Advancing Multilingual Speaker Identification and Verification for Indo-Aryan and Dravidian Languages

Braveenan Sritharan
Dept. of Computer Science & Engineering, University of Moratuwa
Sri Lanka
braveenans.22@cse.mrt.ac.lk

Uthayasanker Thayasivam
Dept. of Computer Science & Engineering, University of Moratuwa
Sri Lanka
rtuthaya@cse.mrt.ac.lk

Abstract

Multilingual speaker identification and verification is a challenging task, especially for languages with diverse acoustic and linguistic features such as Indo-Aryan and Dravidian languages. Previous models have struggled to generalize across multilingual environments, leading to significant performance degradation when applied to multiple languages. In this paper, we propose an advanced approach to multilingual speaker identification and verification, specifically designed for Indo-Aryan and Dravidian languages. Empirical results on the Kathbath dataset show that our approach significantly improves speaker identification accuracy, reducing the performance gap between monolingual and multilingual systems from 15% to just 1%. Additionally, our model reduces the equal error rate for speaker verification from 15% to 5% in noisy conditions. Our method demonstrates strong generalization capabilities across diverse languages, offering a scalable solution for multilingual voice-based biometric systems.

1 Introduction

In today's world, biometric recognition is revolutionizing how we identify and verify individuals. Traditional methods, such as passwords, personal identification numbers, or signatures, are often inconvenient because they can be forgotten, stolen, or forged (Jain et al., 2004). In contrast, biometric traits are unique to each individual, making them difficult to replicate or steal. These systems rely on either physiological characteristics, such as fingerprints, iris patterns, or facial features, or behavioral traits, such as handwriting, voice, or keystroke patterns, to identify a person (Tolba et al., 2006).

Among these biometric traits, voice-based recognition offers clear advantages. Two factors make it a strong choice: First, speech is a natural and easy signal for users to provide. Second, the wide availability of phones and low-cost microphones make voice capture accessible and convenient for many applications (Reynolds, 2002). In voice-based biometric recognition, there are two distinct modes of operation: speaker identification, which typically involves recognizing an individual from a larger pool, and speaker verification, which focuses on validating a specific identity claim (Togneri and Pullella, 2011).

Voice-based recognition systems can be classified by their language handling capabilities into monolingual and multilingual systems (Nagaraja and Jayanna, 2012). Monolingual systems are trained and tested within a single language, offering high accuracy but limited flexibility outside that specific linguistic context. Multilingual systems, on the other hand, are designed to recognize speakers across multiple languages within a single model, eliminating the need for separate models for each language. This versatility makes multilingual systems well-suited for environments where multiple languages are spoken.

Recent advancements in self-supervised learning (SSL) have significantly enhanced the performance and robustness of voice-based recognition systems. SSL models, particularly in the context of the upstream model, play a crucial role in feature extraction. Here, rich speech features are captured and transferred to a downstream model, which is responsible for tasks such as speaker identification and verification (wen Yang et al., 2021). By separating the feature extraction and task-specific components, SSL models offer greater flexibility, improving the performance of voice recognition systems, particularly in multilingual applications.

Despite these advances, multilingual systems still lag behind their monolingual counterparts in terms of accuracy (Javed et al., 2023). This performance gap is particularly significant in multilingual countries such as India, where linguistic diversity presents a unique challenge. India's population speaks languages from four main language fam-

Proceedings of the First Workshop on Natural Language Processing for Indo-Aryan and Dravidian Languages (IndoNLP2025), pages 93–99

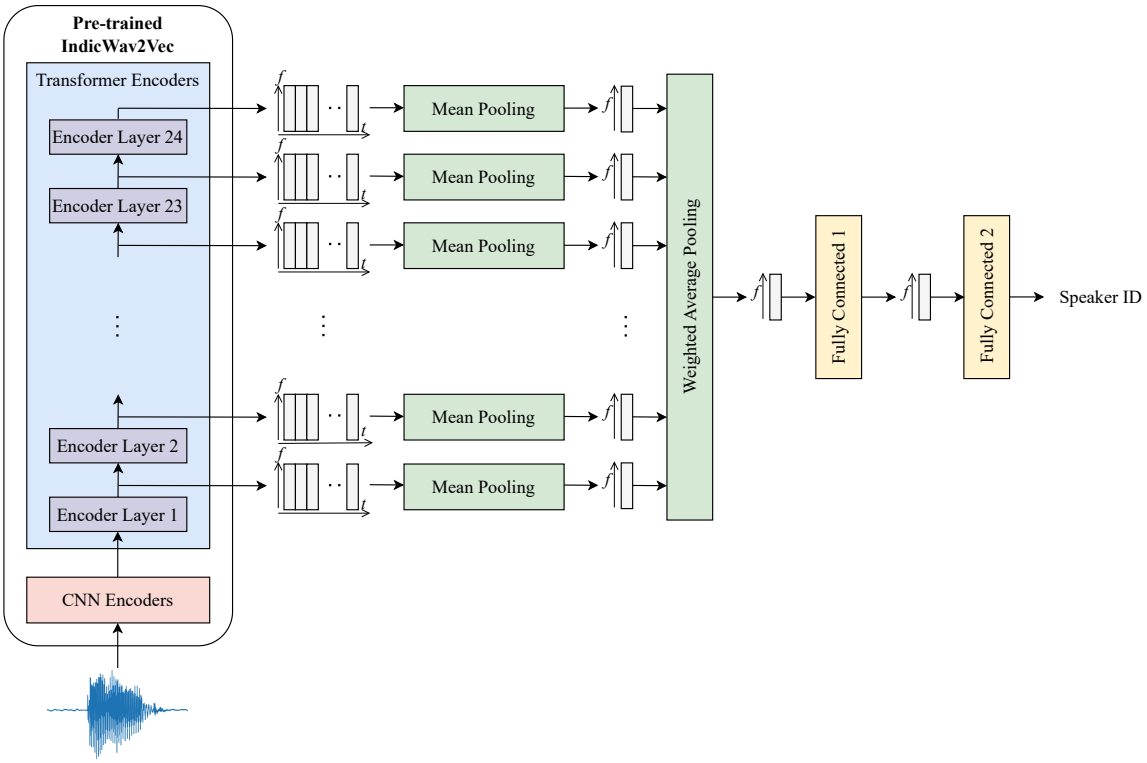

Figure 1: Architecture of our speaker identification model. The model processes an input speech signal in .m4a format (sampled at 16 kHz) using the pre-trained IndicWav2Vec model (Javed et al., 2023) to generate 24 frame-level representations. These are then mean-pooled along the time axis to create utterance-level representations. A weighted average pooling is applied across the 24 utterance-level representations to produce the final representation, which is passed through two fully connected layers to predict speaker identity. Layer dimensions and additional hyper-parameters are detailed in Section 3.

ilies, with approximately 96% of speakers using languages from the Indo-Aryan and Dravidian families, while the remaining languages have smaller speaker bases (Kakwani et al., 2020). In this context, a multilingual voice recognition system capable of handling multiple languages within a single model is crucial. It would eliminate the need for separate models for each language, streamlining speaker identification and verification processes across India's diverse linguistic landscape.

In this paper, we propose a novel architecture for voice-based biometric recognition using the pre-trained IndicWac2Vec model (Javed et al., 2023) to enhance both speaker identification and verification. Our model was tested under two conditions: clean and noisy environments. While there was a slight improvement in monolingual speaker identification accuracy, the major gain was in multilingual speaker identification accuracy, where the performance gap between monolingual and multilingual systems decreased from around 15% to 1%. Additionally, instead of creating a separate speaker verification model, we used the speaker embeddings from our speaker identification model for verification. Compared to the standard approach, our method reduced the equal error rate from 15% to 5% on unknown data in both clean and noisy conditions, demonstrating improved multilingual voice-based recognition.

2 Methodology

Our speaker identification model builds upon the architecture proposed by Javed et al.. To enhance the model's performance on speaker identification and verification tasks, we have introduced two key modifications, as illustrated in Figure 1.

2.1 Weighted Average Pooling Strategy

The original model employs mean pooling, which averages representations from all transformer encoder layers to generate a single vector. While straightforward, this approach assumes equal contribution from all layers, which may not align with the properties of speech representations. Prior stud-

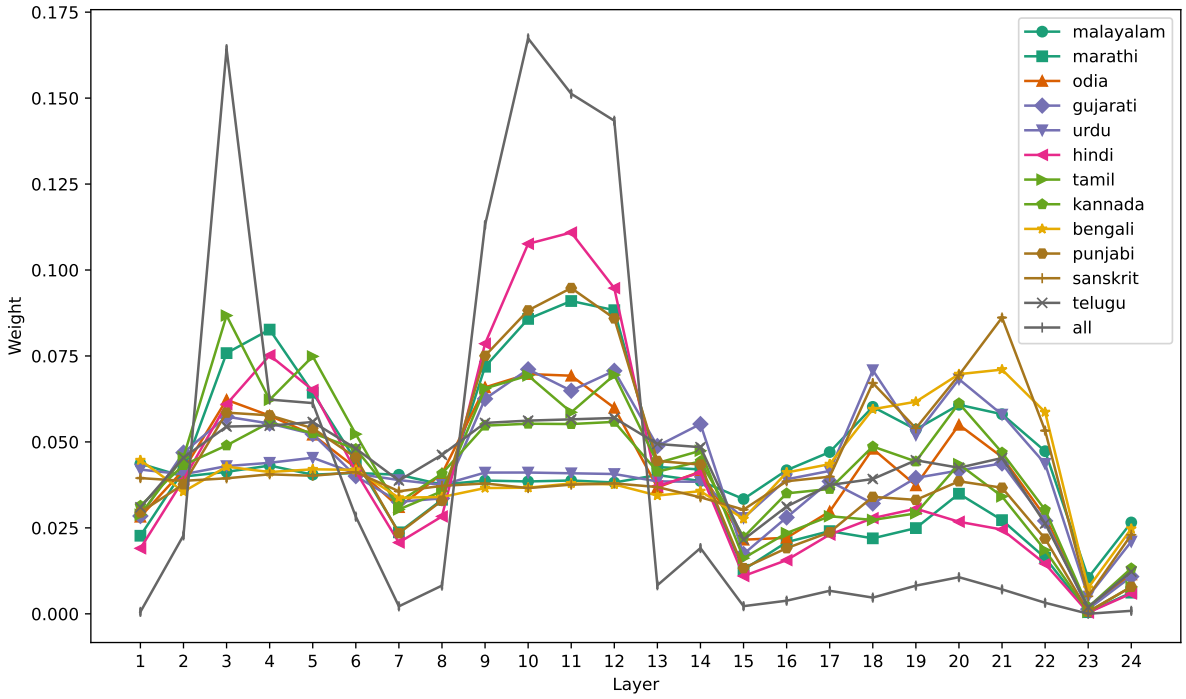

Figure 2: Encoder layer-wise representation weights for speaker identification models trained on specific languages and a multi-language dataset. The figure contains 13 subplots: each of the first 12 subplots shows a model trained exclusively on one language, labeled by the language name. The final subplot, labeled "all," displays results from a model trained on a combined dataset incorporating all 12 languages. This visualization highlights the variation in layer importance across language-specific models and the multi-language model.

ies (Chen et al., 2022) have shown that middle layers of transformer-based models often capture speaker-specific features more effectively than the initial or final layers.

To address this limitation, we employ a weighted average pooling strategy, which assigns learnable weights to each layer's representation. This approach enables the model to emphasize layers that capture speaker-specific features while reducing the contribution of less relevant layers. By prioritizing these layers, the model effectively exploits the hierarchical structure of transformer outputs, as supported by the findings in (Chen et al., 2022), which highlight the importance of middle layers for speaker-related tasks.

2.2 Additional Embedding Layer

In the baseline architecture, the aggregated representation is passed directly to the classifier. To enhance the model's ability to refine speaker-related features, we introduce an additional embedding layer, a linear transformation applied to the pooled representation before classification. This layer refines the pooled representation to better distinguish speaker-specific features, leveraging the hypothesis that increased depth improves feature separability (Shi et al., 2020). Additionally, the refined embeddings support both speaker identification and verification, providing a unified representation that enhances the model's accuracy and robustness.

The resulting model improves accuracy and robustness for speaker identification and verification. Details on layer dimensions and other hyperparameters are in Section 3.

3 Experimental Setup

Our speaker identification model consists of two main components: an upstream model and a downstream model. The upstream model, IndicWav2Vec, is a Wav2Vec-based model pretrained on half a million hours of raw speech data across 128 Indian languages (Javed et al., 2023). Following standard practice in speech processing benchmarks, such as SUPERB (wen Yang et al., 2021), we freeze the upstream model and train only the downstream model. This approach allows us to use the rich, pre-trained representations while reducing computational complexity, as only the downstream model is updated to predict speaker

Language	SID - Mono		SID - Multi					
	Clean	Noise	Dravidian		Indo-Aryan		All	
			Clean	Noise	Clean	Noise	Clean	Noise
Bengali	**99.64**	**99.63**	-	-	99.54	99.44	99.54	99.37
Gujarati	97.73	**95.97**	-	-	97.63	95.68	**97.79**	94.94
Hindi	**99.39**	**99.08**	-	-	99.23	98.68	99.17	98.62
Kannada	98.54	99.86	**99.00**	**100.00**	-	-	98.91	**100.00**
Malayalam	**99.94**	**99.93**	**99.94**	99.65	-	-	99.77	99.58
Marathi	**94.11**	97.97	-	-	**94.11**	98.39	93.82	**98.47**
Odia	98.17	**98.39**	-	-	**98.24**	97.82	98.10	97.37
Punjabi	**99.41**	99.24	-	-	99.13	99.19	98.94	**99.37**
Sanskrit	99.82	99.56	-	-	**99.94**	99.24	**99.94**	**99.62**
Tamil	96.41	96.73	**96.96**	**96.85**	-	-	96.12	96.65
Telugu	93.61	**96.48**	**94.81**	95.83	-	-	94.73	95.66
Urdu	**99.72**	**99.29**	-	-	99.62	99.20	99.27	99.23

Table 1: Performance of different languages on the Speaker Identification (SID) tasks, specifically for the Mono and Multi language settings, evaluated on both clean and noisy datasets. In the SID-Multi task, languages are grouped into three categories: Dravidian (a model trained on all Dravidian languages and tested on each language individually), Indo-Aryan (a model trained on all Indo-Aryan languages and tested on each language individually), and All (a model trained on all languages combined and tested on each language individually)

identity. [1] [2].

As the speaker identification task is framed as a classification problem, we use cross-entropy as the loss function and accuracy as the evaluation metric. In the speaker verification task, we first train the model for multilingual speaker identification, then extract speaker embeddings. These embeddings are compared using cosine similarity, and performance is evaluated using the Equal Error Rate (EER), which represents the point at which the false acceptance rate equals the false rejection rate. Hyper-parameter tuning, performed using grid search, was applied to both tasks to optimize the model's performance [3].

For evaluating our model's performance, we select the Kathbath dataset (Javed et al., 2023), which is particularly well-suited for speaker identification tasks involving Indo-Aryan and Dravidian languages. This dataset is the largest available for Indian languages, making it an ideal choice for multilingual speaker identification. It includes 8 Indo-Aryan languages—Gujarati, Marathi, Bengali, Odia, Hindi, Punjabi, Sanskrit, and Urdu—and 4 Dravidian languages—Kannada, Malayalam, Tamil, and Telugu. All 12 languages are widely spoken, ensuring the model's generalization across a diverse set of linguistic and acoustic features. The dataset is divided into four categories: Clean Known, Noise Known, Clean Unknown, and Noise Unknown, which allows for robust evaluation under varying conditions of noise and speaker familiarity. The "Clean" and "Noise" labels distinguish between clean and noisy audio, while "Known" and "Unknown" indicate whether the speaker is seen or unseen during training. We follow the recommended train-test splits for each dataset.

4 Results and Discussion

A key architectural modification in our model is the use of weighted average pooling for features extracted from the pre-trained IndicWav2Vec model, replacing traditional mean pooling. Figure 2 demonstrates that layer contributions are not uniform; notably, layers 9 through 12 consistently receive higher weights across all models. This suggests that these deeper layers play a substantial role in encoding speaker identity, as they may capture more abstract, speaker-specific features that are essential for accurate identification.

Furthermore, there is a strong correlation between the weight patterns in monolingual and multilingual models. Layers with relatively small weights in monolingual models appear even smaller in the multilingual model, while those with higher

[1] Our model was implemented using PyTorch.

[2] All experiments were conducted on an NVIDIA Quadro RTX 6000 GPU with 30GB of RAM.

[3] The fully connected layer has a dimension of 1500, with a batch size of 32 and a learning rate of 2.5×10^{-3}.

Model	Clean - Known	Clean - Unknown	Noisy - Known	Noisy - Unknown
Speaker Identification Monolingual (SID-Mono) - Accuracy				
XLS-R	94.2	-	92.4	-
IndicWav2Vec	95.6	-	95.2	-
Ours	**98.04**	-	**98.51**	-
Speaker Identification Multilingual (SID-Multi) - Accuracy				
XLS-R	70.71	-	69.22	-
IndicWav2Vec	79.26	-	78.08	-
Ours	**97.96**	-	**98.12**	-
Automatic Speaker Verification - EER				
XLS-R	2.15	12.05	2.83	11.58
IndicWav2Vec	**2.08**	15.33	**2.11**	15.39
Ours	4.61	**5.15**	5.23	**5.55**

Table 2: Performance comparison of different models on various tasks, including Speaker Identification (SID) in both monolingual (SID-Mono) and multilingual (SID-Multi) settings, and Automatic Speaker Verification (ASV). For SID-Mono and SID-Multi, the accuracy is reported for both clean and noisy conditions on known speakers. For ASV, the Equal Error Rate (EER) is reported for clean and noisy conditions on both known and unknown speakers. Ours denotes the model proposed in this work, which outperforms the other models, XLS-R and IndicWav2Vec, in most settings.

weights tend to be accentuated in the multilingual setting. This consistency suggests that the multilingual model captures a generalizable layer-wise structure across languages, reinforcing the importance of weighted pooling in effectively leveraging essential layers for robust speaker representation. These findings demonstrate that our approach preserves key features across languages, enhancing speaker identification accuracy.

Table 1 presents the performance of our model on the SID task across monolingual and multilingual settings, evaluated on both clean and noisy datasets. In the monolingual setting, the model achieves high accuracy on several languages, with Bengali, Hindi, and Malayalam exceeding 99% accuracy. However, languages like Marathi and Telugu show a drop in performance, particularly in noisy conditions. This indicates that noise significantly impacts speaker identification for these languages, potentially due to their unique acoustic characteristics. Overall, the monolingual performance demonstrates the model's capability to accurately identify speakers in controlled environments, though its performance is more sensitive to noise in certain languages.

In contrast, the multilingual setting shows a slight decrease in accuracy compared to the monolingual case, which is expected due to the added complexity of handling multiple languages. Nevertheless, the model trained on the "All" languages category maintains relatively high performance across languages, demonstrating strong generalization. The Dravidian and Indo-Aryan subsets perform similarly, with the Indo-Aryan model slightly outperforming others in some cases. Notably, the multilingual models exhibit better resilience to noise compared to the monolingual models, suggesting that training with multiple languages helps the model learn more robust speaker features. However, noise remains a challenge, and further improvements in noise robustness are needed for better performance in real-world conditions.

Next, Table 2 compares the performance of our model against two baseline models, XLS-R and IndicWav2Vec, across three tasks: SID in both monolingual (SID-Mono) and multilingual (SID-Multi) settings, and Automatic Speaker Verification (ASV). For both SID-Mono and SID-Multi tasks, our model consistently outperforms the baselines in terms of accuracy, particularly in noisy conditions. In the monolingual setting, our model achieves an accuracy of 98.04% for clean and 98.51% for noisy conditions, significantly surpassing the 95.6% and 95.2% accuracy of IndicWav2Vec and the 94.2% and 92.4% accuracy of XLS-R. Similarly, in the multilingual setting, our model shows remarkable performance, achieving 97.96% in clean and 98.12% in noisy conditions, well ahead of both XLS-R and IndicWav2Vec.

However, when it comes to ASV, our model lags behind the baselines in terms of Equal Error Rate (EER). While XLS-R and IndicWav2Vec achieve

EER values ranging from 2.08 to 2.83 for clean conditions and 11.58 to 15.39 for noisy conditions, our model exhibits better EER values, particularly in unknown conditions, with the best value being 5.15 for unknown speakers in clean conditions and 5.55 for unknown speakers in noisy conditions. These results suggest that while our model excels in speaker identification tasks, further improvements in ASV, especially under known conditions, are necessary. Despite the performance gap in ASV, the results highlight the robustness of our model in SID tasks across both monolingual and multilingual settings, making it a promising candidate for practical voice recognition applications.

5 Conclusion

In this work, we presented a novel approach for multilingual speaker identification and verification using a modified IndicWav2Vec-based model. Our model integrates self-supervised learning techniques to extract rich, robust speech features, which substantially improve speaker identification performance, especially in multilingual settings. Key innovations include a weighted average pooling mechanism for better aggregation of transformer layer representations and an additional embedding layer to refine speaker-specific features. These modifications led to significant improvements, reducing the performance gap between monolingual and multilingual systems from 15% to 1%, and lowering the equal error rate for speaker verification from 15% to 5% under noisy conditions. Our experiments, conducted with the Kathbath dataset, demonstrated the model's ability to generalize effectively across multiple languages. The simplicity of the model structure, combined with its robust performance, positions it as an efficient and scalable solution for voice-based biometric recognition.

6 Limitation

Despite the promising results, our model still faces several limitations. Although it excels in multilingual speaker identification and verification, its performance is limited by the diversity of the training dataset, as it relies heavily on the Kathbath dataset. Expanding the training data to cover a wider variety of languages and acoustic conditions will be crucial for enhancing generalization. Additionally, while the model performs well under clean and moderately noisy conditions, its robustness in highly noisy environments remains a challenge. The equal error rate, though reduced in typical scenarios, may degrade in real-world applications with severe noise or poor-quality recordings. Lastly, the model's computational complexity, especially with the added pooling and embedding layers, may limit its suitability for real-time or resource-constrained applications.

Acknowledgment

We would like to express our gratitude to the National Languages Processing (NLP) Center and DataSEARCH Research Center at the University of Moratuwa for providing the GPUs required to carry out the experiments for this research.

References

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, et al. 2022. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518.

Anil K Jain, Arun Ross, and Salil Prabhakar. 2004. An introduction to biometric recognition. *IEEE Transactions on circuits and systems for video technology*, 14(1):4–20.

Tahir Javed, Kaushal Bhogale, Abhigyan Raman, Pratyush Kumar, Anoop Kunchukuttan, and Mitesh M Khapra. 2023. Indicsuperb: A speech processing universal performance benchmark for indian languages. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 12942–12950.

Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, Gokul N.C., Avik Bhattacharyya, Mitesh M. Khapra, and Pratyush Kumar. 2020. IndicNLPSuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for Indian languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4948–4961, Online. Association for Computational Linguistics.

BG Nagaraja and HS Jayanna. 2012. Mono and cross lingual speaker identification with the constraint of limited data. In *International Conference on Pattern Recognition, Informatics and Medical Engineering (PRIME-2012)*, pages 439–443. IEEE.

Douglas A Reynolds. 2002. An overview of automatic speaker recognition technology. In *2002 IEEE international conference on acoustics, speech, and signal processing*, volume 4, pages IV–4072. IEEE.

Yanpei Shi, Qiang Huang, and Thomas Hain. 2020. H-vectors: Utterance-level speaker embedding using a

hierarchical attention model. In *ICASSP 2020-2020 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 7579–7583. IEEE.

Roberto Togneri and Daniel Pullella. 2011. An overview of speaker identification: Accuracy and robustness issues. *IEEE circuits and systems magazine*, 11(2):23–61.

AS Tolba, AH El-Baz, and AA El-Harby. 2006. Face recognition: A literature review. *International Journal of Signal Processing*, 2(2):88–103.

Shu wen Yang, Po-Han Chi, Yung-Sung Chuang, Cheng-I Jeff Lai, Kushal Lakhotia, Yist Y. Lin, Andy T. Liu, Jiatong Shi, Xuankai Chang, Guan-Ting Lin, Tzu-Hsien Huang, Wei-Cheng Tseng, Ko tik Lee, Da-Rong Liu, Zili Huang, Shuyan Dong, Shang-Wen Li, Shinji Watanabe, Abdelrahman Mohamed, and Hung yi Lee. 2021. Superb: Speech processing universal performance benchmark. In *Interspeech 2021*, pages 1194–1198.