environmental condition, network, and phone model. The BUILD set also included a lexicon and a language specification document. Given the low-resource status of the languages, there are varying degrees of spelling variation and inconsistencies to be found in the data - adding an additional level of challenge, despite the fact that transcriptions were conventionalized in a standardized orthography. The data for most of the languages stem from the IARPA Babel program. More details regarding the audio data can be found in the IARPA Babel Data Specifications for Performers.[10] The Somali data sets stem from the IARPA MATERIAL program. For a larger overview of the corpora used in MATERIAL, including ASR and other data, see [11].

Prior to scoring OpenASR20 submissions, the data was preprocessed with a number of normalization steps. These steps are detailed in section 7.3 of the OpenASR20 Evaluation Plan.[12]

Table 1 lists the permissible BUILD (CONSTR and UNCONSTR), DEV, and EVAL data resources by modality, audio vs. text. The same limits held for all ten languages. The audio durations listed refer to conversations, not total length of audio files (which were separate for each speaker). Table 2 lists the number of audio files and the approximate number of words for each of the ten languages' BUILD, DEV, and EVAL data sets.

Table 1: *BUILD, DEV, and EVAL data resources. Hours of audio indicate duration of conversation, not total duration of audio files.*

Data set	**Audio**	**Text**
BUILD, CONSTR	10 hours	Unlimited
BUILD, UNCONSTR	Unlimited	Unlimited
DEV	10 hours	n/a
EVAL	5 hours	n/a

Table 2: *BUILD, DEV, and EVAL data set file and transcribed word counts.*

	BUILD		DEV		EVAL	
Lang.	**Files**	**Words**	**Files**	**Words**	**Files**	**Words**
AMH	122	64,391	123	65,763	64	33,241
CAN	120	96,943	120	95,893	69	50,087
GUA	134	68,984	124	71,285	62	36,199
JAV	122	64,047	122	68,765	62	33,638
KUR	133	82,418	132	77,930	66	38,479
MON	126	90,258	124	90,260	60	44,306
PAS	131	108,509	136	108,713	60	50,693
SOM	132	87,670	126	85,666	66	44,951
TAM	125	70,980	125	71,107	64	36,057
VIE	126	111,952	132	112,029	68	56,048

2.4. Metrics

The primary metric computed on the submitted output was Word Error Rate (WER), as implemented in the sclite tool of the Speech Recognition Scoring Toolkit SCTK available from NIST.[13] WER is computed as the sum of deletion, insertion, and substitution errors in the ASR output compared to a human reference transcription, divided by the total number of words in the human reference transcription:

$$WER = \frac{\#Deletions + \#Insertions + \#Substitutions}{\#ReferenceWords} \quad (1)$$

Character Error Rate (CER) was also computed. CER is calculated in the same way as WER, but at the character level instead of word level.

In addition, participants were required to self-report time and memory resources used by their ASR system(s). The time information was used to compute a run time factor (compared to the real time of the audio data processed) as a secondary metric, while the memory information was to provide the community with information about the resources required to use the ASR system(s).

3. Participation

Interest in the challenge was high. Originally, 28 teams from twelve countries registered to participate. Interest was fairly evenly distributed across the offered languages. In the end, a total of nine teams from five countries participated *fully*, meaning they made at least one valid CONSTR training submission

on at least one language's EVAL data set. The participating organizations, their countries, and their team names as used in the results are listed in Table 3.

Table 3: *Participants. Asterisk (*) indicates team did not submit required system description.*

Organization	Country	Team Name
Catskills Research Co.	USA	Catskills
Centre de Recherche Informatique de Montréal	Canada	CRIM
National Sun Yat-sen University	Taiwan	NSYSU-MITLab
Shanghai Jiao Tong University	China	Speechlab-SJTU
*Tal	China	*upteam
Tallinn University of Technology	Estonia	TalTech
Tencent	China	MMT
Tencent, Tsinghua University collaboration	China	TNT
Tsinghua University	China	THUEE

The total number of valid submissions across all participants, languages, and training conditions was 128. The number of teams per language and valid submissions by language and training condition are listed in Table 4.

Table 4: *Participation by language.*

Language	Teams	CONSTR	UNCONSTR
AMH	4	9	1
CAN	6	15	7
GUA	4	9	1
JAV	4	9	1
KUR	5	14	2
MON	6	14	6
PAS	3	7	0
SOM	5	13	1
TAM	4	8	1
VIE	4	10	0

All languages received interest and submissions. Cantonese and Mongolian had the most teams participating (6) and also received the highest number of submissions overall (22 for CAN, 20 for MON). Regarding training conditions, there were more submissions for the required CONSTR training condition than the optional UNCONSTR training condition, for all languages.

4. Results and Analysis

An overview of the OpenASR20 Challenge results is available online.[14] The following synopsis focuses on results for WER and, to allow for more meaningful comparisons, on the CONSTR training condition.

Table 5 lists the best WER score for each team that participated in that language, as well as the CER score for the same submission. Scores are ordered from best (lowest) to worst (highest) WER, by training condition, though late submissions and submissions by teams who did not submit the required system description are flagged as indicated and listed at the bottom of the respective category. Submissions for UNCONSTR training were overall rare and did not exist for all languages. For those languages with UNCONSTR submissions, the best UNCONSTR score was better than the best CONSTR score only in the case of Cantonese and Mongolian.

Table 5: *WER and CER for each team's best WER submission by language and training condition. Asterisk (*) indicates missing system description, dagger (†) indicates late submission; both cases are listed at the bottom of their category.*

Lang.	Condition	Team	WER	CER
AMH	CONSTR	TalTech	0.45	0.34
	CONSTR	THUEE	0.46	0.35
	CONSTR	Speechlab-SJU	1.02	0.89
	CONSTR	*upteam	*1.38	*1.36
AMH	UNCONSTR	Speechlab-SJTU	1.02	0.89
CAN	CONSTR	TNT	0.40	0.35
	CONSTR	THUEE	0.44	0.38
	CONSTR	TalTech	0.45	0.40
	CONSTR	NSYSU-MITLab	0.61	0.56
	CONSTR	Speechlab-SJTU	0.76	0.70
	CONSTR	*upteam	*1.31	*1.33
CAN	UNCONSTR	TNT	0.32	0.26
	UNCONSTR	Speechlab-SJTU	0.76	0.70
GUA	CONSTR	THUEE	0.46	0.42
	CONSTR	TalTech	0.47	0.43
	CONSTR	Speechlab-SJTU	0.99	0.96
	CONSTR	*upteam	*1.21	*1.21
GUA	UNCONSTR	Speechlab-SJTU	0.99	0.96
JAV	CONSTR	THUEE	0.52	0.52
	CONSTR	TalTech	0.54	0.54
	CONSTR	Speechlab-SJTU	0.94	0.94
	CONSTR	*upteam	*1.35	*1.35
JAV	UNCONSTR	Speechlab-SJTU	0.94	0.94
KUR	CONSTR	TalTech	0.65	0.61
	CONSTR	THUEE	0.67	0.62
	CONSTR	CRIM	0.75	0.71
	CONSTR	*upteam	*1.09	*1.08
	CONSTR	Speechlab-SJTU	1.12	1.05
	CONSTR	*upteam	*1.09	*1.08
KUR	UNCONSTR	Speechlab-SJTU	1.12	1.05
MON	CONSTR	THUEE	0.45	0.33
	CONSTR	MMT	0.45	0.33
	CONSTR	TalTech	0.47	0.35
	CONSTR	Speechlab-SJTU	0.97	0.80
	CONSTR	†TNT	†0.45	†0.35
	CONSTR	*upteam	*1.03	*1.00
MON	UNCONSTR	MMT	0.41	0.30
	UNCONSTR	TNT	0.46	0.34
	UNCONSTR	Speechlab-SJTU	0.97	0.80
PAS	CONSTR	TalTech	0.46	0.32
	CONSTR	THUEE	0.49	0.34
	CONSTR	*upteam	*1.37	*1.35
SOM	CONSTR	TalTech	0.59	0.59
	CONSTR	THUEE	0.60	0.60
	CONSTR	Speechlab-SJTU	1.04	1.04
	CONSTR	Catskills	1.14	1.14
	CONSTR	*upteam	*1.23	*.23
SOM	UNCONSTR	Speechlab-SJTU	1.04	1.05
TAM	CONSTR	TalTech	0.65	0.42
	CONSTR	THUEE	0.66	0.44
	CONSTR	Speechlab-SJTU	1.06	0.80
	CONSTR	*upteam	*1.35	*1.32
TAM	UNCONSTR	Speechlab-SJTU	1.06	0.80
VIE	CONSTR	TalTech	0.45	0.41
	CONSTR	THUEE	0.46	0.41
	CONSTR	NSYSU-MITLab	0.75	0.70
	CONSTR	*upteam	*1.41	*1.41

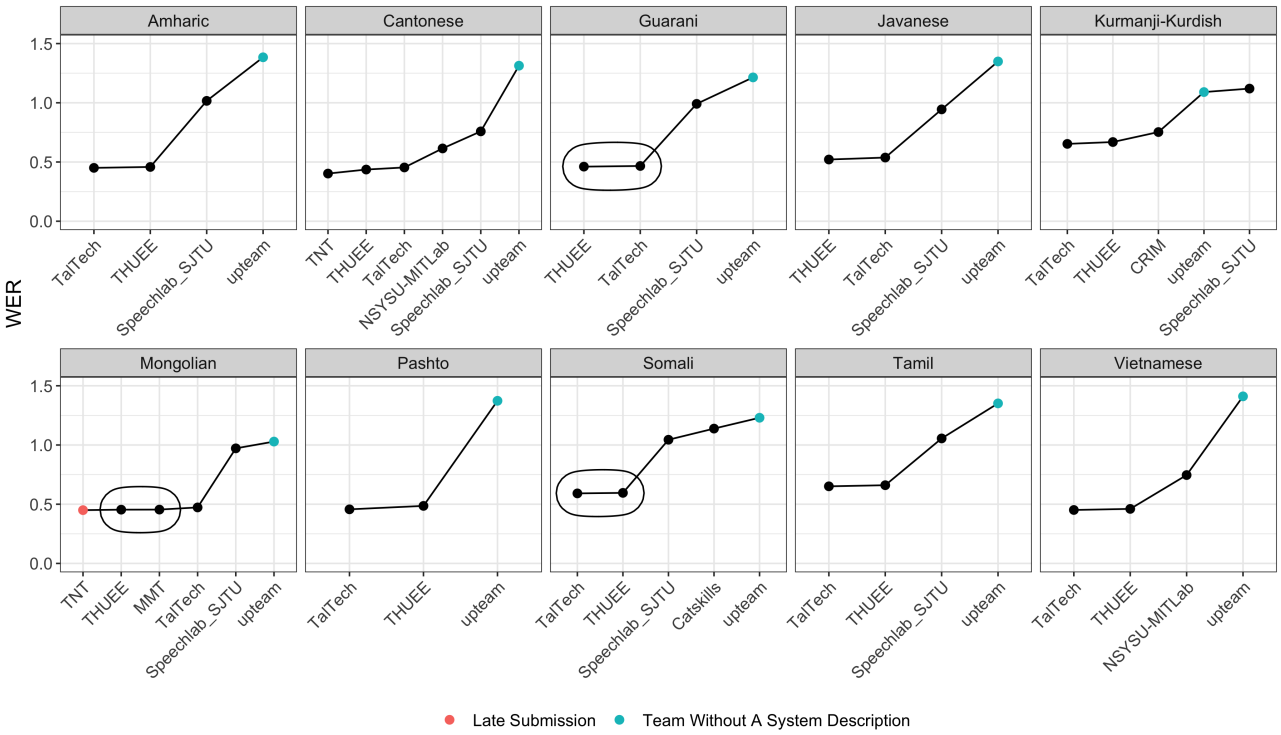

Figure 1: *Best CONSTR WER scores per language and team. Oval enclosing the data points indicates no significant difference at the 95% level.*

The best CONSTR WER scores per team and language were tested for significant differences using Student's t-test. Adjacent data points surrounded by an oval are scores that did not differ significantly according to this test ($p > 0.05$) as shown in Figure 1.

WER was overall high. The best (lowest) results per language range from 0.40 (Cantonese) to 0.65 (Kurmanji Kurdish), with many of the languages in between close to the 0.5 mark for the best result. As mentioned in the Data section, some metadata was available, but some of these factors only had a small amount of data - not enough for comparison. We focused on those with adequate samples (at least 20 audio files). Using the best CONSTR submission for each language, we explored the effect of dialect and gender on WER. For the dialect distinction, not every case had enough data in each category for a meaningful comparison. For those that did, no significant differences between WER for different dialects were found. For the gender distinction, a significant difference ($p < 0.05$) between WER for male vs. female speech was found only for Javanese and Pashto; in both cases the female speech resulted in significantly better WER than the male speech.

5. Discussion and Conclusions

The results indicate that ASR for low-resource languages, and in particular paired with a challenging data type such as conversational telephone speech, remains a difficult problem, as evidenced by the worse WER scores compared to more widely studied languages with large amounts of training data available. Some of the languages tested in OpenASR20 (CAN, KUR, PAS, TAM, VIE) were tested in the aforementioned Babel program from 2013-2016 as well, sourcing materials from the same larger data sets. This allows for some limited comparability, although the evaluation data sets used were not identical, and notably the training data was not limited as severely in Babel as in OpenASR20. While the results were not directly comparable, the best scores achieved in OpenASR20 were close to those achieved in Babel years before with more training data, and in some cases better. This indicates potential progress in that similar results can now be achieved with less resources.

We plan on performing error analyses in relation to challenges specific to the different language data, as well as, in collaboration with participants, in relation to training data usage.

While 28 teams originally registered for the challenge, only nine made valid submissions on at least one language's EVAL set. It will be useful to determine the reasons for this drop, and how to lower the barriers to entry to encourage wider participation. One consideration to become more accessible is to provide a computing platform with the infrastructure for participants to run their systems, instead of them having to rely on their own, which may vary widely between organizations. In addition, it may be useful to examine the usefulness of the UNCONSTR training condition with its low participation and minimal or no performance gain; its presence may represent another factor discouraging wider participation.

6. Disclaimer

These results presented in this paper are not to be construed or represented as endorsements of any participant's system, methods, or commercial product, or as official findings on the part of NIST or the U.S. Government. Opinions, interpretations, conclusions and recommendations are those of the authors and are not necessarily endorsed by the United States Government.

7. References

[1] D. S. Pallett, "The role of the National Institute of Standards and Technology in DARPA's Broadcast News continuous speech recognition research program," *Speech Communication*, vol. 37, no. 1, pp. 3–14, May 2002.

[2] A. F. Martin and J. S. Garofolo, "NIST speech processing evaluations: LVCSR, speaker recognition, language recognition," in *2007 IEEE Workshop on Signal Processing Applications for Public Security and Forensics*, 2007, pp. 1–7.

[3] NIST. (2009) Rich transcription evaluation. [Online]. Available: https://www.nist.gov/itl/iad/mig/rich-transcription-evaluation

[4] L. Besacier, E. Barnard, A. Karpov, and T. Schultz, "Automatic speech recognition for under-resourced languages: A survey," *Speech Communication*, vol. 56, pp. 85–100, Jan. 2014.

[5] D. Beermann, L. Besacier, S. Sakti, and C. Soria, Eds., *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)*. Marseille, France: European Language Resources association, May 2020.

[6] IARPA. (2016) Babel. [Online]. Available: https://www.iarpa.gov/index.php/research-programs/babel

[7] NIST. (2021) OpenASR Challenge. [Online]. Available: https://www.nist.gov/itl/iad/mig/openasr-challenge

[8] IARPA. (2021) MATERIAL. [Online]. Available: https://www.iarpa.gov/index.php/research-programs/material

[9] NIST. (2020) OpenSAT. [Online]. Available: https://www.nist.gov/itl/iad/mig/opensat

[10] ——. (2013) IARPA Babel data specifications for performers. [Online]. Available: https://www.nist.gov/system/files/documents/itl/iad/mig/IARPA_Babel_Performer-Specification-08262013.pdf

[11] I. Zavorin, A. Bills, C. Corey, M. Morrison, A. Tong, and R. Tong, "Corpora for cross-language information retrieval in six less-resourced languages," in *Proceedings of the Workshop on Cross-Language Search and Summarization of Text and Speech (CLSSTS2020)*. Marseille, France: European Language Resources Association, May 2020, pp. 7–13.

[12] NIST. (2020) OpenASR20 Challenge evaluation plan. [Online]. Available: https://www.nist.gov/document/openasr20-challenge-evaluation-plan

[13] ——. (2018) SCTK, the NIST scoring toolkit. [Online]. Available: https://github.com/usnistgov/sctk

[14] ——. (2021) OpenASR20 Challenge results. [Online]. Available: https://www.nist.gov/itl/iad/mig/openasr20-challenge-results