FLEURS: FEW-SHOT LEARNING EVALUATION OF UNIVERSAL REPRESENTATIONS OF SPEECH

Alexis Conneau[†‡1], *Min Ma*[†2], *Simran Khanuja*[†3], *Yu Zhang*[2],
Vera Axelrod[2], *Siddharth Dalmia*[2], *Jason Riesa*[2], *Clara Rivera*[2], *Ankur Bapna*[‡2]

[1]Meta AI Research, [2]Google Research, [3]Carnegie Mellon University

aconneau@fb.com, {minm, ngyuzh, vaxelrod, sdalmia, ankurbpn}@google.com, skhanuja@cs.cmu.edu

ABSTRACT

We introduce FLEURS, the Few-shot Learning Evaluation of Universal Representations of Speech benchmark. FLEURS is an n-way parallel speech dataset in 102 languages built on top of the machine translation FLoRes-101 benchmark, with approximately 12 hours of speech supervision per language. FLEURS can be used for a variety of speech tasks, including Automatic Speech Recognition (ASR), Speech Language Identification (Speech LangID), Speech-Text Retrieval. In this paper, we provide baselines for the tasks based on multilingual pre-trained models like speech-only w2v-BERT [1] and speech-text multimodal mSLAM [2]. The goal of FLEURS is to enable speech technology in more languages and catalyze research in low-resource speech understanding.[1]

Index Terms— Massively Multilingual Speech Recognition, Low-Resource Language Dataset, Speech Language Identification, Speech Information Retrieval, Few-/Zero- Shot Learning

1. INTRODUCTION

Speech technology has been rapidly evolved in the past few years, with striking achievements nourished from self-attention models [3, 4], pre-training approaches [5, 6, 7, 2], and massively multilingual speech models [8, 9]. Methods such as wav2vec 2.0 [5] have demonstrated strong performance on the multilingual LibriSpeech dataset [10], in particular in the few-shot learning scenario with only 10 minutes of labeled data [11]. The recent scaled-up multilingual wav2vec 2.0 model, XLS-R [12], has expanded similar few-shot capabilities to many more languages, including low-resource ones. By leveraging large-scale pre-training datasets like Multilingual LibriSpeech (MLS) [13] and VoxPopuli [14], XLS-R also provides representations that can be used across downstream tasks, with significant gains over previous baselines on speech recognition, translation and classification. More recently, mSLAM [2], a joint speech and text multilingual pretrained model, outperformed XLS-R on speech translation and ASR and improved over speech-only baselines on Speech-LangID.

Such recent progress has been made possible with the release of both large-scale pre-training and evaluation datasets like Multilingual LibriSpeech [13], VoxPopuli [14], CoVoST-2 [15], CommonVoice [16], and the re-use of existing datasets like BABEL [17]. However, there are a couple of shortcomings of the existing corpora: First, many datasets only contain a small and often disparate set of languages, as shown in Table 1. The majority of human spoken languages are often not covered, yet many of them have millions of active speakers. Second, since it requires great efforts to obtain high-quality human transcriptions, the amount of supervised speech data is usually limited. For example, only 1.8k hours out of 400k hours in VoxPopuli [14] were transcribed by human. Alternatively, even we can produce transcripts by ASR models, it may reinforce the system errors. Moreover, corpora may lack of diversity in the content domains, speakers, etc.

Fleurs builds on Flores - a multi-way parallel text translation dataset across 102 languages, which allows us to collect spoken utterances for each language directly and build a multi-way parallel speech corpus to enable research on ASR, LangID and Speech-to-speech translation. This is in contrast to most existing Speech translation corpora that use a pipelined approach - taking spoken speech, transcribing it, then translating it, then collecting spoken utterances in the target language and before finally aligning these segments back to speech segments in the original language - which is often a noisy process. Instead, we reached out to vendors who recruited native speakers in each language. We ensured at least 7 speakers per language, and each sentence was spoken by 3 speakers. Speakers access the recording remotely from their homes and were instructed to record audio in a quiet environment either on their Android phone or a desktop/laptop computer. We removed spoken utterances that did not match the transcript or were too noisy in the quality validation phase with a separate set of evaluators.

[1]We publicly released the FLEURS dataset via TFDS at `https://tensorflow.org/datasets/catalog/xtreme_s` and Huggingface at `https://hf.co/datasets/google/fleurs`. [†]Equal Contributions. [‡]Equal Advising Contributions. Work done while Alexis and Simran were at Google.

978-1-6654-7189-3/22/$31.00

With FLEURS, we aim to address these issues and to catalyze research towards building massively multilingual speech and text representations and their evaluation on a variety of tasks. While there are a few other datasets containing n-way parallel speech and text, including Europarl-ST [18], MuST-C [19], mTEDx [20] and the CVSS corpus [21], to the best of our knowledge, FLEURS is the only dataset spanning over 100 languages enabling research on a diverse set of languages and domains. FLEURS is well-suited for several downstream tasks including ASR, Speech-to-Text and Speech-to-Speech Translation, Speech LangID, and Multilingual Speech-to-Speech and Speech-to-Text Retrieval. We compare FLEURS to existing common public multilingual corpora in Table 1.

A few key properties of FLEURS to note:

- FLEURS contains n-way parallel speech and text in 102 languages, with a particular focus on the low-resource languages ($> 80\%$ data), across seven geographical groups.
- FLEURS provides natural human speech and high quality transcripts for each language with strong quality control. Three speakers spoke each sentence, and then evaluators validated whether each of the spoken utterances matched the transcript.
- FLEURS spans a wide range of language groups, writing systems, and linguistic families.
- FLEURS uses a bottom up approach of collecting spoken utterances for aligned segments, while most other datasets are aligned at a document level with automatic segmentation and alignment for segments; we applied strict quality control to deliver high-quality supervised parallel data.

In addition to describing the dataset, we provide baselines for Speech-LangID, ASR and Speech-Text retrieval (both Speech-to-Text and Text-to-Speech retrieval) by fine-tuning the multilingual w2v-BERT [1] and the mSLAM [2] models on these tasks.

2. DATASET

2.1. Speech Data Collection

We start with the FLoRes-101 dataset.[2] FLoRes-101 contains 3001 sentences extracted from English Wikipedia and these sentences have been translated in 101 languages by human translators. Because the test set of FLoRes-101 is not publicly available, we only use the dev and devtest sets, which contain 2009 sentences in total. The data is split into train, development (dev) and test sets with disjoint speakers, with a target ratio of utterance numbers of 7:1:2. For each sentence in the 102 languages (101 counted in FLoRes plus English), we collected three recordings by three different native speakers, with at most 70% from any one gender where possible.

We apply careful quality control for the data: each recording was evaluated by additional workers to assess whether it corresponded to the input sentence. Invalid recordings were discarded, leaving zero to three recordings per sentence in the final dataset. In the first version of the dataset, 21.5% of the sentences are missing because none of the three recordings were validated. We plan to fill these gaps in the future versions of the dataset. All recordings are kept as they-are, from quiet or noisy environments, without data augmentations. The speech recordings use a sampling rate of 16kHz, the sample encoding is 32-bit float PCM. All the utterances are within 30 seconds.

2.2. Textual Data

For source transcripts, we reuse the transcripts produced by human annotators from [24]. We maintain the English translated transcripts, which are useful for tasks such as multimodal speech translation evaluations.

The variety of orthographic symbols of languages complicates the tokenization process. For example, Chinese text in both traditional and simplified scripts does not have space between tokens. Depending on the transcribers, Japanese and Korean may or may not contain space irregularly. To ease the pain for other researchers and facilitate apple-to-apple comparisons and reproducibility, we provide the tokenized versions of the sentences. We apply NFC (`https://en.wikipedia.org/wiki/Unicode_equivalence#Normalization`) and then FST [25] normalization to each sentence, lower-case, normalize and remove punctuations. We also split words into characters, and use the symbol | to indicate word boundaries. For each sentence, three versions are provided: the original raw transcript (SRC_RAW), the preprocessed version (SRC_NORM) and its character-based version (SRC_CHAR), which should be used for ASR.

To establish the baseline, we use a universal vocabulary of characters as our modeling and evaluation unit in this paper. Among the possible modeling units ($e.g.$ character, wordpiece, sentence-piece. etc.) for massively multilingual ASR, this requires the least resources to build, and matches a common evaluation metric ($i.e.$ character error rate (CER)).

2.3. Taxonomy and Statistics

By construction, FLoRes sentences also cover a diversity in domains from Wikipedia, including nature, politics, science, travel, sports etc. Each sentence also has an associated integer "index" between 1 and 2009, which can be used to recover the n-way parallelism from one language to another ($i.e.$ sentence i in language A is the translation of sentence i in language B).

There are multiple ways to categorize languages. The languages of FLEURS cover 16 language families (distribution

[2]Note: For clarity we have renamed FLoRes "Chinese (Simp)" to "Mandarin Chinese" (code "cmn") and "Chinese (Trad)" to "Cantonese Chinese" (code "yue").

Data	Languages	Duration	Domains	Speech Type	Transcripts	Parallel text	Parallel speech
Europarl-ST [18]	6	0.5k hrs	Parliament	Spontaneous	Yes	Yes	No
MLS [13]	8	50.5k hrs	Audiobook	Read	Yes	No	No
MuST-C [19]	9	0.4k hrs	TED talks	Spontaneous	Yes	Yes	No
mTEDx [20]	9	1k hrs	TED talks	Spontaneous	Yes	Yes	No
CVSS [21]	22	1.1k h	Open domain	Read/Synthetic	Yes	Yes	Yes
CoVoST-2 [15]	22	2.9k hrs	Open domain	Read	Yes	Yes	No
VoxPopuli [14]	24	400k hrs*	Parliament	Spontaneous	Partial	Partial	Partial
BABEL [17]	25	2k hrs	Conversational	Spontaneous	Yes	No	No
CommonVoice [16]	93	15k hrs	Open domain	Read	Yes	No	No
Voxlingua-107 [22]	107	6.6k hrs	YouTube	Spontaneous	No	No	No
CMU Wilderness [23]	700	14k hrs	Religion	Read	Yes	Yes	Yes
FLEURS (this work)	102	1.4k hrs	Wikipedia	Read	Yes	Yes	Yes

Table 1. A comparison of commonly used datasets for multilingual speech representation learning, ASR, Speech Translation and Speech-LangID. CommonVoice statistics as on 24th May 2022. *VoxPopuli only has 1.8k hours transcribed speech.

Data Statistics	WE	EE	CMN	SSA	SA	SEA	CJK	All
train speech hours	231h	134h	116h	237h	124h	112h	32h	987h
dev speech hours	29h	18h	14h	24h	16h	14h	4h	120h
test speech hours	68h	43h	33h	58h	37h	35h	9h	283h
train transcript tokens	1475k	772k	630k	1072k	699k	525k	405k	5578k
dev transcript tokens	184k	107k	75k	116k	93k	65k	51k	692k
test transcript tokens	443k	260k	181k	272k	210k	158k	116k	1640k

Table 2. Statistics for speech and transcript data in FLEURS.

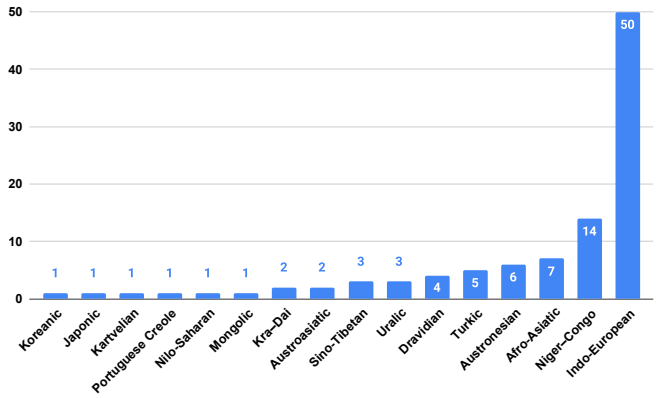

Fig. 1. Distributions of language families in FLEURS (y-axis is the count).

shown in Figure 1), and 27 unique writing systems (distribution shown in Figure 2). We grouped languages along with their geographical areas: Western Europe (WE), Eastern Europe (EE), Central-Asia/Middle-East/North-Africa (CMN), Sub-Saharan Africa (SSA), South Asia (SA), South-East Asia (SEA) and China/Japan/Korea (CJK). In Table 2, we present the basic statistics of FLEURS data per geographical group. See the supplement material for a full list of meta information (ISO codes, language families, estimated numbers of speakers, etc.)

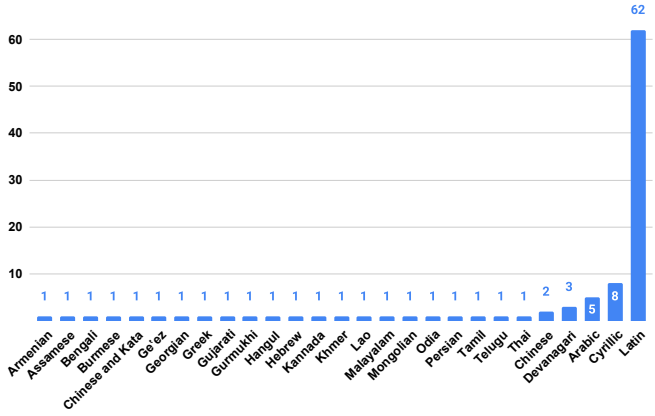

Fig. 2. Distributions of writing systems in FLEURS (y-axis is the count).

3. TASK BASELINES

3.1. Experimental Setup

FLEURS enables evaluations for several core speech tasks. In this paper, we focus on speech recognition, speech language identification and speech-text retrieval. Training a giant model from scratch only on the FLEURS dataset will easily overfit. Instead, we adopt the trendy pre-training and fine-tuning methodology [7, 1, 2] to build the massively multilingual baselines.

Multilingual Speech-only Pre-training: Multilingual pre-trained models have achieved significant gains in a range of NLP and ASR tasks. We initialize fine-tuning from a 600M parameter wav2vec-BERT [1] model, which had been pre-trained on 429k hours unlabeled speech data in 51 languages pooling from VoxPopuli [14], MLS [13], CommonVoice [16] and BABEL [17]. Pre-training for this baseline (dubbed w2v-bert-51 (0.6B)) is *speech-only*.

Multilingual Multimodal Pre-training: In addition to pre-

train on speech data, incorporating textual and speech-text data into pre-training allows for transfer learning across the two modalities [13]. We explore fine-tuning from a multilingual model which has been pre-trained with the same speech data, and 10TiB of unlabeled text from mC4 corpus which includes 101 languages [26]. The *speech-text* pre-trained model consisted of 600M parameters, dubbed mSLAM (0.6B). We followed [2] for pre-training and fine-tuning configurations.

3.2. Seen Languages and Unseen Languages

The languages for which speech data was available during pre-training are referred to as *seen* languages. There are 54 seen languages:

- **WE (17 languages)**: American English (en), Catalan (ca), Croatian (hr), Danish (da), Dutch (nl), Finnish (fi), French (fr), German (de), Greek (el), Hungarian (hu), Irish (ga), Italian (it), Latin American Spanish (es), Maltese (mt), Portuguese (pt), Swedish (sv), Welsh (cy)
- **EE (12)**: Bulgarian (bg), Czech (cs), Estonian (et), Georgian (ka), Latvian (lv), Lithuanian (lt), Polish (pl), Romanian (ro), Russian (ru), Slovak (sk), Slovenian (sl), Ukrainian (uk)
- **CMN (7)**: Arabic (ar), Kazakh (kk), Kyrgyz (ky), Mongolian (mn), Pashto (ps), Persian (fa), Turkish (tr)
- **SSA (3)**: Ganda (lg), Swahili (sw), Zulu (zu)
- **SA (7)**: Assamese (as), Bengali (bn), Hindi (hi), Oriya (or), Punjabi (pa), Tamil (ta), Telugu (te)
- **SEA (5)**: Cebuano (ceb), Indonesian (id), Lao (lo), Thai (th), Vietnamese (vi)
- **CJK (3)**: Cantonese (yue), Japanese (ja), Mandarin (cmn)

Languages which do not have any pre-training speech data are referred to as *unseen* languages. There are 48 unseen languages:

- **WE (8)**: Asturian (ast), Bosnian (bs), Galician (gl), Icelandic (is), Kabuverdianu (kea), Luxembourgish (lb), Norwegian (nb), Occitan (oc)
- **EE (4)**: Armenian (hy), Belarusian (be), Macedonian (mk), Serbian (sr)
- **CMN (5)**: Azerbaijani (az), Hebrew (he), Sorani-Kurdish (ckb), Tajik (tg), Uzbek (uz)
- **SSA (17)**: Afrikaans(af), Amharic (am), Fula (ff), Hausa (ha), Igbo (ig), Kamba (kam), Lingala (ln), Luo (luo), Northern-Sotho (nso), Nyanja (ny), Oromo (om), Shona (sn), Somali (so), Umbundu (umb), Wolof (wo), Xhosa (xh), Yoruba (yo)
- **SA (7)**: Gujarati (gu), Kannada (kn), Malayalam (ml), Marathi (mr), Nepali (ne), Sindhi (sd), Urdu (ur)
- **SEA (6)**: Filipino (fil), Javanese (jv), Khmer (km), Malay (ms), Maori (mi), Burmese (my)
- **CJK (1)**: Korean (ko)

Text training data utilized in pre-training includes most languages, except the following 20 *text-unseen* languages:

- **WE (6)**: Asturian (ast), Bosnian (bs), Kabuverdianu (kea), Luxembourgish (lb), Norwegian (nb), Occitan (oc)
- **CMN (1)**: Hebrew (he)
- **SSA (9)**: Fula (ff), Ganda (lg), Kamba (kam), Lingala (ln), Luo (luo), Northern-Sotho (nso), Oromo (om), Umbundu (umb), Wolof (wo)
- **SA (2)**: Assamese (as), Oriya (or)
- **SEA (1)**: Lao (lo)
- **CJK (1)**: Cantonese (yue).

4. DOWNSTREAM TASKS

4.1. Speech Recognition

To build ASR baselines, we add a decoder which consists of two LSTM[27] layers to fine-tune from pre-trained models, using a Connectionist Temporal Classification (CTC) [28] loss. The baselines use a 6100-character vocabulary which was built from the SRC_NORM field of the data. We multilingually fine-tune on all 102 locales, and report results for both speech-only and speech-text pre-trained models. Our finetuning parameters follow [2]. We evaluate the fine-tuned ASR models for all locales in terms of character error rate. The language identification labels are not included in ASR modeling, and there is no language model used for hypothesis scoring.

4.1.1. Correlation with Language Geographical Groups

As shown in Table 3, European language groups (WE and EE) obtain better CER than the other groups, which was expected in part due to the larger amounts of unlabeled data in the two groups of languages from MLS and VoxPopuli. CMN, SSA, SA and SEA observe moderate CERs, while CJK gets the highest group average CERs.

Reducing the recognition error rates for other geographical groups is a key direction for future work, and understanding the differences arising from pre-trained models can be helpful. It is observed that fine-tuning from a speech-text pre-trained model leads to 0.5% regression in CER as compared to fine-tuning from a speech-only pre-trained model. Most degradation is observed in SA, SSA and CJK, which are three languages groups consist of rich writing systems. In other geographical groups, mis-recognized characters from a different writing system occur less frequently. Further breaking down the distributions of error types, substitution errors are dominating across all the groups, which is a common error

Model	WE	EE	CMN	SSA	SA	SEA	CJK	Avg.
w2v-bert-51 (0.6B)	9.5	9.1	13.0	13.6	17.4	12.4	30.5	12.9
mSLAM (0.6B)	9.5	9.1	13.2	14.3	19.0	12.7	32.5	13.4

Table 3. **Speech recognition** - FLEURS Massively multilingual ASR baselines, reporting % CER (↓), by geographical group.

Model	WE	EE	CMN	SSA	SA	SEA	CJK	Avg.
Speech recognition CER for speech seen languages								
Number of languages	17	12	7	3	7	5	3	54
w2v-bert-51 (0.6B)	9.9	8.7	11.6	10.6	11.7	14.1	34.0	12.9
mSLAM (2B)	9.7	9.0	10.9	11.2	12.7	14.5	36.4	13.4
Speech recognition CER for speech unseen languages								
Number of languages	8	4	5	17	7	6	1	48
w2v-bert-51 (0.6B)	8.8	10.1	15.0	14.2	23.0	11.0	20.1	14.0
mSLAM (0.6B)	8.9	9.4	16.4	14.9	25.3	11.2	20.6	14.8

Table 4. **Speech recognition** on speech seen and unseen languages, reporting % CER (↓), by geographical group.

pattern in multilingual ASR [29], especially when no explicit language id information was incorporated. For SA, the group seeing the second highest substitution error rate: most substitutions come from Urdu. Urdu is acoustically similar to Hindi, so many Urdu utterances were predicted in Devanagari script, while the reference texts are in the Perso-Arabic script. CJK languages are known for the vast number of homophones in speech, which adds difficulties in selecting the correct character without aid from language models. There are a couple of ways to improve, such as: to include the language specific information, to utilize language model fusion, and to apply automatic transliteration to normalize the output writing systems before calculating CER [30].

4.1.2. Differences in Seen and Unseen Speech Languages

Experimental results in Table 3 show that fine-tuning from multimodal pre-training is overall slightly worse than fine-tuning from speech-only pre-training (similar to the patterns observed in [2]). Particularly, it lags behind more for the *unseen* languages (Table 4). Most gaps come from SA, SSA, and CMN groups. The exception is EE group, where fine-tuning from multi-modal pre-training outperforms the speech-only baseline. The differences indicate that multimodal and speech-only pre-training can be more beneficial for certain languages. Specifically, for languages which were not seen in pre-training, a large fraction of them observes a test CER worse than global average due to fine-tuning on very limited amount of supervised data. These observations align with previous findings in [12]. Both CJK groups observed the highest error rates, likely caused by the need for a larger vocabulary of characters to reduce substitution errors.

In addition, for the unseen languages which achieved a lower CER than average, most of them use Latin or Cyrillic script based writing systems. Interestingly, unseen languages which use scripts other than the two systems can still obtain good CER: for example, ml_in (Malayalam script), kn_in (Kannada), gu_in (Gujarati), ne_np (Devanagari). The success in recognizing unseen Malayalam, Kannada, Gujarati, Nepali can potentially be attributed to other Indian language data (bn_in, te_in, pa_in, as_in, ta_in) presented in the pre-training.

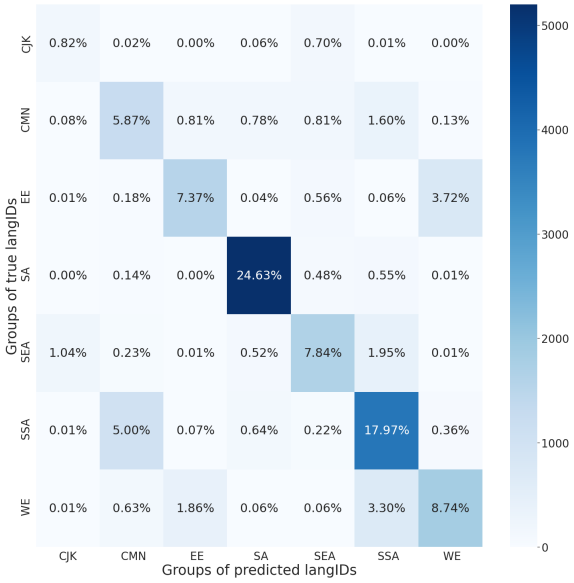

Fig. 3. Matrix of geographical groups of true language labels *vs.* groups of languages predictions, on the test utterances mis-classified by mSLAM (0.6B) speech langID model.

4.2. Speech Language Identification

For the LangID task we ensure that the train / dev / test splits have different speakers. We fine-tune our models on Speech LangID classification following [2]. As shown in Table 5, fine-tuning from mSLAM obtains 73.3% macro-average accuracy on FLEURS LangID, while fine-tuning from w2v-bert-51 (0.6B) obtains 71.4% respectively.

The group average accuracy decreases in the order of: CJK > WE > EE > CMN > SEA > SSA > SA. This could be due to the following reasons: (1) there are only four languages in the CJK group, they are relatively easy to distinguish from each other and from languages in the other groups; (2) Most of the data seen during pre-training is from Western European and Eastern European languages; (3) CMN, SEA, SSA and SA are geographical regions which are known for language diversity, but with limited amounts of publicly available pre-training data. We also observed that it is important to have a language's speech data presented in the pre-training, in order to achieve a good identification accuracy during speech LangID fine-tuning. As shown in Figure 3, there are less mis-identifications across different geographical

Model	WE	EE	CMN	SSA	SA	SEA	CJK	Avg.
w2v-bert-51 (0.6B)	85.3	78.4	72.9	59.1	52.0	65.7	89.7	71.4
mSLAM (0.6B)	84.6	81.3	75.9	62.2	51.7	73.4	87.8	73.3

Table 5. **Speech identification** - FLEURS speech LangID baselines, reporting % accuracy (↑) at the language level, and aggregated by geographical group.

groups ($e.g.$ 67.6% incorrect predictions of SEA languages were classified as other languages in SEA group), indicating that separating languages from the others which are in proximity in geography is more challenging.

4.3. Cross-modal Speech-Text Retrieval

In the Speech-to-Text Retrieval task, given an audio sample, the task is to retrieve its most probable transcription amongst the text samples in that language's test set. In the Text-to-Speech Retrieval task, given a text sample, the task is to retrieve the audio sample that most closely corresponds to it, amongst all the audio samples in that language's test set. As user interactions with machines move beyond textual queries, multi-modal retrieval of documents across the web is of growing interest [31, 32, 33]. The ability to build fixed-sized vector representations for queries be it speech, text or images, will enhance such content retrieval models [34, 35]. FLEURS is a rich multi-lingual, multi-modal, n-way parallel dataset which we hope will act as a benchmark to accelerate research in this field. FLEURS can act as a test-bench for various kinds of speech-text retrieval scenarios like speech-to-speech, text-to-text, speech-to-text, and text-to-speech retrieval along with testing cross-lingual and zero-shot capabilities.

As part of our multilingual baseline, we study the efficacy of pre-trained models towards learning fixed sized representations for both speech-to-text and text-to-speech retrieval. Speech-to-Text Retrieval is given an audio sample, the task is to retrieve its most probable transcription amongst a set of text samples. Text-to-Speech Retrieval is given a text, the task is to retrieve the audio sample that most closely corresponds to it, amongst a set of audio samples. Given the multi-modal nature of the task, we only fine-tune the multi-lingual multi-modal pretrained model, $i.e.$ mSLAM (0.6B), on the training set from all languages in FLEURS. Following [36, 37], cross-modal embeddings are trained using the additive margin softmax loss with in-batch negative sampling. We add bi-directional loss for retrieving speech given a text query and vice-versa [36]. We obtain embeddings for the normalized text (SRC_NORM) for all languages.

For evaluating the speech-to-text retrieval task, we report the % Precision at 1 (P@1) retrieval score of retrieving the correct text segment given a speech query from a database of in-domain textual keys collected from the FLEURS test set. Similarly, for text-to-speech retrieval task, we report the P@1 retrieval score of retrieving any of the speakers who speaks the correct textual query. We report results for both retrieving text segments from speech queries and speech segments from textual queries. The summary tables for each geographical group can be found in Table 6 for both speech-to-text and text-to-speech retrieval. The detailed retrieval score for each language can be found in supplement material.

Task	WE	EE	CMN	SSA	SA	SEA	CJK	Avg.
Speech-to-Text Retrieval	87.6	91.1	79.4	83.9	67.7	54.8	4.7	76.9
Text-to-Speech Retrieval	83.7	88.3	77.1	83.5	61.4	55.4	4.7	74.4

Table 6. **Cross-modal Speech-Text Retrieval** - FLEURS massively multilingual Speech-to-Text and Text-to-Speech retrieval baselines, reporting % P@1 (↑) score, by geographical group.

We observe an average P@1 of 76.9% for speech-to-text retrieval and a P@1 of 74.4% for text-to-speech retrieval. We observe that P@1 for seen languages in almost all geographical groups (except SA and SEA) is higher than their unseen counterparts, as is the case with speech recognition and language identification. In particular, we notice a steep degradation in the retrieval performance on CJK languages. We anticipate this to be the result of tokenization mismatch between the fine-tuning and the pre-training regime.

We also observe some interesting language specific peculiarities in the retrieval performance. For example, while Odia (or) is seen in speech, it is unseen in text since it is not present in the mc4 corpus [26] on which the mSLAM model was trained. This is exacerbated by Odia's unique script [38], which leads to the language being unrepresented in the tokenizer. On the other hand, Urdu (ur) is seen in the text pre-training but unseen in speech. This is interesting because Urdu performs considerably worse in Text-to-Speech retrieval compared to Speech-to-Text. We believe this is because Urdu is phonetically close to other SA languages like Hindi, consistent with our observations in Section 4.1.1, making it hard for the model to disambiguate speech without pre-training data in the speech modality.

5. CONCLUSION

We introduced FLEURS, a new dataset for Few-shot Learning Evaluation of Universal Representations of Speech, in 102 languages. FLEURS is an n-way parallel speech dataset that can be used to evaluate speech recognition, classification and retrieval methods. By building up baseline ASR, language identification and retrieval systems on FLEURS, we show that it is especially suited to evaluate data-efficient multilingual pre-trained representations of speech (and text). We hope this dataset will catalyze research in few-shot tasks in many languages, enabling progress towards building speech technologies for the world.

6. REFERENCES

[1] Yu-An Chung, Yu Zhang, Wei Han, Chung-Cheng Chiu, James Qin, Ruoming Pang, and Yonghui Wu, "W2v-BERT: Combining contrastive learning and masked language modeling for self-supervised speech pre-training," 2021.

[2] Ankur Bapna, Colin Cherry, Yu Zhang, Ye Jia, Melvin Johnson, Yong Cheng, Simran Khanuja, Jason Riesa, and Alexis Conneau, "mSLAM: Massively multilingual joint pre-training for speech and text," *CoRR*, vol. abs/2202.01374, 2022.

[3] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," in *Proc. of NIPS*, 2017.

[4] Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, and Ruoming Pang, "Conformer: Convolution-augmented transformer for speech recognition," *Proc. of Interspeech*, 2020.

[5] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Proc. of NeurIPS*, 2020.

[6] Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli, "Unsupervised cross-lingual representation learning for speech recognition," in *Proc. of Interspeech*, 2021.

[7] Yu Zhang, Daniel S Park, Wei Han, James Qin, Anmol Gulati, Joel Shor, Aren Jansen, Yuanzhong Xu, Yanping Huang, Shibo Wang, et al., "Bigssl: Exploring the frontier of large-scale semi-supervised learning for automatic speech recognition," *IEEE Journal of Selected Topics in Signal Processing*, 2022.

[8] Bo Li, Ruoming Pang, Tara N. Sainath, Anmol Gulati, Yu Zhang, James Qin, Parisa Haghani, W. Ronny Huang, Min Ma, and Junwen Bai, "Scaling end-to-end models for large-scale multilingual asr," *arXiv*, 2021.

[9] Wenxin Hou, Yue Dong, Bairong Zhuang, Longfei Yang, Jiatong Shi, and Takahiro Shinozaki, "Large-Scale End-to-End Multilingual Speech Recognition and Language Identification with Multi-Task Learning," in *Proc. Interspeech 2020*, 2020, pp. 1037–1041.

[10] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *Proc. of ICASSP*. IEEE, 2015, pp. 5206–5210.

[11] Qiantong Xu, Alexei Baevski, Tatiana Likhomanenko, Paden Tomasello, Alexis Conneau, Ronan Collobert, Gabriel Synnaeve, and Michael Auli, "Self-training and pre-training are complementary for speech recognition," in *Proc. of ICASSP*, 2020.

[12] Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, Alexei Baevski, Alexis Conneau, and Michael Auli, "XLS-R: Self-supervised cross-lingual speech representation learning at scale," *arXiv preprint arXiv:2111.09296*, 2021.

[13] Vineel Pratap, Qiantong Xu, Anuroop Sriram, Gabriel Synnaeve, and Ronan Collobert, "MLS: A large-scale multilingual dataset for speech research," in *Proc. of Interspeech*, 2020.

[14] Changhan Wang, Morgane Riviere, Ann Lee, Anne Wu, Chaitanya Talnikar, Daniel Haziza, Mary Williamson, Juan Pino, and Emmanuel Dupoux, "VoxPopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation," in *Proc. of ACL*, 2021.

[15] Changhan Wang, Anne Wu, and Juan Pino, "CoVoST 2 and massively multilingual speech-to-text translation," *arXiv*, 2020.

[16] Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M. Tyers, and Gregor Weber, "Common Voice: A massively-multilingual speech corpus," *Proc. of LREC*, 2020.

[17] Mark JF Gales, Kate M Knill, and Anton Ragni, "Low-resource speech recognition and keyword-spotting," in *International Conference on Speech and Computer*. Springer, 2017, pp. 3–19.

[18] Javier Iranzo-Sánchez, Joan Albert Silvestre-Cerda, Javier Jorge, Nahuel Roselló, Adria Giménez, Albert Sanchis, Jorge Civera, and Alfons Juan, "Europarl-st: A multilingual corpus for speech translation of parliamentary debates," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 8229–8233.

[19] Roldano Cattoni, Mattia Antonino Di Gangi, Luisa Bentivogli, Matteo Negri, and Marco Turchi, "Must-c: A multilingual corpus for end-to-end speech translation," *Computer Speech & Language*, vol. 66, pp. 101155, 2021.

[20] Elizabeth Salesky, Matthew Wiesner, Jacob Bremerman, Roldano Cattoni, Matteo Negri, Marco Turchi,

Douglas W Oard, and Matt Post, "The multilingual TEDx corpus for speech recognition and translation," *arXiv preprint arXiv:2102.01757*, 2021.

[21] Ye Jia, Michelle Tadmor Ramanovich, Quan Wang, and Heiga Zen, "Cvss corpus and massively multilingual speech-to-speech translation," *arXiv preprint arXiv:2201.03713*, 2022.

[22] Jörgen Valk and Tanel Alumäe, "VoxLingua107: a dataset for spoken language recognition," in *Proc. of SLT*, 2020.

[23] Alan W Black, "Cmu wilderness multilingual speech dataset," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 5971–5975.

[24] Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzman, and Angela Fan, "The FLORES-101 evaluation benchmark for low-resource and multilingual machine translation," *arXiv preprint arXiv:2106.03193*, 2021.

[25] Mehryar Mohri, "Weighted finite-state transducer algorithms. an overview," *Formal Languages and Applications*, pp. 551–563, 2004.

[26] Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel, "mT5: A massively multilingual pre-trained text-to-text transformer," *arXiv preprint arXiv:2010.11934*, 2020.

[27] Sepp Hochreiter and Jürgen Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[28] Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *Proceedings of the 23rd international conference on Machine learning*, 2006, pp. 369–376.

[29] Xinjian Li, Siddharth Dalmia, Juncheng Li, Matthew Lee, Patrick Littell, Jiali Yao, Antonios Anastasopoulos, David R Mortensen, Graham Neubig, Alan W Black, et al., "Universal phone recognition with a multilingual allophone system," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 8249–8253.

[30] Jesse Emond, Bhuvana Ramabhadran, Brian Roark, Pedro Moreno, and Min Ma, "Transliteration based approaches to improve code-switched speech recognition performance," in *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2018, pp. 448–455.

[31] Michael Witbrock and Alexander G Hauptmann, "Speech recognition and information retrieval: Experiments in retrieving spoken documents," in *Proceedings of the DARPA speech recognition workshop*, 1997, vol. 97.

[32] Shin-ya Ishikawa, Takahiro Ikeda, Kiyokazu Miki, Fumihiro Adachi, Ryosuke Isotani, Ken-Ichi Iso, and Akitoshi Okumura, "Speech-activated text retrieval system for multimodal cellular phones," in *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*. IEEE, 2004, vol. 1, pp. I–453.

[33] Lin-shan Lee, James Glass, Hung-yi Lee, and Chun-an Chan, "Spoken content retrieval—beyond cascading speech recognition with text retrieval," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 9, pp. 1389–1420, 2015.

[34] Aashi Jain, Mandy Guo, Krishna Srinivasan, Ting Chen, Sneha Kudugunta, Chao Jia, Yinfei Yang, and Jason Baldridge, "MURAL: Multimodal, multitask representations across languages," in *Findings of the Association for Computational Linguistics: EMNLP 2021*, Punta Cana, Dominican Republic, Nov. 2021, pp. 3449–3463, Association for Computational Linguistics.

[35] Paul-Ambroise Duquenne, Hongyu Gong, and Holger Schwenk, "Multimodal and multilingual embeddings for large-scale speech mining," *Advances in Neural Information Processing Systems*, vol. 34, 2021.

[36] Yinfei Yang, Gustavo Hernández Ábrego, Steve Yuan, Mandy Guo, Qinlan Shen, Daniel Cer, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil, "Improving multilingual sentence embedding using bi-directional dual encoder with additive margin softmax," in *IJCAI*, 2019.

[37] Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang, "Language-agnostic BERT sentence embedding," in *Proceedings of ACL*, Dublin, Ireland, May 2022, pp. 878–891, Association for Computational Linguistics.

[38] Ramesh Kumar Mohapatra, Tusar Kanti Mishra, Sandeep Panda, and Banshidhar Majhi, "Ohcs: A database for handwritten atomic odia character recognition," in *2015 Fifth National Conference on Computer Vision, Pattern Recognition, Image Processing and Graphics (NCVPRIPG)*. IEEE, 2015, pp. 1–4.