Multilingual Speech Dataset for ASEAN Languages

Hay Mar Soe Naing, Win Pa Pa

Natural Language Processing Lab.
University of Computer Studies, Yangon, Myanmar

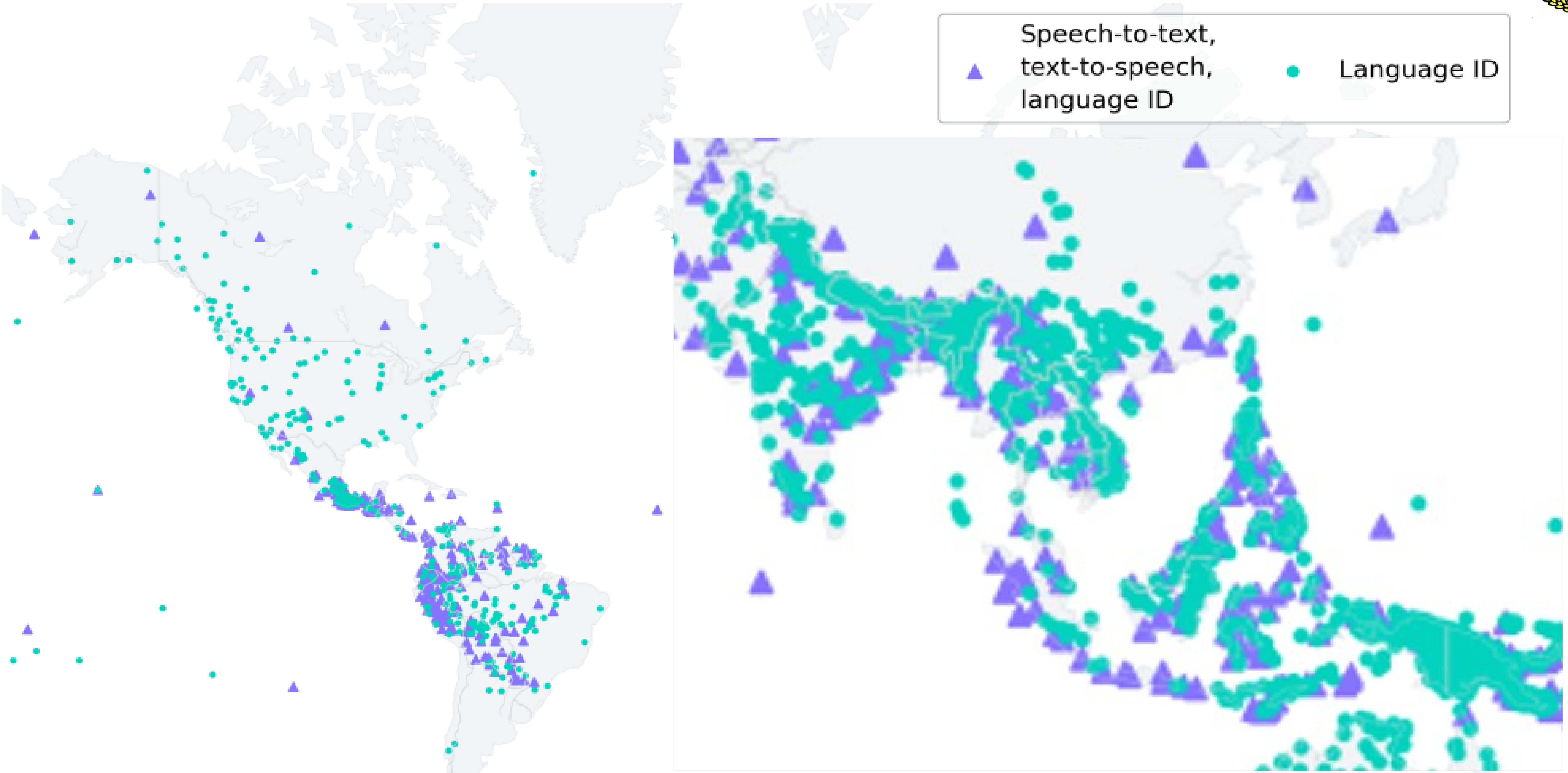
Multilingual Speech Dataset for ASEAN Languages

- Multilingualism need not only textual, but also spoken form, when information services are to extend beyond national boundaries, or across language groups.
- Multilingual spoken services are a growing industry, and relied heavily on human operators.
- Few commercial multilingual speech services.
- Large amount of parallel speech-text data not available in most languages (especially Myanmar)

Problem Statement (1) - MMS

- **Massively Multilingual Speech (MMS)** project was launched by Meta AI (over 1,000 languages).
- MMS Based on the Bible readings to be a largest speech dataset.
- MMS cannot cover the daily usage, commonly spoken style and the usage of official Myanmar language especially in speech recognition and translation area.

The World's Language Diversity Through Meta AI (MMS)



- Most of the people who live in central region of Myanmar use the official Myanmar (Burmese) language

https://ai.meta.com/blog/multilingual-model-speech-recognition/

Problem Statement (2) - Whisper

- Small portions of ASEAN language (mainly Myanmar, Khmer and Laos) are included in most multilingual speech corpus.
- **In Whisper launched by OpenAI**, the training portion of Myanmar language in multilingual speech recognition is approximately 0.1 hours of 117,113 hours of audio.
- Multilingual transcription error rate is too high (around 120% of WER % on FLEURS)
- Speech translation BLEU scores is too small (less than 0.5 BLEU scores for Myanmar language)

Ref : https://github.com/openai/whisper.

OpenAI: Whisper

Speech Translation on FLEURS (BLEU scores)

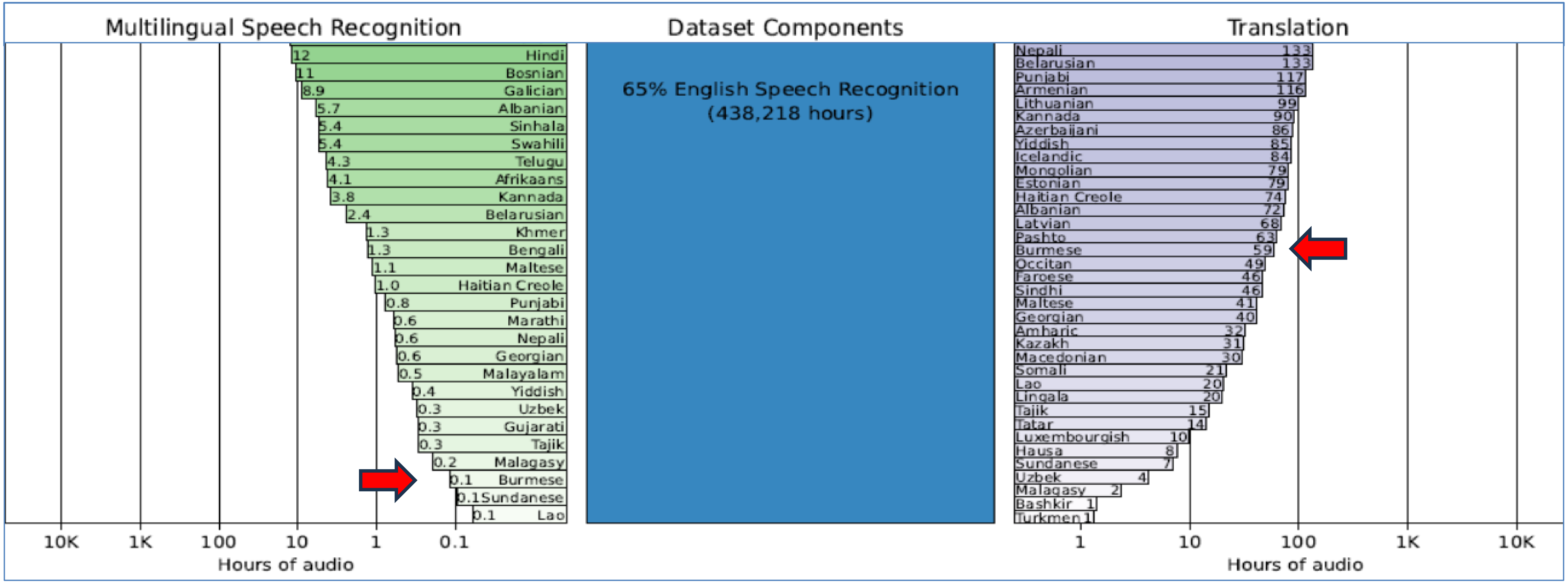
Model	Croatian	Hungarian	Armenian	Indonesian	Icelandic	Italian	Japanese	Javanese	Georgian	Kazakh	Khmer	Kannada	Korean	Luxembourgish
Whisper tiny	0.6	0.1	0.1	0.3	0.4	5.3	0.2	0.2	0.1	0.1	0.1	0.8	0.5	0.8
Whisper base	3.7	0.2	0.1	2.6	0.4	11.3	1.5	0.2	0.2	0.2	0.1	0.9	3.7	1.7
Whisper small	14.6	4.8	0.7	16.4	1.8	17.8	9.6	1.4	0.2	0.8	0.5	2.3	12.2	5.7
Whisper medium	23.0	15.5	10.4	24.1	6.8	21.6	14.9	5.0	1.3	4.3	3.3	8.5	19.2	13.6
Whisper large	25.4	18.3	13.2	27.2	6.6	23.5	17.0	5.1	2.7	6.3	5.2	9.9	20.0	15.4
Whisper large-v2	27.0	21.2	16.0	29.1	9.1	23.6	18.9	6.2	2.4	5.4	6.1	11.6	21.3	16.8

Model	Lingala	Lao	Lithuanian	Latvian	Maori	Macedonian	Malayalam	Mongolian	Marathi	Malay	Maltese	Myanmar	Norwegian	Nepali
Whisper tiny	0.1	0.2	0.1	0.2	0.3	1.0	0.8	0.1	0.2	0.3	0.6	0.1	1.4	0.1
Whisper base	0.1	0.3	0.3	0.4	1.0	5.4	1.4	0.1	0.9	2.1	1.4	0.1	8.4	0.3
Whisper small	0.5	2.0	1.9	1.5	3.9	15.3	5.7	0.1	3.8	14.1	4.9	0.0	22.0	2.9
Whisper medium	0.9	8.1	9.6	10.0	8.5	23.5	13.8	0.5	10.9	23.2	11.2	0.2	29.1	12.7
Whisper large	1.2	9.3	12.0	12.5	9.4	26.4	16.5	1.0	13.1	25.5	12.8	0.5	30.5	12.9
Whisper large-v2	1.0	11.0	14.0	14.3	10.2	27.7	16.7	1.0	12.9	27.3	13.5	0.4	31.4	16.1

Ref : https://github.com/openai/whisper.

OpenAI: Whisper

Training Dataset Statistics



Proposed Solution

- Voice enabled services are rapidly growing and high margin opportunity.
- Very difficult to have one speech recognizer/synthesizer for each language.
- The focus is
  1) To develop common multilingual corpora with support for multiple languages.
  2) To build appropriate language specific linguistic analysis modules.

Proposed Solution

- Create the multilingual speech corpus of most commonly spoken language for each country in their most spoken style.
- Pre-train wav2vec 2.0 models supporting more ASEAN languages (low resources languages)
- Fine-tune these models to build multilingual speech processing tasks.

Impact:

- Multilinguality
  - Goal: high quality cross-modality, cross-lingual generation at low cost.
  - Utilize knowledge transfer across languages, and alleviate data requirement.
  - One model for multiple translation directions.

Output/Outcome:

- An open, large-scale, multilingual speech corpus for various tasks.
- Validated usefulness
  - Language Identification, Speech recognition, Speech to Text, Speech Translation
- Pre-trained checkpoints for wav2vec, ASR and S2T translation.

Conclusion:

- Building a multilingual speech corpus and speech processing model contributes to ASEAN languages.
- Combining data from all available languages during pre-training can also improve performance compared to using multiple languages during fine-tuning.
- Analyzing how wav2vec works with ASEAN languages.
- Evaluating pretrained models on different languages can impact performance.

Welcome Collaboration!!

Thank You!