



A Study of Speech Embedding Similarities Between Australian Aboriginal and High-Resource Languages

Eliathamby Ambikairajah¹, Jingyao Wu², Ting Dang³, Vidhyasaharan Sethu¹

¹University of New South Wales, Australia

²Massachusetts Institute of Technology, USA

³University of Melbourne, Australia

e.ambikairajah@unsw.edu.au, jingyaow@mit.edu, ting.dang@unimelb.edu.au,
v.sethu@unsw.edu.au

Abstract

Low-resource languages, such as Australian Aboriginal Languages, are underrepresented in the AI landscape due to limited availability of digital data, which in turn hinders speech processing model development. Leveraging sufficiently similar high-resource languages may help bridge this gap. This study examines the similarities between speech embeddings of aboriginal languages and 107 high-resource languages, including English, Spanish, and Mandarin, using Wav2Vec2 and VoxLingua107-ECAPA-TDNN. Through three language identification tasks, we analyze Warlpiri, Dalabon, and Light Warlpiri alongside 107 other languages. Our results reveal that aboriginal languages are most frequently identified as Māori, suggesting phonetic or structural similarities, while showing significant differences from globally dominant languages. Additionally, we also observe that Warlpiri and Dalabon exhibited closer matches with Hindi and Malayalam, than with other languages. **Index Terms:** Australian aboriginal language, Spoken language identification, Wav2Vec2, ECAPA-TDNN

1. Introduction

Australia, with its rich linguistic diversity, is home to over 250 Indigenous languages [1]. However, the AIATSIS 2018–19 survey reveals that only 123 Aboriginal and Torres Strait Islander languages are actively spoken today, and this number continues to decline [2, 3]. Additionally, modern advances in Natural Language Processing (NLP) and speech processing technologies have been predominantly developed for high-resource languages such as English, Spanish, and Mandarin [4, 5, 6]. This focus has led to a technological divide where low-resource languages, including Aboriginal languages, remain underrepresented.

Understanding the linguistic differences and similarities between Aboriginal and high-resource languages is crucial for effective AI model development. Figure 1 demonstrates the embeddings in the latent space [7] of four different languages: English, Mandarin, and two Australian Aboriginal languages (AALs), Warlpiri and Dalabon. It should be noted that Warlpiri is the most extensively studied Australian Aboriginal language and has a well established dictionary [8]. It is evident that Warlpiri and Dalabon (green and yellow) distinctly represent the unique nature relative to widely studied languages (blue and purple). This indicates that Aboriginal languages possess unique characteristics that set them apart from high-resource languages such as English and Mandarin, therefore, existing AI models developed for high-resource languages will find it difficult to generalize to Aboriginal languages, necessitating the analysis and model development specifically for AALs.

Additionally, studies have found that several Aboriginal

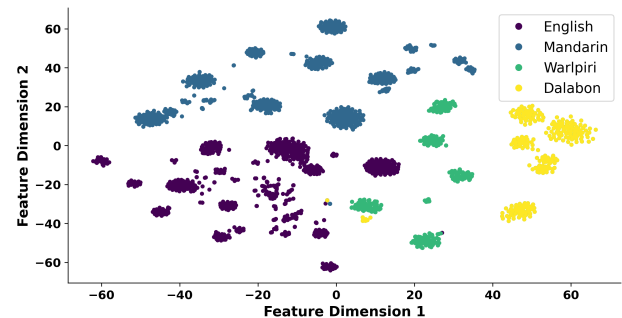


Figure 1: Distributions of speech embeddings in the latent space for four different languages: English, Mandarin, and two Australian Aboriginal languages, Warlpiri and Dalabon.

languages, including Warlpiri, exhibit the use of retroflex consonants, a feature also found in high-resource languages like Hindi and Tamil, indicating shared articulatory properties [9]. This suggests that despite their distinctiveness, there are cross-linguistic articulatory overlaps that can inform comparative linguistic research and preservation efforts. Therefore, identifying languages that are similar to Aboriginal ones is essential for advancing AI technologies to benefit Aboriginal communities. For example, a high-resource language similar to an Aboriginal language can aid in transfer learning or augment data for low-resource languages.

Existing studies on Australian Aboriginal Languages primarily focus on linguistic aspects, including the analysis of phoneme distributions, phonology, grammar, and lexical characteristics [10, 11, 12, 13]. These studies aim to understand the linguistic uniqueness and assist in the annotation or comprehension of the AALs themselves. However, there has been little research analyzing these languages from the perspective of AI modeling, especially when compared to high-resource languages, to explore how such analysis could aid in the development of AI models for AALs. In this paper, we conduct a comparative analysis of three AALs with 107 high-resource languages as integrated into multilingual AI models, to identify which of these languages exhibit the most similarities with AALs. This research enhances our understanding of the complexities inherent in Aboriginal languages and supports efforts to preserve and revitalize them for future generations.

In Section 2, we discuss the language similarity identification tasks, introducing three different approaches. Section 3 details the experimental setups including datasets, implementation details and evaluation metrics, while Section 4 presents the results.

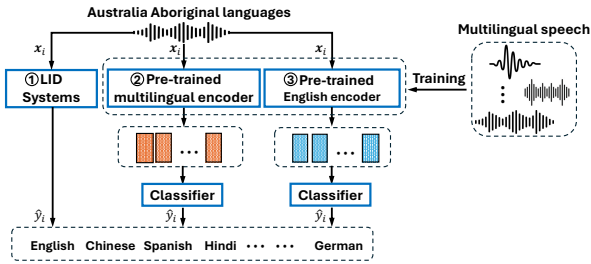


Figure 2: Three approaches for similarity identification tasks for AAL speech utterances: i) using a LID system to output language identification results; ii) using or fine-tuning a pre-trained multilingual encoder with a classifier for language identification; iii) fine-tuning a pre-trained English encoder with classifier for language identification.

2. Language similarity identification

The limited availability of speech data and the significant lack of annotations make the analysis of AAL speech particularly challenging. Furthermore, the absence of AI systems specifically designed for AALs further complicates speech analysis. However, existing AI systems trained on other languages are expected to misclassify AALs as those languages. This misclassification suggests that AALs share certain linguistic and acoustic characteristics with the languages they are misclassified as. Consequently, the misclassification rate could serve as an indicator of similarity, where a higher misclassification rate for a particular AAL suggests that it shares similar properties with those languages. Building on this concept, we propose leveraging pre-trained multilingual models to analyse AALs by identifying and ranking their similarities to other commonly used languages based on misclassification rates.

2.1. Similarity identification tasks

As illustrated in Fig. 2, three approaches are employed for language similarity identification. For an AAL speech utterance represented by x_i , the first approach utilizes a LID system $f(\cdot)$, which outputs the probabilities of classifying the utterance into N languages, represented as \mathbf{y}_i . The predicted language label is determined by $\hat{y}_i = \arg \max_i \mathbf{y}_i$. The second approach (Fig. 2) involves a pre-trained multilingual encoder, which requires additional training of classifiers for language identification. Similarly, the third approach involves a pre-trained English language encoder and also requires additional training of language identification classifiers. The output of these classifiers is also an N -dimensional vector, represented as $\mathbf{y}_i = g(\mathbf{z}_i)$, where $g(\cdot)$ is the classifier function, and the final language is selected based on the maximum probability. The objective is to analyse the misclassification rate by providing AALs as input to the LID system or pre-trained encoders (Fig. 2). The misclassification rate is defined as the number of AAL utterances incorrectly labeled as one of the known languages in the LID or encoder systems, as shown in Fig. 2.

2.2. Approach 1: LID for similarity identification

For the first approach, we utilized the VoxLingua107-ECAPA-TDNN model [7], which was specifically developed for LID tasks. This model was trained using 107 languages, providing comprehensive global language coverage. Such extensive coverage facilitates detailed comparative analyses between AALs and other languages. Given an AAL speech utterance, the out-

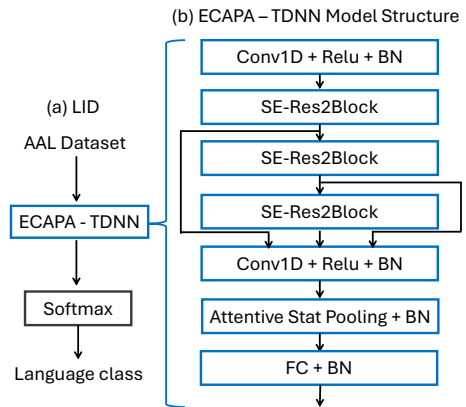


Figure 3: (a) LID for similarity identification using VoxLingua107-ECAPA-TDNN model. (b) Detailed network topology of the VoxLingua107-ECAPA-TDNN model.

put is a 107-dimensional vector representing the probabilities of each language class. The misclassification rate was calculated based on the predicted language class.

The model uses the ECAPA-TDNN model structure that incorporates channel attention, propagation, and aggregation in Time Delay Neural Network. As shown in Fig. 3(b), the model utilizes convolutional layers followed by Squeeze-Excitation Res2Blocks (SE-Res2Blocks) [14]. These blocks include dilated convolutions within residual structures and a 1D Squeeze-Excitation block, allowing the model to capture temporal dynamics and preserve salient acoustic and linguistic information throughout the network. The latent embeddings from all SE-Res2Blocks are concatenated and passed through a fully convolutional layer with batch normalization. Channel- and context-dependent attentive statistics pooling is employed to extract the most salient feature representations, which are then transformed by fully connected layers to produce the *VoxET embeddings*. These embeddings are subsequently classified into 107 language categories using a softmax layer, as shown in Fig. 3(a).

2.3. Pre-trained encoders for similarity identification

To allow for flexibility in similarity identification tasks, we also explored pre-trained encoders with additional training. We aim to deepen our understanding of how AALs compare to widely used, resource-rich languages such as English. These high-resource languages can greatly influence AAL due to their abundant resources, which can be leveraged to advance AAL technologies. Consequently, we can select languages of particular interest and train the pre-trained encoders with classifiers to gain a more comprehensive understanding of the similarities.

2.3.1. Language selection

We selected 14 languages out of 107, to conduct a detailed analysis of AALs. This selection includes four widely spoken languages, such as English, Spanish, Mandarin, and Hindi, for which most current systems and datasets have been developed. Additionally, we included the top ten most similar languages identified in LID tasks: Māori, Finnish, Tamil, Javanese, Malayalam, Swahili, Telugu, Tibetan, Pashto, and Assamese. As there are no pre-trained models available for these specific 14 languages, we trained the LID system using the dataset for this selection. During the inference phase, we evaluated AAL speech utterances to determine which of the 14 languages each

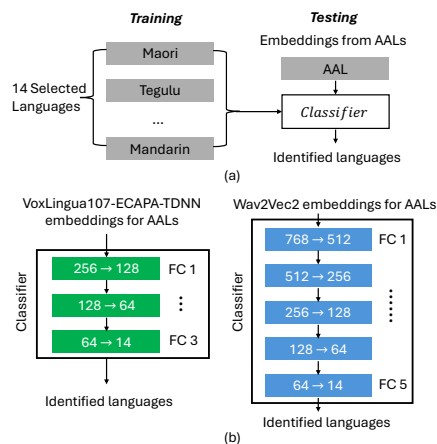


Figure 4: (a) System overview of models trained on 14 selected languages and tested using AALs. (b) Network structures of the classifier for VoxET and Wav2Vec2 embeddings respectively.

utterance is closest to, as illustrated in Fig. 4(a).

2.3.2. Approach 2: Multilingual encoder

We also utilized VoxLingua107-ECAPA-TDNN as the pre-trained encoder. In addition to its primary function of performing LID tasks, it can serve as a pre-trained multilingual encoder that provides 256-dimensional *VoxET embeddings* for subsequent LID tasks. As illustrated in Fig. 4(b), we employed a three-layer fully-connected (FC) network, with each layer using the *tanh* activation function, to classify 14 languages. The model was trained on these 14 languages and tested on AAL speech embeddings extracted using VoxLingua107-ECAPA-TDNN.

2.3.3. Approach 3: English-based encoder

To determine whether different model structures and pre-trained encoders affect language similarity identification and to support more reliable and well-grounded conclusions, we also explored the use of Wav2Vec2 [15] as a pre-trained encoder. This encoder is originally trained on English data.

Wav2Vec2 is a self-supervised model designed for speech representation learning. It processes raw audio waveforms to learn contextualized speech features, ultimately producing speech embeddings. We utilized the Wav2Vec2 base model, which consists of 12 Transformer layers, to extract 768-dimensional feature embeddings and trained classifiers for the classification of 14 language classes. As depicted in Fig. 4(b), five FC layers with *tanh* activation function are used.

3. Experimental setups

3.1. Dataset

DoReCo dataset. AAL speech recordings are obtained from DoReCo (Language DOCumentation REference CORpus) [16], which consists of more than 50 low-resourced languages, totaling over 100 hours of recordings. Despite this broad linguistic coverage, only a few languages originate from Australia. Therefore, three AALs are analyzed in this work due to their relatively higher availability of recordings: *Warlpiri* [17], *Dalabon* [18], and *Light Warlpiri* [19]. The speech recordings are first pre-processed by removing low-quality and irrelevant speech, such as removing recordings with background noises (e.g., moving tables) or coughing. The remaining audio signals are then downsampled to 16 kHz. Each recording is seg-

mented into 10-second clips to fit the model input length and maintain consistency with the utterance length in the VoxLingua107 dataset. The final number of 10-second audio clips in Warlpiri, Dalabon, and Light Warlpiri are 786, 284, and 507, respectively.

VoxLingua107. To obtain the 14 languages for model training, we selected the corresponding speech recordings from the VoxLingua107 dataset. It is a popular world LID dataset, featuring realistic speech collected from YouTube videos. Audio recordings are extracted from these videos, resulting in 6.6k hours of content. This dataset consists of short speech segments (4-10 seconds) of 107 languages spread across the world. We only select 14 out of 107 languages for model development, with 1,000, 100, and 300 utterances for the training, validation, and test sets, respectively, from each language.

3.2. Implementation details

Parameters. To train the systems for identifying 14 languages, the Adam optimizer was employed with an initial learning rate optimized to 0.001 over 70 epochs. Cross-entropy was used as the loss function. The model was run nine times using different seed numbers to ensure robust results, and the average performance was reported. We utilized the Facebook [20] and SpeechBrain implementation [21] for Wav2Vec2 and VoxET.

Evaluations. Misclassification rate is used as the evaluation metric to determine the similarity as:

$$\sigma_i = \frac{N_{misclassified}^i}{N_{total}} \quad (1)$$

where $N_{misclassified}^i$ refers to number of AAL utterances misclassified to language i and N_{total} is the total number of test AALs utterances. The σ_i values will be ranked across all languages to identify the order of similarity.

For the tasks with pre-trained encoders, the trained model will first be evaluated using the trained 14 languages to guarantee its reliability. The performance is evaluated using the classification accuracy ρ as $\rho = 1 - \frac{N_{misclassified}}{N_{total}}$.

4. Results

4.1. LID performance

Fig. 5 depicts the misclassification rates of Warlpiri, Dalabon, and Light Warlpiri to the other 107 language classes. For clarity, only the top 20 languages are displayed. The results show that Warlpiri utterances are misclassified as Māori which is the most frequent misclassification, followed by Japanese and Finnish. Similarly, Dalabon is also most frequently misclassified into Māori, followed by Assamese and Tibetan. Light Warlpiri is most frequently misclassified into Assamese, Māori and Pushto. The high misclassification rates indicate that there may be phonetic or acoustic similarities between these AALs and other languages, particularly Māori, contributing to the model’s confusion. It is possibly due to shared regional or typological features between AAL and Māori, as Māori is the indigenous language of the Māori people of New Zealand which is geographically close to Australia [22]. These similarities can potentially assist in the annotation of AAL datasets, enhance linguistic understanding, and aid in AI model development.

4.2. Pre-trained encoders for similarity identification

To compare the similarities between AALs and the other 14 languages, it is essential that the models trained on these 14 languages are reliable and accurate. Therefore, we first evaluated

Table 1: LID accuracy for different languages using VoxET and Wav2Vec2 embeddings.

	Māori	Finnish	Tamil	Javanese	Malayalam	Swahili	Telugu	Tibetan	Pushto	Assamese	English	Mandarin	Spanish	Hindi
VoxET	0.990	0.996	0.995	0.988	0.997	0.999	0.990	0.998	0.996	0.990	0.991	0.998	0.993	0.994
Wav2Vec2	0.652	0.665	0.586	0.583	0.538	0.608	0.470	0.666	0.651	0.550	0.922	0.766	0.679	0.627

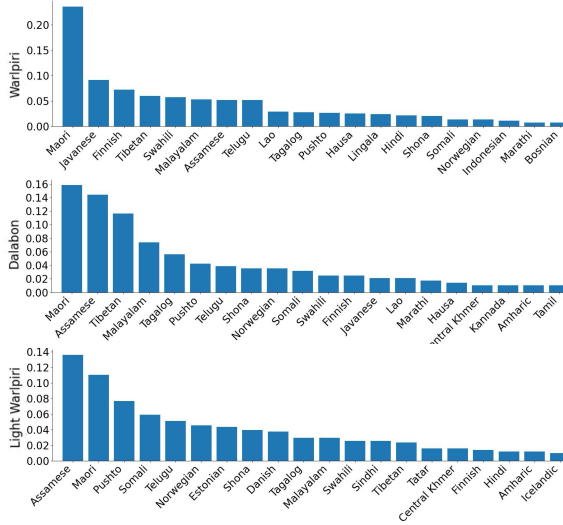


Figure 5: Misclassification rates of (a) Warlpiri (b) Dabalon and (c) Light Walpiri into other languages that the models are trained using VoxLingua107-ECAPA-TDNN model.

the performance of the models on these 14 languages to validate their effectiveness. Subsequently, we used the trained models to assess the similarities of AAL.

Performance for 14 languages. The classification accuracy ρ for 14 languages is listed in Table 1. It is observed that the model performs very well when trained and tested using VoxET embeddings for all 14 languages, validating the effectiveness of the trained models for the 14 languages for subsequent analysis.

AALs similarities to 14 languages. Fig. 6(a) reports the misclassification rates using multilingual embeddings VoxET. It is observed that the most frequently misclassified languages are similar to those in Section 4.1, with Māori standing out at the top of the list for all three Aboriginal languages. However, it is noted that AALs are rarely misclassified as English, Spanish, or Mandarin. This finding indicates that AALs do not share close phonetic or acoustic similarities with these widely spoken languages. Interestingly, among the high-resource languages analyzed, AALs are more frequently misclassified as Hindi. For instance, in the case of Warlpiri, Hindi ranks fifth in misclassification frequency as in Fig. 6(a). Similar findings are also observed for Dalabon and Light Warlpiri. This indicates AALs and Hindi may share common linguistic characteristics, consistent with [9].

Comparison between pre-trained encoders. We further evaluated the performance of Wav2Vec2 embeddings and compared them to VoxET embeddings to assess the impact of multilingual encoders on the similarity identifications. Initially, we examined the performance across 14 languages, as shown in Table 1, and compared with VoxET embeddings. We observed that Wav2Vec2 showed worse performance across all languages except English. This is because Wav2Vec2 is pre-trained exclusively on English, causing the embeddings for non-English languages to lack relevant information and potentially convert distinctive language features into English-related information.

Subsequently, we analyzed the similarity of AALs to the

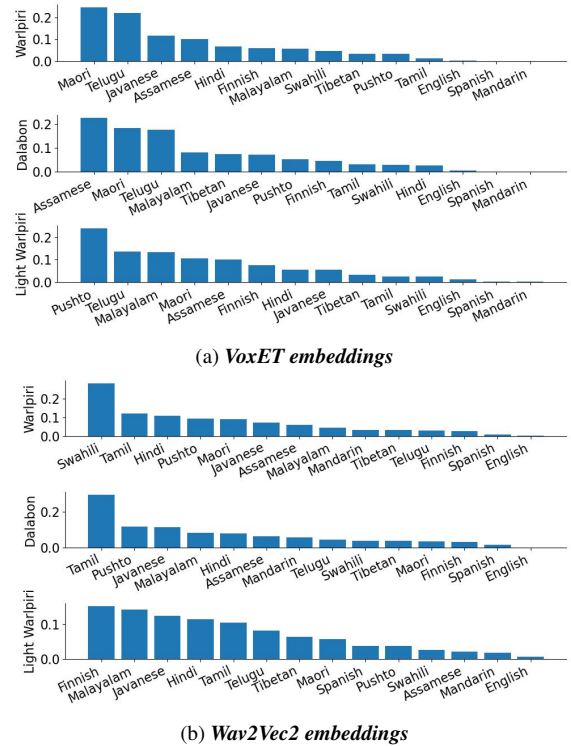


Figure 6: Misclassification rates using (a) VoxET embeddings and (b) Wav2Vec2 embeddings.

other 14 languages using this model, as in Fig. 6(b). Despite the differences in similarity identification between the two embeddings, they still share some commonalities. For Warlpiri, both models identify Māori and Hindi among the top five most similar languages. Similarly, for Dalabon and Light Warlpiri, Malayalam ranks in the top five for both models. This further confirms that AALs share strong similarities with certain other languages, regardless of the embeddings or models.

5. Conclusion

This study investigated which high-resource languages are most similar to Australian Aboriginal Languages (AALs) based on automatic language identification systems. Additionally, we analyzed the latent representations of AALs using an English-trained encoder and a multilingual encoder to further examine their linguistic relationships. Our experiments focused on three AALs—Warlpiri, Dalabon, and Light Warlpiri—and revealed that while these languages exhibit distinct characteristics that differentiate them from widely studied languages like English and Mandarin, they share certain properties with Māori, Hindi, and Malayalam. These findings provide valuable insights into the positioning of AALs in the speech embedding space and suggest potential pathways for leveraging high-resource languages in AI-driven speech processing. Future work includes validating these findings through linguistic analysis, exploring their implications for AI model development, and working toward greater inclusivity and representation of AALs in speech technology.

6. Acknowledgment

The authors would like to thank the School of Electrical Engineering and Telecommunications at UNSW Sydney, Australia, for providing funding for this research initiative.

7. References

- [1] R. M. Dixon, *Australian languages: Their nature and development*. Cambridge University Press, 2002.
- [2] Australian Government, Department of Infrastructure, Transport, Regional Development, Communications and the Arts, “National indigenous languages report,” <https://www.arts.gov.au/what-we-do/indigenous-arts-and-languages/indigenous-languages-and-arts-program/national-indigenous-languages-report>, 2020, accessed on 30/09/2024.
- [3] R. D. Australian Government Department of Infrastructure, Transport and Communications, “National indigenous languages report – chapter 3,” 2020, retrieved February 13, 2025. [Online]. Available: https://www.arts.gov.au/sites/default/files/documents/4national-indigenous-languages-report-pdf-chapter3_0.pdf
- [4] V. Pratap, A. Tjandra, B. Shi, P. Tomasello, A. Babu, S. Kundu, A. Elkahky, Z. Ni, A. Vyas, M. Fazel-Zarandi *et al.*, “Scaling speech technology to 1,000+ languages,” *Journal of Machine Learning Research*, vol. 25, no. 97, pp. 1–52, 2024.
- [5] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, “Robust speech recognition via large-scale weak supervision,” in *International conference on machine learning*. PMLR, 2023, pp. 28 492–28 518.
- [6] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, “Hubert: Self-supervised speech representation learning by masked prediction of hidden units,” *IEEE/ACM transactions on audio, speech, and language processing*, vol. 29, pp. 3451–3460, 2021.
- [7] J. Valk and T. Alumäe, “VoxLingua107: a dataset for spoken language recognition,” in *Proc. IEEE SLT Workshop*, 2021.
- [8] M. Laughren, K. Hale, J. E. Nungarrayi, M. P. P. Jangala, R. Hoogenraad, D. Nash, and J. Simpson, *Warlpiri Encyclopaedic Dictionary*. Aboriginal Studies Press, 2022.
- [9] S. R. Hamann, *The phonetics and phonology of retroflexes*. lot Utrecht, The Netherlands, 2003, vol. 75.
- [10] P. A. Busby, “The distribution of phonemes in australian aboriginal languages,” *Pacific Linguistics. Series A. Occasional Papers*, no. 60, p. 73, 1980.
- [11] A. Butcher, “Linguistic aspects of australian aboriginal english,” *Clinical linguistics & phonetics*, vol. 22, no. 8, pp. 625–642, 2008.
- [12] I. G. Malcolm, “Australian creoles and aboriginal english: phonetics and phonology,” *Varieties of English*, vol. 3, pp. 124–141, 2008.
- [13] H. Koch and R. Nordlinger, *The languages and linguistics of Australia*. Berlin, Germany: De Gruyter Mouton, 2014.
- [14] B. Desplanques, J. Thienpondt, and K. Demuyne, “Ecapadnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification,” *arXiv preprint arXiv:2005.07143*, 2020.
- [15] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, “wav2vec 2.0: A framework for self-supervised learning of speech representations,” *Advances in neural information processing systems*, vol. 33, pp. 12 449–12 460, 2020.
- [16] L. Paschen, F. Delafontaine, C. Draxler, S. Fuchs, M. Stave, and F. Seifart, “Building a time-aligned cross-linguistic reference corpus from language documentation data (doreco),” in *Proceedings of the Twelfth Language Resources and Evaluation Conference*, 2020, pp. 2657–2666.
- [17] C. O’Shannessy, “Warlpiri doreco dataset,” Language Documentation Reference Corpus (DoReCo) 1.2. Berlin & Lyon: Leibniz-Zentrum Allgemeine Sprachwissenschaft & laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), 2022, accessed on 30/09/2024. [Online]. Available: <https://doreco.huma-num.fr/languages/war1254>
- [18] M. Ponsonnet, “Dalabon doreco dataset,” Language Documentation Reference Corpus (DoReCo) 1.2. Berlin & Lyon: Leibniz-Zentrum Allgemeine Sprachwissenschaft & laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), 2022, accessed on 30/09/2024. [Online]. Available: <https://doreco.huma-num.fr/languages/ngal1292>
- [19] C. O’Shannessy, “Light warlpiri doreco dataset,” Language Documentation Reference Corpus (DoReCo) 1.2. Berlin & Lyon: Leibniz-Zentrum Allgemeine Sprachwissenschaft & laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), 2022, accessed on 30/09/2024. [Online]. Available: <https://doreco.huma-num.fr/languages/ligh1234>
- [20] F. AI, “wav2vec 2.0 base model,” <https://huggingface.co/facebook/wav2vec2-base>, accessed: 2024-11-08.
- [21] H. Face, “Speechbrain lang id model (voxlingua107 ecapa),” <https://huggingface.co/speechbrain/lang-id-voxlingua107-ecapa>, accessed: 2024-11-08.
- [22] R. A. Benton, *The Maori Language: Dying or Reviving?*. ERIC, 1997.