ECAPA-TDNN Embeddings for Accented Speech Classification

Juan Pablo Zuluaga, Sara Ahmed, Danielius Visockas, Francielle Vargas

Mentor: Cem Subakan

🤔 Are large-scale pre-trained acoustic models changing the paradigm of research on ASR and other acoustic downstream tasks?

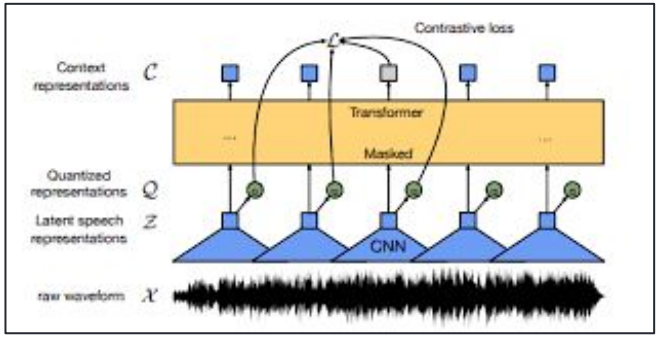
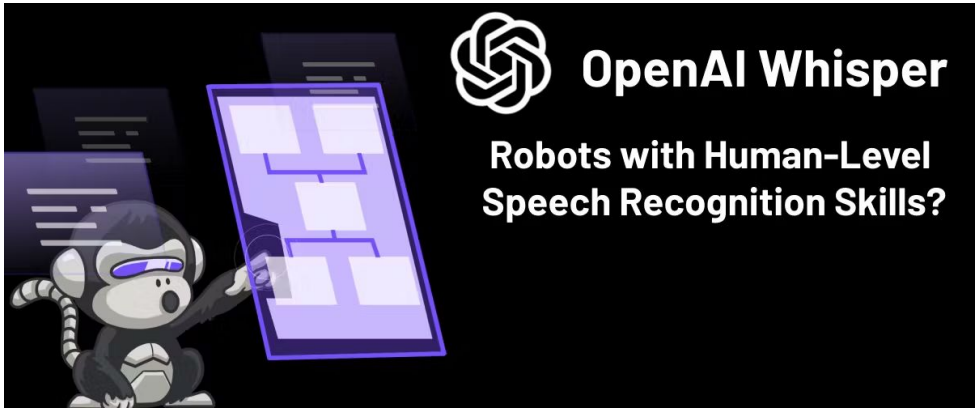
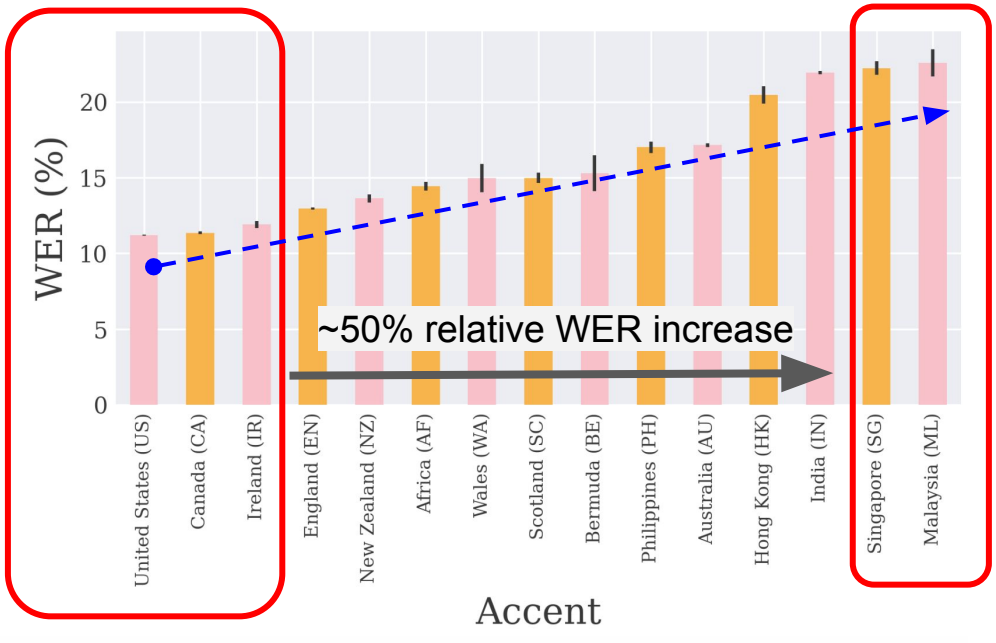


Figure 1: **Self-supervised cross-lingual representation learning.** We pre-train a large multilingual wav2vec 2.0 Transformer (XLS-R) on 436K hours of unannotated speech data in 128 languages. The training data is from different public speech corpora and we fine-tune the resulting model for several multilingual speech tasks.

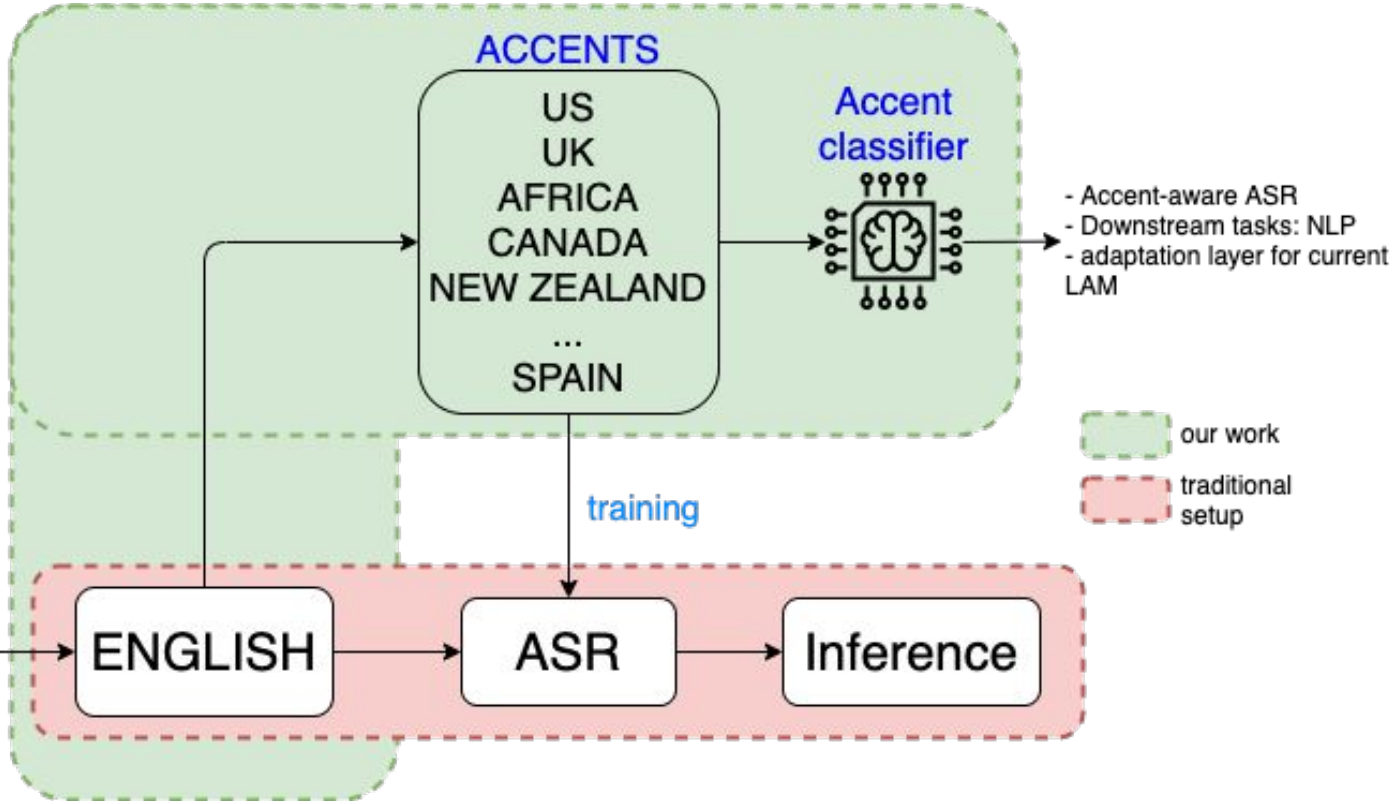


Background



Cámbara, G., Peiró-Lilja, A., Farrús, M. and Luque, J., 2021. English Accent Accuracy Analysis in a State-of-the-Art Automatic Speech Recognition System. arXiv preprint arXiv:2105.05041.

Suggested Framework



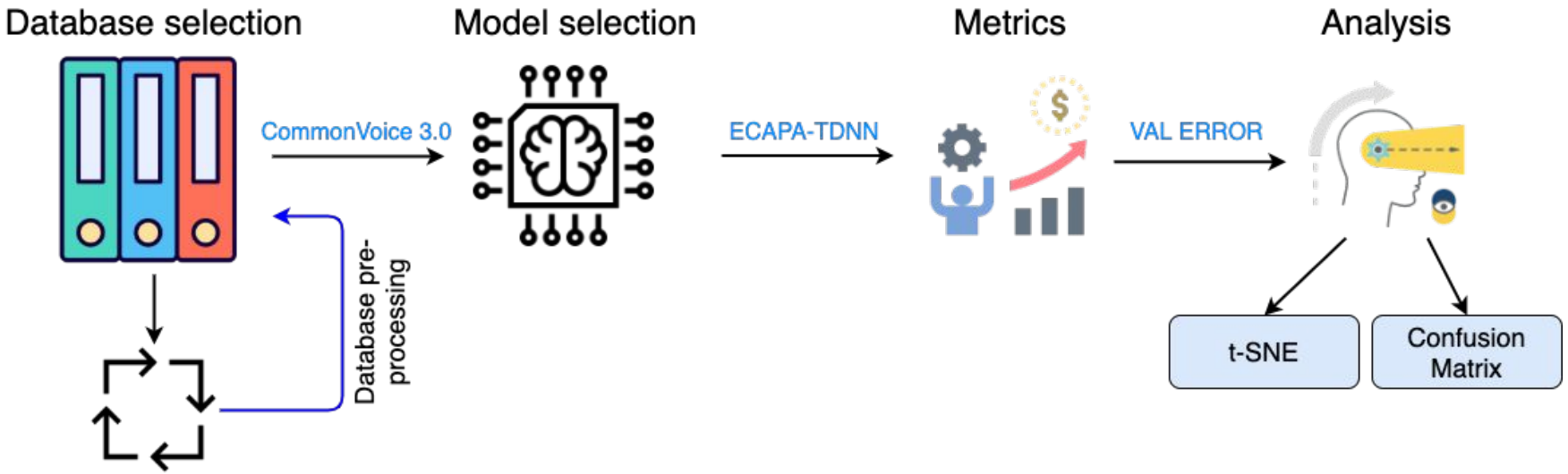
Setup

- Dataset: **CommonVoice 3.0** → 16 accents from the EN set
  - **Train set**: ~50 hrs / 45k samples
  - **Dev set**: 1.24 hrs / 1062 samples
  - **Test set**: 1.15 hrs / 972 samples
- Accents: **African, Australian, Bermuda, Canada, England, Hong Kong, India, Ireland, Malaysia, New Zealand, Philippines, Scotland, Singapore, South Atlantic, US, Wales**

- Recipe in **SpeechBrain** (based on CommonLanguage but with Accents)
  - ECAPA-TDNN model (fine-tuned and trained from scratch models)
  - 🏋️ Training: 20 epochs, same batch size
  - 💿 ~2 days on 1 RTX3090 GPU

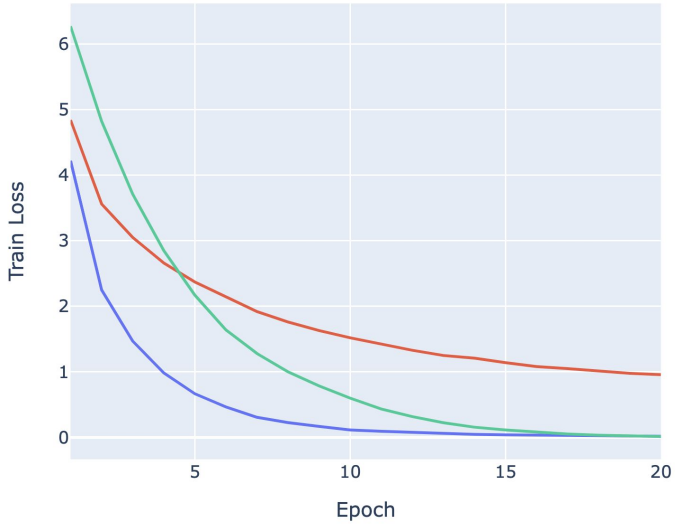
SpeechBrain

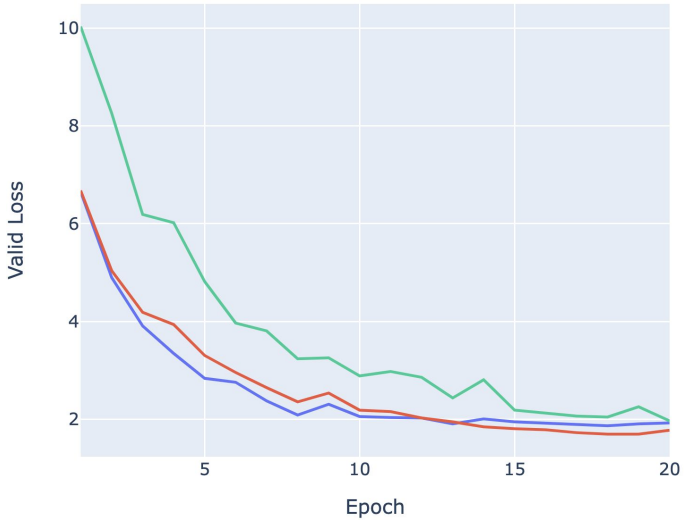
Pipeline

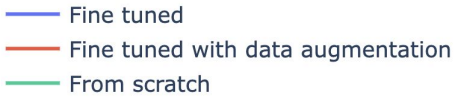

Database selection
Model selection
Metrics
Analysis
CommonVoice 3.0
ECAPA-TDNN
VAL ERROR
Database pre-processing
t-SNE
Confusion Matrix

Training Analytics

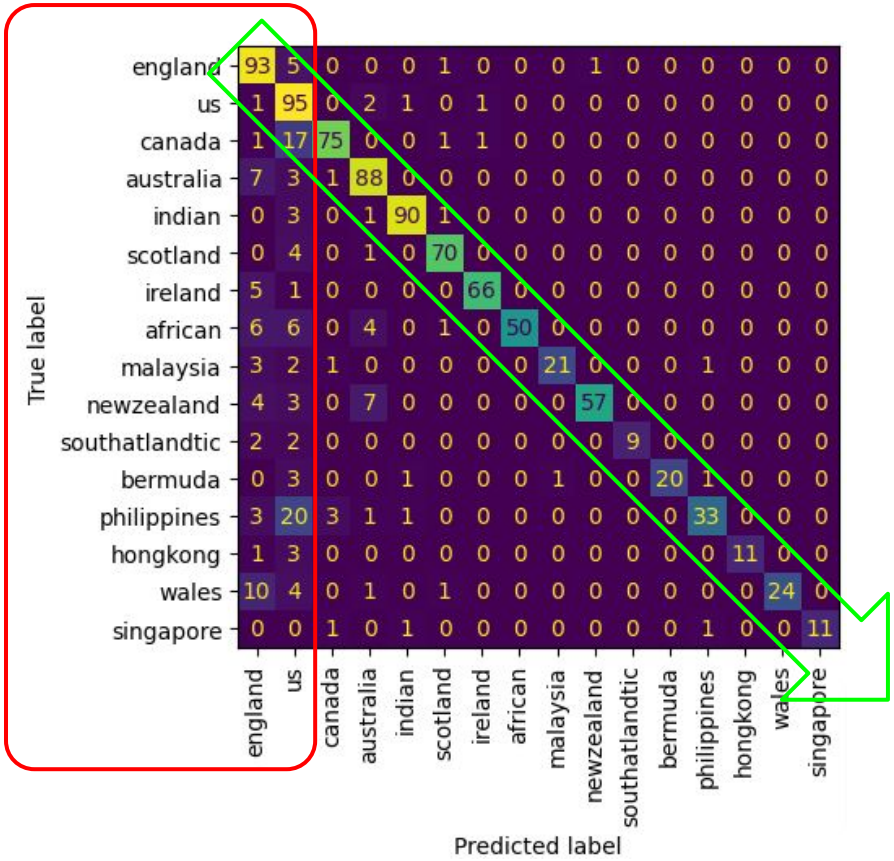
SpeechBrain	Trained from scratch	Fine-tuned	**Fine-tuned w/ data augmentation**
Test set Accuracy (not weighted)	82%	85%	💥 **87%**





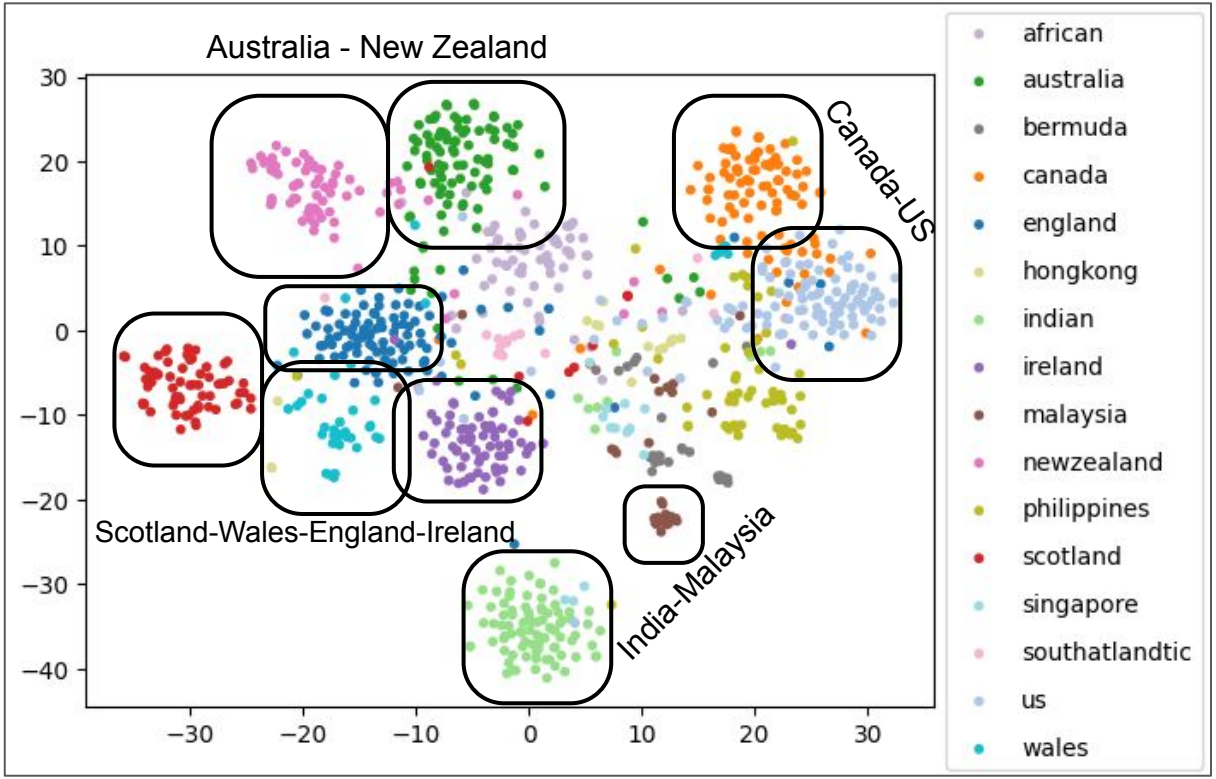


Analysis of Accent Classification with fine-tuned model (with data augmentation)



Analysis of Accent Classification with fine-tuned model (with data augmentation)

- 📊 **t-SNE** shows a level of clustering based on **phonologically similar accents**
- 👎🏽 misclassifications: **England-Wales** and **US-Canada**



Conclusion

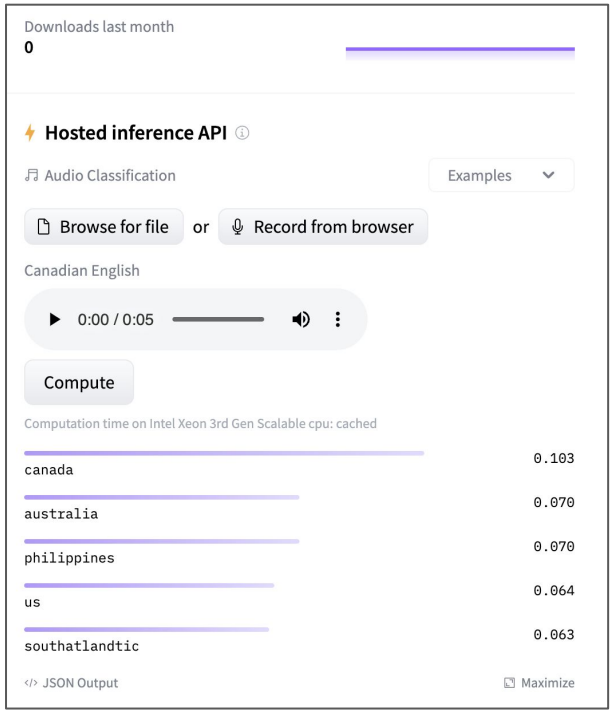
✅ English speech was classified based on 16 accents using the ECAPA-TDNN architecture.

- 🤖 3 systems: from-scratch 🥉, fine-tuned 🥈, and *fine-tuned with data augmentation* 🥇.
- 🚫 Misclassifications: level of structure, e.g., between England-Wales and US-Canada.
- 🎯 internal categorization of embeddings: t-SNE → **level of clustering** based on phonological similarity.

🔮 Future work:

- ASR frameworks can be more inclusive to accented speech → beginning of new era
- Implement proposed accent classification system → improve ASR:
    - Contextual biasing of ASR? Accent-aware LM swap at decoding time?
    - One layer of 'adaptation' like adapters in NLP

THANK YOU!
Hugging Face 🤗 Demo



SpeechBrain

🤗: bit.ly/SLT_accents

github.com/JuanPZuluaga/accent-recog-slt2022

References

Desplanques, B., Thienpondt, J., & Demuynck, K. (2020). Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification. *arXiv preprint arXiv:2005.07143*.

Huang, C., Chen, T., & Chang, E. (2004). Accent issues in large vocabulary continuous speech recognition. International Journal of Speech Technology, 7(2), 141-153.

Ravanelli, M., Parcollet, T., Plantinga, P., Rouhe, A., Cornell, S., Lugosch, L., Subakan C., ... & Bengio, Y. (2021). SpeechBrain: A general-purpose speech toolkit. *arXiv preprint arXiv:2106.04624*.

Backup slides

Introduction

- Statistical analysis has shown that accent is one of the key factors in speaker variability that affects the performance of ASR (Huang, Chen, and Chang 2004).
- Pretraining of large acoustic models, such as Whisper, do not take into account accented speech. This leads to mitigating its performance for more low-resource accents, despite the language spoken being high-resource.