# A Survey of Multilingual Models for Automatic Speech Recognition

LREC 2022 Marseille

Hemant Yadav, Sunayana Sitaram

Microsoft Research India, Bangalore {t-hyadav, sunayana.sitaram}@microsoft.com

## Motivation

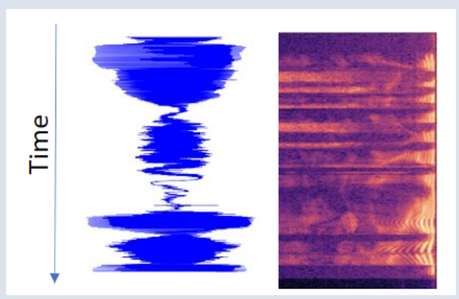
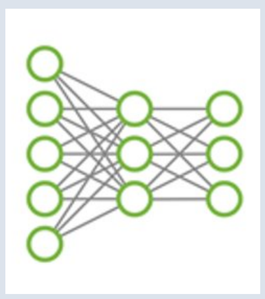
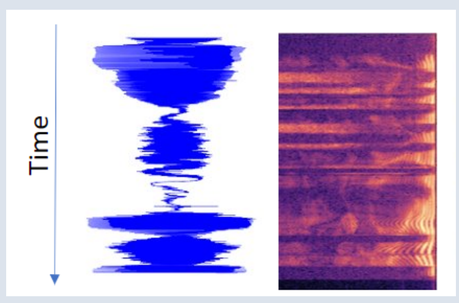
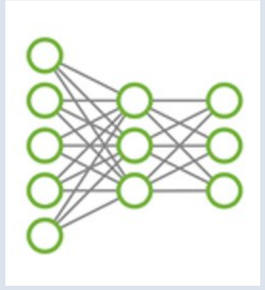
- The majority of the world's languages do not have usable systems due to the lack of large speech datasets.
- Recent advances in cross-lingual transfer and Self Supervised Learning (SSL) may help improve the situation for low-resource languages.
- We survey the SOTA on multilingual models, cross-lingual transfer and SSL for ASR, highlight findings and pose open research challenges.

## Research questions

- Are multilingual models better in terms of performance compared to monolingual models for high and low-resource languages?
- Can multilingual ASR models exploit unlabeled data for improving performance on low-resource languages?
- What are the best practices for building multilingual ASR models?
- Where should resources be invested for improving low-resource ASR in the era of multilingual models?
- What are the open questions and main challenges that need to be addressed going forward?

## General ASR pipeline

Automatic Speech Recognition converts a raw speech signal to a text transcription.

| Input | Architecture | Text-pair | Technique |
|---|---|---|---|
|   |  | My name is Hemant | Supervised |
|   |  | Nil | Unsupervised |

## Open problems

- More research needed to determine whether these techniques can be used for very low-resource languages that have only a few hours of data, and on languages other than Indo-European.
- Building evaluation benchmarks to cover diverse language families and datasizes would be a crucial investment for improving low-resource language ASR.
- Multilingual ASR models are currently trained on less than 10 languages, need to study the implications of larger models.
- Transferring SSL representations across different speech tasks is an interesting future direction.

## Findings and Recommendations

- Seq2seq training on labelled data performs better than CTC due to better language modeling.
- Multilingual models perform better than monolingual counterparts trained with the same amount of data for a single language.
- Combining the data of all languages available during pre-training also improves performance compared to using multiple languages only during fine-tuning.
- Fine-tuning plays an important role in the accuracy of multilingual models. Techniques that can be used for further improvements include phone-mapping, using a feature extractor trained on multiple languages and using a common decoder for related languages.
- When building a multilingual model that can be re-used across different languages, it is desirable that the performance on high-resource languages does not degrade while making improvements over low-resource languages. Some studies show that this degradation indeed occurs, however, there are strategies such as sampling and using language ID information that can be used to alleviate the problem.
- Many factors affect the overall accuracy on target languages, including the choice of features and the languages that the features are trained on and the choice of architecture while building models trained on multiple languages.
- The choice of pre-training data matters. Pre-training from a diverse set of languages, or languages related to the target language is better than restricting it only to one language such as English even if the total pre-training data remains the same.
- Pre-training with a diverse set of languages has been shown to improve performance on languages that are not present in the training data, though more experimentation is needed to study this further.
- The size of pre-training and fine-tuning data matters - although it is possible to exploit unlabeled data for pre-training, the fine-tuning data needs to be of a reasonable size to get performance improvements.
- SSL and the use of unlabeled data are exciting directions for low-resource ASR, particularly when labeled and unlabeled data are from different domains.