INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Efficient Spoken Language Recognition via Multilabel Classification

Oriol Nieto, Zeyu Jin, Franck Dernoncourt, Justin Salamon

Adobe Research, San Francisco, CA, USA

onieto@adobe.com

Abstract

Spoken language recognition (SLR) is the task of automatically identifying the language present in a speech signal. Existing SLR models are either too computationally expensive or too large to run effectively on devices with limited resources. For real-world deployment, a model should also gracefully handle unseen languages outside of the target language set, yet prior work has focused on closed-set classification where all input languages are known a-priori. In this paper we address these two limitations: we explore efficient model architectures for SLR based on convolutional networks, and propose a multilabel training strategy to handle non-target languages at inference time. Using the VoxLingua107 dataset, we show that our models obtain competitive results while being orders of magnitude smaller and faster than current state-of-the-art methods, and that our multilabel strategy is more robust to unseen non-target languages compared to multiclass classification.

Index Terms: spoken language recognition, efficient architectures, multilabel classification

1. Introduction

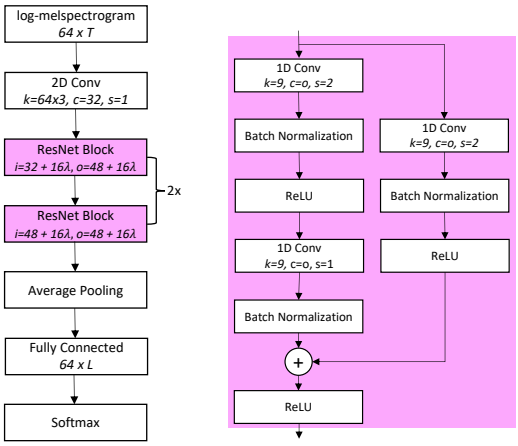
Automatic speech recognition (ASR) techniques have achieved near human accuracy for various languages [1, 2], and enable applications such as text-based audio/video editing [3, 4], voice translation [5], and virtual assistants [6]. When the language is not known in advance, Spoken Language Recognition (SLR) is often a necessary first step before running ASR, as most ASR systems require this information (the target language) to correctly transcribe the speech signal. Past research in SLR has been organized around challenges such as NIST Language Recognition Evaluations [7], focusing on improving accuracy with large neural networks and extensive data. Recent advances in large-scale and self-supervised models have achieved impressive generalization across hundreds of languages with near-perfect accuracy [8, 9, 10].

Despite the significant progress, there are two important limitations that make real-world deployment of such models challenging in resource-constrained scenarios: (1) models are too large or resource intensive for devices with limited compute power, (2) they operate on a closed set of target languages, i.e., they are not designed to handle the scenario where an unknown language is presented to the model at inference time. Most modern SLR systems are based on deep architectures, all of which contain millions if not tens or hundreds of millions of parameters. One of the smallest models is based on X-vectors [11], which were initially introduced for the task of speaker recognition [12]. Such vectors are computed using blocks of Time-Delayed Neural Networks (TDNNs) [13] and an aggregation layer, allowing for a large receptive field with small kernels, yielding good performance with a reduced amount of parameters. A larger and more powerful version of this model was recently introduced for the same task of speaker recognition, and it uses higher capacity TDNNs with an emphasized channel attention, propagation, and aggregation (ECAPA) set of mechanisms [14]. Such a model was later used successfully for SLR [15]. ECAPA-TDNNs have around four times the number of parameters as the original X-vector model. More recent and larger systems are capable of jointly recognizing a given language and solving additional tasks in a single pass. For example, Whisper [8] can perform multi-lingual speech to text transcription, having an implicit SLR system embedded in the model. This architecture is an encoder-decoder Transformer [16] trained on 680k hours of speech. The smallest version of this model has almost 2x the parameters of ECAPA-TDNN, and its medium version is over an order of magnitude larger than its smallest counterpart. Finally, XLS-R [9] is another self-supervised model whose smallest version has around 300 million parameters. It is based on Wav2Vec 2.0 [17] and is pre-trained with over half a million hours of speech covering 128 languages. When fine-tuned on the SLR task, it achieves state-of-the-art results on most test datasets. While self-supervised models lead the charts in terms of accuracy, they require significant compute power to operate.

There is a growing need for robust SLR models that can run efficiently embedded on-device. SLR on-device eliminates computation and networking costs involved with running SLR in the cloud, and can prevent tracking and other potential threats to user privacy. However, large-scale models such as those discussed above are impractical to use on, e.g., a mobile device, due to size and runtime constraints. There are use cases where a user only requires on-device SLR for a limited number of languages (e.g., the languages they speak), not hundreds, providing an opportunity to trade-off the language set size for model efficiency. Furthermore, the models in the studies discussed above are incapable of detecting unsupported languages: they assume the input is always one of the target languages, and would incorrectly classify a non-target language as one of the target languages the model was trained on. For real-world deployment, a model should be able to identify this scenario and "fail gracefully" by classifying the input as "unknown" or "other".

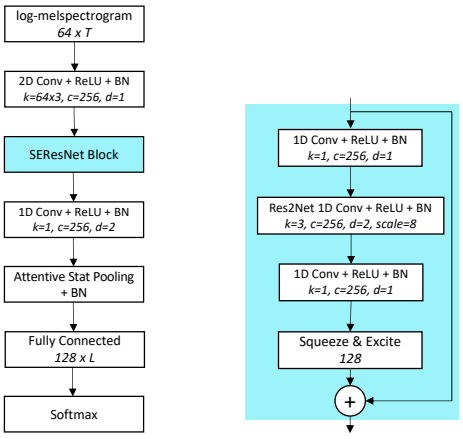
In this paper we address these two aforementioned limitations. We investigate small-footprint models for SLR that can run effectively on-device. We propose novel variants of widely-used architectures, including TC-ResNets [18] and ECAPA-TDNNs [13], and compare them to top-performing large models using VoxLingua107 [19], a sizeable speech dataset recently introduced for SLR research [9]. We show that our proposed lightweight architectures achieve competitive error rates with models that are two orders of magnitude larger in terms of pa-

10.21437/Interspeech.2023-1986



(a) TC-ResNet10 (b) ResNet Block

Figure 1: *Proposed TC-ResNet10 architecture, where T is time frames, c number of channels, s stride, L number of languages, and i and o input and output to the ResNet Block, respectively.*



(a) LECAPAT (b) SEResNet Block

Figure 2: *Proposed LECAPAT architecture. d is dilation size, BN is Batch Normalization [22], and* scale *is the scale for the Res2Net model [23]. All convolution strides are set to $s = 1$.*

rameters. To handle non-target (unseen) languages, we propose a multilabel classification approach that is novel in the context of SLR, and show that it produces models that are more robust compared to modeling all non-target languages via a single "Other" class in the commonly used multiclass setup. To the best of our knowledge, this is the first work to study SLR from an efficiency viewpoint and address the problem of non-target languages at inference time.

2. Models for Efficient SLR

We explore two families of architectures for efficient SLR: Temporal Convolution ResNets (TC-ResNets) [18], and ECAPA-TDNN [15]. For the latter, we propose a modification that makes it significantly lighter, which we call LECAPAT. For all the proposed models, our input is a log-melspectrogram with 64 mel-frequency bins computed from an input audio signal sampled at 16 kHz. The mel-spectrogram is computed using a 25 ms Hann window, FFT size of 64 ms, and a 10 ms hop size.

2.1. Temporal Convolution Residual Networks

Temporal convolution networks apply a 2D convolution to the input spectrogram with a kernel whose height matches the number of frequency bins, compacting the frequency dimension into a set of 1D time representations (one per channel). Then 1D convolutions can be applied in the subsequent layers of the architecture, making it highly efficient in both number of parameters and inference time. These networks were employed for early approaches to music recommendation using deep learning [20], and more recently added residual connections [21] for efficient keyword spotting [18].

We propose two different flavors of TC-ResNets for efficient SLR: TC-ResNet10 is a slightly larger version of the model introduced in [18], but with fewer layers (10 as opposed to 14), depicted in Figure 1. The two ResNet blocks are repeated (i.e., four ResNet blocks in total), where their size changes based on the repetition number $\lambda \in \{0, 1\}$. This model has 200k parameters. The second model is TC-ResNet14, introduced by Choi et al. [18], with 100k parameters.

2.2. LECAPAT: Light ECAPA Time-Delayed NNs

TDNNs [13] were introduced to efficiently model long temporal contexts. They use dilated convolutions to increase the receptive field, and are thus capable of modeling long-term structures without adding extra complexity in terms of computation or model size. ECAPA-TDNNs [14] added the following set of mechanisms: (i) per-channel attention, (ii) squeeze-excitation (SE) residual blocks that contain an SE mechanism [24] and a Res2Net model [23], and (iii) an aggregation technique that attends to several levels of feature statistics across the architecture. While more computationally expensive, ECAPA-TDNNs achieve better performance on speaker recognition, and have been shown to be highly effective for SLR [15].

We propose a more efficient and lightweight version of ECAPA-TDNN, LECAPAT (for Light ECAPA-TDNN). Our main intuition is that, while the original model was designed to recognize thousands or millions of speakers, SLR typically involves recognizing dozens or at most hundreds of languages, suggesting a model with reduced capacity may still perform well on this task. We reduce the complexity of ECAPA-TDNN making it two orders of magnitude smaller: from 21 million parameters down to 600k. LECAPAT employs a single SE residual block instead of three, and we reduce the number of parameters in each of the main blocks, as depicted in Figure 2.

2.3. Multilabel Classification

ASR applications typically only support a subset of languages. Thus, an SLR model should recognize when a non-target language, i.e., a language that is not one of the languages the model is trained to recognize, is provided as input, to avoid sending it to be transcribed by ASR. A common approach in multiclass tasks for handling non-target classes is to add a single "Other" class. Our early experiments revealed that this approach, i.e., adding a single class for all non-target languages, resulted in poor performance for SLR. The "Other" class must capture many different languages, including some that may be similar to a target language. Our conjecture is that the model struggles to simultaneously group all non-target languages *and* separate them from the target languages. For example, in our experiments Spanish and German are target languages, while Catalan and Yiddish are not. The multiclass setup has to group Cata-

lan and Yiddish and at the same time separate Catalan from the relatively close Spanish and Yiddish from the relatively close German, a challenging proposition.

Instead, we propose to train a *multilabel* model for SLR, a novel approach in this context. Unlike the multiclass setup which is forced to always return a single positive class, multilabel models allow for (1) several classes to be positive at the same time and (2) no positive classes for a given input. Thus, the model can focus on representing the target languages and separating them from non-target languages, without explicitly modeling all possible non-target languages. Implementation-wise, we swap the final softmax layer in our models with sigmoids, such that zero positives are allowed. We return the language with the highest output activation as the model's prediction, unless all activations are below a threshold (0.5) in which case the "Other" class is detected. Note that this adds no extra complexity to the models during inference.

3. Experiments

3.1. Data

We use the VoxLingua107 dataset [19] for all of our experiments. The dataset consists of a training set and a "development" set for model evaluation. The training set comprises 6,628 hours of speech from 107 languages, with an average of 62 hours per language, though it is heavily imbalanced. The development set contains 1,609 manually validated speech segments from 33 languages.

Our focus is on real-world scenarios where the target language set is limited, as is often the case for speech applications, for example due to localization constraints. This common scenario is one for which efficient models are particularly promising as, we hypothesize, the reduced language set suggests we should be able to train performant models with significantly lower capacity compared to models targeting hundreds of languages. To this end, we construct a new test set taking a subset of $L = 11$ languages from the VoxLingua107 development (evaluation) set, which are the languages for which speech-to-text (ASR) is supported in a commercial video editing tool. We call this test set VoxLingua11, with the distribution of seconds per language depicted in Figure 3. We maintain the class imbalance for these languages from the VoxLingua107 development set, to make VoxLingua11 easy to reproduce.

We use VoxLingua11 to evaluate performance of our models and baselines under the multiclass setup where only target languages are included in the test set. Since our final goal is to evaluate our models when non-target languages are present in the test set, we also create VoxLingua11+O, which expands VoxLingua11 with an "Other" class containing 128 samples with at least one sample from each of the non-target languages in the VoxLingua107 development set. VoxLingua11+O can be used to evaluate models under both the multiclass and multilabel setups, where in the former models have 12 output neurons (11 target languages + "Other"), and in the latter they have 11 output neurons and "Other" is predicted as described in Section 2.3.

To make the models robust to real-world recordings, we augment the training set with noise, reverb, and random equalization following the work of Su et al. [25]. Since the training set is imbalanced, we use the augmentation to balance our training data, such that each training epoch contains a balanced sampling of the target languages. The "Other" class is treated as a 12th language by randomly sampling across the non-target languages during training. All the models evaluated in this work take log-melspectrograms of 10-second audio clips as input. If the input recording is shorter than 10 s, we center it and zero pad to the left and right to obtain a 10 s clip. If it is longer, we take a random 10 s subclip during training. At test time, for longer recordings we slide a 10 s window over the input with a 5 s hop size and average the predictions over time.

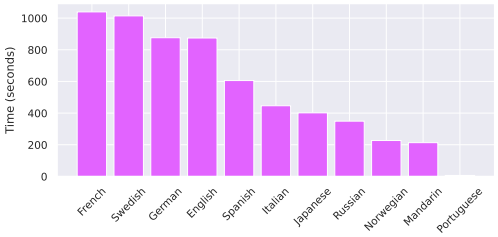


Figure 3: *Seconds per language in the VoxLingua11 test set.*

3.2. Baselines

We use the SLR models reviewed in Section 1 as baselines. The open source implementations of the X-Vector[1] and ECAPA-TDNN [26] are trained from scratch on the VoxLingua107 training set. We verify that we obtain similar results to those reported in the original publications. For the larger baselines, XLS-R and Whisper, we rely on their published weights, as these higher capacity models were trained on a diverse conglomerate of datasets–including VoxLingua107–that make up much larger amounts of data. We use the 300M parameter (i.e., smallest) version of XLS-R fine-tuned on VoxLingua107[2] and the "Medium" and "Tiny" versions of Whisper [8].

3.3. Metrics and Optimization

We report the models' error rate err, the standard for SLR evaluation, where $err = 100(1 - acc)$ and acc is the multiclass classification accuracy. We report the inference runtime of each model on VoxLingua11 using either a CPU (Intel Cascade Lake on a g4dn.4xlarge G4 AWS instance) or a V100 GPU as a real-time factor (rtf) coefficient, i.e., the total test set audio duration divided by the total model inference time (larger rtf = faster).

Our proposed models are trained on the augmented VoxLingua107 training set using a V100 GPU, the Adam optimizer (β_1=0.9 and β_2=0.999), early stopping, and hyperparameter Bayesian optimization to determine learning rates and batch sizes, which fluctuate between $[10^{-3}, 10^{-5}]$ and $[32, 128]$, respectively. We minimize the categorical and binary cross-entropy loss for multiclass and multilabel models, respectively.

4. Results

4.1. Accuracy versus size and runtime

We start by examining the trade-off between model capacity and error rate, depicted in Figure 4, with our proposed models in bold. Results are shown for multiclass classification on VoxLingua11 so that we can compare to all baselines. The models trained from scratch only see the 11 selected languages from VoxLingua107 during training. The error rates are also reported in Table 1. Note the log scale of the parameters (y) axis. Expectedly, we see a correlation between model size and error rate

[1] https://github.com/KrishnaDN/x-vector-pytorch

[2] https://huggingface.co/TalTechNLP/voxlingua107-xls-r-300m-wav2vec

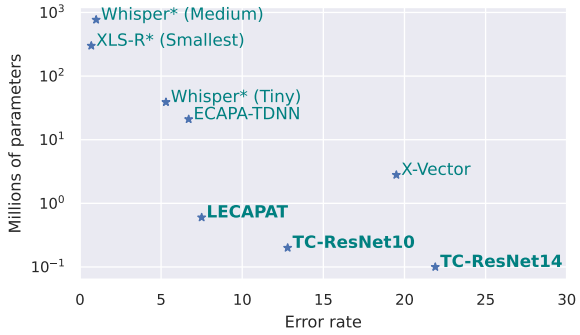


Figure 4: *Model size versus error rate, multiclass classification on VoxLingua11. Proposed models in bold. * indicates original weights used, all other models were trained from scratch.*

Model	*err* ↓	rtf (CPU) ↑	rtf (GPU) ↑
XLS-R (Smallest)	0.7	11.02	13.03
Whisper (Medium)	1.0	3.43	28.31
Whisper (Tiny)	5.3	39.54	65.56
ECAPA-TDNN	6.7	83.95	522.40
X-Vector	19.5	524.2	656.54
LECAPAT (Ours)	7.5	678.42	810.93
TC-ResNet10 (Ours)	12.8	966.91	821.56
TC-ResNet14 (Ours)	21.9	932.12	824.26

Table 1: *Multiclass error rate and rtf on VoxLingua11.*

in both the baselines and our proposed models. For the baselines, we see Whisper (Tiny) and ECAPA-TDNN provide an order-of-magnitude (OOM) reduction in model size compared to Whisper (Medium) and XLS-R (20-40M vs 300-800M) with a relatively low drop in error rate of 4-5 points. X-Vector is a further OOM smaller but its error rate is significant.

Most interestingly, we see that our proposed LECAPAT model performs very closely to the Whisper (Tiny) and ECAPA-TDNN baselines with an error rate of 7.5, but is a whole two OOM smaller (and four OOM smaller than the largest baselines), with just 0.6M parameters. The TC-ResNet models, though smaller still, are notably less accurate, thus providing a less desirable tradeoff. LECAPAT (and the ResNets) are also significantly faster to run inference with, as shown in Table 1, running over 800 times faster than real-time on GPU, which is 12 times faster than Whisper (Tiny) and 1.5 times faster than ECAPA-TDNN. It is interesting to note that despite their similar size and accuracy, the latter is around 8 times faster than the former, which shows that model size is not a good proxy for model runtime and highlights the importance of reporting runtime metrics when considering models for deployment. It is also noteworthy that LECAPAT remains fast on CPU, whereas ECAPA-TDNN becomes 6 times slower. This makes our proposed model a strong candidate for deployment scenarios where a GPU may not be available. While our best proposed model is outperformed by the strongest baselines in terms of accuracy, it trades off 6 accuracy points for a dramatic four OOM reduction in size and almost two OOM increase in speed.

4.2. Dealing with non-target languages

Finally, we examine how our proposed multilabel training strategy for handling non-target languages at inference time (cf. Section 2.3) compares to a multiclass setup which explicitly models them via an additional "Other" class. For this experiment, we train both multiclass and multilabel models with the full VoxLingua107 training set. The results for our proposed models on VoxLingua11+O using both strategies are reported in Table 2. We note the error rates are generally higher than those for VoxLingua11. This is expected, as the classification problem is more challenging now given the presence of the additional, highly heterogeneous, class "Other". It also shows that studies which only evaluate SLR against a known set of target languages are most likely not representative of how models would perform under deployment conditions where non-target languages may be presented to the model. We see that our multilabel strategy significantly outperforms the multiclass setup for all three models, with the improvement being most notable for our top-performing LECAPAT model. We hypothesize that the multiclass model struggles to learn an embedding space that can model a large number of non-target languages as a single "Other" class while keeping it disjoint from the target languages. To explore this, we plot the normalized confusion matrices for LECAPAT trained with each of the two strategies in Figure 5. We see the multiclass model (left) is able to better classify non-target languages as "Other," but suffers from notable leakage of target classes into the "Other" class. The multilabel model (right) alleviates this considerably, e.g., improving the accuracy for Spanish, Italian, Russian, and Norwegian by over 10 percentage points each.

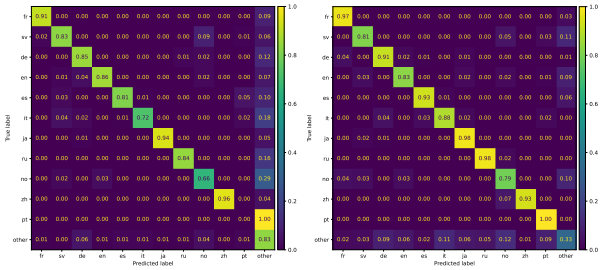


Figure 5: *Normalized confusion matrices for LECAPAT on VoxLingua11+O for multiclass (left) and multilabel (right).*

Model	Multiclass	Multilabel
LECAPAT	22.56	**14.85**
TC-ResNet10	29.02	**23.93**
TC-ResNet14	32.07	**31.01**

Table 2: *Error rates for our proposed models using the multiclass and multilabel training strategies on VoxLingua11+O.*

5. Conclusion

In this paper we addressed the problem of efficient Spoken Language Recognition (SLR) in the presence of non-target languages. To the best of our knowledge this is the first study on efficient SLR, and the first to propose multilabel training to handle non-target languages at inference time. Our experiments show that our top-performing proposed model, LECAPAT, performs almost on par with models that are two OOM larger, and only 6 accuracy points lower than the largest models while being four OOM smaller and almost two OOM faster. Additionally, we show that our proposed multilabel training strategy outperforms the multiclass setting by a considerable margin when non-target languages are present at inference time. We hope this study will stimulate further work on the topic of efficient SLR under real-world conditions.

6. References

[1] D. Wang, X. Wang, and S. Lv, "End-to-end mandarin speech recognition combining CNN and BLSTM," *Symmetry*, vol. 11, no. 5, 2019. [Online]. Available: https://www.mdpi.com/2073-8994/11/5/644

[2] D. Amodei, S. Ananthanarayanan, R. Anubhai, J. Bai, E. Battenberg, C. Case, J. Casper, B. Catanzaro, Q. Cheng, G. Chen *et al.*, "Deep speech 2: End-to-end speech recognition in english and mandarin," in *Proc. International Conference on Machine Learning*, 2016, pp. 173–182.

[3] Z. Jin, G. J. Mysore, S. Diverdi, J. Lu, and A. Finkelstein, "Voco: Text-based insertion and replacement in audio narration," *ACM Transactions on Graphics*, vol. 36, no. 4, pp. 1–13, 2017.

[4] O. Fried, A. Tewari, M. Zollhöfer, A. Finkelstein, E. Shechtman, D. B. Goldman, K. Genova, Z. Jin, C. Theobalt, and M. Agrawala, "Text-based editing of talking-head video," *ACM Transactions on Graphics*, vol. 38, no. 4, pp. 1–14, 2019.

[5] M. Federico, R. Enyedi, R. Barra-Chicote, R. Giri, U. Isik, A. Krishnaswamy, and H. Sawaf, "From speech-to-speech translation to automatic dubbing," in *Proc. International Conference on Spoken Language Translation*, 2020, pp. 257–264. [Online]. Available: https://aclanthology.org/2020.iwslt-1.31

[6] V. Kepuska and G. Bohouta, "Next-generation of virtual personal assistants (microsoft cortana, apple siri, amazon alexa and google home)," in *Proc. IEEE Computing and Communication Workshop and Conference*, 2018, pp. 99–103.

[7] S. O. Sadjadi, C. Greenberg, E. Singer, L. Mason, and D. Reynolds, "The 2021 NIST Speaker Recognition Evaluation," in *Proc. The Speaker and Language Recognition Workshop*, 2022, pp. 322–329.

[8] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," OpenAI, Tech. Rep., 2022.

[9] A. Babu, C. Wang, A. Tjandra, K. Lakhotia, Q. Xu, N. Goyal, K. Singh, P. von Platen, Y. Saraf, J. Pino, A. Baevski, A. Conneau, and M. Auli, "XLS-R: Self-supervised Cross-lingual Speech Representation Learning at Scale," in *Proc. Interspeech*, 2022, pp. 2278–2282.

[10] H. Liu, L. P. G. Perera, A. W. H. Khong, E. S. Chng, S. J. Styles, and S. Khudanpur, "Efficient self-supervised learning representations for spoken language identification," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1296–1307, 2022.

[11] D. Snyder, D. Garcia-Romero, A. McCree, G. Sell, D. Povey, and S. Khudanpur, "Spoken Language Recognition using X-vectors," in *Proc. The Speaker and Language Recognition Workshop*, 2018, pp. 105–111.

[12] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, 2018, pp. 5329–5333.

[13] V. Peddinti, D. Povey, and S. Khudanpur, "A time delay neural network architecture for efficient modeling of long temporal contexts," in *Proc. Interspeech*, 2015, pp. 3214–3218.

[14] B. Desplanques, J. Thienpondt, and K. Demuynck, "ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification," in *Proc. Interspeech*, 2020, pp. 3830–3834.

[15] M. Ravanelli, T. Parcollet, P. Plantinga, A. Rouhe, S. Cornell, L. Lugosch, C. Subakan, N. Dawalatabad, A. Heba, J. Zhong, J.-C. Chou, S.-L. Yeh, S.-W. Fu, C.-F. Liao, E. Rastorgueva, F. Grondin, W. Aris, H. Na, Y. Gao, R. D. Mori, and Y. Bengio, "SpeechBrain: A general-purpose speech toolkit," 2021, arXiv:2106.04624.

[16] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, vol. 30. Curran Associates, Inc., 2017, pp. 5998–6008. [Online]. Available: https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[17] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Advances in Neural Information Processing Systems*, vol. 33. Curran Associates, Inc., 2020, pp. 12 449–12 460. [Online]. Available: https://proceedings.neurips.cc/paper/2020/file/92d1e1eb1cd6f9fba3227870bb6d7f07-Paper.pdf

[18] S. Choi, S. Seo, B. Shin, H. Byun, M. Kersner, B. Kim, D. Kim, and S. Ha, "Temporal Convolution for Real-Time Keyword Spotting on Mobile Devices," in *Proc. Interspeech*, 2019, pp. 3372–3376.

[19] J. Valk and T. Alumäe, "VoxLingua107: a dataset for spoken language recognition," in *Proc. IEEE Spoken Language Technology Workshop*, 2021.

[20] A. van den Oord, S. Dieleman, and B. Schrauwen, "Deep content-based music recommendation," in *Advances in Neural Information Processing Systems*, vol. 26. Curran Associates, Inc., 2013, pp. 2643–2651. [Online]. Available: https://proceedings.neurips.cc/paper/2013/file/b3ba8f1bee1238a2f37603d90b58898d-Paper.pdf

[21] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[22] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. International Conference on Machine Learning*, 2015, p. 448–456.

[23] S.-H. Gao, M.-M. Cheng, K. Zhao, X.-Y. Zhang, M.-H. Yang, and P. Torr, "Res2net: A new multi-scale backbone architecture," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 2, pp. 652–662, 2021.

[24] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7132–7141.

[25] J. Su, Z. Jin, and A. Finkelstein, "HiFi-GAN-2: Studio-quality speech enhancement via generative adversarial networks conditioned on acoustic features," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2021, pp. 166–170.

[26] M. Ravanelli, T. Parcollet, P. Plantinga, A. Rouhe, S. Cornell, L. Lugosch, C. Subakan, N. Dawalatabad, A. Heba, J. Zhong, J.-C. Chou, S.-L. Yeh, S.-W. Fu, C.-F. Liao, E. Rastorgueva, F. Grondin, W. Aris, H. Na, Y. Gao, R. D. Mori, and Y. Bengio, "SpeechBrain: A general-purpose speech toolkit," 2021, arXiv:2106.04624.