OpenASR21: The Second Open Challenge for Automatic Speech Recognition of 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 2021, the National Institute of Standards and Technology (NIST), in cooperation with the Intelligence Advanced Research Project Activity (IARPA), conducted OpenASR21, the second cycle of an open challenge series of automatic speech recognition (ASR) technology for low-resource languages. The OpenASR21 Challenge was offered for 15 low-resource languages. Five of these languages were new in 2021. OpenASR21 also introduced a case-sensitive scoring track on a wider set of data genres for three of the new languages, as a proxy for assessing ASR performance on proper nouns. The paper gives an overview of the challenge setup and results. Fifteen teams from seven countries made at least one required valid submission. 504 submissions were scored. Results show that ASR performance under a severely constrained training condition is still a challenge, with the best Word Error Rate (WER) ranging from 32% (Swahili) to 68% (Farsi). However, improvements over OpenASR20 were made by augmenting training data with perturbation and text-to-speech techniques along with system combination.

Index Terms: automatic speech recognition, evaluation, low-resource, conversational speech, news broadcast, topical broadcast, case sensitivity, IARPA MATERIAL, Amharic, Cantonese, Farsi, Georgian, Guarani, Javanese, Kazakh, Kurmanji Kurdish, Mongolian, Pashto, Somali, Swahili, Tagalog, Tamil, Vietnamese

1. Introduction

ASR technology performance is a long-standing human language technology (HLT) research area. NIST began conducting ASR tests in the 1980s with English read speech in limited domains. A series of Large Vocabulary Continuous Speech Recognition (LVCSR) tests were conducted in the 1990s in collaboration with the Spoken Language Program of Defense Advanced Research Projects Agency (DARPA). Over time, more data and data genres were added, as well as other high-resource languages such as Arabic and Spanish. [1] and [2] provide synopses of these tests over time. The DARPA Effective, Affordable, Reusable Speech-to-Text (EARS) program ran from 2002 to 2004 and also was the start of the 2002-2009 Rich Transcription (RT) evaluation series.[3] The 2006-2011 DARPA Global Autonomous Language Extraction (GALE) program included an ASR component that continued to further ASR evaluation.[4] Performance improved over years of repeat testing in these programs, reaching near human transcription accuracy for some languages and genres.

HLT applications, including ASR (in its own right and also as a feeder to downstream technologies such as machine translation), are becoming more widely available to more of the world's native speakers of low-resource languages. A need for improved HLT performance in low-resource settings, along with HLT generally maturing and expanding into more challenging areas, have led to increased research interest in the area of HLT for low-resource languages in recent years, as laid out in [5], for example. Several programs and workshops reflect this increased interest. From 2012 to 2016, IARPA conducted the IARPA Babel program, which tested rapid development of ASR and keyword search technologies for languages with little transcribed data available.[6] The Workshop on Spoken Language Technologies for Under-Resourced Languages (SLTU) has been 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.[7] The series is slated to continue in 2022 as the 1st Annual Meeting of the ELRA/ISCA Special Interest Group on Under-Resourced Languages (SIGUL).[1] ASR challenges for specific low-resource language groups have recently been held in the context of such series, for example for low-resource Indian languages.[8],[9] The Workshop on Resources and Technologies for Indigenous, Endangered and Lesser-resourced Languages in Eurasia (EURALI)[2] is another recent initiative, scheduled for 2022.

Against this background, NIST hosts the OpenASR Challenge series.[10] It originated as a spin-off of the IARPA Machine Translation for English Retrieval of Information in Any Language (MATERIAL) program, which encompassed more tasks, with an overall goal of effective triage and analysis of large amounts of data in less-studied languages.[11] Every year of MATERIAL was accompanied by a simplified smaller evaluation open to all, with a focus on a particular technology aspect relevant to MATERIAL. The open challenges aim to test core technology capabilities, such as ASR, that are expected to ultimately support the overall MATERIAL task.

The first OpenASR Challenge, OpenASR20, focused on assessing the state of the art of ASR for low-resource languages.[12],[13] The task was to perform ASR on speech data in these languages, producing written text output. This challenge motivated the special session "OpenASR20 and Low Resource ASR Development" at INTERSPEECH 2021.[14] Due to the interest in OpenASR20, a second cycle of this challenge was conducted a year later as OpenASR21. Fifteen low-resource languages were offered. All ten OpenASR20 languages were repeated with the same datasets; five were new additions for 2021. Also new in 2021 were dedicated datasets from more varied genres that were offered with a case-sensitive scoring track for some of the languages, as a proxy for assessing ASR performance on proper nouns. The challenge series continues to make use of only existing data, thus offering an afford-

---

[1]https://sigul-2022.ilc.cnr.it
[2]https://sites.google.com/view/eurali/

able evaluation option for multiple languages. The OpenASR Challenges are implemented as a track of NIST's Open Speech Analytic Technologies (OpenSAT) evaluation series.[15]

2. Challenge Setup

This section gives an abbreviated overview of the protocols used to conduct OpenASR21. For a more detailed description, please refer to the OpenASR21 Challenge Evaluation Plan.[16]

2.1. Languages and Tracks

The OpenASR21 Challenge was offered for 15 low-resource languages (shorthand in parentheses); all ten OpenASR20 languages as well as five new ones: Amharic (AMH), Cantonese (CAN), Farsi (FAR) (new), Georgian (GEO) (new), Guarani (GUA), Javanese (JAV), Kazakh (KAZ) (new), Kurmanji Kurdish (KUR), Mongolian (MON), Pashto (PAS), Somali (SOM), Swahili (SWA) (new), Tagalog (TAG) (new), Tamil (TAM), and Vietnamese (VIE). Teams could attempt as many of the 15 languages as they wished.

OpenASR21 had two scoring tracks, each with its own datasets. The case-insensitive scoring (CIS) track was offered for all 15 languages. System output on the CIS EVAL datasets was scored using case-insensitive scoring. For the ten languages repeated from OpenASR20, the datasets remained identical, allowing for comparability over time. The case-sensitive scoring (CSS) track was offered for three of the new languages - KAZ, SWA, and TAG. System output on these three datasets was scored using case-sensitive scoring; words capitalized differently from the reference transcript were not counted as a match. For any language attempted, processing the CIS EVAL dataset was mandatory. Processing the additional CSS EVAL dataset, if available, was optional.

2.2. Training Conditions

The OpenASR21 Challenge offered three different training conditions: Constrained (CONSTR), Constrained-plus (CONSTR+), and Unconstrained (UNCONSTR). For any language processed, a CONSTR training condition was mandatory. The other conditions were optional.

The CONSTR condition severely limited training data resources. The only speech data permissible for training under this condition was a specific ten-hour set provided by NIST for the language being processed. Additional text data from any language was permissible for training if provided under the challenge or publicly available. The CONSTR condition allowed for comparison across teams.

The CONSTR+ condition, new in 2021, followed the same training data restrictions as CONSTR, but additionally allowed publicly available and previously existing speech pretrained models, created from unlabeled speech data in any language, and created from labeled speech data in any language except the language being processed.

In the UNCONSTR condition, teams were allowed to use speech data outside of the provided 10-hour training set and additional publicly available speech and text data from any language. UNCONSTR training allowed for evaluating the effect of additional training data on performance.

Teams were required to declare any permissible additional data used in their system description.

Table 1: *BUILD, DEV, and EVAL data resources.*

Dataset	Audio	Text
BUILD, CONSTR and CONSTR+	10 hours	Unlimited
BUILD, UNCONSTR	Unlimited	Unlimited
DEV	10 hours	n/a
EVAL	5 hours	n/a

Table 2: *BUILD, DEV, and EVAL dataset file and transcribed word counts. datasets are CIS unless specified as CSS.*

	BUILD		DEV		EVAL	
Language	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
FAR	120	62,909	120	67,122	384	38,899
GEO	123	68,870	124	69,192	70	31,862
GUA	134	68,984	124	71,285	62	36,199
JAV	122	64,047	122	68,765	62	33,638
KAZ	130	61,005	140	60,086	66	32,793
KAZ CSS	154	66,924	155	68,211	212	32,703
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
SWA	128	63,016	142	62,247	74	31,341
SWA CSS	159	66,623	159	69620	188	36,557
TAG	132	64,298	146	64,334	96	32,412
TAG CSS	169	79,324	167	82,836	264	39,349
TAM	125	70,980	125	71,107	64	36,057
VIE	126	111,952	132	112,029	68	56,048

2.3. Data

The data used in the challenge consisted of speech in three different genres: conversational speech (CS), news broadcast (NB), and topical broadcast (TB). The CIS datasets consisted of only CS data while the CSS datasets consisted of a mix of all three genres. Teams received distinct datasets for system training (BUILD), development (DEV), and evaluation (EVAL) for each of the languages. The data were sampled at 8kHz, 44.1kHz, or 48kHz and provided in .sph or .wav format, depending on the language and genre. The BUILD set also included a lexicon and a language specification document. The data for most of the languages were originally collected for the IARPA Babel program and are described in more detail in the IARPA Babel Data Specifications for Performers.[17] The Somali data stemmed from the IARPA MATERIAL program. [18] gives a more detailed overview of the MATERIAL corpora.

Table 1 lists the BUILD, DEV, and EVAL audio and text data amounts per language. Table 2 lists the number of audio files and the approximate number of words in the BUILD, DEV, and EVAL datasets for each language.

2.4. Metrics

The submitted text output was scored by computing Word Error Rate (WER) as the primary metric, as implemented in the sclite tool of NIST's Speech Recognition Scoring Toolkit SCTK.[19] WER is the sum of errors (deletions, insertions, and substitutions) in the ASR output compared to a human reference transcription, divided by the total number of words in the reference transcription:

$$WER = \frac{\#Deletions + \#Insertions + \#Substitutions}{\#ReferenceWords}$$

$$(1)$$

For the CSS evaluation datasets, WER was calculated case-sensitively. Character Error Rate (CER) was also computed. CER is computed like WER, but at the character instead of word level.

Teams also had 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. The memory resources provided insight into the resources required to use the ASR system(s). CER and self-reported time and memory resources are not reported in this paper.

3. Participation

Originally, 26 teams from 13 countries registered to participate. Ultimately, 15 teams from seven countries made at least one valid CONSTR submission on at least one language's EVAL dataset, as required. The total number of valid submissions was 504. A list of participating organizations with their team names as used in the results is provided on NIST's OpenASR21 Challenge Results page.[20]

4. Results and Analyses

In this section, we present OpenASR21 key results and analyses for teams with at least one valid CONSTR EVAL submission. Figure 1 shows the best WER score for each team in the CIS track. The languages are ordered by best WER in the CON-STR training condition. Within each language, teams are ordered from best to worst WER. We note performance varies by language. In general, performance improves for training conditions with more data, as expected, with CONSTR+ having better WER than CONSTR, and UNCONSTR having better WER than CONSTR+. The CONSTR+ condition has the largest gain, which is encouraging since the additional data came from existing pretrained models at no additional cost.

As noted in the setup section, the purpose of the CSS track is to serve as a proxy for ASR performance on proper nouns, which we equate to capitalized words. Figure 2 shows the best WER score for each team in the CSS track. We see trends similar to those in the CIS track, in that ASR performance differs by language and by training condition, but the language and training effect are less pronounced in the CSS track.

As mentioned in the data section, in addition to CS speech, the CSS dataset included NB and TB genre data to provide the test data with enough proper noun coverage for the results to be meaningful. As a result, the dataset used in the CSS track is not the same as that used in the CIS track, and thus we cannot make a direct comparison on the CS data portion to see the effect of casing. Instead, we scored the CSS submissions with two scoring settings, preserving vs. not preserving case. The absolute difference in WER between preserving case and not for the best CONSTR track submissions ranges between 0.82% and 2.50%.

To gain a clearer understanding of ASR performance on proper nouns, we computed four additional statistics: (1) the number of capitalized words the system got correct, (2) the number of capitalized words the system got wrong, (3) the number of capitalized words the system got wrong due to case, and (4) the number of capitalized words the system got wrong not due to case. These counts are shown as percentages in Figure 3
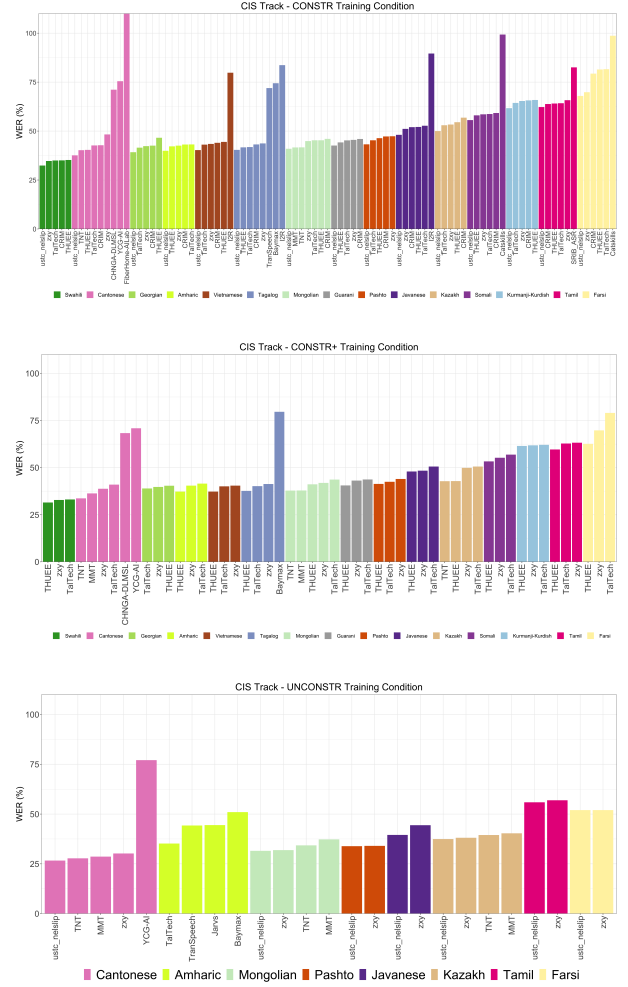

Figure 1: *Best WER in the CIS track achieved by each team for the language(s) and training condition(s) they participated in (CONSTR top, CONSTR+ middle, UNCONSTR bottom). Languages ordered by best WER in the CONSTR condition. Y-axis is limited to 100% for legibility.*

and Figure 4, respectively. We note that the percentages of capitalized words that the systems got wrong are higher than those they got right. Of those they got wrong, the majority was due to getting the word wrong rather than mixing up the word casing, which suggests that recognizing proper nouns is still challenging even if case is not a concern.

Because the CSS dataset had a variety of genres, we also looked at genre effect. Figure 5 shows the WER for the three genres. We note that CS is more challenging than NB and TB, which presumably is due to the less structured nature of CS speech.

Ten of the 15 languages in OpenASR21 are repeats from OpenASR20, allowing us to track performance over time. Figure 6 shows the best WER achieved for each language across the two OpenASR challenges. The sign test was performed and found that the difference between the two challenges is significant at the 95% confidence level for all languages. The gain, however, was achieved by a different team, so there could also be a team effect. The three best scoring teams from OpenASR20 also participated in OpenASR21, and the results of
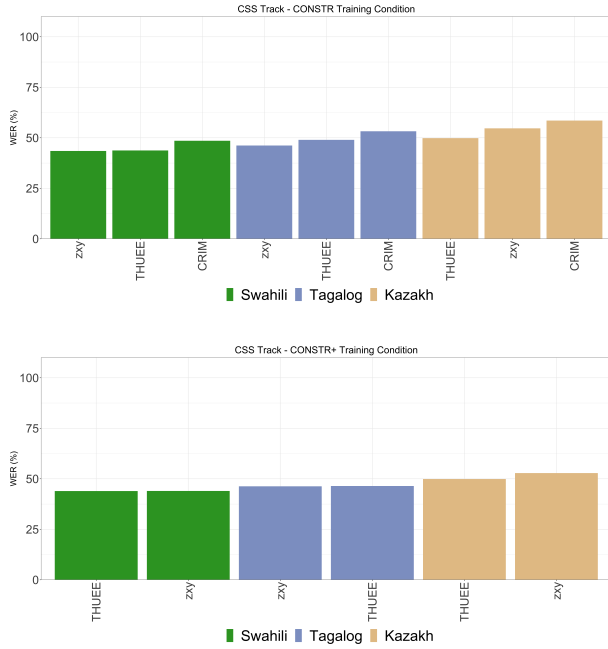
CSS Track - CONSTR Training Condition

CSS Track - CONSTR+ Training Condition

Figure 2: *Best WER in the CSS track per language and training condition (CONSTR top, CONSTR+ bottom, no participation for UNCONSTR). Languages ordered by best CONSTR WER.*
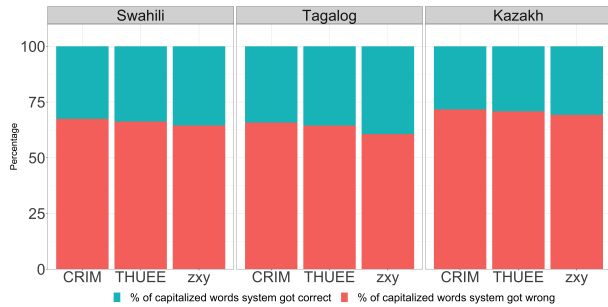

Figure 3: *Percentage of capitalized words that systems got right (cyan) and wrong (red) from the best scoring CSS submissions in the CONSTR training condition.*

Figure 4: *Percentage of capitalized words that systems got wrong due to case (pink) or not (red) from the best scoring CSS submissions in the CONSTR training condition.*

Figure 5: *WER across the three genres for CONSTR training in the CSS track.*

Figure 6: *Best WER achieved for each language across OpenASR20 and OpenASR21 for CONSTR training in the CIS track.*

the sign test on their 2020 vs. 2021 WERs indicate that the gains from TalTech in Amharic and Vietnamese and THUEE in Guarani are significant at the 95% confidence level. The top scoring team in OpenASR21 attributes the key gain to be effective use of additional data generation using text-to-speech and data perturbation techniques, along with system combination.

5. Conclusions

The OpenASR21 Challenge was the second OpenASR Challenge NIST conducted in collaboration with IARPA to assess current ASR performance for languages with low training data resources. Fifteen languages with three training conditions were offered. OpenASR21 also introduced a case-sensitive scoring track as a proxy to ASR performance on proper nouns for three of the 15 languages. 15 teams from seven countries participated meeting the submission requirements. The results show that ASR performance in the CONSTR training condition, in which all teams were provided with only ten hours of speech data for
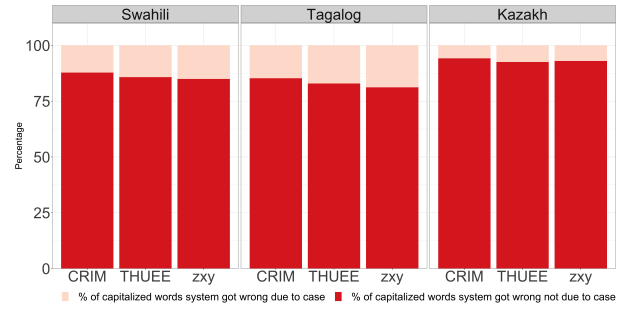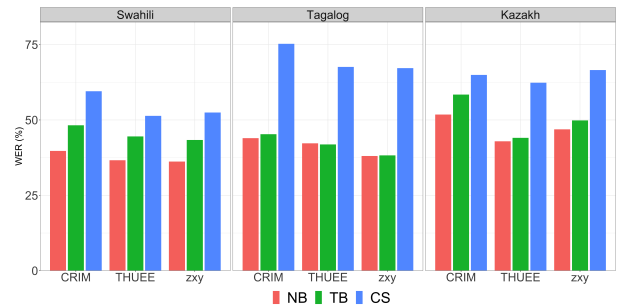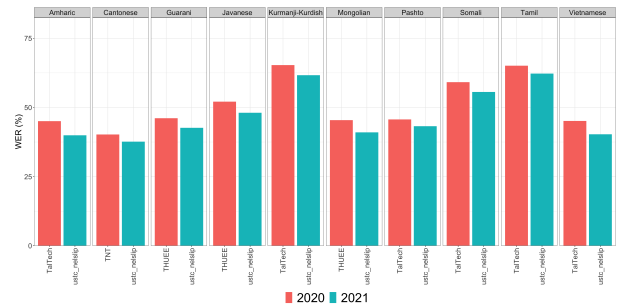
training, is still a challenge, with some languages presenting more difficulty than others. Moreover, improvement was seen over the best WER for all repeat languages from OpenASR20, and is attributed to augmenting the data using text-to-speech and perturbation techniques as well as system combination.

6. Disclaimer and Acknowledgment

These results presented in this paper are not to be construed or represented as endorsements of any team's system, methods, or commercial product, or as official findings on the part of NIST, IARPA, 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] J. Olive, C. Christianson, and J. McCary, *Handbook of Natural Language Processing and Machine Translation: DARPA Global Autonomous Language Exploitation*, 1st ed. Springer Publishing Company, Incorporated, 2011.

[5] 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.

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

[7] 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.

[8] B. M. L. Srivastava, S. Sitaram, R. Kumar Mehta, K. Doss Mohan, P. Matani, S. Satpal, K. Bali, R. Srikanth, and N. Nayak, "Interspeech 2018 Low Resource Automatic Speech Recognition Challenge for Indian Languages," in *Proc. The 6th Intl. Workshop on Spoken Language Technologies for Under-Resourced Languages*, 2018, pp. 11–14.

[9] A. Diwan, R. Vaideeswaran, S. Shah, A. Singh, S. Raghavan, S. Khare, V. Unni, S. Vyas, A. Rajpuria, C. Yarra, A. Mittal, P. K. Ghosh, P. Jyothi, K. Bali, V. Seshadri, S. Sitaram, S. Bharadwaj, J. Nanavati, R. Nanavati, and K. Sankaranarayanan, "MUCS 2021: Multilingual and Code-Switching ASR Challenges for Low Resource Indian Languages," in *Proc. Interspeech 2021*, 2021, pp. 2446–2450.

[10] NIST. (2021) Open Automatic Speech Recognition (OpenASR) Challenge. [Online]. Available: https://www.nist.gov/itl/iad/mig/openasr-challenge

[11] IARPA. (2021) Machine Translation for English Retrieval of Information in Any Language (MATERIAL). [Online]. Available: https://www.iarpa.gov/index.php/research-programs/material

[12] K. Peterson, A. Tong, and Y. Yu, "OpenASR20: An Open Challenge for Automatic Speech Recognition of Conversational Telephone Speech in Low-Resource Languages," in *Proc. Interspeech 2021*, 2021, pp. 4324–4328.

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

[14] ISCA. (2021) INTERSPEECH 2021. [Online]. Available: https://www.interspeech2021.org/

[15] NIST. (2020) Open Speech Analytic Technologies Evaluation Series (OpenSAT). [Online]. Available: https://www.nist.gov/itl/iad/mig/opensat

[16] ——. (2021) OpenASR21 Challenge Evaluation Plan. [Online]. Available: https://www.nist.gov/document/openasr21-challenge-evaluation-plan

[17] ——. (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

[18] 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.

[19] NIST. (2018) Speech Recognition Scoring Toolkit (SCTK), the NIST Scoring Toolkit. [Online]. Available: https://github.com/usnistgov/sctk

[20] ——. (2022) OpenASR21 Challenge Results. [Online]. Available: https://www.nist.gov/itl/iad/mig/openasr21-challenge-results